

Using Lexical Bundles to Discriminate between Fraudulent and Non-fraudulent Financial Reports

Abstract

This is the first study to analyze language at the phraseological level of fraudulent and non-fraudulent MD&As. Specifically, we analyzed lexical bundles, phrases that are at least four words in length and occur in text at a minimum pre-specified rate. In this paper we used Natural Language Processing (NLP) techniques to extract lexical bundles from 202 Management's Discussion and Analysis (MD&A) sections of annual 10-K reports, 101 of which were fraudulent. We found which lexical bundles occurred most often in each set of MD&AS those bundles that were used at a significantly different rate. We then provided a theoretical basis for the difference in the use of bundles with loaded connotations. In sum, we propose the technique of analyzing language at the lexical bundle level as a potential auditing tool for assessing risk in audit engagements.

Keywords: (fraudulent financial reporting, lexical bundles, phraseology, 10-K, natural language processing)

Introduction

Recent research has investigated deceptive language in fraudulent annual reports and quarterly earnings conference calls as an auditing tool for assessing engagement risk. The linguistic indicators used for the analyses have generally consisted of readability cues (Li, 2008), psycho-social dictionaries (Larcker & Zakolyukina, 2010) such as those used in Linguistic Inquiry and Word Count (LIWC) (Pennebaker & Graybeal, 2001), and

cues indicating word and sentence complexity (Humpherys, Moffitt, Burns, Burgoon, & Felix, 2011; Moffitt & Burns, 2009). Several researchers (Bournois & Point, 2006; Merkyl-Davies & Brennan, 2007a, 2007b; Rutherford, 2005) have studied external financial reports as a separate genre with distinctive linguistic properties. For example, Presidents' Letters are constructed with long words and sentences, few pronouns except for a high number of first person plural pronouns ("we", "our"), more affect and colorful words and phrases, and an extremely high proportion of positive words and phrases. These studies share a common methodology because they have examined instances of single words in a "bag-of-words" manner in which relative position in a sentence or phrase, as well as context, is ignored. What is lacking is an understanding of the genre-specific semantics of connected words or common phrases that are used to describe the financial health and future outlook of a company. Understanding how phrases are used in deceptive corporate reports could lead to new techniques for auditors to assess risk.

In past research, dictionary-based analyses using LIWC, for example, extracted words and put them in a pre-defined category, such as words related to "money", regardless of the way that word is used or the context of the phrase or sentence. There are some problems with this approach. First, there is ambiguity because many words have more than one meaning. Second, the dictionaries are general-purpose so the word categories may not be appropriate or adequate for a very specific genre like financial reports. Finally, context of individual words cannot be considered since each word is handled separately without regard to its place in the document. Researchers have called for studies that addresses the context issue via natural language processing (NLP) techniques (Larcker & Zakolyukina, 2010).

This research attempts to fill that gap by considering the use of particular type of phrase known as a lexical bundle (Biber & Barbieri, 2007). In this paper we take an approach to

extracting units from text that have less semantic ambiguity than context-free single words (unigrams). This approach is particularly appropriate for formal genres that have a fairly rigid writing style such as financial reports. In this project we extract entire phrases (e.g., “the fair value of”) that are more semantically unambiguous and provide context to the individual words. Because of this we can provide more reliable interpretations of what an author means compared to interpreting the use of a single word.

This paper contributes to research streams in Accounting and Information Systems in the following ways: 1) We discuss the literature on phraseology and lexical bundles with respect to financial statements; 2) Using a sample of 202 MD&As, we identified lexical bundles that might be used to discriminate between fraudulent and non-fraudulent financial statements; and, 3) From an accounting standpoint, we discuss a subset of these lexical bundles to clarify why they differentiated fraudulent and non-fraudulent financial statements.

The rest of this paper consists of the following sections: first we define lexical bundles and previous research on that topic, next we review previous research in Fraudulent Financial Reporting and our research question, then the methodology is set forth followed by the results, a discussion of the results, and a conclusion.

Lexical bundles

The variability in patterns and usage of words and phrases in natural language is much lower than would be predicted by grammar and lexicon alone (Wray & Perkins, 2000) In fact, language, whether written or spoken, is up to 70% formulaic (Sinclair, 1991). Written and spoken language composition has been compared to stitching a quilt together, the patches being pre-constructed phrases (Marco, 2000). Phrasal constructions that have been investigated over the years include collocations, and lexical bundles.

Collocations have been defined as “fixed, non-idiomatic, identifiable phrases or constructions” (Benson, Ilson, & Benson, 1986). Strictly speaking, collocations are any sequence of two or more words that occur within a specified window length more frequently than by chance alone. Collocated words do not need to be directly adjacent to each other: when *build* and *momentum* are the collocated words they can exist as “*build momentum*”, or “*build a lot of momentum.*” They are arbitrary in their construction; however, to be considered collocations, they must recur at a pre-specified rate.

Lexical bundles are a specialized type of collocation. They are the most frequent multi-word sequences in a given register (e.g., financial reports, biology journals, history journals). Operationally, lexical bundles generally have been studied as four-word sequences that occur at least 20 times per million words in a given register (Biber & Barbieri, 2007; Cortes, 2004; Hyland, 2008; Wray & Perkins, 2000). Lexical bundles are domain specific (Hyland, 2008; Smadja, 1993) For example, Cortes (2004) found that 64.2% of the lexical bundles identified in History research journals did not meet the criteria to be classified as lexical bundles in Biology research journals. Moreover, 82.6% of bundles identified in the biology literature were not identified as bundles in history journals. In this project, phrases were considered lexical bundles if they occurred at least 20 times per million words in either a fraudulent or non-fraudulent MD&A corpus. Bundles must have also appeared in at least 15% of either the fraudulent or non-fraudulent documents. This last measure prevented a phrase from qualifying as a lexical bundle based on frequent usage in just a few MD&As.

The role of tools to aid auditors in detecting fraud

Much of the past research in fraudulent financial reporting has focused on analyzing the numbers found in financial reports for inconsistencies and anomalies that might indicate fraud (Beneish, 1997; Dechow, et al., 1996; Lee, Ingram, & Howard, 1999; Summers &

Sweeney, 1998) as well as concentrating on developing tools to help auditors analyze the quantitative data.

To make audit processes more effective, the American Institute of Certified Public Accountants' (AICPA) Auditing Standards Board released Statement on Auditing Standards (SAS) No. 99 in 2002. SAS 99 (AICPA, 2002) identifies three ways Financial Fraudulent Reporting can be committed by overstating earnings or understating losses: 1) supporting documents can be altered, falsified, or manipulated, 2) significant events or transactions can be misrepresented or omitted from financial statements, and 3) accounting principles can be intentionally misapplied. In the MD&A, fraud would be perpetuated by presenting a false version of past performance and an unrealistic outlook for the future, misrepresenting the significance of key events, omitting significant facts, and/or providing misleading information about the current health of the company.

SAS 99 gives guidance to auditors for fulfilling their responsibility of attesting that financial statements are free from material misstatements "whether they are caused by error or fraud" (AICPA, 2002). Loebbecke et al. (1989) make the distinction between errors and fraud, aka irregularities. Errors are not purposefully concealed which should make them more discoverable by auditors. When financial statement errors are detected, they are reported routinely to management and fixed immediately. In contrast, purposely concealed irregularities are more difficult to discover. When auditors interview managers about irregularities, managers are forced to lie to perpetuate the concealment. Since irregularities are difficult to detect and it is not in management's interest to reveal them. Unfortunately, assessing risk is a non-intuitive, humanly-biased, cognitively difficult task. Because managerial fraud happens so infrequently, most auditors have little direct experience to detect it effectively (Fanning, Cogger, & Srivastava, 1995). Behavioral accounting researchers (Eining, Jones, & Loebbecke, 1997; Pincus, 1989;

Zimbelman, 1997) report the difficulty auditors have synthesizing large amounts of information properly when predicting engagement risk.

Risk assessment can improve with experience, knowledge, training, reasoning skills, and tools (Loebbecke, et al., 1989). Without adequate exposure to certain cues that indicate fraud, it can be difficult for auditors to develop their own heuristics to discern problems in financial statements. To mitigate this problem, the AICPA suggests the use of Analytical Procedures (APs), for auditing (AICPA, 1988). APs are methods used to understand a company. According to SAS 56 (AICPA, 1988), APs “range from simple comparisons to the use of complex models involving many relationships and elements of financial and non-financial data”. Any computational tool, including statistical modeling and machine learning algorithms, used to understand a company’s profile as well as its engagement risk, uses APs.

Therefore, in today’s financial reporting environment, both the increased volume of financial data and the need for timely analyses call for efficient, automated techniques to augment auditors’ manual approaches. Recently, advances in computer classification techniques have enabled researchers and audit experts to use various types of data mining techniques to highlight possible instances of computer fraud. There are several key types of data mining which can be used against a variety of data types: a) Associative Rule Mining, often referred to by Market Basket Analysis, which reveals patterns of data items that occur frequently together; b) Classification and Prediction, which discovers a set of common cues or features that can discriminate among classification categories; c) Cluster Analysis, which slices a data set into smaller clusters that contain similar data items; and d) Sequential Pattern and Time-Series Mining, which looks for relationships among data that occurs in succession (Han & Kamber, 2001).

Key advantages of statistical and machine learning tools are that they eliminate human bias in decision making and consistently weigh and combine risk factors (Lin, Hwang, & Becker, 2003). Furthermore, adopting statistical and machine learning tools can mitigate the natural conflict that exists between the goal of audit effectiveness and the market pressures to attain audit efficiency (Green & Choi, 1997). When auditors fail to correctly assess risk initially, both audit efficiency and audit effectiveness suffer. Assessing engagement risk too low will reduce audit effectiveness by increasing the chance of undetected fraud while assessing engagement risk too high will reduce audit efficiency with unnecessary tests and costly investigations. Using a tool during the planning stage of the audit to properly assess risk should boost both audit efficiency and audit effectiveness.

To date, researchers (Calderon & Cheh, 2002; Fanning & Cogger, 1998; Fanning, et al., 1995; Gaganis, Pasiouras, & Doumpos, 2007; Kirkos, Spathis, & Manolopoulos, 2007; Kotsiantis, Koumanakos, Tzelepis, & Tampakas, 2006; Kovalerchuk & Vityaev, 2005; Lin, Hwang, & Becker, 2003; Spathis, 2002) and auditing professionals have applied data mining techniques to quantitative financial data to identify patterns of manipulation. The texts that accompany the financial data in 10-Ks, annual reports, etc., largely have been overlooked in this type of data mining. This is a significant oversight because it is estimated that unstructured text represents over 80% of current data (Zhang & Zhou, 2004). Fortunately, natural language processing (NLP) data mining techniques, including text mining, linguistic feature mining, and classification by text features, can be used to analyze the texts in financial statements. Text mining refers to looking for hidden patterns or cues in texts; linguistic feature mining refers to dissecting texts with respect to specific linguistic categories, such as words associated with positive affect. These analyses, such as providing word count of words with more than three syllables or

categorizing verb type, are far more complex than humans can perform practically. NLP is a multi-disciplinary research area that combines progress in computer science, linguistics, mathematics, communication, and psychology. NLP focuses on using computing power to process unstructured human language in spoken or written form (Zhou, Burgoon, Twitchell, Qin, & Nunamaker, 2004). Supporting NLP, high-performance computing systems can process text data to discover linguistic cues that can be used to classify the texts into categories, such as fraudulent vs. non-fraudulent financial statements (Humpherys, et al., 2010, Forthcoming; Moffitt & Burns, 2009) or deceptive vs. truthful statement in non-financial documents (Fuller, Biros, & Wilson, 2008; Hancock, Curry, Goorha, & Woodworth, 2008). A careful analysis of features of written texts can reveal which linguistic cues discriminate documents containing deceit from those documents that are truthful.

Using NLP, previous research (Moffitt & Burns, 2009) identified linguistic cues in MD&As that may highlight financial fraud. This study extends that previous research by identifying the most frequent and differing lexical bundles in fraudulent and non-fraudulent MD&As. Importantly, our current study of using automated approaches to extract and analyze lexical bundles complements past research using a “bag-of-words” approach by evaluating language at the phrase level. Our research questions for this study are:

RQ1: *What are the most frequently used lexical bundles in fraudulent and non-fraudulent MD&As?*

RQ2: *Which lexical bundles are used at significantly different rates in fraudulent and non-fraudulent MD&As?*

Methodology

Fraudulent 10-Ks were identified by searching for AAERs that included the term '10-K'. Companies named in AAERs are assumed to be guilty of earnings manipulations (Dechow, et al., 1996). After excluding 40 companies and their associated 10-Ks from the 141 initially identified (see Table 1), 101 company 10-Ks were left for analysis.

Table 1: Sample selection criteria for fraudulent 10-Ks	
Count of companies identified as fraudulent by searching through AAERs	141
Count disqualified because fraud did not involve 10-Ks	(20)
Count disqualified because 10-K was not available from the SEC	(10)
Count disqualified because 10-K did not contain management discussion section	(10)
Final count of qualifying 10-Ks used in the final sample	101

101 comparable non-fraudulent 10-Ks were chosen by selecting companies with Standard Industrial Classification (SIC) codes that exactly matched the companies that filed fraudulent 10-Ks. Each matching company's 10-K was also filed in the same year or in the previous/following year and had no amendments. The purposes of these criteria are to minimize potential confounds because of differing economic conditions or differences between non-comparable industries. The non-fraudulent companies have no AAERs attached to them, which suggests a history of compliance to SEC regulations. MD&As were extracted from each 10-K.

The Lexical Bundles were extracted from the MD&As using a program written in the Python programming language. The program identifies lexical bundles and exports their counts to a Comma Separated Value (csv) file. We identified lexical bundles that were four to ten words long that met the following criteria: bundles had to occur at a rate of at

least 20 times per million words and in at least 15% unique fraudulent or non-fraudulent MD&As. The rate of lexical bundles in each corpus are reported at a normalized rate of bundles per million words in order to make the bundle data comparable and to match previous research investigating lexical bundles.

Many of the smaller lexical bundles are sub-components of larger bundles. Table 2 shows the frequency per million words of the constituents of a 6-word bundle. The phrase “to continue as a going concern” accounts for 66 of the 91 uses of the phrase, “as a going concern”. For this study we focused more on reporting the results from the four-word bundles.

Table 2: Frequency of the components of a six-word bundle					
4-word bundles	N	5-word bundles	N	6-word bundle	N
as a going concern	91				
		continue as a going concern	74		
continue as a going	76			to continue as a going concern	66
		to continue as a going	68		
to continue as a	68				

Results

Table 3 includes the twenty-six most frequently encountered 4-word lexical bundles from non-fraudulent MD&As. Table 4 shows the twenty-six most frequently encountered 4-word lexical bundles from fraudulent MD&As. The seven most frequent lexical bundles are both in the top seven for each list. The percentage difference column in Tables 3 and 4 indicates the difference in the rate of usage for each phrase. For this paper we report this percentage for the top 26 bundles and discuss the theoretical reasons for the differences for additional bundles in the next section.

Table 3: Top 26 non-fraudulent 4-word lexical bundles ranked by frequency

Lexical Bundle	NonFraud Bundles Per Million Words	NonFraud Rank	Fraud Bundles Per Million Words	Fraud Rank	% diff.
the year ended December	1365	1	1195	2	14%
for the year ended	1223	2	1294	1	6%
as a result of	907	3	856	3	6%
as a percentage of	571	4	791	4	38%
general and administrative expenses	499	5	350	6	42%
million for the year	482	6	422	5	14%
a result of the	441	7	332	7	33%
selling general and administrative	313	8	249	11	26%
in connection with the	305	9	248	12	23%
during the year ended	291	10	78	120	274%
the fourth quarter of	287	11	211	19	36%
years ended december and	278	12	176	28	58%
there can be no	272	13	294	8	8%
was primarily due to	268	14	205	20	31%
the years ended december	252	15	163	32	55%
can be no assurance	245	16	264	9	8%
year ended december compared	245	17	142	42	73%
liquidity and capital resources	243	18	150	39	62%
the consolidated financial statements	239	19	179	27	34%
for the years ended	239	20	159	34	50%
ended december compared to	235	21	127	52	85%
of financial condition and	233	22	176	29	32%
in the fourth quarter	221	23	189	24	17%
be no assurance that	212	24	248	13	17%
the company believes that	210	25	127	53	66%
the first quarter of	206	26	160	33	29%

Many of the 4-word lexical bundles occur most often as constituents of larger lexical bundles. Table 5 lists longer lexical bundles that are comprised of top fraudulent lexical from Table 4. The numbers within parentheses next to the bundles in Table 4 identify the larger bundle it is part of.

Table 4: Top 26 fraudulent 4-word lexical bundles ranked by frequency

Lexical Bundle	Fraud Bundles Per Million Words	Fraud Rank	NonFraud Bundles Per Million Words	NonFraud Rank	% diff.
for the year ended (1)	1294	1	1223	2	6%
the year ended December (1)	1195	2	1365	1	14%
as a result of	856	3	907	3	6%
as a percentage of	791	4	571	4	38%
million for the year	422	5	482	6	14%
general and administrative expenses	350	6	499	5	42%
a result of the	332	7	441	7	33%
there can be no (3)	294	8	272	13	8%
can be no assurance (3)	264	9	245	16	8%
the fair value of	257	10	171	30	50%
selling general and administrative	249	11	313	8	26%
in connection with the	248	12	305	9	23%
be no assurance that (3)	248	13	212	24	17%
have a material adverse (4)	232	14	155	33	50%
in the year ended	226	15	72	156	213%
a percentage of net	224	16	148	37	51%
a material adverse effect (4)	221	17	151	36	47%
material adverse effect on (4)	218	18	138	42	58%
the fourth quarter of	211	19	287	11	36%
was primarily due to	205	20	268	14	31%
primarily due to the	199	21	192	27	4%
in process research and (2)	199	22	78	130	153%
process research and development (2)	199	23	76	138	160%
in the fourth quarter	189	24	221	23	17%
in the united states	185	25	148	38	25%
could have a material (4)	183	26	74	149	146%

Table 5: Lexical bundles derived from top 26 4-word fraud bundles

Fraudulent Lexical Bundles	ID	Frequency per million words in fraud corpus
for the year ended december	1	995
in process research and development	2	199
There can be no assurance that	3	245
could have a material adverse effect on	4	146

Overall, 564 four-word phrases met the criteria to be called lexical bundles. In addition, 220 five-word, 96 six-word, 42 seven-word, 13 eight-word, 4 nine-word, and 1 ten-word lexical bundles were identified. According to the results, phrases in MD&As are used more frequently than in previously analyzed registers. The most frequent bundles occurred nine times more often than the most frequent bundles found in biology or history scholarly journals (Cortes, 2004)

Discussion

In addition to the most frequent lexical bundles, many bundles are interesting because of their meaning and the large difference in their rate of usage between fraudulent and non-fraudulent MD&As. Table 6 contains the bundles that will be discussed in this section. We begin by exposing the environmental setting and motivations that may drive the use of these phrases.

Table 6: A subset of the lexical bundles that differed between fraudulent and non-fraudulent MD&As			
Lexical Bundle	Fraud Bundles Per Million Words	NonFraud Bundles Per Million Words	% diff.
the fair value of	257	171	50%
in foreign currency exchange	41	21	97%
in process research and development	199	76	160%
goodwill and other intangible assets	121	82	47%
long lived assets and	49	21	139%
purchase method of accounting	44	21	115%
to continue as a going concern	15	91	513%
disclosures about market risk	85	115	36%
material impact on the	38	52	35%
a material effect on	95	70	35%

Due to market-driven and executive compensation pressures, some companies have resorted to “earnings management” accounting which is strongly decried by the

Securities and Exchange Commission (SEC). The demand by investors, analysts, and management to “make the numbers” creates a cycle that is difficult to sustain. A company, facing a predicament, makes puffed up announcements about earnings, growth, or new opportunities to attract the attention of Wall Street analysts who, based on this hyperbole, predict hard-to-achieve earnings for that company. The company then must meet or beat analysts’ projections. Next, the company’s management compounds the problem by making other misleading announcements which also capture the analysts’ attention, causing them to raise earnings projections for the quarter. Auditors, balancing their financial oversight responsibilities with client retention, do not want to be viewed as setting up barriers as the company manipulates earnings to satisfy Wall Street. Involved in this cycle, the outcome of a company, and the reputation of its associated auditors, can teeter between great opportunity and devastating downfalls.

When organizations manipulate earnings, the trustworthiness of their financial statements is called into question. One of the key historical strengths of the U.S. capital markets has been the scrupulousness of the U.S. standards for financial reporting. Therefore, threatening the credibility of financial reporting can weaken the market overall and harm our system of financial disclosure (Munter, 1999).

There are several gray areas of accounting to which vigilant auditors should pay attention: creative acquisition accounting, including purchase accounting and in process research and development; “big bath” charges to current period, including disposition of long-lived assets; and, fair value accounting. In our analyses, we discovered significantly more lexical bundles in fraudulent financial statements that would point to these “earnings management” gray areas.

Creative acquisition accounting, including purchase accounting and in-process research and development

This section describes why fraudulent MD&As might include the language “purchase method of accounting,” “goodwill and other intangible assets,, and “in process research and development” at a higher rate than non-fraudulent MD&As. In a business combination, pooling-of-interests accounting is preferred by companies when they want more leeway in accounting for the acquisition. However, there are many cases in which a company is forced to apply the purchase method of accounting to recording the net assets. The purchase method of accounting refers to the way that net assets of the acquired business are recorded by the purchaser. According to the purchase method, the net assets are recorded at the fair market value of the consideration given by the acquiring company. If the purchase price exceeds the fair market value of the net assets, that excess is recorded under goodwill. Goodwill must be amortized or “dragged” against future earnings which is a problem for companies that want to report a solely upward trend in earnings quarter-to-quarter, year-to-year. To avoid the amortization of goodwill so that the excess can be charged against current earnings, some companies record part of the purchase price as “in-process” research and development. Because the expense, sometimes a quite substantial portion of the acquisition price (Munter, 1999), is taken in the current period, there is no future “dragging down” of earnings. One example of this practice of misusing the purchase method of accounting for acquisitions is Jamaica Water Properties (Knapp, 2009).

“Big bath” charges to current period, including disposition of long-lived assets

This section describes why fraudulent MD&As might include the language “long lived assets and” at a higher rate than non-fraudulent MD&As. Companies in financial distress may perform wholesale aggressive restructuring to improve cost and expense structure for the future. Companies may choose to take big cuts in one fell swoop to avoid

impairment to earnings in subsequent years. Though this may be a legitimate effort, this may also be a last-ditch effort to manage earnings. The estimation of the restructuring accrual, which will be accounted for in analysts' estimates of future earnings, may end up being much larger than the actual amount.

Fair value accounting

This section describes why fraudulent MD&As might include the language “the fair value of,” and “in foreign currency exchange” at a higher rate than non-fraudulent MD&As. The definition of fair value accounting under Statements of Financial Accounting Standard (SFAS) 157, *Fair Value Measurements* is “the price in an orderly transaction between market participants to sell the asset or transfer the liability in the market in which the reporting entity would transact for the asset or liability, that is, the principal or most advantageous market for the asset or liability” is. It is defined by International Financial Reporting Standards (IFRS) as “the amount for which an asset could be exchanged between knowledgeable, willing parties in an arm’s length transaction”. For financial reporting, there has been and continues to be a shift from applying cost-basis, using historical costs, to an increasing use of fair value accounting. In the debate about cost-basis vs. fair value accounting, one chief problem with fair value accounting is that it is not objective. High levels of expertise and judgment must be exercised when setting fair market value to assets (Zack, 2009).

There are several ways in which a company in a bind may choose to use fair value accounting to further their agenda to demonstrate growth in earnings. These include changing the value of available-for-sale investments to a deceptive highest or best use of an asset, anchoring the fair value estimate of a transaction to other, non-‘orderly’ transactions, understating a debt obligation, and misrepresenting foreign currency exchange adjustments.

Fair value accounting in financial statements may be used as a secondary fraudulent measure by an organization to cover up a first occurrence of fraud, such as asset misappropriation. In preparing financial statements, management may fail to disclose appropriate fair value information in footnotes.

Lexical bundles used in more frequently non-fraudulent MD&As:

Of considerable interest were the lexical bundles used more frequently by non-fraudulent companies, perhaps in an attempt to signal their lack of deception. Non-fraudulent companies may take special care in using more conservative language and accounting practices. Across 4-word, 5-word, and 6-word bundles (see Table 2), all of which occurred significantly more frequently in non-fraudulent statements, the theme of continuing as a 'going concern' was emphasized. Another phrase that was more common in non-fraudulent MD&As indicated that the company would have 'disclosures about market risk'. Interestingly, non-fraudulent companies used the phrase "material impact on the" more often; the phrase 'a material effect on' was favored much more frequently by fraudulent companies.

Conclusion

The lexical bundles analyzed for this study may be highly specific to the sample of MD&As used. As a result, the study may suffer from weak diagnostic power when applied to another data set (Li, 2010, In Press). Because our sample used MD&As filed with the SEC, our data set was restricted to mostly large, publicly traded companies. Thus, we are limited when trying to generalize to the population of all types of companies.

On the other hand, this study is the first to analyze the phraseological differences between fraudulent and non-fraudulent MD&As. Studying language at the lexical bundle level provides researchers a more intuitive and reliable sense as to the author's meaning

and intent compared to analyzing individual words. The contextual cues offered by groups of words are far richer than those that can be drawn from single words. Moreover, in this study we found non-trivial differences in the rate certain phrases were used. Due to space limitations, we could not expose them all in this paper. However, we did select a few key phrases and provide theoretical explanations to their rate of usage difference.

This type of research could be valuable to auditors as they seek for ways to assess engagement risk a priori and during the audit. Using a tool that extracts key phrases, auditors can focus their attention on language that accompanies high-risk companies. Future research should conduct a more extensive analysis of longer lexical bundles, and those bundles that differentiate high- and low-risk companies.

References

- AICPA (1988). SAS No. 56: Analytical Procedures.
- AICPA (2002). SAS No. 99: Consideration of Fraud in a Financial Statement Audit.
- Beneish, M. (1997). Detecting GAAP Violation: Implications for Assessing Earnings Management Among Firms with Extreme Financial Performance. *Journal of Accounting and Public Policy*, 16(3), 271-309.
- Benson, M., Ilson, R., & Benson, E. (1986). *The BBI Combinatory Dictionary of English*. Amsterdam/Philadelphia: Benjamins.
- Biber, D., & Barbieri, F. (2007). Lexical Bundles in University Spoken and Written Registers. *English for Specific Purposes*, 26, 263-286.
- Bournois, F., & Point, S. (2006). A Letter from the President: Seduction, Charm, and Obfuscation in French CEO Letters. *Journal of Business Strategy*, 27(6), 46-55.
- Calderon, T. G., & Cheh, J. J. (2002). A Roadmap for Future Neural Networks Research in Auditing and Risk Assessment. *International Journal of Accounting Information Systems*, 3(4), 203-236.
- Cortes, V. (2004). Lexical Bundles in Published and Student Disciplinary Writing: Examples from History and Biology. *English for Specific Purposes*, 23, 397-423.
- Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1996). Causes and Consequences of Earnings Manipulation: An Analysis of Firms Subject to Enforcement Actions by the SEC. *Contemporary Accounting Research*, 13(1), 1-36.
- Eining, M. M., Jones, D. R., & Loebbecke, J. K. (1997). Reliance on decision aids: An examination of auditors' assessment of management fraud. *Auditing: A Journal of Practice & Theory*, 16(2), 1-19.
- Elliot, R., & Willingham, J. (1980). *Management Fraud: Detection and Deterrence*. New York, NY: Petrocelli.
- Fanning, K., & Cogger, K. O. (1998). Neural Network Detection of Management Fraud Using Published Financial Data. *International Journal of Intelligent Systems in Accounting, Finance & Management*, 7(1), 21-41.
- Fanning, K., Cogger, K. O., & Srivastava, R. (1995). *Detection of Management Fraud: A Neural Network Approach*. Paper presented at the Proceedings of the 11th Conference on Artificial Intelligence for Applications.
- Financial Accounting Standards Board. Summary of Statement No. 157, Fair Value Measurements, from <http://www.fasb.org/summary/stsum157.shtml>
- Fuller, C. M., Biros, D. P., & Wilson, R. L. (2008). Decision Support for Determining Veracity via Linguistic-Based Cues. *Decision Support Systems*, 46(3), 695-703.

- Gaganis, C., Pasiouras, F., & Doumpos, M. (2007). Probabilistic Neural Networks for the Identification of Qualified Audit Opinions. *Expert Systems with Applications*, 32(1), 114-124.
- Green, B. P., & Choi, J. H. (1997). Assessing the Risk of Management Fraud through Neural Network Technology. *Auditing: A Journal of Practice & Theory*, 16(1), 25-28.
- Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann.
- Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2008). On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication. *Discourse Processes*, 45(1), 1-23.
- Humpherys, S., Moffitt, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of Fraudulent Financial Statements using Linguistic Credibility Analysis. *Decision Support Systems*, 50(3), 585-594.
- Hyland, K. (2008). As Can Be Seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes*, 27, 4-21.
- International Financial Reporting Standards (2010). from <http://www.ifrs.com/index.html>
- Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data Mining Techniques for the Detection of Fraudulent Financial Statements. *Expert Systems with Applications*, 32(4), 995-1003.
- Knapp, M. C. (2009). *Contemporary Auditing: Real Issues and Cases* (7th ed.): South-Western Cengage Learning.
- Kotsiantis, S., Koumanakos, E., Tzelepis, D., & Tampakas, V. (2006). Forecasting Fraudulent Financial Statements Using Data Mining. *International Journal of Computational Intelligence*, 3(2), 104-110.
- Kovalerchuk, B., & Vityaev, E. (2005). Data Mining for Financial Applications. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 1203-1224): Springer US.
- Larcker, D. F., & Zakolyukina, A. A. (2010). Detecting Deceptive Discussions in Conference Calls. *Rock Center for Corporate Governance Working Paper Series*, 83.
- Lee, T. A., Ingram, R. W., & Howard, T. P. (1999). The Difference Between Earnings and Operating Cash Flow as an Indicator of Financial Reporting Fraud. *Contemporary Accounting Research*, 16(4), 749-786.
- Li, F. (2008). Annual Report Readability, Current Earnings, and Earnings Persistence. *Journal of Accounting & Economics*, 45, 221-247.

- Li, F. (2010, In Press). The Information Content of Forward-Looking Statements. *Journal of Accounting Research*.
- Lin, J. W., Hwang, M. I., & Becker, J. D. (2003). A Fuzzy Neural Network for Assessing the Risk of Fraudulent Financial Reporting. *Managerial Auditing Journal*, 18(8), 657-665.
- Loebbecke, J. K., Eining, M. M., & Willingham, J. J. (1989). Auditors' Experience with Material Irregularities - Frequency, Nature, and Detectability. *Auditing: A Journal of Practice & Theory*, 9(1), 1-28.
- Marco, M. J. L. (2000). Collocational Frameworks in Medical Research Papers: A Genre-Based Study. *English for Specific Purposes*, 19(1), 63-86.
- Merkyl-Davies, D. M., & Brennan, N. (2007a). *The Chairman's Report as a Corporate Genre: Employing a Corpus-based Genre Analysis Approach*. Paper presented at the BAA's Annual Conference, Royal Holloway, University of London.
- Merkyl-Davies, D. M., & Brennan, N. (2007b). Discretionary Disclosure Strategies in Corporate Narratives: Incremental Information or Impression Management? *Journal of Accounting Literature*, 27.
- Moffitt, K., & Burns, M. B. (2009). *What Does That Mean? Investigating Obfuscation and Readability Cues as Indicators of Deception in Fraudulent Financial Reports*. Paper presented at the Fifteenth Americas Conference on Information Systems, San Francisco, CA, August 6th - 9th, 2009.
- Munter, P. (1999). SEC Sharply Criticizes "Earnings Management" Accounting. *The Journal of Corporate Accounting and Finance*(Winter).
- Pennebaker, J. W., & Graybeal, A. (2001). Patterns of Natural Language Use: Disclosure, Personality, and Social Integration. *Current Directions in Psychological Science*, 10(3), 90-93.
- Pincus, K. V. (1989). The Efficacy of a Red Flags Questionnaire for Assessing the Possibilities of Fraud. *Accounting Organizations and Society*, 14(1/2), 153-163.
- Rutherford, B. (2005). Genre Analysis of Corporate Annual Report Narratives: A Corpus Linguistics-based Approach. *Journal of Business Communication*, 42(4), 349-378.
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Smadja, F. (1993). Retrieving Collocations from Text: Xtract. *Computational Linguistics*, 19(1), 143-177.
- Spathis, C. T. (2002). Detecting False Financial Statements Using Published Data: Some Evidence from Greece. *Managerial Auditing Journal*, 17(4), 179-191.

- Summers, S. L., & Sweeney, J. T. (1998). Fraudulently Misstated Financial Statements and Insider Trading: An Empirical Analysis. *The Accounting Review*, 73(1), 131-146.
- Wray, A., & Perkins, M. R. (2000). The Functions of Formulaic Language: An Integrated Model. *Language & Communications*, 20, 1-28.
- Zack, G. M. (2009). *Fair Value Accounting Fraud: New Global Risks and Detection Techniques*: John Wiley & Sons.
- Zhang, D., & Zhou, L. (2004). Discovering Golden Nuggets: Data Mining in Financial Application. *IEEE Transactions on Systems, Man, and Cybernetics -- Part C: Applications and Reviews* 34(4), 513-522.
- Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., & Nunamaker, J. F., Jr. (2004). A Comparison of Classification Methods for Predicting Deception in Computer-Mediated Communication. *Journal of Management Information Systems*, 20(4), 139-165.
- Zimbelman, M. F. (1997). The Effects of SAS No. 82 on Auditors' Attention to Fraud Risk Factors and Audit Planning Decisions. *Journal of Accounting Research*, 35(Supplement).