Linguistic Correlates of Deception in Financial Text
A Corpus Linguistics Based Approach

Saliha Minhas[1] and Prof A. Hussain
University of Stirling, szm@cs.stir.ac.uk
University of Stirling, ahu@cs.stir.ac.uk

Abstract: In the present era, linguistic imprints are everywhere. Blogs, tweets, and texting all leave traces of our intentions and emotions. Some call this our linguistic output – akin to a fingerprint.[2] Consequently, its use by those given to lies and deception would be distinct from truth-tellers. Can this uniqueness be harnessed to battle criminality, given the rising level of financial fraud this link is testing empirically in the financial reporting domain? A corpus of 6.3m words is constructed, the composition being narrative sections of 102 annual reports/10-K from firms formally indicted for financial statement fraud juxtaposed with the corresponding narratives from 306 firms of the same industry, time period and size. Language use is examined using techniques from the Corpus Linguistics toolkit. This embraces frequency counts and keyword identification. The latter is undertaken using custom-built wordlists for the financial domain. Additionally, Linguistic Inquiry Word Count (LIWC) 2015, a dictionary-based tool, is also executed over the corpus to further aid in identification of the linguistic correlates of deception. A statistical procedure, Principal Component Analysis, is then run over the keywords and LIWC variables uncovered to further highlight those words that show up the greatest difference in use between the fraud and non-fraud reports. Finally, Multidimensional Scaling is employed to enable visualisation of the differences in the use of linguistic features between the two reports. The results indicate that the linguistic constructs examined are distinctively different when the two sets of narratives are compared.

[1] The authors gratefully acknowledge patience shown by editorial team as the article underwent development and for providing critical feedback.
[2] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, 'Lying Words: Predicting Deception from Linguistic Styles', Personality and Social Psychology Bulletin vol. 29, pp. 665-675, 2003

1. Introduction

A recent report[3] puts the annual cost of fraud to the UK economy at £193 billion a year – equating to more than £6,000 lost per second every day. Such costs are then mitigated by business through higher costs on products and services borne by the unsuspecting consumers unconnected to these financial scams. Therefore, it is a self-evident truth that fraud in all its guises negatively impacts quality of life. In the financial fraud arena, detection of the different types of financial fraud as depicted by Ngai[4] is tackled using quantitative variables. Models built on such a premise have failed spectacularly. The 2008 financial crash was in hindsight attributed to models that narrowly focused on features that did not capture the full scale of the risks faced by individuals and firms on the investments they made[5,6]. In this paper, in order to contribute to the search for alternatives that can be an additional aid to predicting catastrophic financial events, financial narratives are examined. Specifically, the aim is to show that those who have engaged in deception, such as financial statement fraud (FSF), have different language patterns from truth-tellers. To demonstrate this, a corpus is constructed from 102 narratives from annual reports/10-K of fraud firms aligned with 306 narratives from similar non-fraud firms. This unbalanced composition of the reports is an attempt to reflect the real-world scenario where there are more truthful narratives than deceptive ones. This corpus will be examined using the techniques commonly applied within the methodological discipline of corpus linguistics. Specific wordlists are also be deployed in an attempt to determine keyness differences between the two types of reports. A tool commonly applied in deception research (Hauch et al. provide a comprehensive list),[7] LIWC is also executed over the corpus to pick out 'cognitive, and

[3] J. Croft, "Fraud Costs the UK up to £193bn per year report says," in Financial Times, ed, 2016
[4] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," Decision Support Systems, vol. 50, pp. 559-569, 2011
[5] P. Omerod, 'Ostrich Economics', 2009
[6] J. Stiglitz, Freefall America, Free Markets and Sinking of the World Economy, New York: W.W. Norton & Company Inc, 2010
[7] V. Hauch, I. Blandon-Gitlin, J. Masip, and S. L. Sporer, "Are Computers Effective Lie Detectors? A Meta-Analysis of Linguistic Cues to Deception," Personality and Social Psychology Review, vol. 19, pp. 307-342, 2014

structural components'[8] present in reports that could differentiate fraud from non-fraud firms. This is the first study in deception research that executes the new version of LIWC (2015) replete with updated dictionaries over the corpus. The paper is structured as follows: Section 2 covers the rationale for using a corpus to examine deceptive texts and overviews similar pertinent work. Section 3 presents the corpus linguistics methodology and the approach taken to execute LIWC over the reports. It also introduces Principal Component Analysis (PCA) and Multi-Dimensional Scaling (MDS) that are executed over the results from corpus analysis and LIWC variables to determine the most distinguishing features and to enable visualisation in 2D. Section 4 presents the final results and Section 5 concludes the paper.

2. Background and Literature Review

Corpus Linguistics is a fast growing methodology in contemporary linguistics.[9] This entails the construction of a corpus: 'a body of naturally occurring language'.[10] Analysis of the corpus is then performed with the help of specialised software. This approach is rooted in the empirical school of thought, originating from the scientific method. It argues that the use of a corpus provides insights into the patterns of language use, the extent to which they are used, and the contextual factors that influence variability.[11] According to Sinclair (2001), the raison d'etre of corpus-based language study is to identify differences: 'the distinguishing features of one type of text only come to the forefront when contrasted to another type of text'.[12] This is particularly needed in deception-based research where the key is to be able to recognise a lie. In the past,

[8] J. Pennebaker, R. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC 2015", University of Texas, Austin 2015

[9] S. T. Gries, "What is corpus linguistics?", Language and Linguistics Compass vol. 3, pp. 1–17, 2009

[10] T. McEnery and A. Wilson, Corpus Linguistics, An Introduction: Edinburgh University Press, 2005

[11] D. Krieger, "Corpus Linguistics: What it is and how it can be applied to teaching." The Internet TESL Journal, 2003

[12] B. C. Camiciottoli, *Rhetoric in financial discourse*. The Netherlands: Rodopi, 2013

researchers[13,14] have set up controlled experiments to aid in distinguishing a liar from a truth-teller. However, such studies are hampered by poor reproducibility of results; subjects have no personal loss or gain at stake and the motivation to lie is weak. Fitzpatrick and Bachenko propose the 'construction of standardised corpora that would provide a base for expanding deception studies, comparing different approaches and testing new methods'.[15] They recommend using publicly available data as it is likely to be a rich source of ground truth evidence. A perfect example of this is the Enron e-mail corpus [15].[16] This has been extensively interrogated and linguistic features put through algorithms to pick up patterns that could be indicative of fraud and workplace behavioural cues. This kind of empirical data would be very hard to attain in a laboratory setting. Further in the arena of high stakes deception there is a 'sparsity of ground truth verification for data collected from real world sources' (Fitzpatrick and Bachenko, 2012). To address such a short supply of 'ground truth' a new

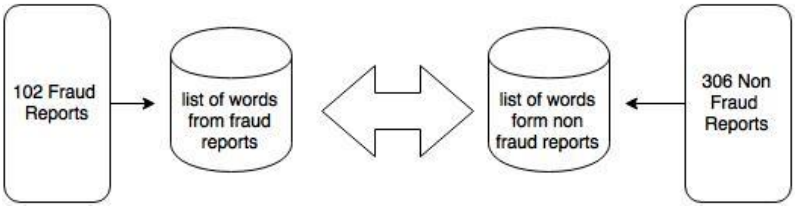

Figure 1: Reports set-up in AntConc to perform keyword analysis

[13] J. T. Hancock, L. E. Curry, S. Goorha, and M. Woodworth, "On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication," *Discourse Processes*, vol. 45, pp. 1-23, 2007

[14] N. D. Duran, C. Hall, P. M. McCarthy, and D. S. McNamara, "The linguistic correlates of conversational deception: Comparing natural language processing technologies," *Applied Psycholinguistics*, vol. 31, pp. 439-462, 2010

[15] E. Fitzpatrick and J. Bachenko, "Building a Data Collection for Deception Research," in *Proceedings of the EACL 2012 Workshop on Computational Approaches to Deception Detection,*, Avignon, France, 2012, pp. 31-38

[16] J. Hardin and G. Sarkis, 'Network Analysis with the Enron Email Corpus', *Journal of Statistics Education*, vol. 23, No. 2, 2015

corpus of 6.3 million words is constructed. The narratives were collected from firms known to have committed FSF and matched with narratives from similar firms/same time period.

McNamara et al. outline pre-requisites for building a corpus which stipulates that the language must be of a particular genre and be thematically related. It must also be balanced and representative. A corpus is said to be balanced 'if the relative sizes of each of its subsections have been chosen with the aim of adequately representing the range of language that exists in the population of texts being sampled'.[17] A representative corpus is one sampled in such a way that it contains all the types of text, in the correct proportions, that are needed to make the contents of the corpus an accurate reflection of the variety of language that it samples (McNamara, Graesser, McCarthy, and Cai, 2014).

The AR/10-K are financial texts of a particular genre and fulfils the representative and balanced criteria. However, as McNamara et al. point out, it does not need to be a 'perfect corpus; we just need one that gets the ball rolling'. This perfect corpus would be time consuming and expensive to collect. The practical aspects of corpus compilation are underappreciated.[18] The results from corpus-based studies should be 'practical and suggestive rather than exhaustive and definitive' (McNamara, Graesser, McCarthy, and Cai, 2014).

The alternative, rationalist school of thought led by Noam Chomsky refutes the validity of using corpora to adequately represent language. Chomsky argues that all empirical collections of language samples are skewed and incomplete.[19] They are skewed in that they favour particular uses of language at the expense of others, and incomplete because the number of sentences in a language is infinite; no finite collection of text could ever fully represent all possible configurations of words (McEnery and Wilson, 2005). Empiricists like McEnery argue that the use of a corpus enables 'good real world performance' by assigning probabilities to linguistic events so that they can say which sentences are 'usual' and 'unusual' (McEnery and Wilson, 2005). They concede that corpora cannot provide complete accounts of language use but maintain that it enables key insights into language use that would be otherwise difficult to grasp. They emphasise that our language capacity is infinite, and our language use is limited. Largely,

[17] D. S. McNamara, A. C. Graesser, P. M. McCarthy, and Z. Cai, *Automated Evaluation of Text and Discourse with Coh-Metrix*. New York: Cambridge University Press 2014

[18] P. Rayson, 'Computational Tools and Methods for Corpus Compilation and Analysis', in *Cambridge Handbook of English Corpus Linguistics*, ed., 2015

[19] "Corpus Linguistics" Research Starters eNotes.com, Inc. eNotes.com 24 Nov, 2016

people speak in preformed phrases that are repeated over and over again without knowing it. This is well-captured in a corpus.

A corpus has been used to differentiate liars from truth-tellers. Some recent research is now briefly described. Burgoon et al. built a corpus of 1114 statements made by a CEO formally indicted for fraud.[20] From this corpus they extracted key linguistic markers of deception, first introduced by Zhou et al.[21] The results from these markers were then put through hypothesis testing. They find fraud-related utterances differed systematically from non-fraud utterances. Specifically they state that 'consistent with recent evidence

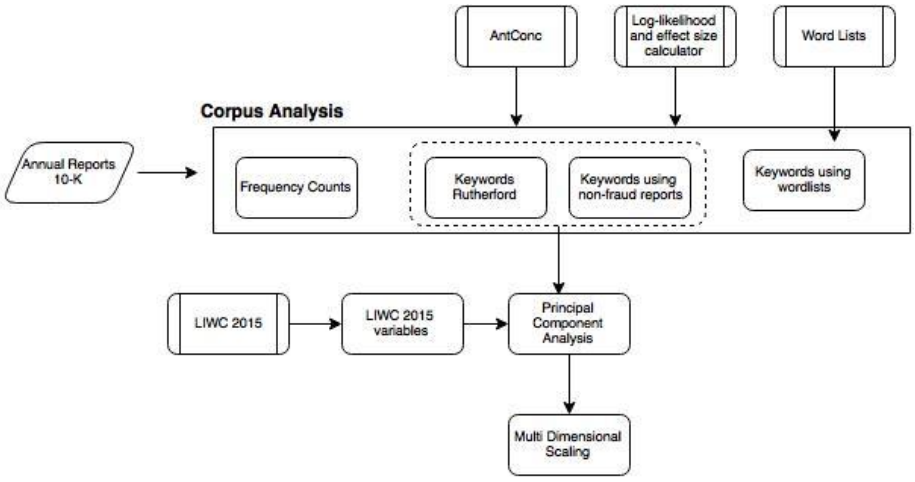

Figure 2: Proposed Corpus Analysis based Approach

[20] J. Burgoon, W. J. Mayew, J. S. Giboney, A. C. Elkins, K. Moffitt, B. Dorn, "Which Spoken Language Markers Identify Deception in High-Stakes Settings? Evidence From Earnings Conference Calls," *Journal of Language and Social Psychology,* vol. 35, pp. 123-157, 2015

[21] L. Zhou, J. Burgoon, J. Nunamaker, and D. Twitchell, "Automating Linguistics-Based Cues for Detecting Deception in Text-based Asynchronous Computer Mediated Communication," *Group Decision and Negotiation,* vol. 13, pp. 81-106, 2004

in the political arena by Braun et al. (2015) that fraud utterances were longer and more laden with details than non-fraud ones' (Burgoon, Mayew, Giboney, Elkins, Moffitt, and Dorn, 2015).

Fuller at al. also extracted linguistic-based cues from 367 written statements prepared by suspects and victims of crimes on military bases.[22] They found that linguistic markers related to length and details of messages, quantity of emotive language used; language that distanced the speaker from the message were significantly different between liars and truth-tellers.

Burns et al. used 50 transcribed 911 calls (25 truthful and 25 deceptive calls) and executed LIWC (2007) over the data.[23] They found that truthful callers display more negative emotion and anxiety than deceivers. They also referred to others in third-person singular form and gave more details. Deceivers used third-person plural at a higher rate, perhaps to deflect blame. They also demonstrated more immediacy than truth-tellers by using more first person singular and first person plural pronouns.

Fornaciari and Poesio[24] also used LIWC 2007 over a corpus of court transcripts containing both truthful and deceptive testimonies and found marked differences between the two types of testimonies. Larcker and Zakolyukina[25] also used LIWC 2007 over narratives of CEOs and CFOs' conference calls. The analysis indicates that deceptive executives make more references to general knowledge, fewer non-extreme positive emotions, and fewer references to shareholders value and value creation. In addition, deceptive CEOs use significantly fewer self-references, more third-person plural and impersonal pronouns, more extreme positive emotions, fewer extreme negative emotions, and fewer certainty and hesitation words.

Bacheko et al.,[26] McCarthy et al.,[27] Hancock et al., and Duran et al. all built up corpora in a domain of interest and checked linguistic style for differences between liars and truth-teller. The

[22] C. M. Fuller, D. P. Biros, J. Burgoon, and J. Nunamaker, "An Examination and Validation of Linguistic Constructs for Studying High-Stakes Deception," *Group Decision and Negotiation,* vol. 22, pp. 117-134, 2012

[23] M. B. Burns and K. C. Moffitt, "Automated deception detection of 911 call transcripts," *Security Informatics,* vol. 3, p. 8, 2014

[24] T. Fornaciari and M. Poesio, "On the use of homogenous sets of subjects in deceptive language analysis," presented at the Proceedings of the Workshop on Computational Approaches to Deception Detection, Avignon, France, 2012

[25] D. F. Larcker and A. A. Zakolyukina, "Detecting Deceptive Discussions in Conference Calls," *Journal of Accounting Research,* vol. 50, pp. 495-540, 2012

[26] J. Bachenko, E. Fitzpatrick, and M. Schonwetter, "Verification and implementation of language-based deception indicators in civil and criminal narratives," presented at the

findings are all in the affirmative. There is a marked difference that can be detected. For example, McCarthy et al. found that deceivers employ distancing strategies. Liars produce more words, more sense-based words (for example seeing, touching) and used fewer self-oriented but more other-oriented pronouns when

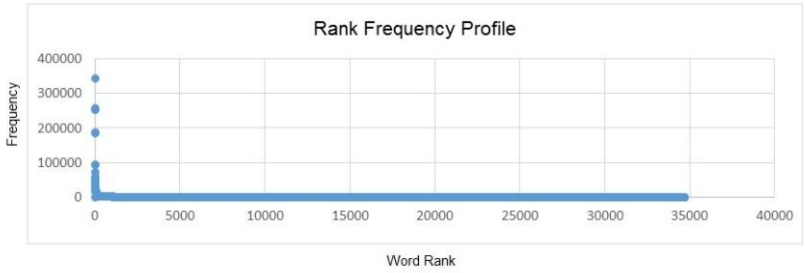

Figure 3: Zipf law in action over the corpus, a plot of word rank versus frequency

Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1, Manchester, United Kingdom, 2008

[27] P. M. McCarthy, N. D. Duran, and L. M. Booker, "The Devil Is in the Details: New Directions in Deception Analysis," in *Twenty-Fifth International Florida Artificial Intelligence Research Society Conference*, Florida, 2012

lying than when telling the truth (Hancock, Curry, Goorha, and Woodworth, 2007). Duran et al. find that the total word count, negation, and personal pronouns are variables able to distinguish narratives of liars from truth-tellers. More directed studies using annual report/10-K as a corpus to pick up linguistic features were conducted by Humphreys et al.,[28] Goel et al.,[29] Glancy and Yadav,[30] Purda and Skillcorn,[31] Throckmorton et al.,[32] Cecchini et al.[33] These studies primarily picked up known linguistic cues to deception, or extracted features that were more pronounced in fraud reports and then applied data mining algorithms, such as classification and clustering. The results all indicate that linguistic features are able to differentiate between the narratives of fraud and non-fraud firms.

3. Methodology

The annual reports/10-Ks of firms formally indicted for FSF were collected from 1989 to 2012. Only the narrative sections were kept (sections that dealt with corporate social responsibility and corporate governance were also removed, in keeping with past research). The narratives were stripped of all formatting and put into .txt files. This resulted in 102 files of narratives from fraud firms matched with 306 files from similar non-fraud firms. These reports were then loaded into AntConc, a freeware corpus analysis toolkit for text analysis.[34] The reports were then put through the following two methods commonly applied in corpus linguistics.

Set up of frequency lists

[28] S. L. Humpherys, K. C. Moffitt, M. B. Burns, J. K. Burgoon, and W. F. Felix, "Identification of fraud financial statements using linguistic credibility analysis," *Decision Support Systems,* vol. 50, pp. 585-594, 2011

[29] S. Goel, J. Gangolly, S. R. Faerman, and O. Uzuner, "Can Linguistic Predictors Detect Fraud Financial Filings?," *Journal of Emerging Technologies in Accounting,* vol. 7, pp. 25-46, 2010

[30] F. H. Glancy and S. B. Yadav, "A computational model for financial reporting fraud detection," *Decision Support Systems,* vol. 50, pp. 595-601, 2011

[31] L. Purda and D. Skillicorn, "Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection," *Contemporary Accounting Research,* vol. 32, pp. 1193-1223, 2015

[32] C. S. Throckmorton, W. J. Mayew, M. Venkatachalam, and L. M. Collins, 'Financial fraud detection using vocal, linguistic and financial cues', *Decision Support Systems,* vol. 74, pp. 78-87, 2015

[33] M. Cecchini, H. Aytug, G. J. Koehler, and P. Pathak, "Making words work: Using financial text as a predictor of financial events," *Decision Support Systems,* vol. 50, pp. 164-175, 2010

[34] L. Anthony. (2014). *AntConc (Version 3.4.3). Available from* http://www.laurenceanthony.net/

These lists: 'record the number of times that each word occurs in the text. It can therefore provide interesting information about the words that appear (and do not appear) in a text'.[35] The frequency information gives an indication of the vocabulary composition of the text. Sinclair noted that 'anyone studying a text is likely to need to know how often each different word form occurs in it'.[36] Additionally, according to McEnery and Hardie, a full appreciation of the frequency of a token in the text is only possible through a normalised frequency which answers the question: 'how often might we assume we will see the word per x words of running text?'[37] In this study *x* is 1000 words, a typical base of normalisation for density scoring.

Keyword Analysis

This is one of 'the most widely-used methods for discovering significant words, and is achieved by comparing the frequencies of words in a corpus with frequencies of those words in a (usually larger) reference corpus' (Baron, Rayson, and Archer). The measure used to determine keyness is a log-likelihood score and/or a log ratio score. The log-likelihood is a *statistical significance* measure – it tells us how much evidence there is for a difference between two corpora. The higher the log-likelihood value, the more significant is the difference between two frequency scores. A score of 3.8 or higher is significant at the level of $p < 0.05$. A negative value indicates underuse in the fraud corpus in relation to the non-fraud reports. However, log-likelihood does not indicate how big or how important a given difference is. The log ratio calculation would show up this difference.[38]

Keyword analysis was performed in 3 alternative ways:

1. The 102 fraud reports were loaded into AntConc. The 306 non-fraud reports were also loaded and set up as the reference corpus. The keyword generation method was set to log-likelihood.

[35] A. Baron, P. Rayson, and D. Archer, "Word frequency and key word statistics in historical corpus linguistics," *International Journal of English Studies*, vol. 20, pp. 41-67

[36] J. Sinclair, *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991

[37] T. McEnery and A. Hardie, *Corpus linguistics: method, theory and practice*. Cambridge: Cambridge University Press, 2012

[38] C. Gabrielatos and A. Marchi, "Keyness: Appropriate metrics and practical issues", CADS International Conference 2012. Corpus-assisted Discourse Studies: More than the sum of Discourse Analysis and computing?, 13-14 September, University of Bologna, Italy, 2012

This produced a list of keywords sorted by keyness scores. Figure 1 shows the approach taken by AntConc to determine keyness for words in the fraud reports. The output produced is a list of word types that are more salient in fraud reports and those that are more salient in non-fraud reports.

2. A study using corpus analysis methods on annual report narratives conducted by Rutherford[39] uncovered words that were deemed indicative of company health. These words were also put through a log-likelihood calculation to determine if there is a difference in usage of these words between fraud and non-fraud reports. For this task, log-likelihood and effect size calculator as devised by Rayson[40] is used. This calculator also computes a log ratio score.

3. Loughran and McDonald[41] developed wordlists customised for the financial domain. They showed that wordlists developed for other disciplines misclassify words in financial texts. The wordlists that they developed included negative, positive and uncertainty bearing words. As indicated by Pollach[42] such words can point to differences in 'themes and attentional foci' between the two sets of reports. These wordlists were loaded into AntConc and raw frequencies were noted. These frequencies are then passed to the log-likelihood calculator to determine keyness scores. A log ratio score is again computed.

In a further bid to pick up differences in linguistic style between the two reports, LIWC 2015 is executed over each text file in the corpus. Tausczik and Pennebaker[43] cite a number of reasons that give weight to using LIWC 2015 to take a closer look at language use. LIWC employs a simple yet intuitive way to measure language use in a variety of settings. LIWC reads written

[39] B. Rutherford, "Genre Analysis of Corporate Annual Report Narratives: A Corpus Linguistics Based Approach," *Journal of Business Communication,* vol. 42, pp. 349-378, 2005
[40] P. Rayson and R. Garside, "Comparing corpora using frequency profiling " in *Proceedings of the workshop on Comparing Corpora*, Hong Kong, 2000
[41] T. Loughran and B. McDonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *The Journal of Finance,* vol. 66, pp. 35-65, 2011
[42] I. Pollach, "Taming Textual Data: The Contribution of Corpus Linguistics to Computer-Aided Text Analysis," *Organizational Research Methods,* vol. 15, pp. 263-287, 2011
[43] Y. R. Tausczik and J. W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," *Journal of Language and Social Psychology,* vol. 29, pp. 24-54, 2009

text in .txt files. Its text analysis module compares each word in the text against the programme's user-defined dictionary. Once the processing module has read and accounted for all words in a given file, it calculates the percentage of total words that match each of the dictionary categories. The new 2015 version of LIWC uses a new updated master dictionary. It is composed of almost 6,400 words, word stems, and selected emoticons. A dictionary word can belong to one or more word categories. An example given by Pennebaker illustrates this point: 'the word 'cried' is part of five word categories: Sadness, Negative Emotion, Overall Affect, Verb, and Past Focus. Hence, if the word 'cried' was found in the target text, each of these five sub-dictionary scale scores would be incremented' (Pennebaker, Boyd, Jordan, and Blackburn, 2015). LIWC 2015 is run over each report in the corpus which results in 35 LIWC variables for each report. Examples of variables include function words, total pronouns, affective processes, cognitive processes, perceptual processes, drives, time orientations and relativity (comprehensive details on variables given by Pennebaker [(Pennebaker, Boyd, Jordan, and Blackburn, 2015)]).

In a bid to determine keywords and LIWC categories that lend most weight to the discrimination task, PCA is executed over these features. Principal component analysis (PCA) is a technique used to bring out strong patterns in a dataset. It simply finds the principal components of data which are those data points that show the greatest variability.[44]

First, the keywords unearthed from step 1 in keyword analysis are gathered. The tf-idf score for each of these keywords are obtained. The tf is the normalised term frequency (number of times a word appears in a report divided by the total number of words in that document). The second term the IDF is computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears. A tf-idf score denotes how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.[45]

The tf-idf scores for keywords unearthed by Rutherford [38] is also similarly calculated.

This produced 2 matrices of 408 rows long denoting the reports in the corpus with columns being the keywords and cells being the tf-idf scores. Another matrix was set up, again 408 rows long.

[44] I. T. Jolliffe, *Principle Component Analysis*. 2nd edition, England, Springer, 2002

[45] C. D. Manning and H. Schutze, Foundations of Statistical Natural Language Processing: MIT Press, 1999.

This time the columns are 35 LIWC variables (described in Pennebaker et al. [2009]) being the columns with the cells being the count of words in the LIWC category found in the report divided by the total number of words in the report. PCA is executed over these matrices.
The principal components or the features (keywords/LIWC variables) that contribute most to the variability are then used to construct matrices of smaller dimensions. Now, in order to show up the difference between the two categories of reports as defined by the reduced set, Multi-Dimensional Scaling (MDS) is executed. This 'provides a visual representation of the pattern of proximities (i.e., similarities or distances) among a set of objects'.[46]

The multidimensionality of the data (the number of features, in other words) is scaled down to a 2D representation which is cognizant of the initial distances between the features. Closer points indicate that the reports are more similar, as defined by the features chosen than some that are further apart. The corpus analysis methodology as described so far is further illustrated in Figure 2.

4. Results and Discussion

Once all the reports (fraud and non-fraud) were loaded into AntConc, frequencies of word types and their rank were plotted (see Figure 3). The corpus follows the natural law observed in all languages and in all corpora: 'systematic frequency distribution such that there are few very high-frequency words that account for most of the tokens in text (e.g. 'a', 'the', 'I', etc.), and many low-frequency words'.[47] This simple pattern is often referred to as 'few giants and many dwarves'.[48] This relationship obeys a power law known as Zipf's law. The rth most frequent word has a frequency f(r) that scales according to formula shown in Eq. 1, r is called the 'frequency rank' of a word, and f(r) is its frequency in a corpus, with $\alpha \approx 1$ (de Gryter, 2009).

Eq.1 $$f(r)\alpha\frac{1}{r^{\alpha}}$$

[46] S. Borgatti (1997). *Multidimensional scaling. Retrieved from* *http://www.analytictech.com/borgatti/mds.htm*.
[47] S. T. Piantadosi, "Zipf's word frequency law in natural language: A critical review and future directions," *Psychonomic Bulletin & Review,* vol. 21, pp. 1112-1130, 2014
[48] Walter de Gruyter, *Corpus Linguistics An International Handbook* vol. 2. Berlin: GmbH and Co, 2009

Next, lemma frequencies were examined. This is where a single item (lemma) is deemed a canonical representative for a set of related (inflected) word forms. For example, variations of the word type 'interest' include 'interested', 'interesting', 'interests'. Their frequencies are all added together and put under the word 'interest'. Table 1, shows the top 20 lemmas in the fraud reports with normalised frequencies. The corresponding frequencies for lemmas in the non-fraud reports are also listed. Figure 4 shows graphically the nature and strength of this relationship for the top 300 lemmas.

On first inspection, there seems to be homogeneity in the words used, and in some cases a similarity in frequency of word usage. As Rutherford argues this stability supports 'the contention that narratives constitute an identifiable genre and implies that where differences do arise, significance can be attached to them' (Rutherford, 2005).

To further check for differences in the mean frequencies of lemmas between the fraud and non-fraud reports, significance testing was performed. A preliminary test for the equality of variances indicates that the variances of the two groups (fraud with non-fraud) were significantly different. Therefore a two-sample t-test was performed that does not assume equal variances.

The hypotheses are as follows:

(Null) Ho: m1 = m2 (means of the fraud and non-fraud reports are equal)

(Alternative) Ha: m1≠m2 (means are not equal)

The mean of the normalised frequencies for all 14066 lemmas in fraud reports were compared with 24441 lemmas in the non-fraud reports. The observed difference (Table 2) is significant (p value < 0.05) and the t stat value is greater than the t critical 2 tailed value. Therefore the null hypothesis can be rejected and the alternative accepted that there is significant difference between how lemmas are used between fraud and non-fraud reports. As noted by Kilgariff: 'any difference in the linguistic character of two corpora will leave its trace in a difference between their word frequency lists'.[49]

[49] A. Kilgarriff, "Using word frequency lists to measure corpus homogeneity and similarity between corpora", *Proceedings 5th ACL workshop on very large corpora*. Beijing and Hong Kong, 1997

The keyword analysis using the non-fraud reports as a reference corpus for comparison revealed the results shown in Figure 5 and 6. The tf-idf score for the top 200 keywords obtained and the matrix constructed was then put through PCA to find features that show the greatest variability between the two categories of reports. These keywords are shown in Figure 7 (top section). The reduced matrix is then put through MDS computation and the results revealed are shown in Figure 7, lower section. It appears that fraud firms are more concerned with bureaucratic issues –'procedures', 'division', 'agreement' and have cash flow issues: 'borrowers', and 'acquisition' (Figure 5 and 6). However once PCA is conducted and the features selected, it seems based on their tf-idf scores the MDS computation indicates a close proximity between the two categories of reports (Figure 9). It appears that for these features, the differences can be quite subtle.

The keyword analysis using keywords unearthed by Rutherford as potential markers of concern with respect to company health were also brought into use. The raw frequencies of all these words outlined by Rutherford were input into the log-likelihood calculator devised by Rayson and Garside and log ratio scores were calculated (Rayson and Garside, 2000). The results are depicted in Figure 8. The log ratio scores for the words above the x axis are more prominent in fraud reports. Conversely the log ratio score that is below the x axis depicts those words more prominent in non-fraud reports.

The fraud firms seem again to be concerned with operations: 'division', 'programme', 'management'. Cash flow issues seem to be coming to the surface again: 'sterling', 'liability', 'asset', 'risk'. Whereas the non-fraud firms use language that seeks to relay details on firm performance, e.g. 'profit', 'growth', 'investment', 'net'. It also shows more confidence by using stronger adjectives such as 'exceptional', 'strong'.

The PCA-selected Rutherford keywords are shown at the top of Figure 9. The tf-idf score for each of these PCA selected Rutherford keywords are then put into a matrix which is then put through MDS. The results shown at the bottom of Figure 9 clearly show that fraud and non-fraud reports can be well separated using terms from the Rutherford study which are then further reduced by PCA to produce the results shown in Figure 9.

The LIWC features chosen by PCA are shown in Figure 10 (top half). Complete descriptions of variables given in Pennebaker et al. (Pennebaker, Boyd, Jordan, and Blackburn, 2015). However it seems that the use of pronouns, adjectives, adverbs, tone, and perceptual and

cognitive processes all contribute to causing variability between the two categories of reports. Once MDS is applied it can be seen that there is a visible distinction in distance that can show up a fraud firm.

The other remaining analysis performed on the corpus was through the use of words denoted as key by Loughran and McDonald (Loughran and McDonald, 2011) in the financial reporting domain. From the log ratios calculated, it can be seen that for the negative words (shown in Figure 11), the underlying themes surrounding cash flow problems has resurfaced. Lemmas associated with 'bankruptcy', 'loss', 'problem', 'shortage', 'fail' are more pronounced. For example, keywords in context reveal statements like: 'The bankruptcy court approved this application', 'reduced gross margins and loss of market share', 'if we fail to cultivate new or maintain existing', 'a result of cash flow shortage'. Whereas the non-fraud firms seem to concentrate more on issues in the external environment with terms such as: 'unfavourable', 'disrupt', 'challenge', 'negative', 'volatile'. For example examination of keywords in context reveal statements like 'extremely unfavourable stock market environment', 'we anticipate could disrupt our business and could result in', 'in a more challenging economic environment', 'economic crises and other challenging market factors', 'demand may be particularly volatile and difficult to predict'.

The positive wordlist produced the results shown in Figure 12. The fraud firms seem again to be concerned with issues surrounding liquidity, for example: 'ability to generate revenues and sustain profitability', 'improve profitability in existing and acquired operations'. The term 'exceed' is used to highlight limitations, e.g. 'operating costs that exceed', 'clinical trials that exceed the capacity of our pilot facility', 'actual costs could significantly exceed these estimates'. In some cases there seems to be some over-optimism. For example: 'Enron has a solid portfolio of asset-based businesses', 'In Q4 (ENRC) 2008 production volumes achieved a solid performance compared to the prior year' (ENRC - Eurasian Natural Resources Corporation). Often the term 'solid' is used with respect to company operations for example: 'New York city solid waste', 'non-hazardous solid waste'. The term 'attractive' is often used with terms that denote acquiring. For example: 'acquiring attractive parcels of land', 'stock may make us a less attractive takeover target' or there is mention of 'attractive assets', 'attractive prices'. Whereas the non-fraud firms use the term 'gain' with a quantifiable result: 'store sales gain of 4.3%', 'unrealized gain of $1.1 million', 'the gain included a pre-tax gain of $570

million'. The term 'strength' is used in a very positive and upbeat manner: 'growth was led by the continued strength', 'we have financial strength', 'far-East and the continued strength of sales'. The term 'improve' is used with reference to products and services, company performance. For example: 'find new customers, improve service', 'help automate and improve a company's business processes', 'maintain and improve manufacturing yields'. The term 'excellent' also used as performance measure: 'excellent profitability', 'Peroni had an excellent year', 'providing excellent customer service'.

The uncertainty wordlist produced the results shown in Figure 13. For fraud firms, the term 'pending' dominates; it is often used in connection with 'pending mergers', 'pending patents', 'pending application', 'pending claims', 'pending litigation', 'pending acquisitions'. The term 'believes' again is used to explain a stance: 'The company/corporation/management/the board believes that…', The term 'rather' is used to highlight an unfavourable alternative for example: 'growth of the company rather than distribute earnings', 'evolutionary change rather than revolutionary disruptions', 'revenues and cash flow for us rather than being sold on a...'. The term 'can' is often used to affirm a point made by the firm: 'cumulative costs can be enormous', 'we can reduce these hidden costs', 'no assurance can be given', 'nor can we predict'. The term 'likely' is used in an attempt to quantify an uncertain outcome: 'values more likely to be eroded', 'income is more likely to', 'common stock likely would decline'.

For the non-fraud firms using the uncertainty wordlists, it can be seen from Figure 13 that the term 'nearly' dominates. This is primarily used to quantify a result: 'We created nearly 1 million', 'customers and nearly 8300 broadband', 'reaching nearly 2.3 million'. The term 'sometimes' is used to refer to a challenging situation: 'numerous and sometimes conflicting', 'sometimes in ways that adversely impact demand', 'sometimes competing industry standards'. The term 'revise' is used in reference to 'revise any forward-looking statements', 'revise procedures', 'revise agreement'. The term 'differ' again is used to highlight differences from expectations: 'actual results could differ from these estimates', 'the costs differ because of higher costs', 'actual results differ from assumptions'.

5. Conclusion

This study has shown that investigating a corpus in a contrastive way can show up patterns of word usage and linguistic style that can alert one to anomalies such as deception.

The approach taken here was to align the investigation along the generally accepted corpus analysis methods. One of these methods is examination of frequency counts. This showed that both fraud and non-fraud firms used similar terminology. However, there is a significant difference in usage as noted by the t-tests conducted. Without such an examination of frequency counts such a distinction would have been difficult to detect.

Another key technique used to examine the corpus were keywords. They are markers of the 'aboutness' (McEnnery and Wilson, 2005) and the style of a text. Keyness was established using AntConc's built in log-likelihood keyword generation method with the non-fraud reports as a reference corpus and through using the alternative method of using a log-likelihood effect size (LL) calculator (Rayson and Garside, 2000). The latter also returned a log ratio score, which showed up more prominently the strength of the difference in keyness between the two types of narratives. Using the former method to determine keyness, liquidity/cash flow concerns could be discerned amongst the fraud firms, with the non-fraud firms giving a more descriptive picture of their operations. This general picture was reinforced by using words unearthed by Rutherford (Rutherford, 2005). The frequencies of these words in the corpus were passed to the LL calculator. The log ratio scores calculated showed clearly the difference in linguistic emphasis. The negative wordlist again brought to fore the concern over liquidity in the fraud reports. The keywords also seem to forewarn of the latent poor company health. The log ratio scores for the positive and uncertainty wordlist again show up the marked difference in emphasis that can alert to anomalies such as fraud and bankruptcy.

The new version of LIWC (2015) was also executed over the reports. In line with previous studies in deception research this showed that some LIWC categories can highlight differences in the narrative style used by fraud/non-fraud firms.

The multivariate nature of the keywords/LIWC variables produced render deception detection a more arduous task. In a bid to determine the features that show the greatest difference between the two types of reports, Principal Component Analysis (PCA) is deployed. The chosen features were then put through Multi-dimensional scaling (MDS) to enable visualisation of the

differences between the two narratives. Again this showed the strength of the differences using keywords/LIWC between the two types of reports.

In a bid to further reinforce the linguistic differences noted, the corpus could be tagged with part of speech. This has been noted to give 'added value'.[50] This would allow a clearer definition of the concepts in the corpus and sharpen any distinctions. A corpus that included more narratives of fraud firms would also strengthen the analysis and the findings. Using wordlists that related to risk and uncertainty could also be used to show up further the latent concerns on productivity/profitability between the types of reports.

However, it seems clear that examination of linguistic styles can be used as an additional armoury by law enforcement agents and auditors to alert to a possible misdemeanour in financial reporting.

[50] G. Leech, "Developing Linguistic Corpora: a Guide to Good Practice Adding Linguistic Annotation", 2004 (web).

Table 1: Top 20 lemmas in fraud reports with corresponding frequency in non-fraud reports

		Normalised Frequencies	
	Lemma	Fraud	Non Fraud
1	company	7.7453	7.4395
2	million	5.8301	7.1380
3	service	5.0957	4.8402
4	product	4.7069	5.0045
5	business	4.5451	4.5467
6	increase	4.3935	5.2396
7	market	4.3745	4.4825
8	result	3.9999	4.0022
9	year	3.9767	3.7335
10	other	3.9166	3.9633
11	may	3.8173	3.2351
12	sale	3.5938	3.6922
13	include	3.5462	3.9395
14	revenue	3.3976	3.3497
15	cost	3.2388	3.4475
16	not	3.0913	2.7420
17	operation	2.7798	3.0798
18	customer	2.6757	3.0049
19	system	2.6168	2.1517
20	financial	2.5592	2.6844

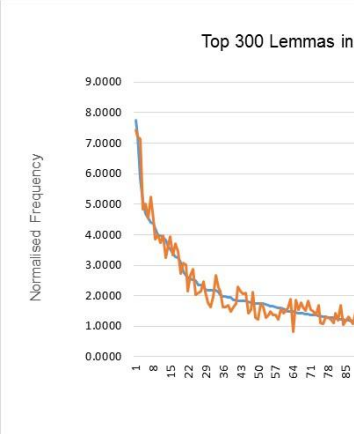

Figure 4: Top 300 Lemmas in fraud and non-fraud reports

t-Test: Two-Sample Assuming Unequal Variances			
		Fraud	*Non-Fraud*
Mean		0.071093417	0.040914856
Variance		0.784882659	0.438103022
Observations		14066	24441
Hypothesized	Mean	0	
df		23176	
t Stat		3.514725315	
P(T<=t) one-tail		0.000220524	
t Critical one-tail		1.644919377	
P(T<=t) two-tail		0.000441049	
t Critical two-tail		1.960066349	

Table 2: Significance testing over lemmas (mean) in fraud and non-fraud reports

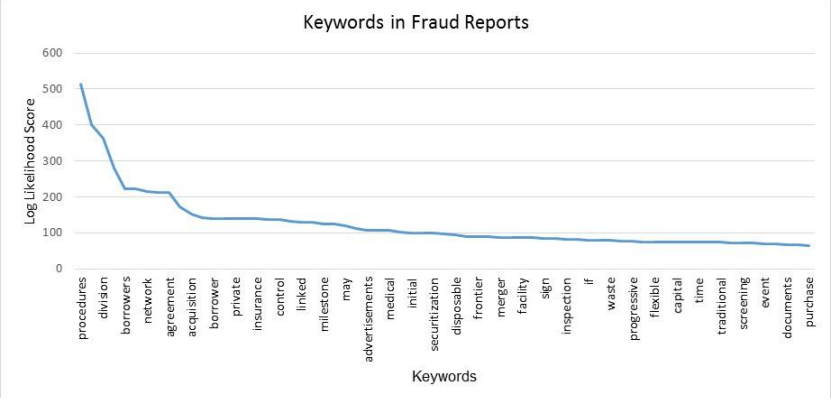

Figure 5: Keywords in fraud reports as identified using log likelihood score in AntConc

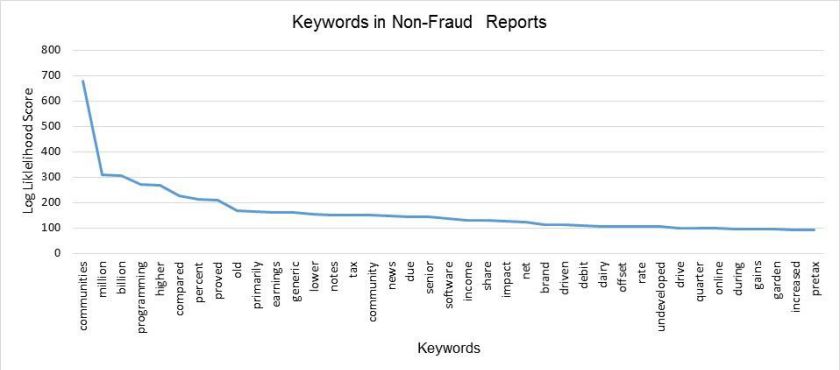

Figure 6: Keywords in non-fraud reports as identified using log likelihood score in AntConc

PCA chosen keywords with MDS applied	
Peer Set: Keywords	'offset', 'rate', 'primarily', 'compared', 'related', 'partially', 'income', 'net', 'due', 'tax', 'lower', 'recorded', 'higher', 'results', 'decreased', 'million', 'losses', 'fair', 'accounting', 'flows', 'quarter', 'currency', 'gains', 'favorable', 'hedges', 'capital', 'fourth', 'unfavorable', 'earnings', 'investments', 'unsecured', 'impact', 'declines', 'pretax', 'pension', 'certain', 'impacted', 'driven', 'gain', 'result', 'notes', 'management', 'adjustments', 'also', 'realized', 'compensation', 'system', 'dollar', 'failures', 'stockholder'
PCA selected keywords, MDS applied	

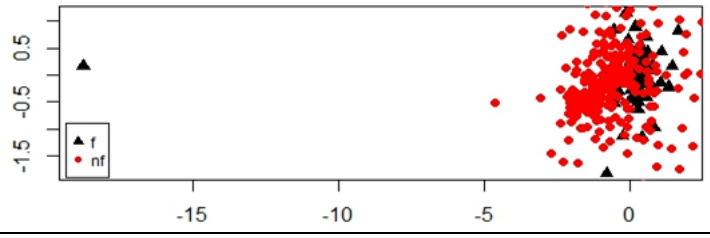

Figure 7: PCA selected keywords

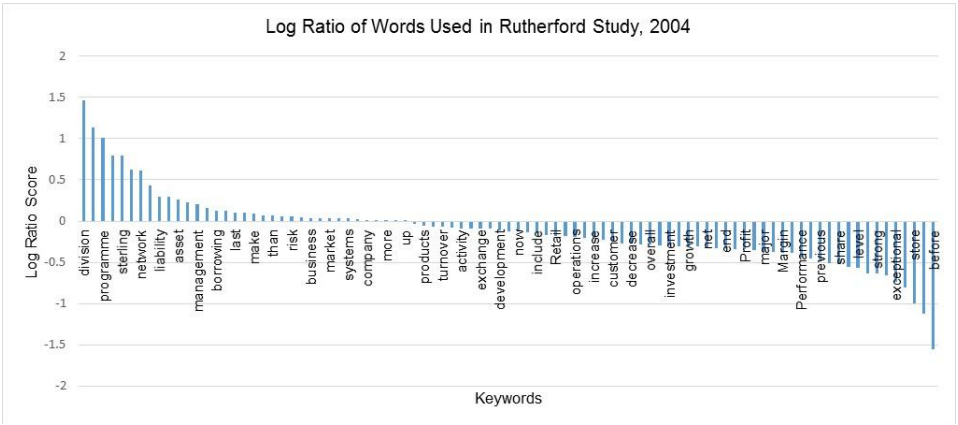

Figure 8: Log ratio scores for keywords used in [199]

PCA chosen Rutherford keywords with MDS applied	
Peer Set: Rutherford	'financial', 'rate', 'capital', 'significant', 'increase', 'interest', 'result', 'cash', 'tax', 'include', 'net', 'years', 'new', 'loss', 'debt', 'management', 'operating', 'operations', 'due', 'company', 'decrease', 'total', 'cost', 'sale', 'store', 'growth', 'lower', 'make', 'revenue', 'higher', 'currency', 'item', 'activity', 'exchange', 'reduce', 'products', 'turnover', 'liability', 'division', 'margin', 'now', 'strong', 'borrowing', 'investment', 'risk', 'services', 'previous', 'property', 'asset', 'trading'

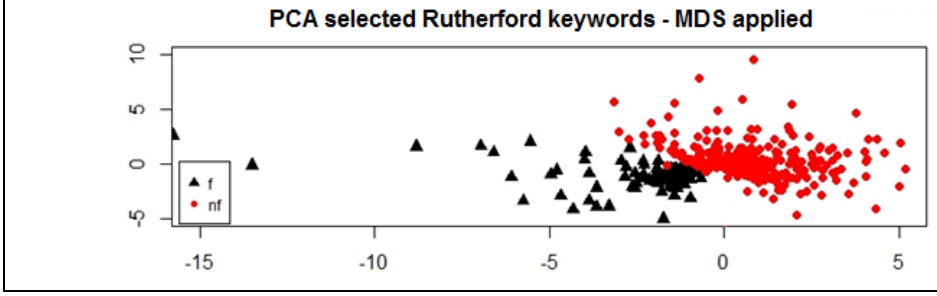

Figure 9: PCA selected Rutherford keywords

PCA chosen LIWC 2015 features with MDS applied	
Peer Set: LIWC 2015	'pronoun', 'Analytic', 'ppron', 'function.', 'drives', 'Clout', 'verb', 'focusfuture', 'auxverb', 'focuspresent', 'cogproc', 'number', 'negate', 'article', 'AllPunc', 'conj', 'relativ', 'WC', 'focuspast', 'interrog', 'quant', 'WPS', 'affect', 'prep', 'Authentic', 'adverb', 'percept', 'compare', 'Tone', 'ipron', 'adj',

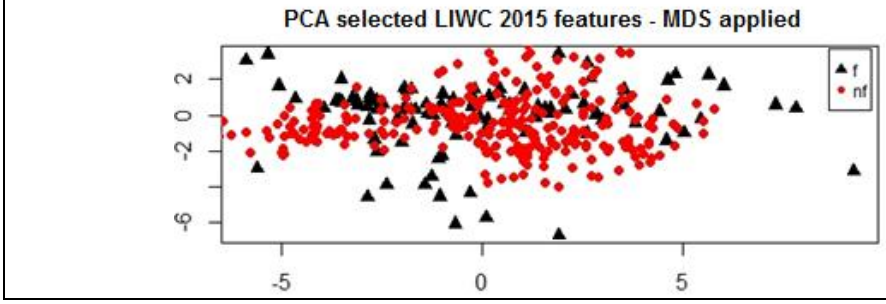

Figure 10: PCA selected LIWC 2015 variables

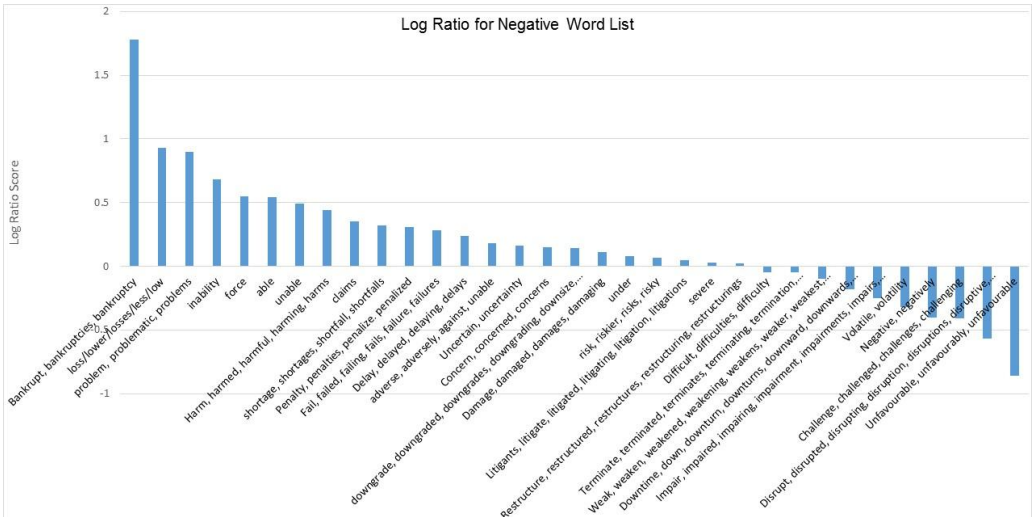

Figure 11: Log ratio scores for negative words from [40]

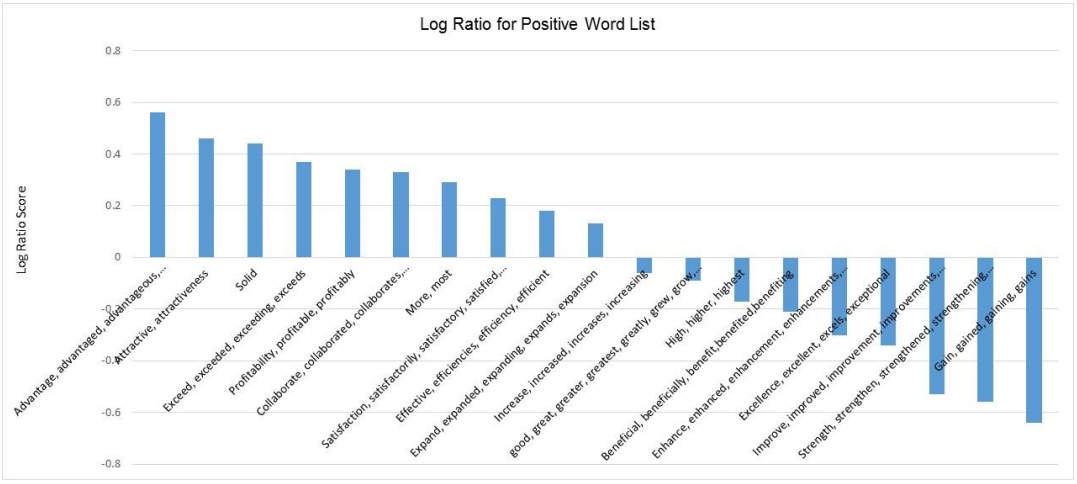

Figure 12: Log ratio scores for positive words from [40]

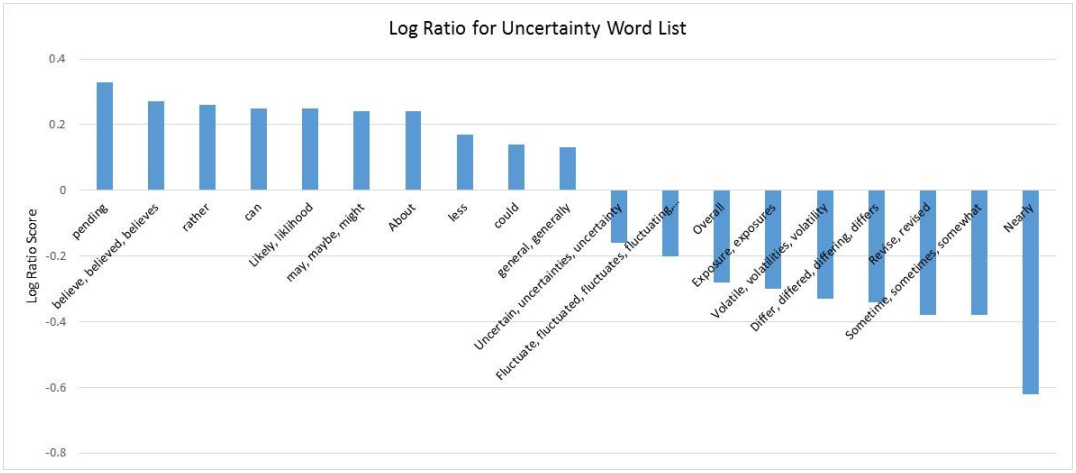

Figure 13: Log ratio scores for uncertainty words from [40]

References

"Corpus Linguistics" Research Starters eNotes.com, Inc. eNotes.com 24 Nov, 2016.

L. Anthony. (2014). *AntConc (Version 3.4.3). Available from* http://www.laurenceanthony.net/

J. Bachenko, E. Fitzpatrick, and M. Schonwetter, "Verification and implementation of language based deception indicators in civil and criminal narratives," presented at the Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1, Manchester, United Kingdom, 2008.

[A. Baron, P. Rayson, and D. Archer, "Word frequency and key word statistics in historical corpus linguistics," *International Journal of English Studies*, vol. 20, pp. 41-67.

S. Borgatti (1997). *Multidimensional scaling. Retrieved from* http://www.analytictech.com/borgatti/mds.htm.

J. Burgoon, W. J. Mayew, J. S. Giboney, A. C. Elkins, K. Moffitt, B. Dorn, "Which Spoken Language Markers Identify Deception in High-Stakes Settings? Evidence From Earnings Conference Calls," *Journal of Language and Social Psychology*, vol. 35, pp. 123-157, 2015.

M. B. Burns and K. C. Moffitt, "Automated deception detection of 911 call transcripts," *Security Informatics*, vol. 3, p. 8, 2014.

B. C. Camiciottoli, *Rhetoric in financial discourse*. The Netherlands: Rodopi, 2013.

M. Cecchini, H. Aytug, G. J. Koehler, and P. Pathak, "Making words work: Using financial text as a predictor of financial events," *Decision Support Systems*, vol. 50, pp. 164-175, 2010.

J. Croft, "Fraud Costs the UK up to £193bn per year report says," in *Financial Times*, ed, 2016.

N. D. Duran, C. Hall, P. M. McCarthy, and D. S. McNamara, "The linguistic correlates of conversational deception: Comparing natural language processing technologies," *Applied Psycholinguistics,* vol. 31, pp. 439-462, 2010.

E. Fitzpatrick and J. Bachenko, "Building a Data Collection for Deception Research," in *Proceedings of the EACL 2012 Workshop on Computational Approaches to Deception Detection,*, Avignon, France, 2012, pp. 31-38.

T. Fornaciari and M. Poesio, "On the use of homogenous sets of subjects in deceptive language analysis," presented at the Proceedings of the Workshop on Computational Approaches to Deception Detection, Avignon, France, 2012.

C. M. Fuller, D. P. Biros, J. Burgoon, and J. Nunamaker, "An Examination and Validation of Linguistic Constructs for Studying High-Stakes Deception," *Group Decision and Negotiation,* vol. 22, pp. 117-134, 2012.

C. Gabrielatos and A. Marchi, “Keyness: Appropriate metrics and practical issues”, CADS International Conference 2012. Corpus-assisted Discourse Studies: More than the sum of Discourse Analysis and computing?, 13-14 September, University of Bologna, Italy, 2012.

F. H. Glancy and S. B. Yadav, "A computational model for financial reporting fraud detection," *Decision Support Systems,* vol. 50, pp. 595-601, 2011.

S. Goel, J. Gangolly, S. R. Faerman, and O. Uzuner, "Can Linguistic Predictors Detect Fraud Financial Filings?," *Journal of Emerging Technologies in Accounting,* vol. 7, pp. 25-46, 2010.

S. T. Gries, “What is corpus linguistics?”, Language and Linguistics Compass vol. 3, pp. 1–17, 2009.

Walter de Gruyter, *Corpus Linguistics An International Handbook* vol. 2. Berlin: GmbH and Co, 2009.

J. T. Hancock, L. E. Curry, S. Goorha, and M. Woodworth, "On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication," *Discourse Processes*, vol. 45, pp. 1-23, 2007.

J. Hardin and G. Sarkis, "Network Analysis with the Enron Email Corpus", *Journal of Statistics Education*, vol 23, No. 2, 2015.

V. Hauch, I. Blandon-Gitlin, J. Masip, and S. L. Sporer, "Are Computers Effective Lie Detectors? A Meta-Analysis of Linguistic Cues to Deception," *Personality and Social Psychology Review*, vol. 19, pp. 307-342, 2014.

S. L. Humpherys, K. C. Moffitt, M. B. Burns, J. K. Burgoon, and W. F. Felix, "Identification of fraud financial statements using linguistic credibility analysis," *Decision Support Systems*, vol. 50, pp. 585-594, 2011.

I. T. Jolliffe, *Principle Component Analysis*. 2nd edition, England, Springer, 2002.

A. Kilgarriff, “Using word frequency lists to measure corpus homogeneity and similarity between corpora”, *Proceedings 5th ACL workshop on very large corpora*. Beijing and Hong Kong, 1997.

D. Krieger, “Corpus Linguistics: What it is and how it can be applied to teaching.” The Internet TESL Journal, 2003.

D. F. Larcker and A. A. Zakolyukina, "Detecting Deceptive Discussions in Conference Calls," *Journal of Accounting Research,* vol. 50, pp. 495-540, 2012.

G. Leech, “Developing Linguistic Corpora: a Guide to Good Practice Adding Linguistic Annotation”, 2004 (web).

T. Loughran and B. McDonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *The Journal of Finance,* vol. 66, pp. 35-65, 2011.

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

P. M. McCarthy, N. D. Duran, and L. M. Booker, "The Devil Is in the Details: New Directions in Deception Analysis," in *Twenty-Fifth International Florida Artificial Intelligence Research Society Conference*, Florida, 2012.

T. McEnery and A. Hardie, *Corpus linguistics: method, theory and practice*. Cambridge: Cambridge University Press, 2012.

T. McEnery and A. Wilson, *Corpus Linguistics, An Introduction*: Edinburgh University Press, 2005.

D. S. McNamara, A. C. Graesser, P. M. McCarthy, and Z. Cai, *Automated Evaluation of Text and Discourse with Coh-Metrix*. New York: Cambridge University Press 2014.

M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying Words: Predicting Deception from Linguistic Styles," *Personality and Social Psychology Bulletin* vol. 29, pp. 665-675, 2003.

E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decision Support Systems,* vol. 50, pp. 559-569, 2011.

P. Omerod, 'Ostrich Economics', 2009.

S. T. Piantadosi, "Zipf's word frequency law in natural language: A critical review and future directions," *Psychonomic Bulletin & Review,* vol. 21, pp. 1112-1130, 2014.

J. Pennebaker, R. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC 2015", University of Texas, Austin 2015.

I. Pollach, "Taming Textual Data: The Contribution of Corpus Linguistics to Computer-Aided Text Analysis," *Organizational Research Methods,* vol. 15, pp. 263-287, 2011.

L. Purda and D. Skillicorn, "Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection," *Contemporary Accounting Research,* vol. 32, pp. 1193-1223, 2015.

P. Rayson and R. Garside, "Comparing corpora using frequency profiling " in *Proceedings of the workshop on Comparing Corpora*, Hong Kong, 2000.

P. Rayson, "Computational Tools and Methods for Corpus Compilation and Analysis," in *Cambridge Handbook of English Corpus Linguistics*, ed, 2015.

B. Rutherford, "Genre Analysis of Corporate Annual Report Narratives: A Corpus Linguistics Based Approach," *Journal of Business Communication,* vol. 42, pp. 349-378, 2005.

J. Sinclair, *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.

J. Stiglitz, *Freefall America, Free Markets and Sinking of the World Economy*, New York: W.W. Norton & Company Inc, 2010.

Y. R. Tausczik and J. W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," *Journal of Language and Social Psychology,* vol. 29, pp. 24-54, 2009.

C. S. Throckmorton, W. J. Mayew, M. Venkatachalam, and L. M. Collins, "Financial fraud detection using vocal, linguistic and financial cues," *Decision Support Systems,* vol. 74, pp. 78-87, 2015.

L. Zhou, J. Burgoon, J. Nunamaker, and D. Twitchell, "Automating Linguistics-Based Cues for Detecting Deception in Text-based Asynchronous Computer Mediated Communication," *Group Decision and Negotiation,* vol. 13, pp. 81-106, 2004.