**Internet Appendix**

Textual Analysis and International Financial Reporting: Large Sample Evidence

Mark H. Lang & Lorien Stice-Lawrence

***Data Sources***

* *Annual Report Data:* Global Reports database from BvD Osiris
* *Accounting, Liquidity, and General Firm-Level Data:* Datastream
* *Analyst Data:* I/B/E/S
* *Institutional Ownership Data:* Thomson Reuter’s International Mutual Fund (TIMF) database

***Annual Report Data Collection and Availability***

The annual reports used in this paper were gathered from the Global Reports database provided within Bureau van Dijk’s Osiris database. We obtained these reports from Osiris’ website using a program called Sikuli which performs automated tasks such as clicking and typing. This program was originally developed to automate and test graphical user interfaces and can be used to automate tasks as varied as typing up an e-mail or playing Angry Birds. Although use of this program greatly decreased the effort involved in obtaining this data, the program operates at roughly the same speed a human would take to perform the same task.[[1]](#footnote-1) Osiris’s database of annual reports covers mostly the glossy copy that is distributed to investors, presumably because this document is widely available on firm websites. The Osiris database also includes regulatory filings in addition to or in place of the annual reports, when those regulatory filings are available.[[2]](#footnote-2)

The Sikuli code that we used to download our annual report sample is available at: <https://marklangresearch.web.unc.edu/download-annual-reports/>. As noted in the comments contained within this code, Sikuli is a screenshot-based coding language, so the script must be customized to match the visual characteristics of the computer that it is running on. Additionally, Osiris no longer includes their Global Reports database in their academic subscriptions, which is why we have not updated our sample period.

***Data Treatment***

We convert the PDF version of the annual reports to text files using the Xpdf and QPDF software programs. The conversion of PDF files to text can sometimes lead to garbled output and makes it difficult to identify tables. To deal with this, we first use the Lingua::EN::Sentence Perl module to break the text of each report into sentences. Then we remove all “sentences” that do not contain at least 50% alphabetic characters, similar to Li (2008) and Miller (2010), and delete sentences where more than 20% of the characters are not alphanumeric (usually because they contain foreign language characters or symbols added by a conversion error). We also exclude lines consisting of fewer than 50 alphabetic characters, for example page headings.

This procedure does not successfully delete all table labels because the PDF conversion process can separate the numbers and labels in tables into separate “sentences.” This adds noise to the fog measure, which is not designed to analyze this type of content, but is appropriate to include in the other textual measures, such as comparability, under the assumption that the text in tables is widely read and includes relevant information for financial statement users.

We use the Lingua::EN::Fathom module to calculate the Fog score, total word count, and document word vectors for the remaining text. We further process these word vectors before using them to calculate comparability by removing all stop words such as “the” and “and” and by stemming the remaining words.[[3]](#footnote-3)

Because of our unique international sample, our documents require an additional level of processing not required when researchers use text files of the regulatory filings of US firms obtained from EDGAR. In order to exclude non-English words or proper nouns, and also to remove any remaining garbled text from the PDF conversion process, we require all of the word stems used in our comparability measures to be present in the Loughran and McDonald business word list (February 2013 version).[[4]](#footnote-4) While this procedure may remove some relevant words, the Loughran & McDonald dictionary is very comprehensive (as it was initially based on a more generally dictionary and includes the addition of business-specific terms) and includes such disparate words as “estrogen,” “barber,” and “genocide.” This additional screen only applies to our comparability measures which use comparisons of individual words, but does not apply to our other measures. Although proper nouns and words in foreign languages are not appropriate to include in comparability measures, we think it would be problematic to exclude them from the other measures because they contribute to total word count and are part of what makes up “foggy” or boilerplate sentences.

The preceding procedures generally do a good job at editing and cleaning English documents but are not able to identify all foreign language documents or all documents with conversion errors. To exclude these, we delete annual reports that have fog scores less than 12 or greater than 30, those that do not have at least 50 sentences, those with fewer than 100 words, those with fewer than 30 unique word stems, and those where the ratio of the total number of words to the number of unique word stems is greater than 60. Next, we delete all observations where the number of words in the final document is above (below) the 99th (1st) percentile within the firm’s country. Lastly, we exclude all firm-years where the firm has an annual report that appears to be an SEC filing (those which include “Form 10-K” or “Form 20-F”, less than 5% of the initial sample) in order to prevent US regulatory language being spuriously linked with determinants or outcomes of disclosure when they are really driven by SEC requirements.

Of the remaining documents, some are duplicates for a given firm-year. This is because Osiris occasionally has multiple documents associated with a single firm-year. For the most part, however, we have duplicates in our data because our download process occasionally required us to restart our Sikuli program, resulting in the same document being downloaded multiple times. We retain only one document per firm-year, giving preference to documents that are longer (most likely to be the annual report itself as opposed to supporting documents) and documents with a later report date (to delete preliminary reports). Osiris also occasionally substitutes the annual report of the parent company for that of a subsidiary when the subsidiary’s annual report is not available, so we identify documents that are exact matches in terms of number of words, number of sentences, and fog, and retain the document only for the firm with the highest level of total assets (most likely to be the parent company). This process deletes 666 potential duplicates.

In order to merge our annual reports with financial data, we match the company name associated with each report with company names in the Osiris and Datastream databases. Most names do not have a perfect match, so we standardize common business terms (such as “Incorporation” to Inc) and delete stop words. For any remaining names we fuzzy match using the Compged command in SAS. We accept matches with a score of 200 or less (for comparison, the score for “Gold Mine Inc” and “Gold Mines Inc” is 100). We hand checked remaining unmatched firms for countries which still had more than five percent of the names unmatched. Most of the remaining unmatched firms were delisted firms or subsidiaries of larger companies.

1. The benefit of a program that mimics the speed of a human is that it does not place an undue burden on the server of the website from which it is extracting data, unlike more efficient programs such as Perl. [↑](#footnote-ref-1)
2. Our Data Treatment section describes how we identify and delete documents that are likely to be US regulatory filings. [↑](#footnote-ref-2)
3. We use the stop words list in the Lingua::Stopwords Perl module. Stemming is the process of removing grammatical endings from words in order to combine words with similar meanings. We stem words using the Lingua::Stem::En Perl module which uses the Porter stemming algorithm. [↑](#footnote-ref-3)
4. For more information on the word list see <http://www3.nd.edu/~mcdonald/Word_Lists.html>. [↑](#footnote-ref-4)