Doing well by talking good? A topic modelling-assisted discourse study of corporate social responsibility

Article

Accepted Version

This is a pre-publication version accepted for publication in Applied Linguistics, 2016. Please refer to the published version of this article if you wish to quote from it.

Doing Well by Talking Good?

A topic modelling-assisted discourse study of corporate social responsibility

Sylvia Jaworska, Anupam Nanda

Abstract

Using the novel technique of topic modelling, this paper examines thematic patterns and their changes over time in a large corpus of corporate social responsibility (CSR) reports produced in the oil sector. Whereas previous research on corporate communications has been small-scale or interested in selected lexical aspects and thematic categories identified *ex ante*, our approach allows for thematic patterns to emerge from the data. The analysis reveals a number of major trends and topic shifts pointing to changing practices of CSR. Nowadays 'people', 'communities' and 'rights' seem to be given more prominence, whereas 'environmental protection' appears to be less relevant. Using more established corpus-based methods, we subsequently explore two top phrases - 'human rights' and 'climate change' that were identified as representative of the shifting thematic patterns. Our approach strikes a balance between the purely quantitative and qualitative methodologies and offers applied linguists new ways of exploring discourse in large collections of texts.

Keywords: topic-modelling, corporate social responsibility, discourse, human rights, climate change

1. Introduction

Language plays a fundamental role in human decision-making and business decisions are no exception. Stakeholders make decisions based on not just numerical data, but also on texts that businesses publish through various channels. They are often subsumed under the term *corporate disclosures* and defined as the public release of economic data (Gibbins et al. 1990). Disclosures can be divided into *mandatory* and *voluntary*. The former are required by law and include highly conventionalised documents such as quarterly (e.g. 10-Q) and annual (e.g. 10-K) reports. The latter encompasses information provided by businesses beyond legal requirements through press releases, conference calls or corporate and social responsibility reports. Given the growing importance of transparency, ethical standards and non-financial information, voluntary reporting is currently on the rise (Beattie 2014). It is also the most dynamic practice of corporate communications and an important means by which corporations attempt to influence public discourse and perceptions (Livesey 2002). For this reason, voluntary disclosures warrant critical scholarly attention.

Corporate disclosures have been of interest to linguists for some time. Drawing on the genre approach, a number of scholars investigated lexical and textual features of various types of

disclosures (Skulstad 1996, 2005; Hyland 1998; Nickerson and De Groot 2005; Crawford Camiciottoli 2010). There is also a growing body of research that adopts discourse analysis, especially Critical Discourse Analysis (CDA) (e.g. Breeze 2012; Merkl-Davies and Koller 2012) sometimes in combination with tools and methods developed in corpus linguistics (e.g. Alexander 1999; Lischinsky 2011) to study discursive constructions of key lexis in corporate communications. However, compared with other genres of professional communication corporate disclosures remain an under-researched area in linguistics. Environmental and social reporting in particular has received little attention to date, with the exception of work by Alexander (1999), Skulstad (2005) and Lischinsky (2011) who have examined selected aspects of environmental reports, but in rather limited contexts, for example, reports from one financial year or one or two companies only.

Studying language use in voluntary disclosures is important not only from a pure descriptive interest; linguistic analysis with a critical edge could offer unique insights into practices and goals that businesses pursue in relation to environment and society. Given a growing awareness of the damaging impact of corporations on the environment and the resulting threats to ecological sustainability, this is now an issue of serious concern to the wider public, media and consumers who increasingly challenge the bottom-line-driven business practices and demand more transparency and higher ethical standards. Leaving disclosures, especially voluntary disclosures unscrutinised might diminish the potential for change and further contribute to the reinforcement of the 'business-as-usual' practice or "the change-but-no-change" rhetoric (Milne and Gray 2012: 14).

By far most studies on disclosures have been conducted in accounting and management studies. Here, the concern is mostly with 'who' reports and how 'much' adopting mostly quantitative methodologies (Tregidga et al. 2007). Language is important to this research too, but it is mostly understood as a unit that can be quantified and correlated with various corporate characteristics to statistically gauge the effects of disclosures on business performance and vice versa.

Understandably, some scholars in management studies see the quantitative research methodologies as being too reductionist, and call for a stronger consideration of discursive approaches to voluntary disclosures (e.g. Tregidga et al. 2007). This has sparked some interest, much inspired by the notions of discourse as social practice originated from work by Foucault (1973) and also widely adopted by (critical) discourse analysts in linguistics. While this research is rich in findings and illuminates some of the key messages communicated in disclosures, it also suffers from a number of limitations. Its empirical base is mostly small, often limited to a few

reports produced by one or two companies (e.g. Livesey 2002; Livesey and Kearins 2002; Milne at al. 2006; Brennan and Merkl-Davies 2014). This makes it difficult to generalise from the research findings and hence, we cannot be certain to what extent revealed patterns reflect general lexical and discursive tendencies of disclosures.

In her extensive overview, Beattie (2014) suggests that future research on disclosures would benefit from methodological pluralism, especially the use of computer-aided tools such as those developed in corpus linguistics that do not necessarily rely on pre-defined categories. Given that the nature of corporate disclosures is now rapidly evolving, most notably due to changing social attitudes and financial shocks, she also stresses the need for large-scale diachronic research that investigates 'narrative dynamics', that is, changes in language choices over time. This would permit the benchmarking of practices and ensure that typical or usual as well as atypical and unusual linguistic choices are identified and adequately interpreted.

This study attempts to respond to the above methodological challenges by undertaking a large-scale, computational and corpus-assisted analysis of the emergent corporate genre of Corporate Social Responsibility reports (henceforth CSR reports). In contrast to previous one-dimensional or small-scale content-analytical research, this study is based on a large corpus of CSR reports produced by 21 major oil companies between 2000 and 2013. This sector was chosen because of its direct involvement in environmental disasters and the resulting public criticism. The main questions which this research addresses are:

Q1: What messages and topics are communicated in the CSR reports?

Q2: How did they change over time and in response to significant shock events (e.g. financial crisis, environmental disasters)?

Q3: Which CSR themes emerged as particularly relevant and which ones have been given less prominence in recent years?

This study employs the tool of topic modelling combined with other established methods of corpus linguistics that are increasingly used to examine large amounts of textual data in humanities and social sciences. In contrast to previous methods that are mostly based on coding schemes set *ex ante*, our approach allows for semantic categories to emerge from the data. Topic modelling is performed on the data to identify key thematic patterns of CSR reports and their changes over time. This part of the analysis is quantitative and enables us to detect themes that have been given more prominence over time as well as those that are in decline across the sector. In doing so, we can capture the evolving nature of the genre reflecting changing business practices and goals. For example, issues surrounding people, community and rights appear to be

on the rise, whereas matters pertaining to environmental protection seem to be of lesser priority nowadays.

Quantitative approaches to disclosures have been criticised for being too mechanistic and reductionist (Tregidga et al. 2007). Our approach attempts to strike a balance between the purely quantitative and qualitative methodologies. Topic modelling offers new ways of extracting semantic domains automatically without imposing predefined categories on the data and appears to be more effective than other established corpus-based methods. It delivers quantitative results in the form of word and phrase lists that point to general thematic patterns that can subsequently be studied qualitatively in order to reveal aspects of discourse, for example, specific linguistic choices used to refer to salient or contentious concepts. We exemplify it by studying collocational profiles of two terms 'human rights' and 'climate change' that were identified as being representative of the major shifting tendencies established quantitatively. In this way, topic-modelling can effectively assist discourse analysts in revealing the overall picture of thematic patterns in a given discursive domain and at the same time, helps zoom in the analysis to salient points to be further explored qualitatively.

The remainder of this paper is structured as follows. Section 2 introduces the concept of Corporate Social Responsibility and the emergent genre of CSR reports, whereas Section 3 offers a literature review of the major studies and approaches to corporate disclosures. In section 4, we discuss the methodological underpinnings of our approach, procedures involved in the corpus compilation and methodological tools that were employed to process the data. Section 5 reports the main corpus findings. The first part focuses on the major topics and their distribution over time, while the second part examines the discourse surrounding 'human rights' and 'climate change'. We conclude our paper with observations regarding the benefits and limitations of the adopted methodology and indicate areas for future research.

2. Corporate Social Responsibility (CSR): concept, practice and reporting

Although the term Corporate Social Responsibility (CSR) is a product of the 20th century, the notion has a long history dating back to philanthropic initiatives during the Industrial Revolution (Carroll 2013). Gradually, the concept was expanded to include different groups of stakeholders and social and ethical matters such as racial discrimination, urban decay and from the 1970s increasingly environmental issues. It is the latter factor that accelerated the development of CSR activities not least because of the growing public criticism regarding the negative impact of businesses on the climate and environment. From the 1980s onwards, the term Corporate Social Responsibility has begun to be widely used to refer to corporate activities initiated in response to

environmental damage, employment discrimination and unethical practices (Carroll 2013). It was also around this time that CSR obtained a distinctive organisational status within companies and businesses began to report on CSR activities, first in form of shorter narratives included in annual reports and from the mid-1990s onwards in independent reports. The first stand-alone reports focused predominantly on environmental matters and were accordingly titled environmental reports. Gradually, companies began to include a wider range of issues and at the beginning of the 21st century, the title of Corporate and Social Responsibility was firmly established (Milne and Gray 2012).

The major feature of the CSR reporting is its relative variability. The main reason for this is its voluntary character. Unlike annual reports, CSR reports are not subject to legally binding standards and it is at the discretion of the company to select what to include and how to write the final report. Having said that, over the last two decades efforts have been made to standardise the CSR reporting leading to the burgeoning of many initiatives and frameworks. Of the many initiatives, the Global Reporting Initiative (GRI) emerged as the dominant player in the field (Waddock 2007). Essentially, the GRI uses a stakeholder approach and is based on the concept of the Triple-Bottom-Line (TBL) (Elkington 2007) and also known as the three Ps 'People, Planet and Profit'. Accordingly, the GRI is divided into three major categories: economic, environmental and social each including a number of indicators on which companies need to report. The GRI is thus a set of reporting norms and its prime intention is to ensure consistency and comparability of reporting.

The existence of reporting initiatives such as the GRI has had some homogenising effects on CSR reporting (Bhatia 2012). Most CSR reports nowadays tend to include issues pertaining to the three Ps, of which organizational governance, human rights, the environment, fair operating practices and community involvement appear to be mostly documented (Bhatia 2012). However, the extent to which the three Ps are covered varies considerably between companies depending on the sector, as well as local and global political goals (Breeze 2012).

Critics argue that CSR reporting gives prominence to the documentation of CSR activities and it is rarely concerned with assessing their impact (Vigneau et al. 2014). Thus, its potential to contribute to the development of sustainable future should not be overestimated. Nevertheless, they document examples of practices (even if only intended) and research shows that by making this knowledge public, voluntary disclosures can be critiqued by stakeholders, the wider public and academic researchers leading potentially to changes in business practices (e.g. Livesey and Kearins 2002). In the absence of other independent and reliable CSR performance indicators, CSR reports remain, however, the only publicly available source of information regarding

companies' goals and actions in relation to the environment and society. Thus, they present a unique case at the interface between business and society revealing ways of how this link and especially the delivery of public goods is conceptualised from the point of view of businesses and communicated to the wider world. As discussed in the Introduction, such conceptualisations require critical linguistic scrutiny. The below examination of the lexical environments surrounding 'climate change' and 'human rights' are good cases in point. Human rights seems to be increasingly emphasised in the context of CSR. However, mere mentions of the concept cannot be equated with pro-active stance and academic researchers argue that despite the increased attention, the CSR domain pursues, if at all, a human rights minimalism (Wettstein 2012). An in-depth corpus linguistic analysis of the use of 'human rights' can offer insights into how human rights are conceptualised in CSR and whether a minimalist approach is indeed practiced. Equally, studying the use of 'climate change' allows us to reveal the changing position of the oil industry in this ever controversial matter.

With a few exceptions, the language of CSR reporting has received little attention despite a considerable amount of research on language of corporate disclosures carried out in accounting and management studies and to some extent in linguistics. The main research perspectives and approaches adopted in this research are summarised in the next section.

3. Corporate disclosures: research perspectives and approaches

Business and accounting scholars have shown a considerable interest in the language of financial disclosures going back at least to the 1960s. Here, the notion of readability proved to be particularly popular producing the largest body of research (Beattie 2014). Studies in this area take the construct of readability (mostly measured by the Fog Index) as an indicator of language complexity and an increase in this complexity is considered to be a sign of concealment. To put it simply, in case of bad performance managers may strategically tend to obfuscate information by 'concealing' bad news in longer and more complex disclosures that are difficult to read (e.g. Li 2008). This, in turn, can have an impact on market responses in that investors and stakeholders would need more time to process the information and extract the bits that mangers prefer to stay hidden. Having said that, studies that focus solely on readability are rather one-dimensional and capture only one side of language, namely the form (Rutherford 2005) while ignoring the content and meanings of words. It is, after all, the meaning of the communicated messages that is more likely to influence stakeholders' and public perceptions.

The contents of corporate disclosures have been of interest to scholars working in the field of disclosure index studies. Disclosure indices are created by quantifying the amount of information

about selected thematic categories. These are mostly identified *ex ante* and the data is subsequently scanned for presence or absence of the identified categories. Indices created in this way are then correlated with company-specific variables. Another large body of research adopts techniques of thematic content analysis to study contents of disclosures. Indicative for this research is, for example, work by Beattie et al. (2008), who examine the changes in annual reports produced focusing especially on the visual material.

While disclosure index studies and research using content analysis show some correlations between topics and specific performance indicators, they also suffer from a number of weaknesses. Most importantly, they are based on categories that are defined *a priori*. Hence, they may not necessarily reflect the diversity and changing nature of CSR reporting. Also, the categories that are coded must be sufficiently exhaustive and consistently applied across the studied data sets so that other researchers replicating the study would arrive at the same or comparable results. Unfortunately, only a very few studies concerned with CSR demonstrate a rigorous reliability (e.g. Unerman 2000).

Parallel to the content-analytical research, a number of scholars in accounting and management studies adopt discourse analysis (Livesey 2002; Livesey and Kearins 2002; Milne at al. 2006). Dissatisfied with the quantitative approaches, they explore the notion of discourse as social practice originated in the work by Foucault (1973). For example, Livesey (2002) investigates the discourse of the first social report 'Profit and Principles' published by Shell in 1998 and reveals the contradictory nature of corporate views on sustainable development and a strong profit orientation. In a similar vein, Livesey and Kearins (2002) compare the CSR Shell report of 1998 with a similar document produced by the Body Shop International. Although the two corporations could not be more different in terms of size and products, the analysis demonstrates considerable similarities in their CSR reporting. Both companies draw heavily on rational notions of transparency and accountability and tend to present themselves as caring institutions.

This discourse-analytical strand of research has been invaluable in revealing aspects of corporate CSR discourse but these studies are mostly conceptual in nature and do not offer systematic linguistic insights, despite the fact that most are based on linguistic concepts such as metaphor or discourse. This is partially due to the fact that business studies have a different take on discourse analysis drawing mostly on the Foucauldian concepts of discourse (Foucault 1973). Discourse is here understood as content that can be analysed in order to reveal mechanisms of power and control. While linguists too engage with the Foucauldian notion of discourse, this is

often the starting point. A linguist undertaking discourse analysis would then drill in and examine the specific discursive, lexical and grammatical choices that are used to convey messages.

Although much smaller in scope, there has been some research on corporate disclosures in (applied) linguistics. Similar to studies in accounting and management studies, annual reports and especially CEOs letters have been given most attention (Skulstad 1996; Hyland 1998; Nickerson and De Groot 2005; Rutherford 2005; Merkl-Davies and Koller 2012; Breeze 2012). Despite the dominant focus on annual reports, there are also a few studies that examine selected linguistic features of CSR reports. Using qualitative and quantitative corpus-based techniques, Alexander (1999, 2009) studies the lexical environment of the terms 'sustainable' and 'sustainability' in environmental reports produced by Shell in 1990 and 2000. The analysis shows that these terms are often followed by nominalisations that conveniently erase the agency and responsibility. In this sense, 'sustainability' becomes an elusive concept used to demonstrate 'commitment' in a non-committal way. Adopting the framework of metadiscourse proposed by Mauranen (1993), Skulstad (2005) compares the use of action markers, previews, connectors and reviews in introductory sections of environmental reports and annual reports issued by British companies. The study shows that the introductions of environmental reports make greater use of metadiscursive devices than their counterparts in annual reports and thus place more emphasis on assisting the reader and the reading process. The author argues that the emerging nature of the genre and a lack of established conventions could be a reason for this increased use of metadiscourse. Combining CDA with corpus linguistic tools and methods, Lischinsky (2011) investigates instances of self-reference in a corpus of 50 CSR reports issued by Swedish companies during 2009. His research suggests that businesses communicate social agendas with a mixture of institutional and affilitative voices. The institutional tone is represented by companies' names and provides the necessary legitimacy and credibility. The frequent use of affilitative pronouns 'we' and 'our', on the other hand, points to group dynamism and unity. In the view of the author "this fosters a view of the organisation as a cooperative whole, while maintaining a level of generality that hampers criticism and falsification." (Lischinsky 2011: 272).

Research on corporate disclosures in accounting and management studies highlights many uses to which mandatory and voluntary disclosures are strategically put by companies. Alongside primary goals such as information, accountability and transparency, companies increasingly use disclosures for the purpose of promotion, legitimation and in some cases obfuscation. This research is mostly underpinned by theoretical notions developed in management studies such as impression management, obfuscation theory and the incremental information theory that offer interesting socio-economic interpretations. However, with a few exceptions, this research is

rarely concerned with a thorough analysis of texts and tells us little about the linguistic choices that are deployed to serve corporate goals. In other words, the *what* of communication and its effects are of more importance here. Linguistic research, on the other hand, focuses primarily on texts and is interested in textual, lexical and discourse features of disclosures. As the above brief overview illustrates, a whole array of linguistic devices are deployed by companies to establish legitimacy, trust and promote a positive corporate image.

Combing both business and linguistics perspectives could offer fruitful synergies. The discourse-analytical framework seems to be a point of intersection and a bridge between the two disciplines. However, with a few notable exceptions (Rutherford 2005; Merkl-Davies and Koller 2012), linguists and business scholars seldom talk to each other. This could be partially due to the perceived differences in methodological approaches. There is a belief that business scholars are only interested in 'hard' quantitative data, whereas linguists work only with 'soft' qualitative data that does not produce generalisable results. The above overview demonstrates that this is not necessarily the case. The proponents of qualitative discourse-analytical research see large quantitative analyses of disclosures as reductionist (e.g. Tregidga et al. 2007). Conversely, the discourse-analytical approaches have been critiqued for being too subjective and a lacking a rigorous and sound empirical methodology (cf. Beattie 2014). Our study proposes a novel methodological approach that attempts to strikes a balance between the purely quantitative and qualitative approaches and offers opportunities of synergies for research in business studies and linguistics. The next section outlines the principles and procedures of our approach and the dataset under examination.

4. Methodology and Data

Our methodology follows a corpus-based approach, but extends it by utilising the computational method of topic modelling. Since the beginnings of corpus linguistics, corpus linguists have been interested in capturing discursive representations and themes surrounding diverse social phenomena (e.g. Gabrielatos and Baker 2008; Jaworska and Krishnamurthy 2012; Baker et al. 2013). Much of this work combines qualitative discourse-analytical approaches with quantitative tools and methods, especially collocations and keywords.

Keywords are generally considered good indicators of texts' aboutness and hence are often studied to reveal themes in a given data set. In corpus linguistics, a keyword is considered a word which occurs unusually often in a given corpus, as compared to another usually larger reference corpus and a test of statistical significance is performed to assess this unusualness (Scott 2010). Although keywords retrieved in this way are useful in signposting main topics, there have some

limitations (Baker 2004; Gabrielatos and Marchi 2012). Firstly, the type of keywords retrieved from the target corpus greatly depends on the selection of the reference corpus, its size and contents. Corpus-based retrieval of keywords often utilises the British National Corpus (BNC) as the reference corpus, because it is regarded as a representative compilation of (British) English. It needs to be borne in mind that the BNC reflects the English usage of the late 1990s and might not include many of the newer words; certain items might be identified as key only because they are rare or non-existent in the BNC. Secondly, to capture the main themes, researchers often scan through the lists manually and group keywords into semantic categories (Gabrielatos and Baker 2008; Baker et al. 2013). Although useful, this procedure becomes problematic when dealing with a large dataset (Gabrielatos and Marchi 2012). For example, a keyword list retrieved from the corpus used in the present study produced a list with 15,000 items even though the cut-off point for statistical significance was set at the level of p=0.000001. It would require a larger team of researchers to manually group this amount of keywords into semantic categories. Hence, we abandoned the corpus-linguistic approach to keywords in favour of a computational data-driven technique known as topic modelling.

Topic modelling has only recently begun to be used beyond computational sciences in (digital) humanities and social sciences. It quickly proved to be a robust tool in exploring large amounts of textual data in political (e.g. Lischinsky 2014), historical and literary research (e.g. Riddell 2014; Goldstone and Underwood 2014). As is demonstrated below, it has considerable potential for research in applied linguistics, especially discourse studies interested in analysing large collections of texts.

The term topic modelling refers to a number of generative probabilistic models, of which the most widely adopted is the Latent Dirichlet Allocation (LDA) developed by Blei et al. (2003). As it is with other statistical measures underlying the widely used corpus tools, scholars who are not trained mathematicians may encounter difficulties in understanding the complex equations on which such algorithms are based. But the practice of the topic modelling can be explained in simple terms.

LDA is driven by two assumptions: 1) each document contains a number of topics (the hidden variables) which are represented by a fixed number of words (the observable variables) and 2) the proportions of topics vary in each document. These two assumptions can be compared to a manual content analysis of CSR reports. A researcher interested in the main topics will go through the texts and will soon spot a number of words or phrases that point to, for example, climate change (e.g. environment, climate change) and highlight them perhaps with a green pen. She or he will also encounter words that point to business performance (e.g. assets,

cash flows) and will mark those with a yellow pen. Thus, each topic will be represented by a different colour and a list of words, some of which will be shared with the other topic(s). For example, the word 'change' can also appear in the topic business performance, but it may occur less frequently in this area. Moreover, the researcher would probably notice that in some reports, words associated with business performance are used more frequently than those denoting aspects of climate change, whereas in others the order may be reversed. This would tell her or him something about the general focus of the studied documents. This is precisely what LDA attempts to do computationally. It translates these two assumptions into an algorithm which, based on word frequencies and probability of occurrence, attempts to re-generate the hidden variables, i.e. topics, from the observable variables, i.e. words. The computational routine involves a three-step process: Step 1 – a number of topics that may be present in the corpus is specified by the researcher; Step 2 - the algorithm assigns every word (W) to a preliminary topic (Z); Step 3 (iterative) - the algorithm checks and refines the topic assignments, looping through each word in every document based on two prevalence criteria – frequency of word W in topic Z elsewhere and prevalence of topic Z in document D. The actual LDA formula as offered by Underwood (2012) is provided in Appendix 1. The results show the probability that word W comes from topic Z. In this sense, 'topics' are collections of words that have a high probability of co-occurrence and not topics as we understand the term in everyday language. Interestingly however, words belonging to such groupings seem to share a number of semantic similarities (see Appendix 2). Thus, some scholars suggest other terms, for example, semantic frames (Rychlý 2014), thematic patterns or even discourses (Goldstone and Underwood 2014) when describing the outputs.

As with any other method, this technique also comes with a number of caveats. Firstly, the number of topics needs to be specified by the researcher and given that there is nothing like an ideal number of topics, this may seem rather arbitrary. Given the probabilistic nature of the technique, a different number of topics will yield slightly different results. Most studies in digital humanities seem to settle on between 50 and 150 topics. The question which ultimately arises is how to choose the most adequate number of topics? In our view and for the present moment, this should depend on the knowledge of the field and how many themes one can reasonably expect to appear in the studied corpus. For our analysis, we decided to choose 80 as the number of topics. The decision was based on the KLD[1] rating scheme, which is widely used in business

[1] KLD stands for the Kinder, Lydenberg, Domini Research & Analytics and it is the most widely adopted rating scheme of CSR performance. It lists six broad categories including community, corporate governance, diversity, employee relations, environment, human rights and product. Each category is divided into a number of sub-

studies to measure CSR performance and includes 80 thematic sub-categories across six main themes. Thus, we assume this number to be a good indicator of the range of topics discussed under the banner of CSR. The second issue involves the process of topic labelling. Topic modelling tools give each topic only a numerical ID and it is up to the researcher to name the topics based on the retrieved list of words and word combinations. Undoubtedly, labelling is an intuitive process which relies on the researcher's knowledge and expertise of the field (Riddell 2014). The insight of the specialist, in this case an economist, was indispensable in adequately labelling the topics.

The CSR-Corpus used for the present study consists of 317 CSR reports produced between 2000 to 2013 by 21 major oil companies. Appendix 3 contains a list with the names of the included companies. The rationale for using these oil companies is two-fold: first, these are the largest companies representing major oil-producing regions; second, they report on CSR activities consistently and make most of the reports available on their websites, which ensured good access to the data. The size of the corpus is 14,806,512 tokens. The data was manually collected from the websites of the companies and converted into text files. It needs to be noted that for some companies, there were no CSR reports available for specific years and hence, gaps were filled with relevant narratives taken from annual reports whenever possible. Also, some companies, for example, Gazprom, produced separate environmental reports and also included sections on social responsibility in the annual reports. Both were included in the analysis and hence, for some companies, we had two documents per year. Since our text files were converted from pdf files, there were a number of 'unwanted' characters and these were removed by using a combination of regular expressions and a python script. Because we were interested in words only, numbers and currency abbreviations were removed too. To retrieve topics, we used the Mallet topic model package (McCallum 2002), which is becoming a standard topic modelling tool used in social sciences and digital humanities. The Mallet package includes a stop list which contains grammatical words of English. Since we were primarily interested in lexical items, the stop list was used too. Subsequently the Mallet tools computed 80 topics by grouping together words and two-word combinations. Subsequently, we studied all lexical items retrieved in each topic and based on the main meanings of the items, assigned a topic label. In cases in which the meaning of an item was not clear, we examined the use of the item in our corpus to detect the major senses in which it was used. Each topic included on average 30 single lexical items and 30 two-word combinations. Appendix 2 shows an example of Mallet outputs. As can be seen, the

categories or issues. The thematic categories included in KLD can be viewed here: http://cdnete.lib.ncku.edu.tw/93cdnet/english/lib/Getting_Started_With_KLD_STATS.pdf

vast majority of the items in this example pertain to issues involved environment and its protection and this topic was consequently labelled 'environmental protection'. It must be said that the knowledge of the field was required to understand some of the technical terms, especially those related to financial performance. Thus, the collaboration with an economist proved to be essential to label the topics adequately. About 42 topics were directly related to CSR and Corporate Financial Performance (CFP). Table 1 presents the list of the most frequent topics with the most frequently occurring words and two-word combinations in each.

[INSERT TABLE 1]

Topic modelling gives each topic an alpha value, which measures the concentration of topics across the corpus. The ratio of a topic's alpha divided by the sum of all topic alphas measures the share of the topic in the corpus. We use the ratios as weights for calculating the topic's importance across the time period. In the process of labelling the topics, we realised that some topics were thematically close in that they contained words pointing to similar issues or concepts. Hence, the decision was made to merge such similar topics. This is not unusual; a prominence of a word in topic A does not prevent it from being also prominent in topic B (Goldstone and Underwood 2012). Hence, Goldstone and Underwood (2012) suggest that researchers need to survey all topics in order to identify their interconnectedness. A single topic might be too small a unit to analyse, especially if we want to say something about discourse. Hence, the decision was made to merge topics that contained at least 5 same or similar lexical items in the top 10 words. For example, topics with the numerical id of 16, 23 and 42 all contained the following items amongst the top 10 words: 'human', 'rights', 'community', 'communities' and 'local'. Hence, the three topics were merged into one and labelled 'people, community and rights'. The alpha values for the topics and each word and phrase were added up accordingly.

Table 2 shows the ten largest topics in our data set. 5 of these topics relate clearly to activities associated with CSR and the other 5 to corporate financial performance (CFP). These 10 largest topics are the focus of the analysis in the next section.

[INSERT TABLE 2]

To demonstrate the potential of this approach for qualitative discourse-analytical insights, we further examine two top word combinations 'human rights' and 'climate change' that were identified as representative of general semantic shifts established in our corpus. The analysis of

these two terms is conducted using the established techniques in corpus-assisted discourse studies, namely collocations and examination of expanded concordance lines. These are performed via Sketch Engine (Kilgarriff et al. 2004). Collocations are very useful pointers of recurrent and typical lexical choices in a given data set. Such recurrent preferences are not just a matter of individual choices, but largely reflect established practices and are often a means by which people and actions are evaluated (Stubbs 2001). Although different definitions and approaches to collocation have emerged in corpus linguistics, most would consider collocation as the co-occurrence of two or more words within a certain span (for example –4 to +4) and established on the basis of significance testing (McEnery and Hardie 2012). Various tests of statistical significance are used, of which the most popular are Mutual Information (MI), T-score and more recently LogDice. Each of the tests yields different results because they favour different types of words. For example, MI tends to emphasise low frequency words, whereas t-scores favour words that have a relatively high frequency such as function words. LogDice, which is based on the Dice coefficient, can be positioned in the middle as it combines the relative frequency of the relation X (headword) + Y (collocate) with frequencies of X in the same syntactic position and with any collocate, and Y in any syntactic position (Rychlý 2008). Some researchers see LogDice as the best method of determining collocations (cf. Baker 2014) and thus, this metric is also adopted in this study. Theoretically, the maximal LogDice value can reach 14; 10 points to a very strong relation (saliency), whereas 0 and negative values to no relation (Rychlý 2008).

5. Results

5.1 Main topics and their distribution over time

Table 2 shows the 10 major topics ranked according to its proportion in the whole corpus. The remaining topics accounting for 2.3% of the data contribute each less than 0.005 (as a ratio of a topic's alpha divided by the sum of all alphas – a low number e.g. 0.005 indicates a very low concentration of that topic in the corpus at 0.05%) and were hence not considered here. Although the main aim of CSR reports is to demonstrate company's actions and activities in relation to society and environment, it is interesting to note that still a larger proportion of the corpus focuses on financial concerns and developments, and some of the core CSR areas such as 'people, community and rights' accounts for just 10%. Equally, 'environmental protection' and 'health and safety' constitute a smaller proportion of the whole corpus. Hence and contrary to the wider assumptions (e.g. Breeze 2013), CSR reports are not just about CSR activities; they also communicate extensively about issues related to corporate financial performance (CFP)

including 'business operations', 'research and development', 'future plans and expansions' as well as 'products'. Core CSR areas identified in our corpus include 'people, community and rights', 'environmental protection', 'human capital', 'corporate governance and citizenship', 'environmental protection' and 'health and safety'. Although the two categories 'people, community and rights' and 'human capital' focus on people, they were kept separately as each includes different groups of stakeholders. The latter contains references to primary stakeholders, that is, internal stakeholders who engage with the business directly including employees, management, shareholders and customers, without whose participation an organisation would not survive (Clarkson 1995). The latter focuses mainly on secondary stakeholders, that is, people and organisations external to the companies, who do not engage with the business directly, but can influence or be influenced by it, positively or negatively. In the context of CSR, this group includes mostly local communities, media and special interest groups (Clarkson 1995).

While the distribution of topics highlights the main themes of CSR reporting and thus, answers our first research question, we need to remember that CSR as a business field and a genre has undergone many changes. In order to understand the evolving nature of CSR practices, we analysed the topic distribution over time. Using the alpha ratios, we calculated the proportion of the 10 major topics for each year starting with 2001. Figure 1 shows the results in two diagrams, the first focusing on CSR and the second on CFP topics. As can been seen, the category 'people, community and rights' has been continually on the rise, which confirms the claim that these aspects are gaining greater importance than other CSR areas (Breeze 2013: 166). The second most prominent category is 'health and safety'. It is not surprising to see a rise of this topic in 2006 and then again in 2011. The years 2005 and 2010 mark some of the worst oil spills in the history of the industry and hence, an increased focus on issues surrounding health and safety in the year following the catastrophes. Conversely, the share of the topic 'environmental protection' seems to be in decline since 2001 reflecting a lesser concern with climate and environmental change. This might be due to the concerted efforts of many conservative organisations and think-tanks to support the climate change counter movement (CCCM) that delegitimises the scientific evidence and justifies the status-quo, that is, further unlimited use of energy resources (Brulle 2014). Links between the CCCM and some major oil companies are well known and some, for example, Exxon Mobile have been publically criticised for their complicity in denial campaigns by funding some CCCM organisations (Goldberg 2015).

[INSERT FIGURE 1]

The area of CFP shows some interesting tendencies too. The most prominent CFP category is that of 'future plans/expansion' followed closely by 'research and development'. Whereas in the period from 2001 to 2006, most of the CFP areas show a declining tendency, a sudden rise can be observed in 2008 especially regarding 'future plans/expansion'. In the same year, there was a drop in 'people, community and rights'. Given that the year 2007 marks the beginning of the global recession, it appears that CSR reporting responded to this event by shifting the focus from people and communities to financial performance, possibly in an attempt to ensure stakeholders about the viability of the business despite financial losses. In this way, CSR can be seen not as a fixed but rather fluid concept amenable to external circumstances. Despite the fluctuations, our data point to some stable patterns of the last 13 years. As Figure 2 shows, core areas of CSR have been steadily on the rise, whereas CFP has been declining. This again provides empirical evidence for claims proposed in previous discourse-analytical research that in CSR social issues including human rights and communities are becoming more prominent than aspects of business performance or environment (Breeze 2013). In the context of the oil industry, the heavier focus on people and communities from 2010 onwards is also possibly due to the increasing efforts of the oil industry to demonstrate a social and ethical commitment following the Deepwater Horizon disaster (Breeze 2012).

[INSERT FIGURE 2]

5.2. Climate change vs. human rights

Whereas the previous section provided a general overview of topical changes over time, this section focuses on two phrases 'human rights' and 'climate change' that are representative of the general thematic shift identified above. 'Human rights' is the top two-word combination retrieved from the topic 'people, community and rights' and accounts for 2% of the topic. 'Climate change' is the top phrase in 'environmental protection' and its share amounts to 1.5% of the topic.

Figure 3 presents the use of 'human rights' and 'climate change' in our corpus. As can be seen, 'human rights' have noted a steady rise accelerated after 2010, whereas 'climate change' appears to be given less prominence, especially in recent years.

[INSERT FIGURE 3]

The notion of anthropomorphic climate change can be referred to by other lexical terms in English, of which most frequently used are 'global warming' and 'greenhouse effect' (Grundmann and Krishnamurthy 2010). Interestingly, the two other terms are very rare in the corpus with 'global warming' occurring 138 times and 'greenhouse effect' only 19 times. 'Climate change' with 2,118 occurrences seems to be the preferred term used in the context of CSR.

An increased attention to climate change after 2003 can be noted and this might have been influenced by a number of political and media factors. The wider media campaign following the release of Al Gore's book and film might have played a role. More important from the point of businesses were probably the ratification of the Kyoto Protocol by the EU in 2002 and the publication of the Stern Review on the Economics of Climate Change in 2006, which shifted the focus from climate change as a science to climate change as economics (Koteyko 2012). Increasingly climate change began to be perceived as an investment opportunity and to a lesser extent as a threat. This increased attention was accompanied by a much more pro-active attitude to climate change, as reflected in the collocational profile of 'climate change' in 2007, at the point when the term reached a peak in our corpus. This result confirms the tendency reported in research by Grundmann and Krishnamurthy (2010) on the media coverage of climate change, who too noted an exponential rise after 2005 and a peak in 2007.

[INSERT TABLE 3]

As Table 3 shows, in 2007 'climate change' was strongly associated with the action verb 'combat' and nouns pointing to goals and actions such as 'approach', 'policy', 'goal' and 'initiative'. We also find here a number of associations that signal specific causes and preventative measures including 'greenhouse', 'fossil', 'carbon' and 'emission'. Studying expanded concordance lines of the collocation pair 'climate change' and 'greenhouse' indicates some of the preventative actions that the oil industry introduced or intended to introduce (see concordance lines 1-3).

(1) We have set out two large practical goals to prevent *climate change*: reduce greenhouse emissions; generate profits from selling emission reduction units (ERU) on hydrocarbon markets. (Lukoil)
(2) We also take proactive steps to address emerging global environmental concerns such as *climate change* and greenhouse gas (GHG) emission. Reducing GHG Emission as an

environmentally responsible corporation, PETRONAS has established GHG accounting and inventorisation across the Group. (Petronas)

(3) Eni considers *climate change* and greenhouse gas emissions as key challenges for the evolution of the energy sector, and as such has adopted a carbon management strategy designed to promote the use of low-carbon-content fossil fuels, such as natural gas, and the development of new mitigation technologies. (ENI)

Not much of this pro-active attitude seems to be visible in 2012. As Table 3 indicates, now 'climate change' is framed mostly as a 'challenge' and a 'risk' and there seems to be less emphasis on specific actions and goals. Interestingly, in 2007 climate change was portrayed as a problem that the industry could do something about ('combat', 'prevent') and thus some responsibility was assumed. The focus in 2012 is on climate change as an uncontrollable or unknown force as shown by the frequent associations with 'induce' and 'risks'. The following expanded concordance lines of the collocation pair 'climate change' and 'induce' are illustrative in supporting this claim (see concordance lines 4-6).

(4) The science of climate change modelling has many uncertainties, which lead to difficulties in designing facilities and operations. Eni maps area by area the risks associated with possible natural disaster induced by *climate change*. (ENI)

(5) In general terms, the understanding of impacts induced by *climate change* at regional level is still in the early stages, but, at a mid to long term perspective, areas considered at very high risk include: the Polar region, Far and Middle East, the Gulf of Mexico and Africa. (Exxon Mobil)

(6) Though physical changes induced by *climate change* represent, primarily, a risk for the oil and gas industry […], on the other hand, it is evident that some extreme events could force oil & gas prices up and therefore increase oil producers revenues. (TOTAL)

Now climate change seems to be linked with uncertainties and little understanding. Whereas in 2007 climate change was mostly used in the position of an object, now it increasingly assumes the role of an active and destructive agent causing natural disasters. It is farmed as a risk to the industry 'forcing' it, for example, to increase oil prices. Thus, a discursive shifting of responsibilities can be observed indicating a lesser commitment to climate change. Climate change is even 'used' to legitimise certain actions that from the point of view of consumers could be seen as detrimental (increasing oil prices). This change in attitude might be attributed to the

increase in anti-climate campaigns financially supported by many conservative organisations and foundations, specifically in the United States (Brulle 2014).

Whereas climate change seems to be overall less important, human rights seems to have gained more prominence in CSR as indicated by the increased frequency of the term over time.

[INSERT TABLE 4]

The collocational profiles of the term in 2007 and 2012 are rather similar (see Table 4). The strongest collocate is 'respect'. However, as extracts 7-9 below exemplary illustrate, 'respect' in the vicinity of 'human rights' is often preceded by mental verbs such as 'strive', 'seek' or 'promise' that as such do not necessarily involve volition or action. Similar to sustainability (Alexander 2009), 'human rights' appear to be 'dropped' in texts to express commitment in a rather non-committal way.

(7) "We strive to respect *human rights* and avoid complicity in abuses." (BP)
(8) "Through effective communication and consultation, ExxonMobil seeks to establish and maintain community relationships while actively promoting respect for *human rights*." (ExxonMobil)
(9) Repsol promises to respect the *human rights* of its employees, and will establish the necessary mechanisms to safeguard these rights in all the countries in which the company operates. (Repsol)

Other top collocates in 2007 and 2012 include 'principles' and 'security'. These terms refer to the *Voluntary Principles on Security and Human Rights* which was mentioned 20 times in 2012 and 16 times in 2007. These principles are a set of guidelines that were established in 2000 specifically for extractive sector companies. The document does not have a juridical power. It is a good practice tool to guide companies in protecting human rights in relations to their operations. The reference to these principles demonstrates that the oil industry promotes itself as compliant with industry good practice rules. Apart from these guidelines, the concept of 'human rights' is mostly mentioned in relation to 'labour', 'employment' and 'training' and does not seem to be used much outside these domains. Apart from child labour and awareness training programmes no other specific activities demonstrating how human rights are protected are mentioned and the international dimension of human rights does not seem to play a role. This presents a rather

limited view of human rights and further supports the claim that corporations pursue, if at all, a human rights minimalism (Wettstein 2012).

6. Conclusions

Our novel approach that combines topic modelling with more established corpus tools and techniques offers a number of benefits. Firstly, topic modelling is a robust tool for identifying key semantic categories in a large amount of textual data. In contrast to other methods that rely on pre-defined categories and dictionaries, it is data-driven and allows for semantic categories to emerge from the data. The outputs in form of word lists and two-word combinations can further be used as signposts to funnel the analysis to selected key phrases that can be critically investigated adopting techniques of corpus-assisted discourse studies. Secondly, the approach is an effective way of revealing 'narrative dynamics' (Beattie 2014: 124), that is, shifts in themes over time. In this way, we were able to show a number of trends that point to changing practices of CSR. For example, one of such strategies seems to be a greater focus placed on people, communities and rights as opposed to environmental protection. Thus, it appears that 'people' are more valued than 'planet'. This might be a reflection of changing social attitudes and the current scepticism towards the scientific evidence of climate change also widely communicated in the mass media (Boykoff 2014). As the analysis of the use of 'climate change' has demonstrated, the oil industry is happy to capitalise on such scepticism in order to diminish its own responsibility. The shift from climate change as an object that can be reduced or managed to climate change as a destructive and uncontrollable agent is a compelling discursive example indicating such changes in attitudes. At the same time, the increased mention of 'human rights' should not be equated with a greater scope of activities in the relation to protecting and maintaining human rights. In the context of CSR in the oil industry, human rights are understood mostly in a minimalist sense and the focus is on the domain of labour rights and employment and a general obligation to merely 'respect' them as opposed to being more proactively involved (cf. Wettstein 2012).

The analysis focused only on two terms and arguably, we did not need to perform topic modelling in order to investigate their use in our corpus. However, prior to the analysis we were not aware of the fact that 'climate change' and 'human rights' are the top phrases in clusters of words that focus on environmental protection and people, community and rights respectively. It is only after we used the tool that we realised the prominence of these terms in the two semantic fields. In this way, topic modelling is a good exploratory tool that can reveal the overall picture

of thematic representations in large text corpora that would be otherwise difficult to capture. It can also signpost important terms for further discourse-analytical investigations.

Despite the benefits of this approach, there are also some limitations, of which researchers need to be aware. The number of topics is a thorny issue with a degree of arbitrariness. It is hoped that future computational research will provide techniques to help researchers establish an optimal number of topics for a given collection. Also, topic modelling is a statistical simplification that relies on human interpretations. It is after all the researcher who has to make a judgement and label a topic. This process is largely subjective and as Riddel (2014: 108) reminds us, it is "essential that those using topic models validate the description provided by a topic model by reference to something other than the topic model itself." Knowledge of the studied field and the corpus is, in this case, indispensable. Researchers in applied linguistics often venture into domains outside linguistics, as in this case, corporate communication. Collaboration with scholars familiar with the studied field would ensure a better validation of topic labelling. Work by Rychlý (2014) on an automated labelling based on the LDA and a thesaurus available on Sketch Engine, a network analysis proposed by Goldstone and Underwood (2012) or topic diversity measures suggested by Tran et al. (2013) could reduce the subjectivity of the so far intuitive process and facilitate a more reliable handling of topic labelling and merging.

The present study used topic modelling to explore a particular collection of texts. Although our corpus is large, the reports come from one sector only. In order to understand the nature of CSR and possibly contribute to the theoretical debate, future research would need to examine CSR reports from other sectors too. Cross-country comparisons would also be helpful in shedding light on the impact of socio-political and cultural context on CSR. On the final note, we hope that the above case study would encourage applied linguists who study large collections of texts to use topic modelling so that we could open a discussion about the suitability of the technique for the purpose of research in applied linguistics more widely.

References

Alexander, R. J. 1999. 'Ecological commitments in business: A computer-corpus-based critical discourse analysis' in J. Verschueren (ed.) *Language and Ideology. Selected Papers from the 6th International Pragmatics Conference*. Antwerp, International Pragmatics Association: 14-24.

Alexander, R. J. 2009. *Framing Discourse on the Environment*. New York, Routledge.

Bailin, A. and **A. Grafstein**. 2001. 'The linguistic assumptions underlying readability formulae: a critique.' *Language & Communication* 21: 285-301.

Baker, P. 2004. Querying keywords: Questions of difference, frequency, and sense in keywords analysis.' *Journal of English Linguistics* 32: 346-359.

Baker, P., **Gabrielatos, C**. and **T. McEnery**. 2013. *Discourse analysis and media attitudes*. Cambridge, Cambridge University Press.

Beattie, V. 2014. 'Accounting narratives and the narrative turn in accounting research: Issues, theory, methodology, methods and a research framework.' *The British Accounting Review* 46: 111-134.

Beattie, V., **Dhanani, A**. and **M. J. Jones**. 2008. 'Investigating presentational change in UK annual reports.' *Journal of Business Communication* 45/2: 181-222.

Bhatia, A. 2012. 'The Corporate Social Responsibility Report: The Hybridization of a "Confused" Genre (2007–2011).' *IEEE Transactions on Professional Communication* 55/3: 221-238.

Blei, D. M., **Ng, A**. and **J. I. Michael**. 2003. 'Latent Dirichlet Allocation.' *Journal of Machine Learning Research* 3: 993-1022.

Boykoff , M. T. 2014. 'Media discourse on the climate slowdown.' Nature Climate Change 4: 156-158.

Breeze, R. 2012. 'Legitimation in corporate discourse: Oil corporations after Deepwater Horizon.' *Discourse & Society* 23/1: 3-18.

Breeze, R. 2013. *Corporate Discourse*. London, Bloomsbury.

Brennan, N. M., and **D. M. Merkl-Davies**. 2014. 'Rhetoric and argument in corporate social and environmental reporting: the dirty laundry case.' *Accounting, Auditing and Accountability Journal* 27/4: 602-633.

Brulle, R. J. 2014. 'Institutionalizing delay: foundation funding and the creation of U.S. climate change counter-movement organizations.' *Climatic Change* 122/4: 681–694.

Carroll, A. 2013. 'Corporate Social Responsibility Evolution of a Definitional Construct.' *Business & Society* 38/3: 268-295.

Clarkson, M. 1995. 'A stakeholder framework for analyzing and evaluating corporate social performance.' Academy Management Review 20/1: 92-117.

Crawford Camiciottoli, B. 2010. 'Earnings calls: exploring an emerging financial reporting genre.' *Discourse & Communication* 4/4: 343-359.

Elkington, J. 1997. *Cannibals with forks: The triple bottom line of 21st century business*. Oxford, Capstone Publishing.

Foucault, M. 1973. *The order of things: An archaeology of the human sciences*. New York, Vintage.

Gabrielatos C. and **P. Baker**. 2008. 'Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996-2005.' *Journal of English Linguistics* 36: 5-38.

Gabrielatos, C. and **A. Marchi**. 2012. 'Keyness: Appropriate metrics and practical issues.' *CADS International Conference* 13-14 September 2012, University of Bologna, Italy: http://repository.edgehill.ac.uk/4196/1/Gabrielatos%26Marchi-Keyness-CADS2012.pdf

Gibbins, M., **Richardson, A**. and **J. Waterhouse**. 1990. 'The management of corporate financial disclosure: Opportunism, ritualism, policies, and processes.' *Journal of Accounting Research* 28/1: 121-143.

Goldberg, S. 2015. 'ExxonMobil under investigation over claims it lied about climate change risks.' In: *The Guardian*, 5 November, http://www.theguardian.com/environment/2015/nov/05/exxonmobil-investigation-climate-change-peabody

Goldstone, A. and **T. Underwood.** 2012. 'What Can Topic Models of PMLA Teach Us About the History of Literary Scholarship?' *Journal of Digital Humanities* 2/1, http://journalofdigitalhumanities.org/2-1/what-can-topic-models-of-pmla-teach-us-by-ted-underwood-and-andrew-goldstone/

Goldstone, A. and **T. Underwood.** 2014. 'The quiet transformations of literary studies: What thirteen thousand scholars could tell us.' *New Literary History* 45/3: 359-384.

Grundmann, R. and **R. Krishnamurthy**. 2010. 'The discourse of climate change: a corpus-based approach.' *Critical Approaches to Discourse Analysis across Disciplines* 4/2: 125-146.

Hyland, K. 1998. 'Exploring corporate rhetoric: Metadiscourse in the CEO's letter.' *Journal of Business Communication* 35/2: 224-245.

Jaworska, S. and **Krishnamurthy, R.** 2012. 'On the F-word: a corpus-based analysis of the media representation of feminism in British and German press discourse, 1990-2009.' *Discourse & Society* 23/4: 1-31.

Kilgarriff, A., **Rychlý, P**., **Smrz, P**. and **D. Tugwell**. 2004. The Sketch Engine. *Proc EURALEX 200*. Lorient, France: 105-116.

Koteyko, N. 2012. 'Managing carbon emissions: A discursive presentation of 'market-driven sustainability' in the British media.' *Language and Communication* 32: 24-35.

Li, F. 2008. 'Annual report readability, current earnings, and earnings persistence.' *Journal of Accounting and Economics* 45: 221-247.

Lischinsky, A. 2011. 'The discursive construction of a responsible corporate self' in A.E. Sjölander and J. Gunnarson Payne (eds.) *Tracking discourses: Politics, identity and social change*. Lund, Nordic Academic Press: 257-285.

Lischinsky, A. 2014. 'Tracking Argentine presidential discourse (2007-2014): a computational approach.' Presented at *CADAAD 2014*, Budapest, Hungary.

Livesey, **S.** 2002. 'Global warming wars: Rhetorical and discourse analytic approaches to Exxonmobil's corporate public discourse.' *Journal of Business Communication* 39/1: 117-148.

Livesey, **S**. and **K**. **Kearins**. 2002. 'Transparent and caring corporations? A Study of sustainability reports by The Body Shop and Royal Dutch/Shell'. *Organisation and Environment* 15/3: 233-258.

Mauranen, A. 1993. 'Contrastive ESP rhetoric: metatext in Finnish-English economic texts.' *English for Specific Purposes* 12: 3-22.

McCallum, A. 2002. 'MALLET: A Machine Learning for Language Toolkit.' http://mallet.cs.umass.edu.

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

Merkl-Davies, D. M., and **V. Koller**. 2012. "Metaphoring' people out of this world: a critical discourse analysis of a chairman's statement of a defence firm.' *Accounting Forum* 36/3: 178-193.

Milne, M. and **R. Gray**. 2012. 'W(h)ither ecology? The triple bottom line, the global reporting initiative, and corporate sustainability reporting'. *Journal of Business Ethics* 118: 13-29.

Milne, M., Kearins, K. and **S. Walton**. 2006. 'Creating adventures in wonderland: The journey metaphor and environmental sustainability.' *Organisation* 13/6: 801-839.

Nickerson, C. and **E. De Groot**. 2005. 'Dear shareholder, dear stockholder, dear stakeholder: The business letter genre in the annual general report' in P. Gillaerts and M. Gotti (eds) *Genre Variation in Business Letters*. Bern, Peter Lang: 325-346.

Riddell, A. 2014. 'How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models' in M. Erlin and L. Tatlock (eds.) *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*. New York, Camden House: 91-114.

Robb, S. W. G., **Single L. E**. and **M. T. Zarzeski**. 2001. 'Nonfinancial disclosures across Anglo-American countries.' *Journal of International Accounting* 10/1: 71-83.

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

Rychlý, P. 2008. 'A lexicographer-friendly association score' in P. Sojka and A. Horák (eds.) *Proceedings of Recent Advances in Slavonic Natural Language Processing RASLAN*. Brno, Masaryk University: 6-9.

Rychlý, P. 2014. 'Finding the best name for a set of words automatically' in P. Sojka and A. Horák (eds.) *Proceedings of Recent Advances in Slavonic Natural Language Processing RASLAN*. Brno, Masaryk University: 77-81.

Scott, M. 2010. 'Problems in investigating keyness, or clearing the undergrowth and marking out trails' in M. Bondi and M. Scott (eds.) *Keyness in texts*. Amsterdam, Benjamins: 43-58.

Skulstad, A. S. 1996. 'Rhetorical organization in chairmen's statements.' *International Journal of Applied Linguistics* 6/1: 43-63.

Skulstad, A. S. 2005. 'The use of metadiscourse in introductory sections of a new genre.' *International Journal of Applied Linguistics* 15/1: 71-86.

Stubbs, M. 2001. *Words and phrases. Studies in lexical semantics*. London, Blackwell.

Tran, N. K., **Zerr, S**., **Bischoff, K**., **Niederée, C**., and **Krestel, R**. 2013. 'Topic cropping: Leveraging latent topics for the analysis of small corpora' in Aalberg, T., Papatheodorou, C., Dobreva, M., Tsakonas, G. and C. Farrugia (eds.) *Research and Advanced Technology for Digital Libraries*. Berlin, Springer: 297-308.

Tregidga, H., Milne, M. and **K. Kearins**. 2007. 'The role of discourse in bridging the text and context of corporate social and environmental reporting.' Auckland, NZ: Asia-Pacific Interdisciplinary Research into Accounting Conference, 7-9 July 2007.

Underwoord, T. 2012. 'Topic modeling made just simple enough.' http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

Unerman, J. 2000. 'Methodological issues - Reflections on quantification in corporate social reporting content analysis.' *Accounting, Auditing & Accountability Journal* 28 (8): 667-680.

Vigneau, L., Humphreys, M. and **J. Moon**. 2014. 'How do firms comply with international sustainability standards? Processes and consequences of adopting the global reporting initiative.' *Journal of Business Ethics*, first published online 22 July 2014.

Waddock, S. 2007. 'On CERES, the GRI and Corporation 20/20'. *Journal of Corporate Citizenship* 26: 38-42.

Wettstein, F. 2012. 'CSR and the debate on business and human rights: bridging the great divide.' Business Ethics Quarterly 22/4: 739-770.

Table 1: The most frequent topics in the CSR-Corpus

Rank	Topic ID	Share (based on Alpha values)	Topic Label	Top Words	Top 2 word combinations
1	38	9.1%	future plan/expansion	future, number, years, impact, ensure	long-term, large scale, negative impact, recent years,
2	22	9.0%	business operation	business, operations, performance, operating, process	supply chain, business operations, business opportunities, continuous improvement,
3	19	6.8%	international project management	development, management, projects, international, industry	international standards, development projects, environmental management, projects implemented, international agency
4	54	6.4%	corporate governance/ citizenship	company, production, stock, merger, income	corporation annual report, common stock, income taxes, percent interest, joint venture, millions dollars
5	3	6.3%	product	products, fuel, oil, refinery, diesel	diesel fuel, petroleum products, gasoline diesel, oil products, quality products
6	23	5.6%	people, community & rights	local, community, human,	local communities, community engagement, lessons learned, stakeholder engagement

Note: Topic modelling gives each topic an alpha value, which measures the concentration of topics across the corpus. The ratio of a topic's alpha divided by the sum of all topic alphas measures the share of the topic in the corpus. We use the ratios as weights for calculating the topic's importance across the time period.

Table 2: The 10 (merged) topics with the largest share in the whole corpus

Rank	Topic Label	Weighted Share (based on Alpha values)	Top words	Top Collocations
1	business operation	17.7%	business, operations, performance, operating, process	supply chain, wide range, business operations, business opportunities, continuous improvement,
2	research and development	13.7%	exploration, reserves, development, engineering, project	exploration production, proved reserves, exploration wells, research development, development projects
3	future plan/expansion	10.6%	future, part, years, number, million, increase, ensure	long term, recent years, reporting year, year ahead, operating profit,
4	people, community & rights	10.1%	local, community, rights, training, human	human rights, local communities, community engagement, lessons learned, stakeholder engagement
5	business & finance	10.0%	million, billion, year, total, financial	million barrels, billion barrels, net income, cash flow, consolidated financial
6	human capital	8.8%	training, support, education, programme, team,	training courses, school students, labour organisation, independent experts, labour HR
7	corporate governance/ citizenship	8.1%	board, management, company, annual, audit	annual report, corporate governance, board directors, general meeting, audit committee, CSR report
8	product	7.4%	products, fuel, oil, stations, product	diesel fuel, service stations, petroleum products, gasoline diesel, oil products
9	environmental protection	5.7%	emissions, performance, environmental, climate, waste	climate change, environmental performance, greenhouse emissions, emissions reductions, wastewater treatment
10	health & safety	5.6%	safety, health, responsibility, improve, promote	health safety, occupational health, quality life, health care, injury frequency

Note: Topic modelling gives each topic an alpha value, which measures the concentration of topics across the corpus. The ratio of a topic's alpha divided by the sum of all topic alphas measures the share of the topic in the corpus. We use the ratios as weights for calculating the topic's importance across the time period.

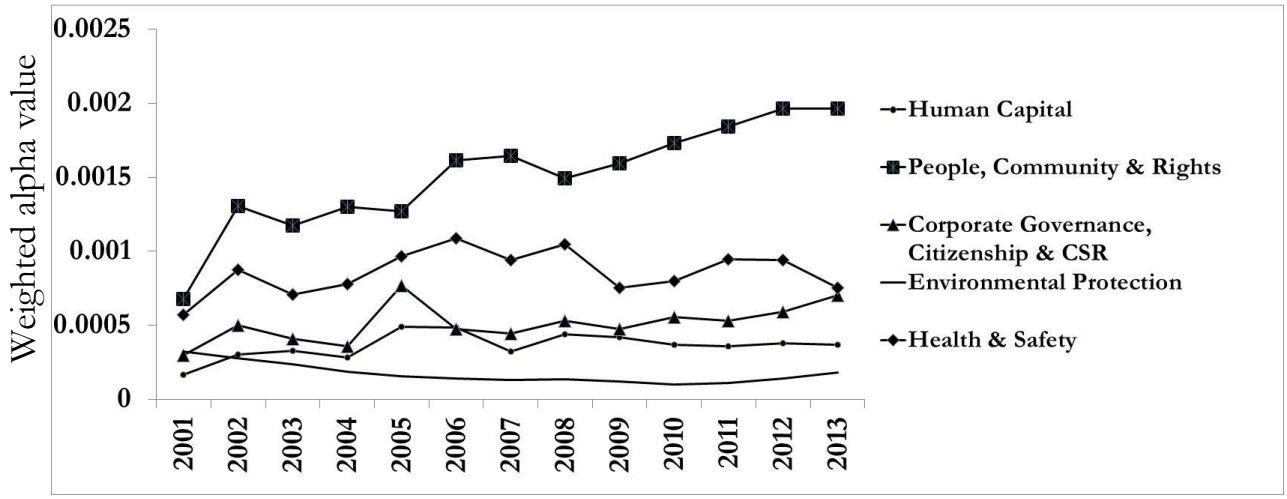
Table 3: The 25 strongest collocations of 'climate change'

2007			2012		
Collocate	Freq.	LogDice	Collocate	Freq.	LogDice
combat	19	10.935	challenge	19	9.933
intergovernmental	9	9.990	mitigation	11	9.845
address	13	9.642	induce	9	9.835
IPCC	7	9.622	risks	9	9.830
approach	13	9.474	address	18	9.828
global	18	9.283	managing	8	9.578
concern	10	9.264	relate	21	9.329
energy	45	9.236	physical	8	9.289
issue	22	9.190	mitigate	8	9.279
greenhouse	9	9.181	framework	11	9.223
biodiversity	8	9.014	convention	6	9.184
nations	5	8.925	risk	39	9.150
action	10	8.877	intergovernmental	5	9.065
policy	17	8.873	IPCC	5	9.050
impact	12	8.770	impact	24	9.049
challenge	7	8.755	strategy	16	8.990
framework	6	8.721	response	8	8.754
goal	7	8.670	nations	5	8.741
fossil	4	8.616	resource	13	8.733
measure	10	8.452	tackle	4	8.715
carbon	6	8.428	future	11	8.710
awareness	4	8.303	against	6	8.613
technology	10	8.297	extreme	4	8.608
initiative	7	8.278	disaster	4	8.595
emission	9	7.960	influence	5	8.427

Table 4: The 25 strongest collocations of 'human rights'

2007			2012		
Collocate	Freq.	LogDice	Collocate	Freq.	LogDice
respect	33	10.533	respect	74	11.399
security	44	10.499	principles	29	10.415
matrix	17	10.349	security	41	10.166
voluntary	19	10.256	practices	23	10.166
universal	15	10.203	labor	31	10.118
principles	17	10.196	chain	28	9.986
abuse	15	10.179	ethical	20	9.802
principle	28	9.971	aspect	20	9.678
declaration	13	9.919	voluntary	19	9.615
violation	12	9.791	training	32	9.547
training	25	9.662	employment	20	9.545
protection	15	9.493	undergo	14	9.443
HR	11	9.457	impact	37	9.405
rights	11	9.338	grievance	14	9.346
internationally	9	9.337	universal	12	9.300
corruption	9	9.290	audits	13	9.276
aspect	10	9.256	principle	20	9.257
transparency	9	9.059	violation	12	9.238
promote	11	9.010	declaration	12	9.214
force	8	8.986	corruption	13	9.177
labor	9	8.953	business	55	9.152
proclaim	6	8.941	screening	11	9.149
complicit	6	8.934	guiding	11	9.147
policy	19	8.903	HR	13	9.125
support	17	8.860	concern	16	9.081

Figure 1: Attributes of Corporate Social Performance (CSP) and Corporate Financial Performance (CFP)



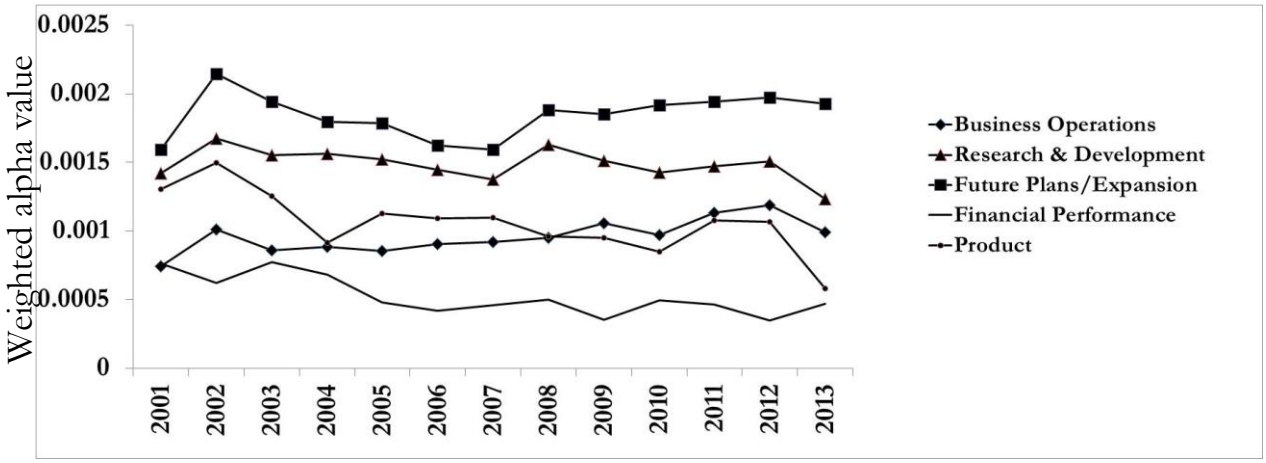
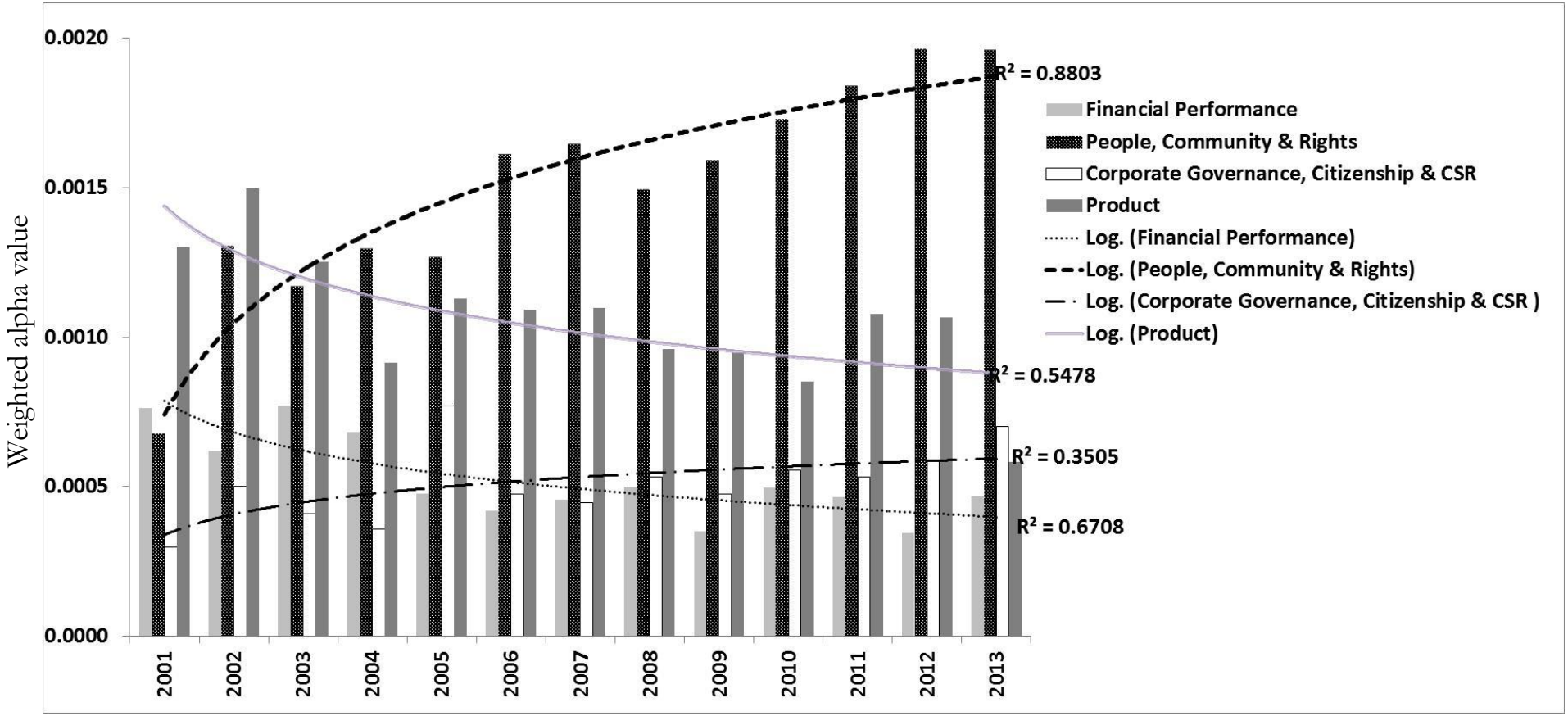
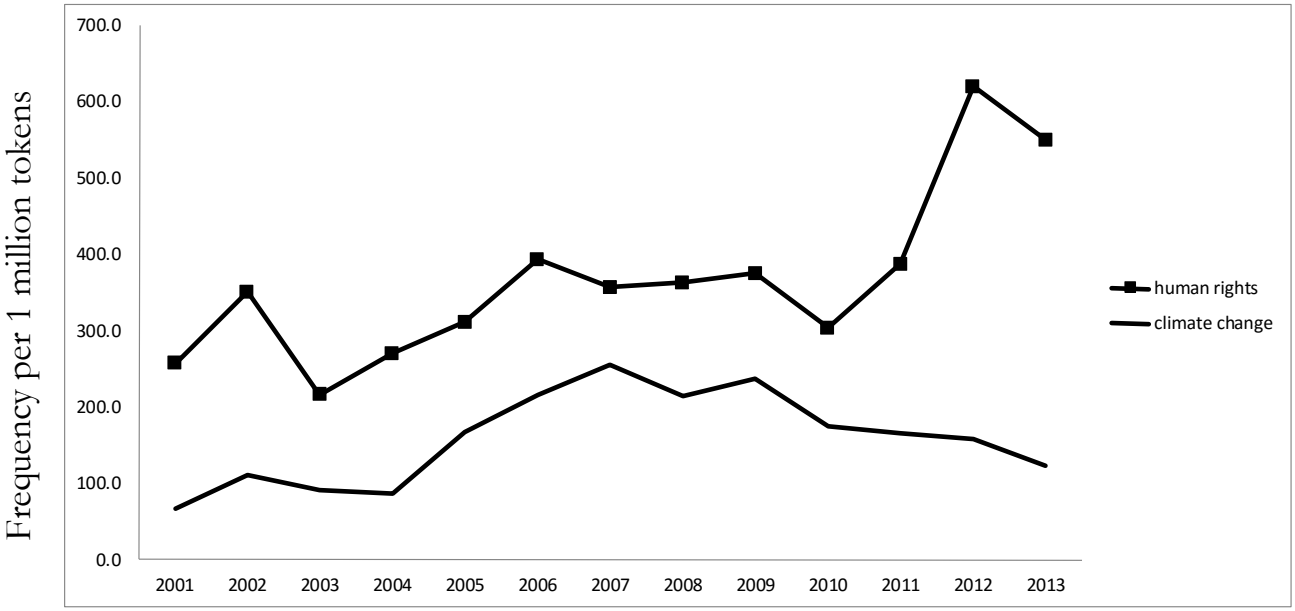


Figure 2: Corporate Social Performance (CSP) vs. Corporate Financial Performance (CFP): Dichotomy in Trends



This is a pre-publication version accepted for publication in Applied Linguistics, 2016. Please refer to the published version of this article if you wish to quote from it.

Figure 3: Distribution of 'human rights' and 'climate change'



Appendix 1: LDA formula (as provided by Underwood 2012)

$$P(Z|W,D) = \frac{\#\ of\ word\ W\ in\ topic\ Z + \beta_w}{total\ tokens\ in\ Z + \beta} * (\#\ words\ in\ D\ that\ belong\ to\ Z + \alpha)$$

Appendix 2: Example of Mallet outputs

<topic id="5" alpha="0.012005268619224773" total Tokens="18557" titles="climate change, scope emissions, principle, fully, reporting, reduction, flaring, initiatives, power, carbon">
<word weight="0.025219593684323974" count="468">emissions</word>
<word weight="0.011909252573152987" count="221">scope</word>
<word weight="0.01066982809721399" count="198">principle</word>
<word weight="0.01061594007652099" count="197">fully</word>
<word weight="0.00948429164196799" count="176">climate</word>
<word weight="0.00921485153850299" count="171">reporting</word>
<word weight="0.008298755186721992" count="154">emission</word>
<word weight="0.007328770814247993" count="136">reduction</word>
<word weight="0.007274882793554992" count="135">flaring</word>
<word weight="0.007274882793554992" count="135">initiatives</word>
<word weight="0.007059330710782993" count="131">power</word>
<word weight="0.006897666648703935" count="128">change</word>
<word weight="0.00673600258662493" count="125">carbon</word>
<word weight="0.006574338524545993" count="122">direct</word>
…..

Appendix 3: List of studied companies

BP, CHEVRON, CNOCC, ENI, EXXON, GAZPROM, LUKOIL, MOL, NORSK, OMV, ONGC, PETROBRAS, PETRONAS, PKN ORLEN, PTT, REPSOL, SASOL, SHELL, SINOPEC, STATOIL, TOTAL.