Diachronic Lexical Changes In Company Reports: An Initial Investigation

Matthew Purver,* **Aljoša Valentinčič,**† **Marko Pahor,**† **Senja Pollak**‡
*School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
†Faculty of Economics, University of Ljubljana, Ljubljana, Slovenia
‡Jožef Stefan Institute, Department of Knowledge Technologies, Ljubljana, Slovenia
m.purver@qmul.ac.uk, {aljosa.valentincic,marko.pahor}@ef.uni-lj.si, senja.pollak@ijs.si

Abstract

We present initial investigations for a diachronic study of lexical changes in financial reporting, looking at methods suitable for analysing semantic associations between financial terms and how these change across time. Our corpus consists of US 10-K annual reports of 30 companies included in the Dow Jones Industrial Average stock index over the years 1996-2015. We grouped the reports by the reported fiscal year and derived word embedding models for each year using both GloVe and a count-based PPMI method; these vectors were then used to calculate cosine similarity between pairs of words. We expect the resulting diachronic patterns of lexical contexts of financial terms to vary with the economic cycle; here we select pairs of terms with strong increasing association over time (e.g. *dividend* and *shareholder*) or strong decreasing association over time (e.g. *dividend* and *gain*), and suggest some qualitative explanations for these changes due to the economic crisis.

Keywords: lexical changes, word embeddings, distributional semantics, financial reporting, 10-K

1. Introduction

The main goal of reporting in the financial system is to ensure that high-quality, useful information about the financial position of firms, their performance and changes in their financial position is available (IASB Framework 2015) to a wide range of users, including existing and potential investors, financial institutions, employees, the government, etc. The central element of the formal system of financial reporting is accounting standards. Common accounting standards increase transparency and comparability of the information that firms communicate to users (investors). Transparency decreases uncertainty about the future prospects of the firm, and the information asymmetry between a firm and external stakeholders; and better understanding of the reporting process assures higher transparency. Here, as part of an ongoing project (FORMICA, 2017), we propose a study of diachronic lexical changes in annual reports; by examining how key terms are used and how this usage changes over time, we hope to gain more insight into how language used in reporting reflects and is affected by the financial cycle.

We collected a corpus of 10-K forms from 30 companies from the Dow Jones Industrial Average (DJIA) index for the period from 1996 to 2015, thus including the period c.2007-8 of the most severe economic and financial crisis since the Great Depression in the 1930s. Here, our initial study focuses on developing suitable methods to automatically characterise word usage and meaning, and track changes over time. We focus initially on a small set of words expected to vary with the economic cycle, and apply methods from distributional semantics, checking the suitability of these methods by deriving year-specific word embeddings and examining diachronic changes in the lexical associations they represent, investigating pairs of terms with strong increases or decreases in association over time.

The paper is structured as follows. After the related work in Section 2, we describe our corpus of annual reports in Section 3. Section 4 presents the selected financial terms for this initial study, and Section 5 the methodology used to discover diachronic changes. After the discussion of results and some tentative qualitative explanations in Section 6), we conclude and present ideas for future work in Section 7.

2. Related Work

2.1. Analysing financial reports

Formal reports contain both strictly regulated, financial sections and unregulated, narrative parts. While the financial aspects have seen a large amount of academic research, studies on narrative parts are relatively scarce. Non-financial information from reports has been used for prediction of financially relevant events (Qiu et al., 2006), such as next year performance (through indicators such as return on equity) (Qiu et al., 2006; Butler and Kešelj, 2009; Kogan et al., 2009; Balakrishnan et al., 2010; Hájek and Olej, 2013; Leung et al., 2015), contemporaneous returns around filing dates (Feldman et al., 2008), stock return volatility (Li, 2010; Loughran and McDonald, 2011), earnings forecast dispersion (Kothari et al., 2009; Loughran and McDonald, 2011), costs of capital (Kothari et al., 2009), financial distress (Hajek et al., 2014), credibility of reports (Athanasakou and Hussainey, 2014) or fraud detection (Goel and Uzuner, 2016).

2.2. Linguistic analysis

Several studies have explored more linguistic aspects, often using a corpus linguistics analysis approach. For example, genre analysis of corporate annual report narratives (U.K. Operating and Financial Review) is proposed by Rutherford (2005), where the authors pay special attention to the "Pollyanna effect" (language biased towards the positive terms). Aiezza (2015) studied the use of verbal markers of forward-looking statements and their contribution to the creation of an ethical image in corporate social responsibility (CSR) reports. Impression management in chairman's statements has been analyzed by Merkl-Davies et al. (2011) and a corpus analysis of stance expression is proposed in Fuoli (2017). Very relevant for our work are analyses with a diachronic aspect, including a diachronic analysis of persuasive language in earnings calls (Camiciottoli, 2017).

2.3. Distributional semantics

One of the most visible trends in the field of natural language processing in recent years is the use of distributed lexical representations in the form of vectors or *word embeddings* learned from observed distributions in raw text. The vectors may be derived directly from observed co-occurrence probabilities, or learned (usually with neural networks) to capture this information implicitly; see e.g. (Baroni et al., 2014; Clark, 2015) for overview and comparison of methods. These representations capture many aspects of word meaning (Firth, 1959), including not only judgements of semantic similarity and relatedness but higher-level regularities including limited kinds of analogy (Mikolov et al., 2013). Word embeddings have been also applied to analyze diachronic semantic changes. For example, Hamilton et al. (2016) use neural network-based embeddings to detect shifts in meaning of words in the Google Books corpus, while Kenter et al. (2015) use similar methods for monitoring shifts in vocabulary over time.

3. Corpus of Annual Reports

We focus on companies from the Dow Jones Industrial Average 30 (DJIA) and use their annual (10-K) reports (FORMICA, 2017). The reports cover the period from 1996 to 2015, but the entire period is not covered for all the companies (depending on the availability of the reports in the EDGAR database). We do not consider the amendments (forms of type 10-K/A and 10-K405/A).

Formal reports contain both strictly regulated, financial sections, and less regulated, narrative parts. In our work we focus on the latter, as our interest is in changes in language used in the reporting process, and therefore extract from the 10-K reports only Part I and Items 7 and 7A from the Part II. For example, Item 7 (Management's Discussion and Analysis (MD&A)), discloses company operations and management in a way that is easy for investors and other interested parties to understand and includes information on what the company does in the face of risks, legislation, competition

For extraction of the selected parts, and cleaning of the dataset, we follow Smailović et al. (2017). In short, the desired document parts are detected by searching for the titles of the sections (e.g., Part I), but taking care that the references to these parts are not considered as titles; we also skip potential .pdf, .xls, .jpg, .zip, .gif objects and tables, and remove html/xml tags to leave plain text (see Smailović et al. (2017) for full details). In total, the dataset contains 528 annual reports, as it can be seen from Table 1.

4. Financial Terms

In this initial methodological investigation for our diachronic study, we manually defined a set of financial terms for examination, rather than attempting to extract them automatically (e.g. on the basis of term relevance or change) so as to avoid domain- or sector-specific terms. Some were very general (*'risk'*, *'profit'*, *'loss'*, *'cash'*); some more specific (*'impairment'*, *'dividend'*, *'repurchase'*, *'residual'*, *'capitalization'*, *'development'*, *'expenditure'*, *'discount'*), and selected as expected to vary with the economic cycle. Our dataset covers arguably the most severe period of economic and financial crisis since the 1930s Great Depression. During this period, past investment mistakes on the part of firms had to be recognized in financial statements, via an accounting procedure called asset *impairment*. Firms must compare the values at which their investments are recorded in statements of financial position (balance sheet) with the value in use and the replacement value; during the crisis, these comparisons result in reporting bottom-line losses. While the procedure is highly discretionary — managers may exploit the resulting write-offs for benefits other than shareholder value maximization — research shows that the signal is viewed as credible by the market in general (Riedl, 2004). Even in empirical environments where the discretionary component may be large, write-offs still indicate declining future performance (Kosi and Valentincic, 2013).

Firms — even the good ones — consequently struggle with signalling their true state to the market. Those in good shape either may return cash to their shareholders, by increasing *dividends* (the financial sector tended to do this during the period in question) or by increasing share *repurchases*. Early research suggested that management prefers to keep dividends stable over time, believing that investors prefer stable dividends and themselves preferring to formulate dividend policies as a fixed percentage of net income (Lintner, 1956). There is a clear asymmetry in dividend increases and dividend decreases: Allen and Michaely (2003) report that only about 5% of dividend changes over 30 years were decreases. From this, various studies report evidence consistent with dividend signalling, e.g. Michaely et al. (1995) report that dividend initiations result in a 3.4% increase in share price, while dividend omissions result in a 7% decrease (note that dividend displacement theory predicts a one to one relation in the same direction (Rees and Valentincic, 2013)). Similar findings have been reported for various other settings, for example large dividend increases and large dividend cuts (Grullon et al., 2002). Share repurchases are also a form of payout and may also be used as signals (Brav et al., 2005). Typically, a firm might repurchase shares on the open market when the managers see the share as undervalued (Brav et al., 2005). However, the commitment to repurchases is less firm. Firms request shareholder permission for the maximum amount they intend to repurchase, but do not then necessarily use the full amount (see e.g. (Berk and DeMarzo, 2014), pp. 610-611). Over time, the prevalence of dividends has been declining both in frequency and in amount, while the importance of repurchases has increased. Before the financial crisis, repurchases for US industrials represented two thirds of firms' aggregate payout to shareholders (Floyd et al., 2015). However, this was reversed during the financial crisis and the importance of dividends has increased again.

5. Method

We divided the corpus into collections for each year, taking the stated fiscal year end in each 10-K report as the year of note; this resulted in the frequencies shown in Table 1. We then used a neural network-based method, GloVe (Pennington et al., 2014), to learn word embedding vectors. We performed simple sentence segmentation based on sentence-final punctuation (`.`/`!`/`?`), and tokenised into words on white space and any non-alphanumeric characters (including remaining punctuation). We used NLTK's WordNet-based lemmatiser to reduce all nouns to their singular version (our selected terms of likely interest were all nouns in this study – other parts of speech were left unchanged). All text was normalised into lower case, embeddings used 100 dimensions, and we trained the models for 50 epochs using a learning rate of 0.05.[1] As a comparison point (see below) we also built a count-based vector space based on positive pointwise mutual information (PPMI), following e.g. (Milajevs et al., 2014), using the 2,000 most common words as the vector dimensions. For both spaces, we experimented with a range of co-occurrence context window sizes of 5, 10 and 20 words; previous work has found that this can affect what is captured by word vector relations (with narrower windows sometimes more likely to capture semantic *similarity* while wider ones reflect semantic *relatedness* (Agirre et al., 2009; Turney et al., 2010)) although this seems dependent on corpus and corpus size (Kiela and Clark, 2014).

Year	N	Year	N	Year	N	Year	N
1996	12	2001	28	2006	29	2011	30
1997	17	2002	29	2007	29	2012	30
1998	24	2003	29	2008	30	2013	30
1999	25	2004	29	2009	30	2014	31
2000	27	2005	30	2010	30	2015	9

Table 1: Document-year counts

We learned word vectors for each year independently; note that this means that vectors cannot be compared directly between years (as the latent dimensions of a GloVe vector space are arbitrary). In future work, we plan to learn transformations to align the vector spaces between years, thus allowing direct comparison, following e.g. (Hamilton et al., 2016). Here, we examine only the similarity between pairs of vectors as measured by cosine distance: this can be compared between years, as GloVe learns vectors whose dot-products correspond to ratios of empirically observed co-occurrence probabilities, and the normalisation in the cosine distance calculation accounts for effects of overall word frequency changes. We confirm our observations by comparing with results from the count-based PPMI vector space; although sparser and harder to interpret without further smoothing, this space is directly comparable between years as dimensions are consistent (being derived from co-occurrence counts with a fixed set of context words).

Our initial method is now to look for diachronic changes in similarity (or association) between words that have a high degree of positive association at some point in time. (Searching for apparent changes in associations with consistently *low* absolute values is of course subject to issues of noise and estimation error, and is harder to interpret intuitively; examining changes in associations with high *negative* values – i.e. dissociations or dissimilarities – is potentially useful and may be investigated in future). Given a candidate word w for investigation (see next section), a lexical neighbourhood L can now be discovered, defined as the set of words which appear in $N_y^{10}(w)$, the set of 10 nearest neighbours of w in any year y:

$$L = \bigcup_{y \in Y} \{w' | w' \in N_y^{10}(w)\}$$

We can now examine changes in similarity S between w and members of L over time, by examining changes in the dot-product (or its length-normalised equivalent, cosine distance) between the vectors $\overline{w}$ and $\overline{w'}$ for any $w' \in L$.

$$S = \frac{\overline{w}.\overline{w'}}{|\overline{w}| \times |\overline{w'}|}$$

6. Results and Discussion

Figure 1 shows an example of the diachronic patterns that can be observed, here for two of our candidate words *'dividend'* and *'repurchase'*. Remembering that the similarity measure here is cosine distance between word vectors, with those vectors derived from observed co-occurrence patterns within 10-word windows, we can interpret these patterns as telling us about words which become more (or less) strongly associated with each other over time.

Changes in context window size make some difference to the measured associations between lexical items, but little difference to the patterns of change in associations. As Figure 1(a-c) show, the 10- and 20-word windows show very similar results, both in terms of level of similarity and pattern of changes over time; the smallest 5-word window diverges from the other two slightly, but shows a similar pattern. As the narrower window is more likely to suffer from data sparsity in this relatively small corpus, we use the wider windows hereafter. Comparing the results using the GloVe method (Figure 1(a-c)) with the equivalent using explicitly co-occurrence-based PPMI where the vector dimensions are fixed across years (Figure 1(d)) again shows similar patterns and magnitude of change over time, but with more noise (probabilities used in the PPMI calculation were not smoothed). We therefore take this as our general method for examining the similarity (or lack thereof) in the usages of words over time.

This is of particular interest in this case, as the words *'dividend'* and *'repurchase'* refer to alternative ways in which firms can distribute profits. Dividends tend to be "fixed" – not necessarily by amount but by a fixed percentage of growth, fixed percentage of profits, or declared to be a residual after investment has been taken care of. Changing this policy can therefore send a strong signal to investors, and is therefore often strenuously avoided. Repurchases, on the other hand, are more flexible — shareholders are not forced to give up their shares in exchange for cash, but only if they wish to do so — and this "un-fixedness" can

[1] Trials with 40 and 60 epochs show similar results; a more comprehensive test will be carried out in future.

make repurchases popular with companies as changing the amounts repurchased does not tend to send strong signals. The increase in association between these words over time is statistically significant (Spearman's R shown in Figure 1, $p < 0.05$ in all cases), and suggests that there is an increasing tendency for firms to use these words in similar ways (i.e. in similar lexical contexts) when reporting. Note that simple direct measures of association do not reveal these patterns: PMI between the two words directly (measured via co-occurrence in the same 20-word context window) shows no significant correlation over time – see Figure 4.

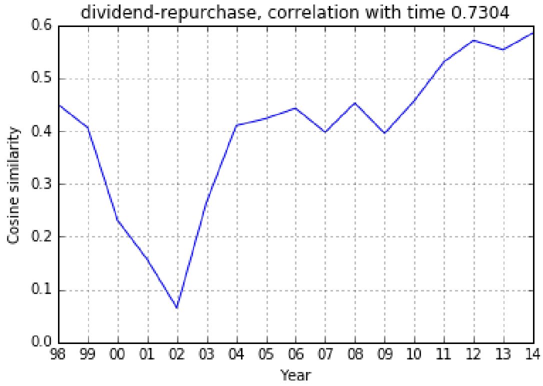
As Figures 2 and 3 show, we can use this method to look for the major diachronic changes across the lexical neighbourhood L more generally, by searching for the words in L whose similarities show large changes over time (here, we used Spearman's R to find the highest correlations with the year ordering). Each figure shows the top 6 positive correlations over time (increases in similarity) and top 4 negative correlations (decreases in similarity) with one of our words of interest; in Figure 2, these are changes in the similarity with the word *'dividend'*; in Figure 3, with *'impairment'*.

Inspecting Figure 2 (*'dividend'*), we can perhaps offer some tentative qualitative explanations. For (a),(b) and (f) (*'quarterly'*, *'shareholder'*, *'paying'*): before the financial crisis, we might expect the association between the two terms to be low, as dividends tended to be replaced by share repurchases (both in frequency and amount). During and after the financial crisis, this association increases significantly. This is possibly due to companies trying to signal to shareholders that their current and expected future profits are sound and can be thus distributed. The same most likely applies to (c) (*'declared'*), although it is unclear why. Dividends have always needed to be declared first with the on-record date, ex-dividend date and payment date (or interval) defined. The positive correlation is possibly due to dividends becoming more prominent in this period and hence the term *'declared'* becoming more frequent as a result.

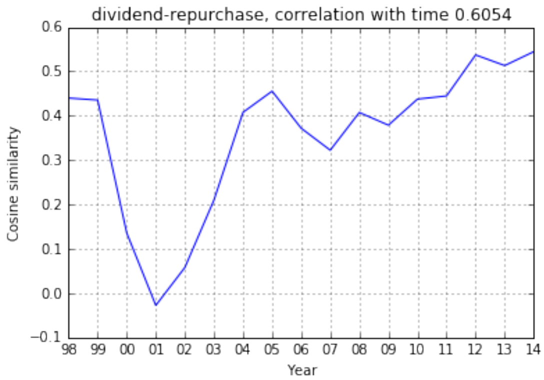
For some cases e.g. (e) (*'production'*) we offer no explanation. Whether production (of physical goods) has actually increased or decreased relative to providing services, or whether *'production'* in this case refers to other concepts will need to be investigated further. Similarly we currently have no insights into the pattern in (i) (*'conversion'*).

For (h) (*'gain'*) and possibly (g) (*'impact'*), as the importance of dividends increased during this period relatively to trends in previous periods (see (Floyd et al., 2015)), the reverse holds for capital gains. If a firm pays out a relatively high proportion of profits as dividends, then share prices will not increase as much as if a firm pays a low proportion of profits in the form of dividends. Hence, the higher the proportion of total return a shareholder receives in the form of dividends, the smaller the proportion of total return in the form of capital gain (in relative terms). Hence the decreasing correlation through time between dividends and capital gains. As dividends were discussed more, gains less.

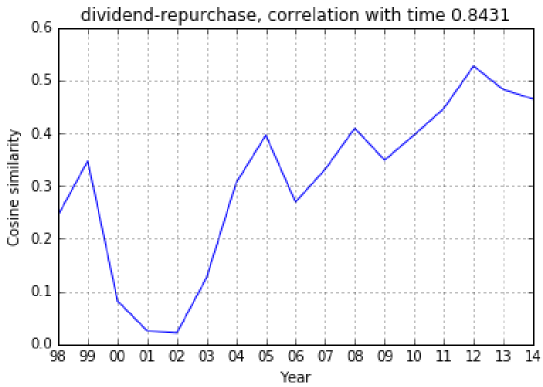
Inspecting Figure 3 (*'impairment'*), we can suggest similar explanatory background. For (a) (*'recognize'*) and (b) (*'testing'*), in the period under study, the first break was the Enron scandal and consequent introduction of the Sarbanes-Oxley act. This brought about an increase in con-



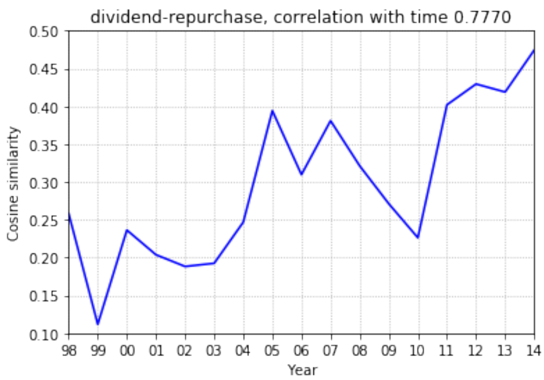
(a) GloVe, context window 20 words



(b) GloVe, context window 10 words

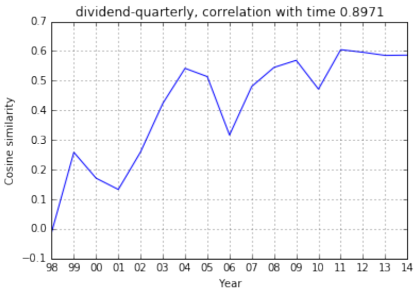


(c) GloVe, context window 5 words

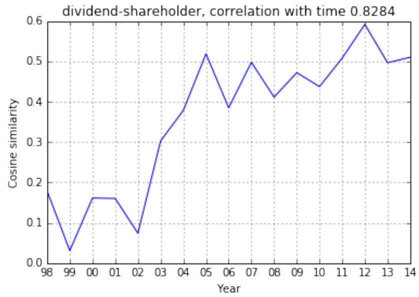


(d) PPMI, context window 20 words

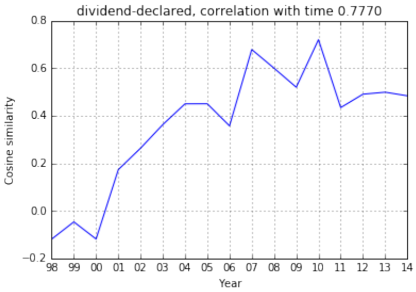
Figure 1: Cosine similarities for the word pair *'dividend'* vs *'repurchase'* over time, using a range of methods and lexical co-occurrence context window sizes.



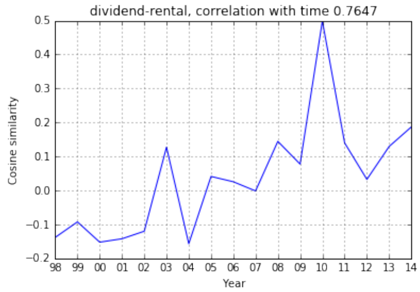
(a) ‘dividend’ vs. ‘quarterly’



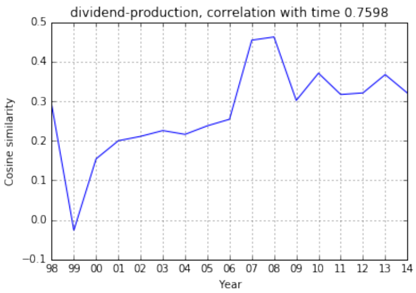
(b) ‘dividend’ vs. ‘shareholder’



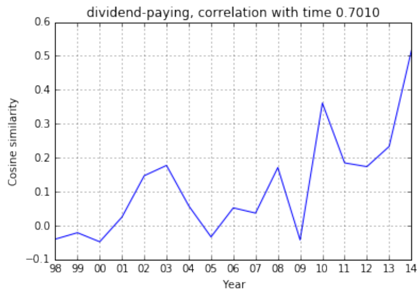
(c) ‘dividend’ vs. ‘declared’



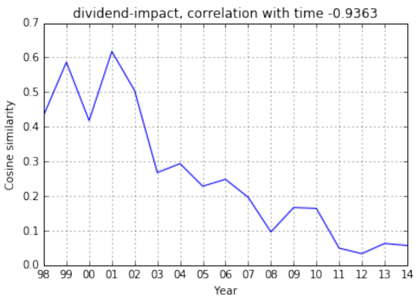
(d) ‘dividend’ vs. ‘rental’



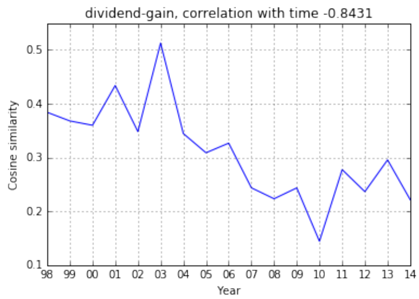
(e) ‘dividend’ vs. ‘production’



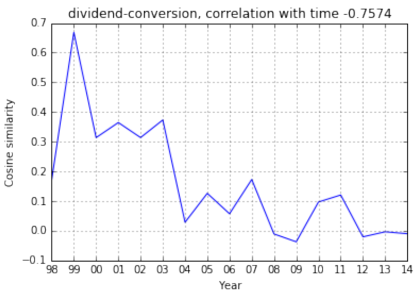
(f) ‘dividend’ vs. ‘paying’



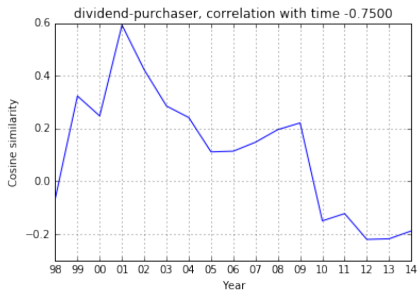
(g) ‘dividend’ vs. ‘impact’



(h) ‘dividend’ vs. ‘gain’

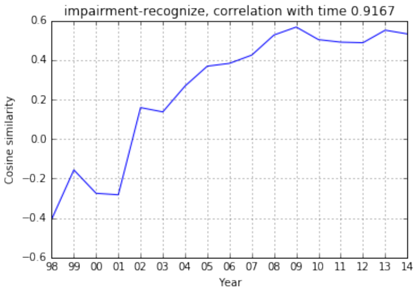


(i) ‘dividend’ vs. ‘conversion’

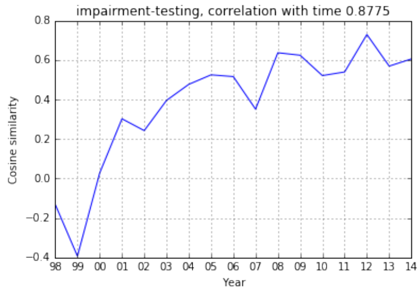


(j) ‘dividend’ vs. ‘purchaser’

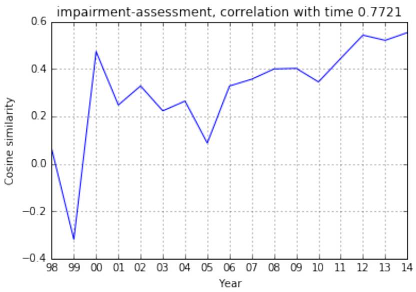
Figure 2: Cosine similarities for word pairs with highest positive and negative correlations over time, for the lexical neighbourhood of ‘dividend’.



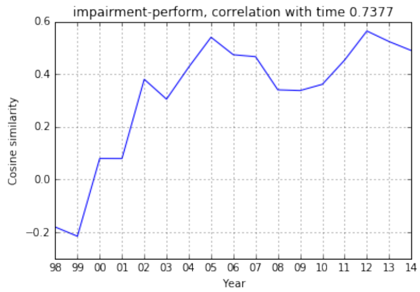
(a) 'impairment' vs. 'recognize'



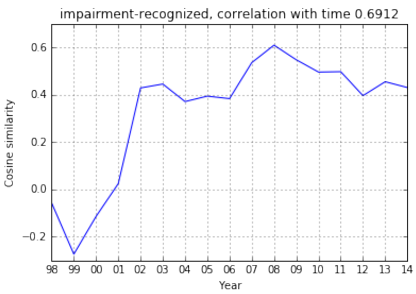
(b) 'impairment' vs. 'testing'



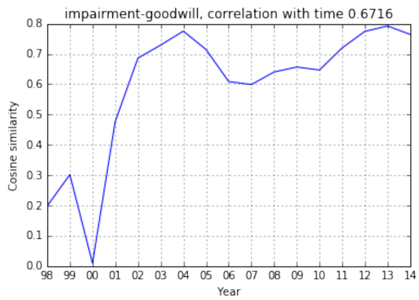
(c) 'impairment' vs. 'assessment'



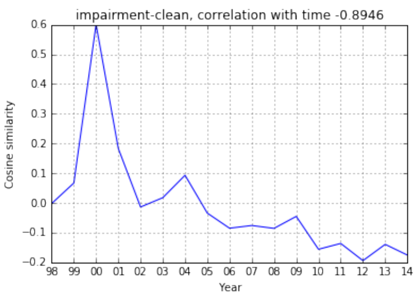
(d) 'impairment' vs. 'perform'



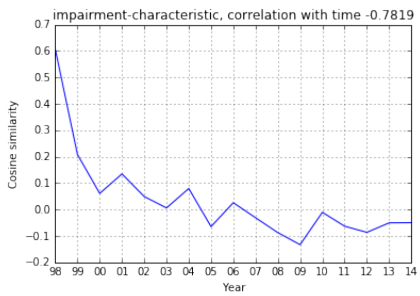
(e) 'impairment' vs. 'recognized'



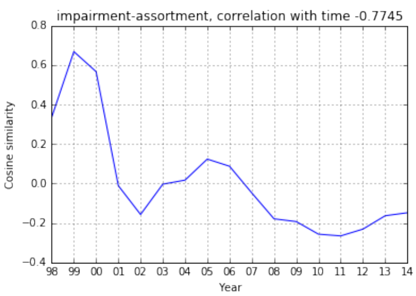
(f) 'impairment' vs. 'goodwill'



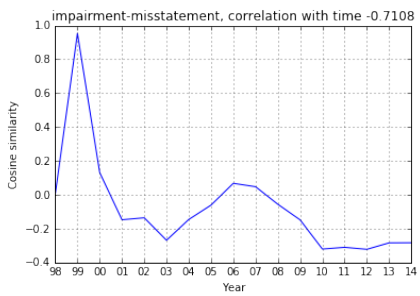
(g) 'impairment' vs. 'clean'



(h) 'impairment' vs. 'characteristic'



(i) 'impairment' vs. 'assortment'



(j) 'impairment' vs. 'misstatement'

Figure 3: Cosine similarities for word pairs with highest positive and negative correlations over time, for the lexical neighbourhood of 'impairment'.

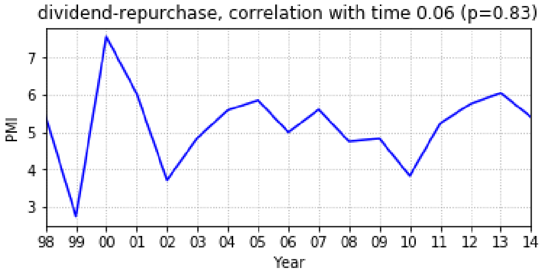


Figure 4: PMI for *'dividend'* vs *'repurchase'* over time.

servatism in preparation of financial statements, including recognizing all possible losses, but only realized gains. If a firm suspected a loss might occur in the future — and during the financial crisis they did suspect this a lot — they would have to recognize the present values of diminished expectations about future cash flows in current financial statements via recording (recognizing) an impairment in financial statements. Assets must be tested for impairment. This is generally done annually, although the regulation is more detailed and worded differently. However, some assets such as goodwill and other intangible assets must be tested for impairment rather than depreciated via "regular" depreciation expense in financial statements. The importance of these assets has generally increased through time, so financial crisis or otherwise, impairment and test would go hand in hand. The same explanation would account for (c) (*'assessment'*) which is an alternative term for *'testing'*; and (e) and (f) (*'recognized'* and *'goodwill'*) which are both related to goodwill and impairment testing.
The final term, (j) (*'misstatement'*) is particularly interesting. A "hump" can be observed with the term association increasing over the pre-crisis years, peaking with fiscal year ends 2006-7, and then decreasing down to a minimum in 2010. This could be due to firms correcting (possibly deliberate) mis-statements in financial statements from the pre-crisis years, as these were dug out by auditors, and recognised in the financial statements. Firms with mis-statements would also often record an impairment, as both are related to firms being too optimistic about their future prospects in the pre-crisis years. After 2010, this effect would therefore not be expected to be as pronounced as before.

7. Conclusion

Although only an initial methodological investigation, this study suggests that the use of word embeddings in a diachronic corpus can give some useful insights into terms used in financial reporting. Using GloVe provides a method to investigate changes in lexical associations which has revealed some intuitive relationships, while discovering others which warrant further investigation in the corpus data to understand the patterns. In future work we plan to extend this study in several ways: first, to use corpus analysis to explore the original context of the terms analysed to help understand the correlations more clearly; second, to explore more specific hypotheses from economic theory and financial research about term relations and changes; and third, to generalise the approach to automatically extend the list of terms of interest, discovering relationships in a more unsupervised fashion.
Our method is currently limited to analysing direct pairwise associations between words: more general properties of the word embeddings, including directions and magnitudes of movements in the general space of word meanings, cannot be derived when training GloVe models separately for each year as here. In future work, we will further investigate the use of the explicitly consistent spaces in the count-based PPMI variant by incorporating more appropriate smoothing, and the use of learned transformations between year-based spaces to make GloVe models consistent, following (Hamilton et al., 2016). Given a suitable model and dataset, it would ultimately be interesting to examine the relationship between terminological usage and companies' financial performance, via descriptive or predictive models.

8. Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency for research core funding (No. P2-0103 and No. P5-0161), as well as for funding of the research project *Influence of formal and informal corporate communications on capital markets* (No. J5-7387). We also thank Jasmina Smailović and Martin Žnidaršič for the corpus preparation phase and Rok Spruk for the discussion during the design of the study.

9. Bibliographical References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of HLT-NAACL*, pages 19–27.

Aiezza, M. C. (2015). "We may face the risks"..."risks that could adversely affect our face." A corpus-assisted discourse analysis of modality markers in csr reports. *Studies in Communication Sciences*, 15(1):68–76.

Allen, F. and Michaely, R. (2003). Payout policy. In G M Constantinides, et al., editors, *Handbook of the Economics of Finance*. Elsevier.

Athanasakou, V. and Hussainey, K. (2014). The perceived credibility of forward-looking performance disclosures. *Accounting and Business Research*, 44(3):227–259.

Balakrishnan, R., Qiu, X. Y., and Srinivasan, P. (2010). On the predictive ability of narrative disclosures in annual reports. *Eur. J. Operational Research*, 202(3):789–801.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, pages 238–247.

Berk, J. and DeMarzo, P. (2014). *Corporate Finance*. Pearson, 3rd edition.

Brav, A., Graham, J. R., Harvey, C. R., and Michaely, R. (2005). Payout policy in the 21st century. *Journal of Financial Economics*, 77(3):483–527.

Butler, M. and Kešelj, V. (2009). Financial forecasting using character n-gram analysis and readability scores of annual reports. In *Canadian Conference on Artificial Intelligence*, pages 39–51.

Camiciottoli, B. C. (2017). Persuasion in earnings calls: A diachronic pragmalinguistic analysis. *International Journal of Business Communication (preprint).*

Clark, S. (2015). Vector space models of lexical meaning. In Shalom Lappin et al., editors, *Handbook of Contemporary Semantic Theory*, pages 493–522. Wiley-Blackwell.

Feldman, R., Govindaraj, S., Livnat, J., and Segal, B. (2008). The incremental information content of tone change in management discussion and analysis. *Available at SSRN: http://ssrn.com/abstract=1126962*.

Firth, J. R. (1959). A synopsis of linguistic theory, 1930-55. In F. R. Palmer, editor, *Selected Papers of J. R. Firth 1952-59*.

Floyd, E., Li, N., and Skinner, D. J. (2015). Payout policy through the financial crisis: The growth of repurchases and the resilience of dividends. *Journal of Financial Economics*, 118:299–316, November.

Fuoli, M. (2017). Building a trustworthy corporate identity: A corpus-based analysis of stance in annual and corporate social responsibility reports. *Applied Linguistics*, page amw058.

Goel, S. and Uzuner, O. (2016). Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports. *Intelligent Systems in Accounting, Finance and Management*, 23(3):215–239.

Grullon, G., Michaely, R., and Swaminathan, B. (2002). Are dividend changes a sign of firm maturity? *The Journal of Business*, 75(3):387–424.

Hájek, P. and Olej, V. (2013). Evaluating sentiment in annual reports for financial distress prediction using neural networks and support vector machines. In *Intl. Conf. Engineering Applications of Neural Networks*, pages 1–10.

Hajek, P., Olej, V., and Myskova, R. (2014). Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making. *Technological and Economic Development of Economy*, 20(4):721–738.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proc. ACL*, pages 1489–1501.

Kenter, T., Wevers, M., Huijnen, P., and de Rijke, M. (2015). Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1191–1200.

Kiela, D. and Clark, S. (2014). A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 21–30.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. (2009). Predicting risk from financial reports with regression. In *Proceedings of HLT-NAACL*, pages 272–280.

Kosi, U. and Valentincic, A. (2013). Write-offs and profitability in private firms: Disentangling the impact of tax-minimisation incentives. *European Accounting Review*, 22(1):117–150.

Kothari, S., Li, X., and Short, J. E. (2009). The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review*, 84(5):1639–1670.

Leung, S., Parker, L., and Courtis, J. (2015). Impression management through minimal narrative disclosure in annual reports. *British Accounting Review*, 47(3):275–289.

Li, F. (2010). The information content of forward-looking statements in corporate filings - a naïve bayesian machine learning approach. *Journal of Accounting Research*, 48(5):1049–1102.

Lintner, J. (1956). Distribution of incomes of corporations among dividends, retained earnings, and taxes. *The American Economic Review*, 46(2):97–113.

Loughran, T. and McDonald, B. (2011). Barron's red flags: Do they actually work? *Journal of Behavioral Finance*, 12(2):90–97.

Merkl-Davies, D. M., Brennan, N. M., and McLeay, S. J. (2011). Impression management and retrospective sense-making in corporate narratives: A social psychology perspective. *Accounting, Auditing & Accountability Journal*, 24(3):315–344.

Michaely, R., Thaler, R. H., and Womack, K. L. (1995). Price reactions to dividend initiations and omissions: Overreaction or drift? *The Journal of Finance*, 50(2):573–608.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, pages 746–751.

Milajevs, D., Kartsaklis, D., Sadrzadeh, M., and Purver, M. (2014). Evaluating neural word representations in tensor-based compositional settings. In *Proceedings of EMNLP*, pages 708–719.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Qiu, X. Y., Srinivasan, P., and Street, N. (2006). Exploring the forecasting potential of company annual reports. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–15.

Rees, W. and Valentincic, A. (2013). Dividend irrelevance and accounting models of value. *Journal of Business Finance & Accounting*, 40(5-6):646–672.

Riedl, E. J. (2004). An examination of long-lived asset impairments. *The Accounting Review*, 79(3):823–852, July.

Rutherford, B. A. (2005). Genre analysis of corporate annual report narratives: A corpus linguistics-based approach. *The Journal of Business Communication*, 42(4):349–378.

Smailović, J., , Žnidaršič, M., Valentinčič, A., Lončarski, I., Pahor, M., Martins, P. T., and Pollak, S. (2017). Automatic analysis of financial annual reports: A case study. *Computación y Sistemas (Special Issue on Advances in Human Language Technologies)*, 21(4):809–818.

Turney, P. D., Pantel, P., et al. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

10. Language Resource References

FORMICA. (2017). *FORMICA Project Corpus (v0.9)*. Project dataset, currently not publicly released.