Stylistic Features in Corporate Disclosures and Their Predictive Power

Philipp Heinrich
Chair of Computational Corpus Linguistics
Friedrich-Alexander-University Erlangen-Nuremberg
Bismarckstr. 6, 91054 Erlangen
philipp.heinrich@fau.de

Abstract

We are concerned with the automatic processing of annual reports submitted to the U.S. SEC's EDGAR filing system. The filings consist of structured as well as unstructured information. One part of the filings, the 10-k forms, contains mostly linguistic data segmented into up to 20 items. We briefly describe what steps have to be taken to extract the relevant linguistic information from the unstructured part of the data. We then present results of a first exploratory corpus analysis and provide descriptive statistical figures for our NLP calculations (sentiment, readability, and further stylistic dimensions) for each item of the 10-k form and point out connections between the semantic content of the analyzed items and the quantitative linguistic observables. The linguistic register both varies across items as well as subject to the standard industrial classification of the company. We conclude by applying a dimensionality reduction algorithm (t-SNE) to the linguistic observables and use the embedding for a qualitative comparison with the company's industry.

Keywords: Stylistic Features, Sentiment Analysis, Corporate Disclosures, Natural Language Processing, Corpus Analysis

1. Introduction

US companies are obliged to file annual financial reports to the U.S. Securities and Exchange Commission (SEC). These filings are stored in the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) on the SEC website and are publicly accessible. The filings contain partly structured information, such as financial information in XBRL format, and information about the industry assignment of the filing company following the standard industrial classification (SIC).[1] The SIC can be used to divide companies into coarse-grained categories such as *finance*, and *manufacturing*, or *mining* (cf. Table 2).[2]

On the other hand, a great part of the financial reports, including the actual 10-k form, consists of relatively free text without semantic mark-up (Loughran and Mcdonald, 2014, 1650). Structuring this textual information is necessary in order to automatically process the data for information retrieval purposes and prediction tasks. Our long-term goal is to facilitate interpretation of the text type at hand, which is pre-defined to a certain extend in content, yet not linguistically standardized. One might e. g. observe increasing sentiment polarity in financial reports when companies risk insolvency, even when the mere quantitative (financial) data is whitewashed in a way that business-administrative key figures do not look conspicuous.

10k-forms consist of up to 20 items of text, which concentrate on topics such as business (item 1), risk factors (1A), properties (2), etc.[3] Most research has focussed on item 7 (the management's discussion of the financial condition and the results of operations), where the operations of the reported year are compared to those of the previous year; this section is deemed most relevant for e. g. stock price prediction and thus has received most attention (Loughran and Mcdonald, 2016). In the present study, we only omit the less prevalent items 1B, 4, 6, 9, 9B, and most of Parts III and IV (items 11–16), see section 2.1. and Table 1 for details on our corpus. We will give a short description of relevant items in section 3.1., where we discuss the influence of the items on the quantitative linguistic features.

The present paper focusses on the analysis of *sentiment*, *readability*, and further *stylistic features* across the items published in 10k-forms. Sentiment conveyed in texts has an obvious impact on stock price performance (see e. g. Bollen et al. (2010; Feuerriegel et al. (2015)). Moreover, also readability affects potential investors' decision-making: (1) more readable 10-k filings increase trading volume (Franco et al., 2015), (2) long filings lead to delays in market reactions (You and Zhang, 2009), and (3) both less readable and longer reports are correlated with lower company earnings persistence (Li, 2008), since poorly performing companies need more and more complicated text to rationalize their situation to investors (see also Bloomfield (2008; Bonsall et al. (2017)). The SEC itself provides guidelines on how to submit filings in plain English and what issues to avoid (Securities and Exchange Commission, 1998) in order to make filings more understandable. Further stylistic features as suggested by Biber (1988) try to capture other dimensions such as the density of information provided, see subsection 2.4.

The contribution of the paper at hand is three-fold: We (1) provide a concise walk-through for pre-processing documents retrieved from the EDGAR file system in section 2.1., (2) give descriptive figures of standard NLP calculations (sentiment polarity, readability, stylistic features) across the different items of 10-k forms in section 3.1., and (3) show how these quantitative linguistic figures can be used in a meaningful way (section 3.2.).

2. Methodology

2.1. Pre-processing

Pre-processing is twofold: first, the files retrieved from the EDGAR file system have to be split into the respective doc-

[1] https://www.sec.gov/info/edgar/siccodes.htm

[2] https://www.osha.gov/pls/imis/sic_manual.html

[3] https://www.sec.gov/about/forms/form10-k.pdf

item	1	1A	2	3	5	7	7A	8	9A	10
total	63,577	53,553	21,224	20,827	51,964	57,863	31,069	33,513	57,265	33,885
filtered	56,977	47,953	18,686	18,619	46,450	51,789	27,693	30,009	51,346	30,334

Table 1: Number of observations in the corpus for all items with more than 20,000 admissable observations; total absolute frequencies are given in the first line, frequencies in the category-filtered corpus in the second line (see section 3.1.).

industry	*finance*	*manufacturing*	*mining*	*services*	*TCEGS*	*other*	*unknown*
total	21,914	23,199	5,856	12,233	5,562	6,635	879
t-SNE	66	265	38	67	53	–	–

Table 2: Distribution of industry categories of the filing companies. *TCEGS* is short for *Transportation, Communications, Electric, Gas, and Sanitary Services*. The first row shows the global distribution in the corpus, the second row the category distribution of the filings used as input for t-SNE (see section 3.2.).

uments[4], then the items have to be retrieved from the actual 10-k forms. While there is software for parsing XBRL readily available (e.g. the open source program Arelle[5]), we focus on the linguistic data provided in the 10-k form. We only parse documents submitted in HTML and retrieve all such 10-k forms between January 2006 and December 2015. This yields a total of 76,278 documents.

As mentioned above, the items are not consistently tagged across filings, and are sometimes merged or left out completetly. This and further formatting errors makes the task of item retrieval non-trivial (Loughran and Mcdonald, 2016, 1191f.). We parse the HTML documents using Python's Beautiful Soup[6] and build a corpus of all items. We omit tables and find beginnings and ends of items by a cascade of simple regular expressions (item *x* is usually preceeded by a headline reading *item x* in various formatting). For the final corpus, we only use items with a word count of more than 200 (we call any such item observation *admissable*), hoping to bypass any parsing errors and to ignore short boilerplate texts such as *not applicable*, etc. Table 1 shows the absolute frequency of all admissable items in our corpus.

2.2. Sentiment Polarity

For calculating document-based sentiment polarity (in the form of numerical sentiment scores ranging from –1 to +1), we use a simplified version of the SentiKLUE algorithm (Evert et al., 2014). Furthermore, we use Python's TextBlob library[7], which in turn builds on the pattern module, in order to calculate subjectivity scores. The library uses a dictionary, averaging the words of a document for gaining absolute figures; subjectivity takes values between 0 and 1, with 0 being completely objective and 1 being completely subjective.

[4]The filings retrieved from EDGAR consist of large files segmented into sections for each original file: information about the filing company, the actual 10-k form that we are concentrating on here (which is usually formatted in HTML but can be in other formats (such as PDF) in some circumstances), structurally rich XBRL data, Excel spreadsheets, figures, etc..

[5]`http://arelle.org/`

[6]`https://www.crummy.com/software/BeautifulSoup/`

[7]`https://textblob.readthedocs.io/en/dev/`

2.3. Readability

Although it might be disputable if readability scores yield reasonable results, a range of standard measures have taken root, see e.g. (Si and Callan, 2001). We use the Flesch-Kincaid grade level originally developed for the US army (Kincaid, 1975), which is based on the average number of words per sentence and the average numbers of syllables per word

$$\ell_{\text{fk}} = 0.39 \left(\frac{n_w}{n_s} \right) + 11.8 \left(\frac{n_y}{n_w} \right) - 15.59, \qquad (1)$$

where ℓ_{fk} is the grade level, n_w the number of words of the item, n_s the number of sentences, and n_y the number of syllables.

Furthermore, we calculate the Fog Index (Gunning, 1952), which takes into account the percentage of complex words:

$$\ell_{\text{fog}} = 0.4 \left(\frac{n_w}{n_s} + 100 \cdot \frac{n_c}{n_w} \right) \qquad (2)$$

Here n_c denotes the number of complex words, i. e. words with more than two syllables. It is noteworthy that business texts in general have a high number of "complex" words according to this definition, although these words are easily understood by the usual audience of these texts (Loughran and Mcdonald, 2014). One way of bypassing this problem when calculating the Fog Index is to exclude these words from the calculation. Although the incorporation of specialized word lists[8] is easily implementable, the measure of plain English proposed in (Loughran and Mcdonald, 2014) correlates negatively with all other measures of readability (see (Bonsall et al., 2017, 335)), and the measure proposed in (Bonsall et al., 2017) in turn is based on proprietary word lists.[9]

For reasons of feasibility and since we are interested in comparing readability between items and companies (and not in an interpretation in terms of absolute readability), we restrict our analysis here to the Fog Index and the Flesh-Kincaid grade level, which show a linear correlation to one

[8]`http://www3.nd.edu/~mcdonald/Data/Plain%20English_LoughranMcDonald.txt`

[9]And one might add that file size itself (as proposed in (Loughran and Mcdonald, 2014, 1644)) is not a very good indicator for readability either.

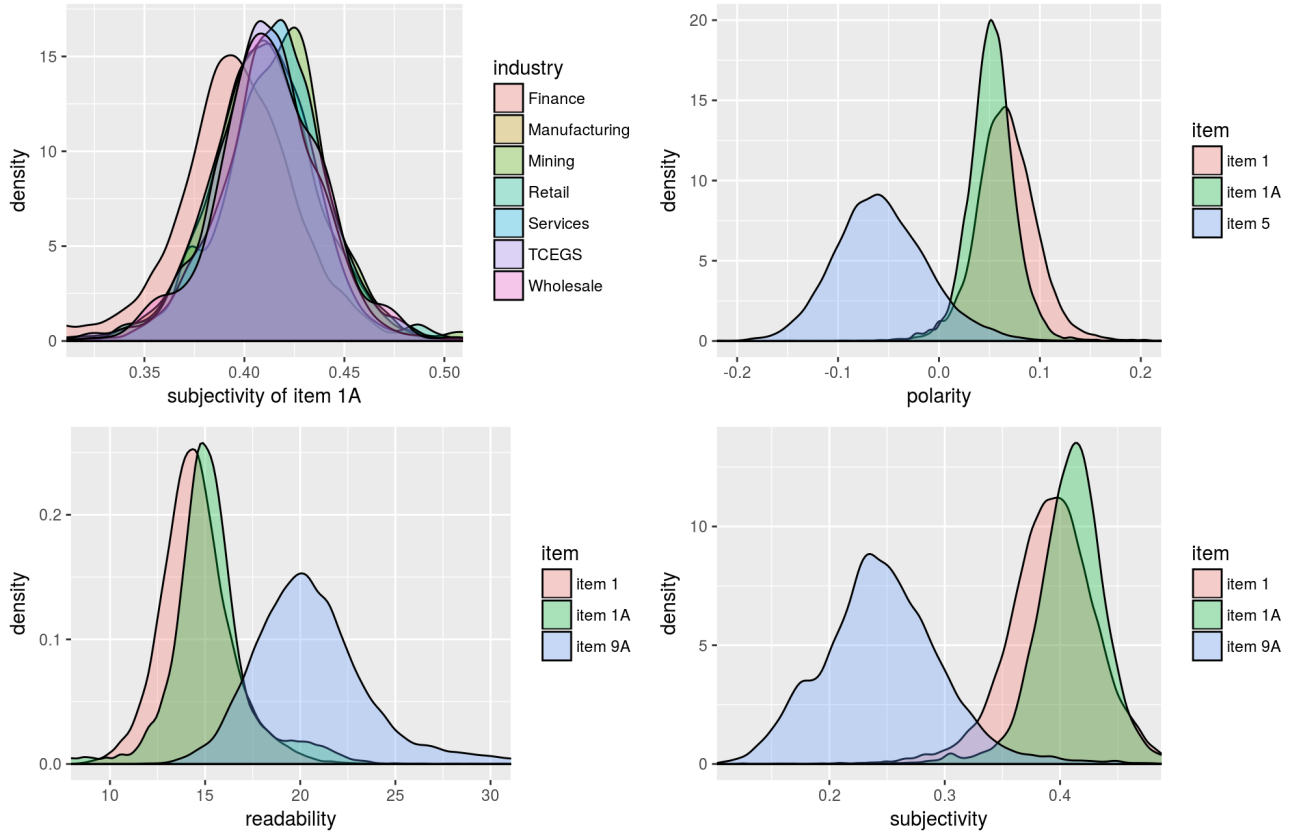

Figure 1: Descriptive figures for selected industries, items, and quantitative linguistic features.

another of more than 99% across our analyses. The original interpretation (both scores used here are supposed to yield an approximate (U.S.) grade level, which can also be transformed into a minimum age required to understand the text at hand) might however not be valid.

2.4. Stylistic Features

A set of stylometric features is generated with the help of the Multidimensional Analysis Tagger (Nini, 2015). It is based on the Stanford Core NLP, extracting 107 features per document (item) as proposed by Biber (1988) – such as the number of verbs, the type-token ratio, the number of passives, and number of used pronouns. These features are then combined into six dimensions reflecting

1. the opposition between involved and informational discours
2. the opposition between narrative and non-narrative concerns
3. the opposition between context-independent context-dependent discourse
4. overt expression of persuasion
5. opposition between abstract and non-abstract information
6. on-line informational elaboration

3. Explorative analysis

3.1. Descriptive figures

There are roughly the same number of documents available for each year (2006–2015) with a peak of 8,796 documents in 2009 and a minimum of 5,955 documents in 2006. The distribution of categories and linguistic observables remains stable over time. For each item we calculate the number of words, number of complex words, readability by means of the the Flesch-Kincaid grade level and the Fog Index (section 2.3.), as well as subjectivity and sentiment polarity scores (section 2.2.) and Biber's stylistic features (section 2.4.) subject to the industry category of the filing company.[10] The distribution of the industry categories can be seen in Table 2. We only list categories with more than 5,000 observed documents and exclude the others from our analyses.[11]

Due to the brevity of the paper at hand we restrict ourselves to two sets of interesting findings with regards to the influence of industry category and the item at hand on sentiment and readability of the reports, see also Figure 1:

1. Industry categories influence ceteris paribus linguistic sentiment and readability. Analyses of Variance (ANOVA) show statistically significant differences in means of subjectivity, polarity, and both readability measures between industry categories ($p < 0.001$). Since we know that the means differ substantially, we can use Tukey's Honestly Significant Test[12], which

[10] The category information is taken from a relatively structured part of the filings.

[11] Category *other* comprises *retail* (3,670 observations), *wholesale* (1,924), *construction* (712), and *agriculture* (329). Companies are classified *unkown* if they are nonclassifiable (288) or if their SIC could not be parsed correctly (591).

[12] Tukey's HSD adjusts the p-values of pairwise t-tests to

shows e. g. that financial companies use the most objective and also the most complex language (for both measures of readability).

2. Readability and sentiment are also very different across items (ANOVAs show again significance). Item 1A (description of risk factors) is written in a very subjective manner compared to the other items; item 9A (conclusions of the company's principal officers) on the other hand is on average written in the most objective manner – however, it is also written in the most complicated manner, with an average Fog Index of approximately 27.3 compared to the second-most complicated item 10 (Fog Index of 19.6). Last but not least, item 5 (explanation of highs and lows of the company's stock and related stockholder matters) is the only item with a negative average polarity score.

Note that there is no uniform scale for the values of Biber's feature dimensions; the numbers are however comparabale to one another across dimensions, cf. Table 3.

dimension	**1**	**2**	**3**	**4**	**5**	**6**
avg. score	-10.5	-4.2	9.0	-4.3	0.7	-1.7

Table 3: Average scores across all items and industries in Biber's six stylistic dimensions.

As for the stylistic features, we summarize our findings as follows:

1. The texts score especially low on dimensions 1, 2, and 4, which means they are – on average – informationally dense, non-narrative, and they do not contain overt expressions of persuasion. On the other hand, they do not depend very much on the context (dimension 3), i. e. they contain many nominalizations and few adverbs. Dimensions 5 and 6 are less salient. With a positive mean in dimension 5, the texts are relatively abstract, a negative score in dimension 6 means that the texts do not contain many post-modifications.

2. Surprisingly, mining companies have on average higher scores in all dimensions, i. e. their texts are less dense in information (mean of -9.5), more narrative (-3.9) and context-dependent (-3.9), and contain more expressions of persuasion (9.5).

3. Item 7A (quantitative information about market risk as of the end of the latest fiscal year) and item 9A score especially low in dimension 1 (means of -16.9 and -14, respectively), which is unsurprising given their very informative character. Moreover, item 3, the legal proceedings, scores noticeably high in dimension 2 (mean of -0.2), making it the item with the most overt expression of persuasion.

3.2. Clustering

Having established the fact that the extraction of the stylistic features yields reasonable results, we now use them as feature vector for a powerful dimensionality reduction technique: t-distributed stoachstic neighbour embedding (t-SNE) transforms high-dimensional data into a lower-dimensional space. The technique uses the Kullback-Leibler divergence (relative entropy) as measure for the faithfulness of the lower-dimensional data. In this method, visualizing map points that are close to one another in the original space far away from one another in the projected space gets a high punishment, whereas "there is only a small cost for using nearby map points to represent widely separated datapoints." (van der Maaten and Hinton, Nov 2008, p. 2581f.) The aim of this experiment is to show that the quantitative linguistic features have predictive power.
We use only documents in which all items are admissable as input for t-SNE, amounting to 489 documents as input. For and each item, we use one of the readability measures (FK), subjectivity, sentiment polarity, and the six stylistic dimensions by Biber. Each document is thus represented by a 90-dimensional vector (10 items times 9 scores), which we transform onto a 2-dimensional plane.
The result of the experiment can be seen in Figure 2. We indicate the industry category of the company of each document by colour in the visualization. We have additionally indicated some of the very obvious clusters by means of circles. Manual inspection of the clusters shows several filings which are very similar on a stylistic level yet not identical. The *manufacturing* cluster slighlty left of the middle of the figure comprises e. g. filings from several companies with strikingly similar phrasings:
KIMBERLY CLARK CORP (CIK 55785) filed a report on 22 February, 2008 with the first item starting as follows:

> *Kimberly-Clark Corporation was incorporated in Delaware in 1928. The Corporation is a global health and hygiene company focused on product innovation and building its personal care, consumer tissue, K-C Professional & Other and health care operations. The Corporation is principally engaged in the manufacturing and marketing of a wide range of health and hygiene products around the world.*

Similarly, *Revett Minerals Inc.* (CIK 1404592) filed a report on March 28, 2012 starting with:

> *Revett Minerals Inc. ("Revett Minerals") was incorporated under the Canada Business Corporations Act in August 2004 to acquire Revett Silver Company ("Revett Silver"), a Montana corporation, and undertake a public offering of its common stock in Canada, transactions that were completed in February 2005.*

And yet another company (*EASTERN CO*, CIK 31107) filed a report on March 13, 2015:

> *The Eastern Company (the "Company") was incorporated under the laws of the State of Connecticut in October, 1912, succeeding a copartnership established in October, 1858. The business of the Company is the manufacture and sale of industrial hardware, security products*

counter the increasing risk of Type I error when performing several comparisons.

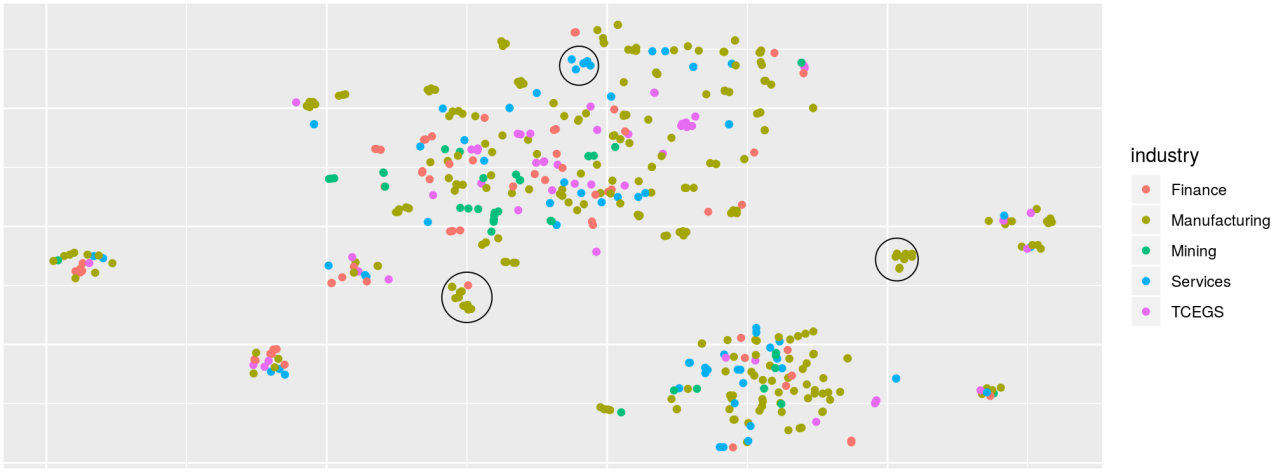

Figure 2: t-SNE projection of quantitative linguistic features (sentiment, polarity, readability, and the Biber features) of 489 documents onto a 2-dimensional plane. The categories of the filing companies are highlighted by colour, and several meaningful clusters can be identified.

> *and metal products from six U.S. operations and seven wholly-owned foreign subsidiaries. The Company maintains thirteen physical locations.*

The resemblance of the paragraphs on a stylistic level is striking. We note that the quantitative linguistic features that we calculated are very superficial, which makes the results of the visual clustering even more interesting. This manual inspection also shows that we did not just detect near-duplicates, but that the filings are similar in style while substantially different in content.

4. Conclusion

The present paper analyzed annual reports submitted to the SEC's EDGAR file system by means item extraction and calculation of standard NLP measures for sentiment polarity, readability, and Biber's genre-specific linguistic dimensions. The work provides proof of concept for the presented observables: polarity scores are high in absolute terms for items concerned with reasons for highs and lows of the company's stock price; the items concerned with general business descriptions are easier to read than disclosures by the financial officer; legal proceedings contain more persuasive text; etc.

The clustering is more than a curious peculiarity: it shows that regularities of quantitative linguistic features can be found on a macroscopic level, which is promising for large-scale event and stock price prediction. In future work, we will therefore (1) include further quantifiable observables such as linguistic uncertainty, and (2) link the quantitative linguistic data with quantitative financial data gained from XBRL and external databases.

We report work in progress; recovery of the items from unstructured data is still not perfect. However, in the long run, we want to contribute to the interested research community by providing a parsed corpus of annual reports in order to facilitate reproducibility. Since the transfer of actual textual data might in fact be a problem due to possible copyright infringement, we will provide our corpus (pre-)processing tools including the items extractor once we have manually checked a larger part of the corpus.

5. Bibliographical References

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press, Cambridge.

Bloomfield, R. (2008). Discussion of "Annual report readability, current earnings, and earnings persistence". *Journal of Accounting and Economics*, 45(2):248 – 252.

Bollen, J., Mao, H., and Zeng, X.-J. (2010). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, October.

Bonsall, S. B., Leone, A. J., Miller, B. P., and Rennekamp, K. (2017). A plain English measure of financial reporting readability. *Journal of Accounting and Economics*, 63(2-3):329–357, April.

Evert, S., Proisl, T., Greiner, P., and Kabashi, B. (2014). SentiKLUE: Updating a polarity classifier in 48 hours. In Preslav Nakov et al., editors, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 551–555, Dublin. Association for Computational Linguistics.

Feuerriegel, S., Ratku, A., and Neumann, D. (2015). Which News Disclosures Matter? News Reception Compared Across Topics Extracted from the Latent Dirichlet Allocation. *News Reception Compared Across Topics Extracted from the Latent Dirichlet Allocation (February 13, 2015)*.

Franco, G., Hope, O.-K., Vyas, D., and Zhou, Y. (2015). Analyst Report Readability. *Contemporary Accounting Research*, 32(1):76–104.

Gunning, R. (1952). *The Technique of Clear Writing. McGraw-Hill International Book Co.* McGraw-Hill International Book Co., New York, NY.

Kincaid, J. (1975). *Derivation of New Readability Formulas: (automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Research Branch report. Chief of Naval Technical Training, Naval Air Station Memphis.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2-3):221–247, August.

Loughran, T. and Mcdonald, B. (2014). Measuring Readability in Financial Disclosures. *The Journal of Finance*, 69(4):1643–1671.

Loughran, T. and Mcdonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, 54(4):1187–1230, September.

Nini, A. (2015). Multidimensional Analysis Tagger (version 1.3).

Securities and Exchange Commission. (1998). *A Plain English Handbook*. SEC Office of Investor Education and Assistance, Washington, District of Columbia.

Si, L. and Callan, J. (2001). A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 574–576, New York, NY, USA. ACM.

van der Maaten, L. and Hinton, G. (Nov 2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

You, H. and Zhang, X.-j. (2009). Financial reporting complexity and investor underreaction to 10-K information. *Review of Accounting Studies*, 14(4):559–586, December.