THE ART OF REPORTING: THE ACCURACY OF MEDIA COVERAGE OF NEW RELEASES BY THE CENTRAL BANK

Bachelor's Thesis
BSc^2 Econometrics / Economics

Rotterdam, July 3, 2016

Written by

J.R. van Berkel
371414

Supervised by

Dr. W.W. Tham
Dr. S. van Bekkum
M.J. van Dieijen

Erasmus School of Economics
Erasmus University Rotterdam


bank
articles
value
distance
communication
positive
based
overallmeasure
different
variable
identical
considered
hand
finding
won
cover
domestic
slightly
additionally
field
please
education
word
vice
european
name
crisis
research
advanced
university
textual
exception
agencies
overview
picture
banks
names
observed
informative
correlations
main
check
history
three
findings
explain
category
year
tuesdays
especially
better
implies
dice
concerning
conjecture
fig:dictionaryapproach
lower
days
either
wall
quality
limitations
constructed
far
colleagues
like
regressions
limited
wsj
obtained
using
harvard
scale
data
dieijen
common
rise
denotes
section
document
smaller
contextjs
database
many
scaled
size
10-k
space
made
results
make
covered
differ
reports
second
still
nature
sources
variables
intervention
ecb
respective
aim
local
uses
amount
set
finds
award
calculate
online
another
might
reporting
total
specifically
selected
r
authors
list
lexical
best
compute
product
modonald
majority
shows
done
appearing
top
occurring
models
term
note
categories
author
written
give
fact
writing
risk
degree
appendix
report
seen
effect
text
covers
displayed
ft
paper
gender
power
full
leads
highest
often
week
explanatory
perform
specific
times
way
scores
favor
relevant
length
series
political
swiss
male
see
printed
strong
others
reserve
feature
economy
lead
true
kl
dictionaries
markets
proxy
case
several
vectors
comparison
longer
refer
investigate
although
cases
smallest
two
countries
show
journalism
take
retrieved
due
uncertainty
compare
much
provided
accurate
texts
dependent
news
eurozone
experience
role
government
fed
documents
conclusion
largest
earnings
given
performed
takes
purpose
shanghai
ranking
months
important
order
equal

I would like to express my gratitude to professor W.W. Tham and M.J. van Dieijen for sharing their dataset, their involvement in the project and their helpful feedback throughout the research period.

Abstract

New information released by central banks is of vital importance in today's economy. By comparing over 80,000 media articles with the statement and speeches they cover, this paper investigates the drivers behind the accuracy of media coverage. Several measures of accuracy are computed and related to a series of article and journalist characteristics. The dictionary based measures are most informative in this context. Evidence is found that news agencies tend to sensationalize; more in printed than in online form. A broad series of journalist characteristics is relevant for predicting reporting accuracy.

Contents

1 Introduction

New information releases by central banks to the market play an important role in the current financial system. The availability of new information enables investors to optimize their investment decisions. Over the last decades central banks have become more transparent (Blinder et al., 2008). The underlying rationale is that central banks which are more open can be held more accountable. Transparency helps managing expectations, an important component of today's central banking. Till recently, the main focus of research was the direct impact of central bank communication on financial markets. This focus implicitly ignores the role of information intermediaries like media on markets. Ahern and Sosyura (2015) find the media to have a large impact on the financial markets in the context of rumors on mergers and acquisitions. A first aim of this paper is to investigate what drives the accuracy of media coverage on central bank communication.

Due to the vast amounts of available central banking communication and corresponding media coverage, it is not feasible to read and classify all texts manually. Instead, machine power could be used to quantify certain aspects of the text. This renders a second aim of this thesis: to design several machine generated measures of lexical similarity. Combining the two aforementioned goals leads to the following research question: "how do article and journalist characteristics relate to the accuracy of media coverage on new releases by the central bank?" The role of the media has, to my knowledge, not yet been subject to investigation in the context of central banking communication.

In order to answer this research question, a large dataset of over 80,000 media articles, together with 3,000 speeches and 800 statements issues by central banks in eight economies will be used. Several measures of lexical distance are examined over the whole dataset. Two main approaches are being employed here: on the one hand several dictionaries will be used, while on the other hand a term based approach is being implemented. Additionally, a series of journalist characteristics on, among others, education and working history is added to the database. By means of regression analysis the effects of the journalist characteristics on the different measures of reporting accuracy are distillated.

The remainder of this paper is structured in the following way. Section 2 elaborates on the history of quantitative textual analysis in economic research. Section 3 discusses how reporting accuracy can be measured in further detail. Section 4 elaborates on the data set which has been used and gives more information on the journalists that have been found and the corresponding characteristics that have been retrieved. Section 5 gives an accessible overview of the accuracy measures over different categories, while section 6 expands the analysis of the previous section with journalist dependent characteristics. The main results of this paper are summed up in the section 7. Finally, section 8 presents limitations to the research and suggestions for further research.

2 Textual Analysis in Economic Research

Textual analysis is a common tool to quantify large amounts of qualitative financial data, such as 10-K reports and newspaper coverage. Li (2006) was one of the first ones to apply textual analysis on financial reports. By measuring the occurrence rate of words related to risk and uncertainty in 10-K fillings, the researcher constructs a proxy for the risk sentiment associated with a certain company. Li (2006) finds that a higher risk sentiment leads to lower earnings in the following year.

Tetlock et al. (2008) measure the direct effect of qualitative verbal information on stock returns. However, they take a more general approach than Li (2006). Instead of focusing on the words related to risk and uncertainty, the researchers use the Harvard-IV-4 Psychosocial Dictionary in order to classify negative words in general, and calculate the proportion of negative words in the financial media to predict firm earnings. A higher proportion of negative words is associated with lower firm earnings. Moreover, stock prices are found to incorporate this information as well. In addition, Tetlock et al. (2008) distinguish between the stories which touch upon firms' fundamentals, and those who do not. The occurrence rate of negative words in stories related to fundamentals is relevant for earnings as well as returns. The researchers overall find support for the conjecture that the media capture qualitative aspects of firms' fundamentals. Gurun and Butler (2012) apply textual analysis in a similar way as Tetlock et al. (2008) did in their research. Gurun and Butler (2012) investigate local newspaper coverage of local companies and compare this to coverage of non-local companies by the same newspapers. The researchers find that news about local companies contains relatively less negative words than news on non-local companies.

A more recent application of textual analysis can be found in the research by Loughran and McDonald (2014). Instead of classifying words as positive or negative, they focus on the difficulty and readability of a text. The authors argue that one of the most common measures used to measure the readability of a text – the Fog Index – is not applicable to financial reports. Two reasons are being advocated. The first one is the occurrence of sesquipedalian words, which are not necessarily difficult to read in financial contexts. The second reason of its limited applicability in a financial context is due to the fact that most texts are written using a similar writing style.[1] Loughran and McDonald (2014) find the length of 10-K documents to be far more informative. This proxy is less prone to measurement errors, easy to calculate and turns out to be more informative for investors.

Ahern and Sosyura (2015) take a different methodology in the way they analyze texts. The researchers analyze the accuracy of media coverage of merger rumors. The researchers deploy textual analysis in order to construct a proxy for the concreteness of a text. They do this by counting the proportion of weak model words occurring in the relevant newspaper article. Together with a series of newspaper and journalist characteristics this measure is being used in order to predict the probability that a rumor comes true. Ahern and Sosyura (2015) find evidence that a higher proportion of weak modal words leads to a lower probability of the rumor coming true.

[1] The Securities and Exchange Commission's plain English initiative, which provides guidelines on how financial disclosures ought to be written, is of particular relevance here.

3 Quantifying Media Coverage Accuracy

In order to allow for a quantitative comparison of texts, two main approaches are deployed in this research. On the one hand, a dictionary based approach will be put into practice. On the other hand, a term based approach will be used as metrics for textual similarity. Both approaches look at reporting accuracy in a different way. The dictionary approach allows for comparison of sentiment: two texts are similar when the proportions of negative, positive and uncertainty related words are equal. For the term based approach, coverage is accurate when the words being used are occurring with the same frequency in coverage and source. Figure 1 gives a graphical overview of the different distance measures which are subject to investigation in this research.

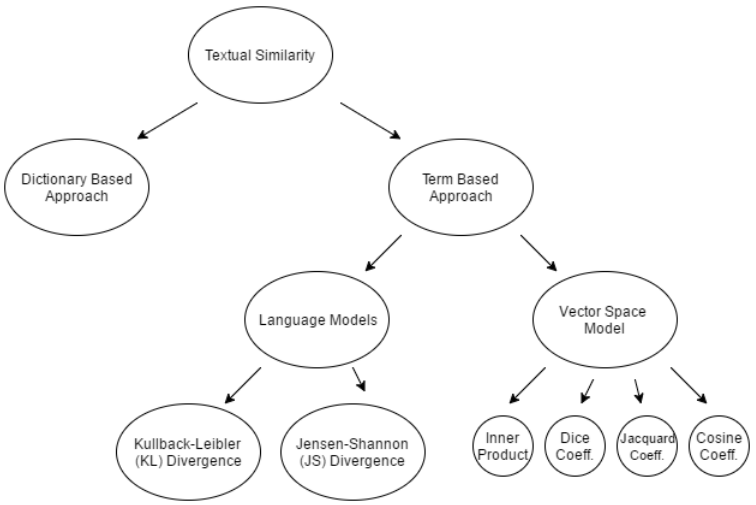

Figure 1: Overview of different measures for textual distance. Please note that the node on the dictionary based approach will be discussed in further detail in the following section.

3.1 Term Based Approach

The term based approach comes from the field of computer science, in which texts are often being analyzed from a different perspective for automated searching algorithms. This approach can be based on a language modeling concept in which probabilistic models of language generation are developed. The very fundamentals of this type of models go back to the early 1900s in which Andrey Markov designed Markov models in order to describe letter sequences (Hiemstra, 2009). A similar approach become common ground during the 1980s, at the beginning of the digital revolution (Hiemstra, 2009). The language modeling approach is often of crucial importance in many information retrieval or machine learning systems in which novelty control is performed (Bigi, 2003). Another approach which can be followed is the vector space model, which is done by performing mathematical operations on the document feature matrix.

3.1.1 Vector Space Model

When using a vector space model, the document or text of concern is represented as a vector of weights (Verheij et al., 2012). In order to compare these two vectors with each other, several measures have been designed. Let y_X denote the vector representation of document X, where $X \in \{A, B\}$ and A and B describe two documents. Several different similarity coefficients have been proposed in the literature; all of which are scaled versions of the inner product and, should hence be independent of document length. There is no broad consensus over which measure performs best overall (Thada and Jaglan, 2013). Thada and Jaglan (2013) and Bullinaria and Levy (2007) find the cosine coefficient to outperform the others in their researches.

- **Cosine Coefficient** One of the measures used for measuring distance between two documents is the cosine coefficient. Equation 1 shows how it is calculated.

$$cos(A, B) = \frac{y_A \cdot y_B}{||y_A||\ ||y_B||} \tag{1}$$

  Since this measure can be interpreted as a cosine, a clear upper and lower bound are present. However, in this specific textual framework, only nonnegative values of the cosine measure can be obtained. A value of 1 indicates that both document vectors are exactly identical, while 0 indicates the opposite. Another property of the cosine coefficient is its relationship with the Euclidean distance (see equation 2).

$$||y_A - y_B||^2 = 2(1 - cos(A, B)) \tag{2}$$

- **Jacquard Coefficient** The Jacquard coefficient is in essence similar to the cosine coefficient. While the numerators are identical, the denominators differ. A value of +1 indicates that both document vectors are exactly identical, while 0 indicates the opposite. The formula for the Jacquard coefficient can be found in equation 3.

$$Jacquard(A, B) = \frac{y_A \cdot y_B}{||y_A||^2 + ||y_B||^2 - y_A \cdot y_B} \tag{3}$$

- **Dice Coefficient** The Dice coefficient is of similar nature as the aforementioned and displayed in equation 4.

$$Dice(A, B) = \frac{2y_A \cdot y_B}{||y_A||^2 + ||y_B||^2} \tag{4}$$

Due to their similar numerators and denominators the Jacquard and Dice distances perform rather similarly. The difference with the cosine distance is in the way repetitions are being treated. The cosine measure is rather indifferent about these, while the Jacquard and Dice distances are both indifferent between repeating the same word another time or a completely new occurrence (please refer to Appendix C for more details on this difference).

3.1.2 Language Models

Besides the vector space model, language models provide a good starting framework for analyzing the similarities between texts as well. Language models essentially compute the probability that a certain word occurs in a piece of text (Hiemstra, 2009). To be more specific, language models compute the probability that a random term T is picked from document D, $P(T|D)$. Let n describe the total number of different terms occurring in a document, then language model Θ is displayed by the vector $(P(T_1|D), P(T_2|D), ...P(T_n|D))$. The language model describes in this way the relative frequency of every word in the text to occur (Hiemstra, 2009).

In order to compare two texts with each other, one should construct two language models Θ_1 and Θ_2, each describing one text. Two commonly used measures in order to compare two texts based on these two functions are the Kullback-Leibler (KL) divergence and the Jensen-Shannon (JS) divergence. Note that Verheij et al. (2012) find the KL divergence measure to outperform the other.

- **Kullback-Leibler Divergence** The Kullback-Leibler divergence measure[2] is an asymmetric distance measure, which implies that the distance from document A to B is not necessarily equal to the distance from document B to A.

$$KL(\Theta_A|\Theta_B) = \sum_{i=1}^{n} \Theta_A(i) log \frac{\Theta_A(i)}{\Theta_B(i)} \tag{5}$$

  In equation 5, Θ_X represents the language model based on document X and $\Theta_X(i)$ denotes the probability of term T_i in document X. One potential problem with this measure is the term $\Theta_B(i)$ which takes value 0 when the respective term does not occur in document B. This leaves a non-finite number for the KL divergence measure. A common solution to tackle this problem is linear interpolation smoothing (Verheij et al., 2012). The mathematical representation for this method is displayed in equation 6.

$$\Theta_B(i) := \lambda\Theta_B(i) + (1 - \lambda)\Theta_{1,...,n}(i) \tag{6}$$

  In equation 6, λ denotes an unknown value between zero and one which must be optimized, while $\Theta_{1,...n}(i)$ describes the probability of term T_i in aggregated documents $1, ...n$. Due to storage limitations I have aggregated all documents for speeches and statements separately. As suggested by Fernández (2007), a λ value of 0.9 has been used for the smoothing.

- **Jensen-Shannon Divergence** A second commonly used measure in order to compare Θ_A with Θ_B, is the Jensen-Shannon (JS) divergence measure[3]. Contrary to the KL distance, the JS distance is symmetric. It can be considered to be a smoothed version of the KL divergence (Verheij et al., 2012).

[2]The programming code used to compute the Kullback-Leibler measures is included in Appendix B.
[3]The programming code used to compute the Jensen-Shannon measures is included in Appendix B.

$$JS(\Theta_A|\Theta_B) = \frac{1}{2}KL(\Theta_A|M) + \frac{1}{2}KL(\Theta_B|M), M = \frac{1}{2}(\Theta_A + \Theta_B) \quad (7)$$

In case of two identical documents, both the JS and the KL divergence measures will take value 0. This implies that they are measured on an inverted scale compared to the measures derived from the vector based approach.

3.2 Dictionary Based Approach

Besides the term based approach, one can also employ a dictionary based approach in order to compare texts with each other. A dictionary is a pre-specified list of words which categorizes all entries (Apel and Grimaldi, 2012). Words can, for example, be categorized as 'negative' or 'positive'. By conducting an automated 'search-and-count-words' approach, a numerical score can be computed.

3.2.1 Relevant Dictionaries

One of the most commonly used word classifications is the Harvard Psychosociological Dictionary (Loughran and McDonald, 2011).[4] In their research on textual analysis, Loughran and McDonald (2011) describe that the quality of textual analysis fully depends on the categories included in the word classification and to what extent these categories match the purpose of the research. The Harvard Dictionary contains 182 tag categories among which two large valence categories containing 'positive' and 'negative' words. It also contains ten semantic dimensions, a broad range of emotions, institutions, roles, people, animals, places, objects, motions and verbs.

One problem with the Harvard Psychosociological Dictionary is that it is not specifically designed for financial purposes (Loughran and McDonald, 2011). This is problematic since many words which are generally considered to be negative in a non-financial context, are not necessarily negative in a financial context. Words like *tax* and *liability* often carry a negative association, while they generally bring a different message across in a financial text. In order to accommodate for this crucial difference, Bill McDonald and Tim Loughran have created their own dictionary[5] based on a large database of 10-K fillings. The word list that is constructed is specific to financial terminology and should allow for a better analysis of the respective texts. The researchers have identified a series of categories which are relevant in a financial context. Whereas the original Harvard dictionary contains as much positive as negative words, the financial dictionary has less positive entries than negative ones: 354 positive versus 2355 negative. Another category that might be relevant for this research is the 'uncertainty' classification, containing 297 words. These three categories all contain relatively many words and render sufficient non-zero proportions.

[4]The relevant Harvard-IV-4 TagNeg (H4N) file can be downloaded via the following link `http://www.wjh.harvard.edu/~inquirer/homecat.htm`.

[5]The complete Loughran and McDonald (2011) dictionary can be downloaded from the web page of Bill McDonald at `http://www3.nd.edu/~mcdonald/Word_Lists.html`. A more detailed description of the dictionary and its development can be found via this link as well.

Another well-known dictionary is that of Laver and Garry (2000)[6]. The researchers quantify policy positions of political parties based on political texts. The dictionary they use has specifically been designed to fit the British political framework. The dictionary consists of nineteen different policy categories in eight different classes. Economy, institutions, values and groups are some of the classes in the dictionary. The economy related class could very well be informative in the central banking context of this research and consists of three categories: one which identifies words in favor of government intervention, one relating to an equal amount of government intervention and one related to less state intervention. Combining these three categories, the total number of policy related words, might be of special interest in this research.

3.2.2 Applying Dictionaries

By using the aforementioned dictionaries, one can calculate the average proportion of positive or negative words and use these measures to quantify the tone of the media coverage. In a similar fashion one could use these proportions to compute the tone of the respective central bank communication. The same thing can be done with the uncertainty and different policy scores scores. The scores of the newspaper articles and the scores of the central bank communication can be compared to each other in order to verify whether the media accurately report on new central bank releases. The following list enumerates some of the measures that will be used to investigate whether the media cover in the same tone as the central bank releases the communication. The proportion of words from a certain category in the central bank communication will be subtracted from the relevant proportion in media coverage.

1. **Difference in proportion of negative words.** A comparison of the proportion of negatively classified words in the central banking communication and the corresponding media coverage, values are stored in `d_prop_neg`. Due to the media's inclination to sensationalize (Ahern and Sosyura, 2015), a larger proportion of these negative words is expected to be present in the media coverage compared to the central banking communication. For this and the following proportions, the proportion of words in central banking communication is subtracted from the proportion in media coverage as displayed in Equation 8.

$$\texttt{d\_prop\_neg} = \texttt{prop\_neg\_media} - \texttt{prop\_neg\_cb} \tag{8}$$

2. **Difference in proportion of positive words.** A comparison of the proportion of positively classified words in the central banking communication and the corresponding media coverage; this difference is stored as `d_prop_pos`. The inclination to sensationalize will draw into two directions here. On one hand the emphasis on negative issues will lead to an overall decrease in positive word usage, while on the other hand media might be likely to exaggerate a positive tone as well.

$$\texttt{d\_prop\_pos} = \texttt{prop\_pos\_media} - \texttt{prop\_pos\_cb} \tag{9}$$

[6]The .cat file containing the Laver and Garry (2000) dictionary can be found at `http://www.kenbenoit.net/courses/essex2014qta/LaverGarry.cat`

3. **Difference in proportion of uncertainty related words.** The Loughran and McDonald (2011) dictionary can also be used to count the number of words related to uncertainty in a financial context. Due to the nuanced and deliberate communication of the central bank, a higher proportion of uncertainty related words is expected there. The difference in this proportion is saved as `d_prop_unc`.

$$\texttt{d\_prop\_unc} = \texttt{prop\_unc\_media} - \texttt{prop\_unc\_cb} \tag{10}$$

4. **Difference in tone.** The tone of an article is computed as the difference between the proportion of negative and positive words (please see Equation 11). The difference in tone between the media coverage and source might be a relevant indicator of reporting accuracy, and is caught as `d_tone`.

$$\texttt{d\_tone} = \texttt{d\_prop\_pos} - \texttt{d\_prop\_neg} \tag{11}$$

5. **Difference in proportion of economic policy words.** For all three different policy related categories a difference in proportion can be calculated (`d_prop_less` for less government intervention, `d_prop_eq` for the same amount and `d_prop_pro` for more). Additionally, the sum of the three is equal to `d_total`.

4 Data and Summary Statistics

The dataset which is being used consists of 82,783 newspaper articles[7]. Two thirds of these articles are covering central bank statements (846 central bank statements are included in the dataset) while one third is on speeches by either presidents (32,427) or vice presidents (2,332) of a series of central banks (3,359 of these speeches are incorporated in this research).[8] The central banks included in the analysis are the Reserve Bank of Australia, the Bank of Canada, the European Central Bank (ECB), the Bank of England, the Bank of Japan, the Swiss National Bank, the Federal Reserve in the United States (Fed) and the Reserve Bank of New Zealand. The distribution of the number of articles over the different geographical entities is shown in Table 1. One can see that the data set contains most articles covering the U.S. The Eurozone also receives a substantial amount of attention. The articles are selected from three media sources: the Financial Times (FT; 16,695 articles), Reuters (59,717 articles) and the Wall Street Journal (WSJ; 6,101 articles). The far majority of the media articles is from Tuesdays, Wednesdays and Thursdays, while some incidental statements have been issued on weekend days (see Table 1 for exact numbers). The number of media articles over the different months of the year is relatively constant (please refer to Table 10 in Appendix A). As shown in Table 1, the database includes more articles around the millennium change and during the financial crisis. The majority of 69,187 of the articles is from an online source, compared to 13,596 articles in printed media. On average the media coverage is much briefer than the central bank communication. Both the number of words and sentences in the media reports is smaller.

[7]Please note that a series of Wall Street Journal articles has been split up based on their relevance.

[8]This dataset was provided by dr. W.W. Tham and M.J. van Dieijen. Media coverage is marked as relevant when the article contains one of the central bank names, a relevant central bank policy word, no mentioning of a company-finance related word and contains the right timing. A series of long Wall Street Journal articles has been split in order to make sure relevant coverage is filtered.

Table 1: Number of media articles by category.

Region	Articles	Day	Articles	Year	Articles	Year	Articles
Aus	2086	Monday	5270	1996	429	2004	5367
Can	1476	Tuesday	24200	1997	2406	2005	4211
Ecb	15469	Wednesday	20295	1998	5192	2006	4686
Eng	7354	Thursday	27350	1999	9265	2007	6054
Jap	7354	Friday	5250	2000	8103	2008	10816
Swi	1019	Saturday	220	2001	5148	2009	7909
Usa	47260	Sunday	198	2002	7085	2010	2411
Zea	765			2003	3701		

The following article characteristics are included in the data set: `speech` denotes whether the article covers a speech (1) or statement (0). Whether speeches are expected to be covered more accurately than statements is an ambiguous question: on one hand speeches might come closer to 'regular' colloquial English, while on the other hand statements are easier for journalists to review; making it possibly easier to cover accurately. The three dummies `FT`, `WSJ` and `Reuters` indicate by which news agency the article is published. A dummy `print` takes value one for printed and value zero for online articles. Based on the rationale behind sensationalizing, one would expect articles in printed media to sensationalize stronger than articles which are only published online. For articles covering speeches, a `president` dummy is also included which distinguishes between speeches given by presidents and vice-presidents. The day of the week, the month and the year an article is published, together with the country it covers are included in the data set as well. During the financial crisis larger proportions of negative and uncertainty related words can be expected to be used by the financial press.

4.1 Journalist Names

The journalists of the articles have been selected by M.J. van Dieijen using named entity recognition. In this way, a list of possible journalist names can be extracted without manually checking all documents. This approach, however, also renders noise in the list of possible journalist names. In order to filter the true journalist names out of the list, a two-phase approach has been performed. First, all results consisting of just one word have been labeled as non-journalist. Most of these concern first names (e.g. Jonathan). Without any further information it will be practically infeasible to find more information on these. Also public figures which have certainly not written any articles are left out (e.g. Saddam Hussein). In case of doubt, these 'names' have been revisited in second stage, in which for every possible journalist I have entered their name in Google together with the relevant news agency. In almost all cases this clearly indicated whether the name was a real journalist (e.g. John Labate), another public figure (e.g. Stephen Jen, one of the world's best-known foreign exchange strategists) or not a name at all (e.g. Aussie Rules).

For 14,923 articles this has rendered one or more journalist names successfully. For the Financial Times this algorithm performs best (of 10,297 articles one or more authors has been found). For Reuters and the Wall Street Journal 2,707 and 1,919 author (teams) have been identified. This is a poor performance for especially Reuters, of which a much larger

number of articles is included in the dataset. In a substantial amount of cases (1,597) two or three authors have been found by the algorithm. This leads to 211 unique author names who have made 16,829 contributions in total. The ten most productive authors, which account for around 40% of the total number of contributions, are displayed in Table 2.

Table 2: The ten journalists which have made most contributions.

Rank	Author's Name	Contributions
1	Ralph Atkins	1005
2	Dave Shellock	992
3	Peter Garnham	985
4	Krishna Guha	760
5	Jennifer Hughes	602
6	Michael MacKenzie	589
7	Neil Dennis	486
8	Steve Johnson	441
9	Chris Giles	414
10	David Turner	358
	All Authors	16829

4.2 Journalist Characteristics

Based on the approach by Ahern and Sosyura (2015), several journalist characteristics have been collected. For this purpose, a wide variety of online sources has been used. LinkedIn is the most important one. For the majority of the journalists, their LinkedIn profile provided sufficient information for assigning values to all relevant variables. In case the LinkedIn profile did not offer enough information, or could not be retrieved, the online encyclopedia Wikipedia has been used. If reliable sources of information were cited, this information is used for this research. In case both LinkedIn and Wikipedia did not yield any information, the journalist's name was entered in Google in quest for more information on their personal and working histories. In some rare cases this provided an informative biography. The biographies found on news agency's websites generally tend to be rather uninformative and do not accurately fit the purpose of this research.

- **Years of Experience as a Journalist** The year the respective journalist started working as a journalist has been included in this dataset. Based on this year, the number of working years as a journalist can be computed per article. Most of the starting years are taken from LinkedIn, an approach which is not without any flaws. In some cases, a gap between their education and experience at one of the big news agencies is observed.[9] For 523 articles, a negative number of years of working experience was observed. In the final dataset, the `experience` variable is floored to zero.[10] The adjusted distribution can be seen in Figure 2. Ahern and Sosyura (2015) find journalists with more experience to report relatively more accurately.

[9]The respective journalist can fill in the information he or she would like to share themselves. It might very well be that they are not always willing to share less honorable jobs at the beginning of their careers. Moreover, LinkedIn is a relatively new platform and it might be that the journalists have not filled in their whole working history.

[10]Since it is not possible to take the logarithm of this variable I have checked what flooring at one and taking the logarithm would do to the analysis. Both times the results are, although minimally, worse than the flooring at zero.

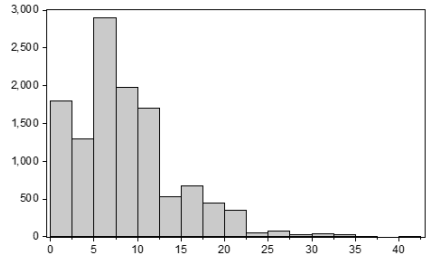

Figure 2: Histogram of the average number of years of experience the journalists have per article.

- **Banking Experience** For the dummy concerning banking experience the relevant working history the journalist has been viewed. In case the respective journalist had experience in the field of banking value one was assigned to this variable, zero otherwise. If the working history could not reliably been retrieved this variable takes value 'NA'. Table 3 displays the distribution of this and following variables. Value one has been assigned to `banking` when one of the authors has relevant banking experience.

- **Central Banking Experience** The dummy `centralb` takes value one when one of the journalists' expertise is particularly relevant to his or her reporting on central bank policy. Experience in central banking or being a professor in the field are strong indicators for this.

- **Banking or Central Banking Experience** Since the two aforementioned dummy variables have a large overlap, it could be of relevance to consider a dummy which takes value one if one of the authors has either banking experience or central banking experience. This assigns value one to `B or C` for 3,864 articles. Having thorough 'inside' knowledge about the banks or central banks is expected to improve reporting accuracy (Ahern and Sosyura (2015) find a similar variable to be borderline significant).

- **Gender** The article dummy `gender` follows the majority gender of the authors which has been determined based on the photos of the journalists and their names. At most three authors have been gathered per article. In case a male-female pair has written the article, this dummy has been assigned value one (male).[11] Ahern and Sosyura (2015) do not find any difference in reporting accuracy between the two genders.

- **Advanced Education** The article dummy `advanced` takes value one if one of the authors has a graduate degree and zero otherwise. In case of no educational information it takes value 'NA'. If no graduate degree has been reported (next to an undergraduate degree), value zero has been assigned. Journalists who have completed a graduate degree might report more accurately due to the possible presence of more accumulated in-depth knowledge.

- **Award-Winner** Winning an award for journalism can be considered as a sign of an exceptional set of skills (Ahern and Sosyura, 2015). In line with Ahern and Sosyura (2015), the Gerald Loeb Award, the SABEW (Society of American Business and Economics Writers) Award and the Pulitzer Prize are into account. To check whether the respective journalist has won any of these awards an overview of all nominees and

[11] In 3.5% of the articles for which authors have been found the contributors formed a mixed pair or trio.

winner of the first two awards and all winners (in the journalism categories) of the Pulitzer Prize has been constructed. The article characteristic `award` takes value one if one of the contributing authors could be found in this overview. Please note that the award winners are almost exclusively reporting for the WSJ. Ahern and Sosyura (2015) do not find this variable to be related to reporting accuracy.

Table 3: An overview of the journalist characteristics by article (all 82,783 articles). If either no journalist could be identified or the relevant information could not be retrieved, this has been marked by 'NA'.

	Banking	Central B.	B OR C	Gender	Advanced	Award	Domestic	Shanghai100	Undergraduate
NA	71012	71042	71008	68019	71911	67856	72062	71820	72272
0	8353	8598	7911	2621	2515	13631	6912	4512	3002
1	3418	3143	3864	12143	8357	1296	3809	6451	7509

- **Domestic** Information on where the respective journalist comes from has been retrieved as well. Based on this, one can determine whether the author is domestic or not. The far majority of journalists comes from the U.S., followed by the U.K. and the other anglophone countries. The remainder of journalists comes from the European continent. LinkedIn is does not report which nationality the respective journalist has. Based on name, high school or working history in many cases a reliable estimation has been made. In case of serious doubts 'NA' has been noted. Since there is no Japanese journalist in our database, media coverage of Japanese central bank communication is solely done by non-domestic journalists. For the European Central Bank, a domestic journalist has been defined as a journalist which is from any of the current Eurozone countries. The variable `domestic` takes value one when either of the journalists is domestic. Domestic journalists might cover the central bank communication from their respective countries more accurately than their non-domestic colleagues. On the other hand, however, this might also result in a larger inclination to sensationalize.

- **Shanghai Ranking 100** In order to check the quality of the undergraduate university, the rank of the respective university in the overall 2015 Shanghai ranking has been retrieved as well. The dummy variable `Shanghai100` takes value one when one of the journalists has obtained an undergraduate degree at a university which is in the top 100 of the 2015 ARWU Shanghai University ranking. A slight majority of the articles falls into this category. Note that Ahern and Sosyura (2015) do not find journalists with a higher quality of their college (as measure by SAT scores) to report more accurately.

- **Relevant Undergraduate Degree** The undergraduate degree is also recorded per journalist. If one of the authors of an article has one or more majors in business, economics, journalism, English, international relations or finance, a dummy variable `undergraduate` takes value one. There is a substantial number of cases in which no information could be retrieved. Ahern and Sosyura (2015) find journalists with a relevant undergraduate degree to report more accurately.

5 Accuracy Measures at a Glance

5.1 Distribution of Term Based Measures

The Jacquard, Dice and cosine textual distance measures have been computed over the full sample of media articles and their corresponding central bank speeches and statements (please refer to Figure 3 for their histograms). Since stop words do not contain much relevant information, these have been sorted out before computing the document feature matrices.[12] For all three measures, a χ^2-like distribution is being observed. The values stay, as predicted within the feasible range between 0 and 1. On average, the values of the cosine measure are larger than the values for the other two.

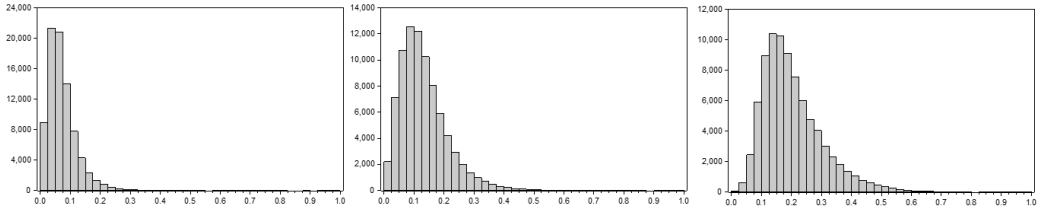

Figure 3: Distribution of the Jacquard (left), Dice (middle) and cosine (right) distance measures.

For all media articles and the corresponding central bank communication, the KL and JS divergence measures have been computed. Again, stop words have been omitted before computing the document feature matrices of each document. The numerical values of both divergence measures are not very informative. Please note that the dimension of the JS divergences is also being found by Grosse et al. (2002).

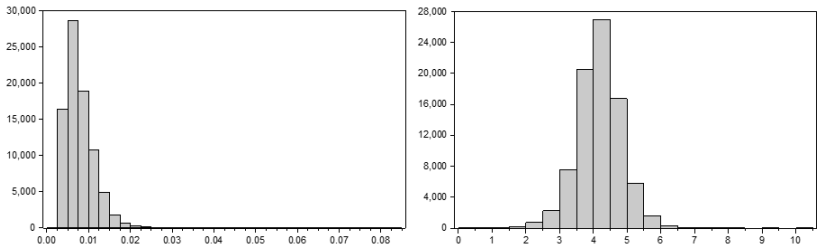

Figure 4: Distribution of the Jensen-Shannon (left) and Kullback-Leibler (right) divergence measures.

In Table 4 an overview of the correlations between the different lexical distance measures is shown. All correlations are significant on a 1% level. The Dice and Jacquard measures have strong positive correlations, which is, given their mathematical similarity, as expected. The cosine measure shows a less strong correlation. The observed inverse relation between the measures derived from the language model and vector space model is due to the inversed

[12]For this purpose, the `stopwords` function in the `quanteda` package for R has been used. For more information on this function, please refer to page 66 of the `quanteda` manual: `https://cran.r-project.org/web/packages/quanteda/quanteda.pdf`.

scale they use. A higher distance between two documents measured by the cosine measure leads, on average, to a higher distance by the KL divergence as well. The strong correlations with the inner product give rise to the hypothesis that also the other measures are not completely independent of text length.

Table 4: An overview of the correlations between the different lexical distance measures within the term based approach. All corelations are signficant on a 1% level.

	Inner	Dice	Jacquard	Cosine	JS	KL
Inner	1.00					
Dice	0.18	1.00				
Jacquard	0.21	0.99	1.00			
Cosine	0.62	0.68	0.68	1.00		
JS	-0.39	-0.04	-0.04	-0.23	1.00	
KL	-0.58	-0.44	-0.44	-0.76	0.41	1.00

5.2 Distribution of Dictionary Based Measures

The two dictionaries have been applied to the full sample of central bank communication and media coverage[13], the distributions of these differences can be seen in Figure 5. For `d_prop_neg`, the majority of the values is positive, while for `d_prop_pos` the average is negative. As a result, the overall difference in tone is negative, confirming the conjecture that media tend to report negatively. Another interesting observation is the larger standard deviation and higher kurtosis values for the `d_prop_neg` series compared to its positive counterpart. The negative averages for the upper row in Figure 5 show that, on average, media tend to use smaller proportions of economic policy related words than the central bank communication. Moreover, the media are inclined to use smaller proportions of uncertainty related words than central banks do in their communication. This might hint on central banks being more nuanced in their communication than its media coverage.

In Table 5, the correlations between the different dictionary proportions are shown. For the three Laver and Garry (2000) categories, a negative correlation between the categories in favor of and against government intervention in the economy is observed. This shows that there seems to be some ground for categorizing articles into according to being in favor or against government intervention. The category for a stable amount of government intervention is positively correlated with both other categories. Additionally, a positive, significant correlation is observed between the proportions of negative and positive words. This implies that articles with a higher proportion of negative words, on average, have higher proportions of positive words as well. The correlations of both categories with the proportion of uncertainty related words is also found to be positive.

[13]The programming code that has been used for this purpose can be found in Appendix B.

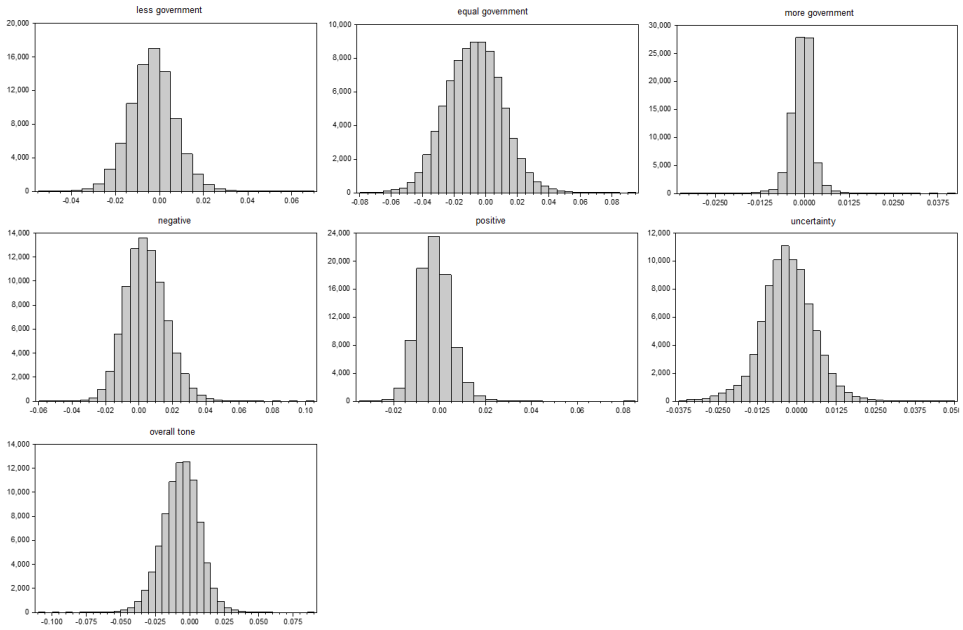

Figure 5: Distribution of the differences in proportion of words in the several categories. The bottom histogram displays the difference in overall tone.

Table 5: An overview of the correlations between the different sentiment differences in the dictionary based approach. The *** and ** indicate significance on 1% and 5% levels respectively.

	d_prop_less	d_prop_eq	d_prop_pro	d_prop_neg	d_prop_pos	d_prop_unc	d_tone
d_prop_less	1						
d_prop_eq	0.143***	1					
d_prop_pro	-0.08***	0.108***	1				
d_prop_neg	0.036***	0.092***	0.081***	1			
d_prop_pos	0.007**	-0.021***	0.158***	0.130***	1		
d_prop_unc	0.100***	-0.086***	0.001	0.265***	0.078***	1	
d_tone	-0.030***	-0.096***	0.008**	-0.852***	0.408***	-0.203***	1

5.3 The Effect of Article Characteristics

On average, the coverage of central bank communication is more negative than the original central bank communication (see Figure 5). This does however not hold to the same extent, if at all for all subcategories in the sample. Regression analysis is performed in order to obtain averages for different accuracy measures over different article characteristics. In none of the regressions the inner product qualifies as a measure of textual distance. As predicted, the inner product depends almost solely on text length and does not show stable outcomes.

5.3.1 Speeches versus Statements

Although media coverage overall is more negative than the source, this difference is much more prominent for statements than for speeches. Table 6 shows regressions on the differences in categorical proportions. A negative value implies that the financial press uses less words of this category than the central bank, while a positive value demonstrates the opposite. As can be seen in Table 6, statements are covered using more negative words than speeches. However, the difference in proportion of positive words for speech coverage is much larger, indicating that speeches tend to be covered by much less positive words than statements. The overall tone of speeches is slightly less negative than the overall tone of statements. Speeches tend to be covered with a slightly higher proportion of financial uncertainty related words, although still less than the central banking communication has. When examining the non-dictionary based textual similarity measures, the Dice and Jacquard distance measures show speech coverage to be less similar than statement coverage, while the cosine and language model measures conclude the opposite.

Table 6: Different textual similarities in relation to speeches or statements (reference category) and to the different news sources (Reuters is the reference category). All estimated coefficients are significant on a 1% level. Please refer to Table 11 and 12 in Appendix A for more numerical details.

	d_prop_neg	d_prop_pos	d_prop_unc	d_tone	Dice	Jacquard	Cosine	JS	KL
constant	0.0063	-0.0004	-0.0031	-0.0067	0.1433	0.0792	0.1920	0.0092	4.3323
speech	-0.0051	-0.0045	0.0005	0.0006	-0.0260	-0.0147	0.0269	-0.0031	-0.3578
	d_prop_neg	d_prop_pos	d_prop_unc	d_tone	Dice	Jacquard	Cosine	JS	KL
constant	0.0021	-0.0027	-0.0037	-0.0048	0.1270	0.0698	0.1968	0.0083	4.2178
FT	0.0083	0.0014	0.0033	-0.0068	0.0190	0.0116	0.0263	-0.0011	-0.1470
WSJ	0.0039	0.0004	0.0019	-0.0035	0.0201	0.0122	0.0153	-0.0021	-0.0757

5.3.2 News Sources

All news agencies use smaller proportions of policy related words than central banks in their communication (please see Table 12). This holds for all three categories from the Laver and Garry (2000) dictionary investigated. Overall, Reuters articles by Reuters have the smallest proportion of policy related words, followed by the WSJ. The FT uses the highest proportion of policy related words. Although all three agencies use more negative words than central banks, the difference is the largest for the FT, followed by WSJ and Reuters respectively. The same holds for the difference in proportion of positive words: the FT uses most of

them, followed by WSJ and Reuters. In the overall tone, the FT deviates most in tone from the source, followed by WSJ and Reuters. These findings give rise to the conjecture that the FT tends to sensationalize more than the other news agencies.[14] The WSJ also uses more uncertainty related words than the FT. Reuters uses least of them, leaving the gap with central bank communication the largest. All term based measures point into the same direction concerning accuracy of reporting: Reuters is least accurate, while the FT and WSJ perform better. Not all measures rank the FT and WSJ in the same order.

5.3.3 Regional Differences

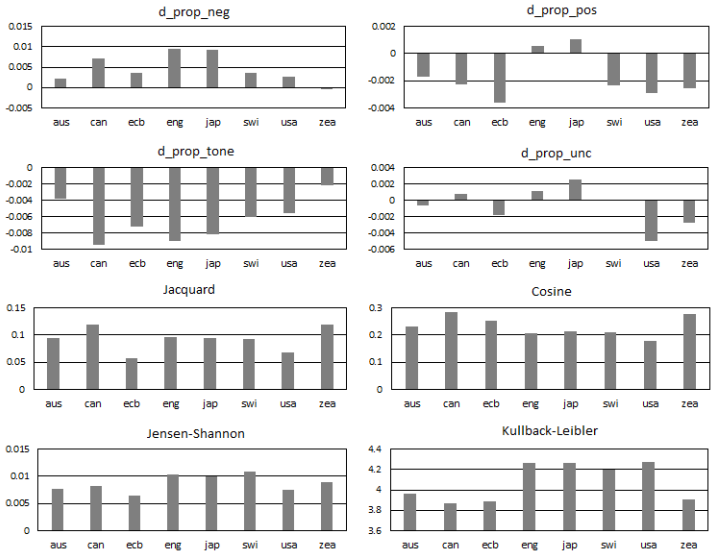

Figure 6: The average values of comparative measures by country. Results are obtained from regressions using the U.S. as a reference category. In four occasions countries are not significantly different from the U.S.: New Zealand in `d_prop_pos`, Switserland in `d_tone`, Canada in `JS` and England in `KL`.

For almost all countries media coverage contains proportionally less policy words than the central bank communication. The notable exception in this framework is the European Central Bank, whose coverage has a higher proportion of policy related words (see Table 13 for the regression results). When it comes to the proportion of negative words used in media coverage only the Reserve Bank of New Zealand is covered accurately by the media (i.e., the difference in proportion of negative words does not significantly differ from zero). As shown in Figure 6, all other central banks receive more negative words, with the English and Japanese central banks worst off. Au contraire, while covering the English and Japanese central banks more positive words are used as well. The ECB is worst off in this aspect. Overall, media coverage has a more negative tone than the central bank communication has in all countries. This effect is strongest for Canada and least strong for New Zealand. The Fed is understated most when it comes to the proportion of uncertainty related used, while

[14] These findings remain significant when correcting for source in which the article is published (print or online). For Reuters, only online published articles are included in the dataset.

the Swiss central bank is covered in the media with the same proportion of uncertainty related words as it uses itself in its communication. The other accuracy measures shed a different light on the picture. The Dice, Jacquard and cosine distance measures find the central banks of Canada and New Zealand to be covered most accurately. while the big European and American central banks are covered least accurately. The KL divergence measure comes to a similar conclusion. The JS divergences are largest for the Swiss central bank however. while the media coverage of the ECB and Fed is found to be most accurate.

5.3.4 President versus Vice-President

Media tend to cover central bank speeches less negatively when they are given by the president (compared to those given by the vice president). The Dice, Jacquard, cosine and KL distance measures all find evidence for the coverage of the president to be more distant from its source than the coverage of the vice president. The JS distance measure comes to the opposite conclusion.

Table 7: Different first part relates textual similarities to whether the speech is given by the president or vice president (reference category) of a central bank, while the second part compares printed and online (reference category) articles. The ***, ** and * indicate significance on 1, 5 and 10% levels respectively. Please refer to Table 14 and 15 in Appendix A for more numerical details.

	d_prop_neg	d_prop_pos	d_prop_unc	d_tone	Dice	Jacquard	Cosine	JS	KL
constant	0.0016***	-0.0050***	-0.0023***	-0.0066***	0.1260***	0.0697***	0.2625***	0.0065***	3.7855***
president	-0.0005**	2.40E-5	-0.0003**	0.0005*	-0.0093***	-0.0055***	-0.0467***	-0.0004***	0.2027***
	d_prop_neg	d_prop_pos	d_prop_unc	d_tone	Dice	Jacquard	Cosine	JS	KL
constant	0.0032***	-0.0025***	-0.0033***	-0.0057***	0.1301***	0.0717***	0.2003***	0.0081***	4.1949***
print	0.0057***	0.0010***	0.0024***	-0.0047***	0.0139***	0.0081***	0.0183***	-0.0012***	-0.0778***

5.3.5 Online versus Printed Media

Moreover, differences are being observed between printed and online media. Table 7 shows a series of regressions with the different accuracy measures as dependent variables and a constant and dummy indicating whether the articles is printed or not as explanatory variables. Printed media are inclined to use more policy related words than online media, but still less policy related words than their source. Moreover, printed media use more negative, positive and uncertainty related words than online media do. This leads to the overall tone of printed media to be more negative, giving rise to the conjecture that articles appearing in newspapers might be sensationalizing news more than those only appearing online. The measures based on the term based approach give a different picture, however. They all find articles appearing in printed media to have a smaller lexical distance to their source than online media.

5.3.6 Days of the Week

A comparison over the different days of the week gives a varying picture when it comes to the several categories of policy words (please refer to Table 16 in Appendix A for the regression results). Overall, the number of policy related words is smallest on Tuesdays and

Wednesdays, while the highest number of policy related words is found to be present during the weekends. As can be seen from Figure 7, the proportion of negative words in media coverage is smaller than that found in central banking communication on Sundays only. On all other days media coverage is more negative, with Friday having the highest proportion of negative words. On Mondays and during the weekends the lowest proportions of positive words are found, whereas Tuesdays and Thursdays see the highest ones. The term based measures do not give a univocal conclusion on distance measures over the days of the week. In all regressions displayed in Table 16, the standard deviation of the estimated coefficients for Saturday and Sunday is much larger than for the other days due to a smaller sample size during the weekend.

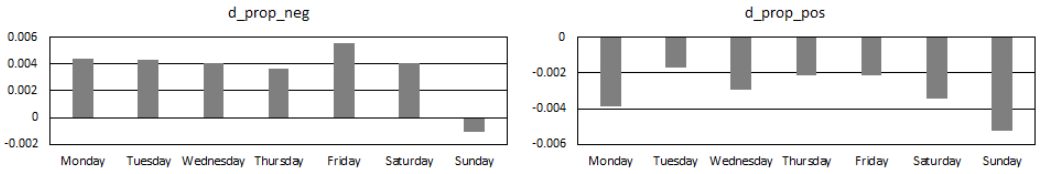

Figure 7: The average differences in proportion rendered by the Loughran and McDonald (2011) dictionary by day of the week. Results are obtained from regressions using Monday as a reference category. For `d_prop_neg`, Wednesday and Saturday and for `d_prop_pos` Tuesday and Saturday do not differ significantly from Monday.

5.3.7 Months of the Year

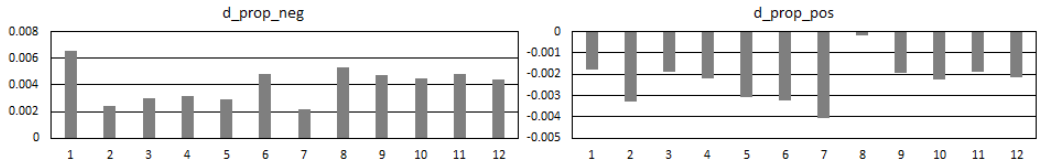

Figure 8: The average differences in proportion rendered by the Loughran and McDonald (2011) dictionary by month. Results are obtained from regressions using January as a reference category. For `d_prop_neg`, for all months the estimated coefficients differ significantly from zero. For `d_prop_pos`, March, September and November do not differ significantly from Monday. Please refer to Table 17 in Appendix A for the regression outcomes.

Monthly analysis shows January to be the month to have the highest proportion of negative words occurring in media coverage compared to central bank communication. All other months still have higher proportions of negative words than the central bank communication. The proportion of positive words does not exhibit any strong seasonal patterns over the year. For all months but August the smaller proportions of positive words are found to be present in the media coverage than the central bank communication. The overall tone difference is most negative in January. The difference in proportion of uncertainty related words and different term based distance measures do not exhibit any systematic monthly developments (please refer to Table 17 for more details).

5.3.8 Yearly Differences

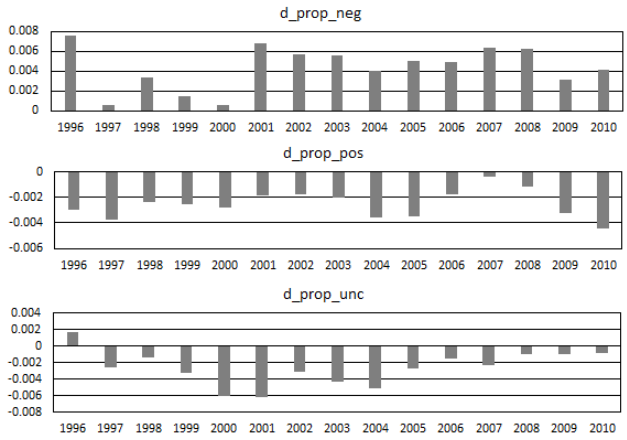

Figure 9: The average differences in proportion rendered by the Loughran and McDonald (2011) dictionary by year. Results are obtained from regressions using 2008 as a reference category. Please refer to the Table 18 in Appendix A for more ifnormation on the signficance of differences.

When comparing the difference in proportions of negative words over the years in the data set, one finds that during 1996, the early 2000s and the crisis years the media to use substantially higher proportions of negative words than the central bank. As can be seen in Figure 9, during 2007 and 2008, the proportions of positive words, on the other hand, is much closer to the proportions of positive words the central banks use (although still significantly different). Overall, no strong changes in tone are being observed during the crisis years. 1996 and the years after the crisis see higher proportions of uncertainty related words to be used. In all cases, these proportions are still smaller than in the central bank communication. The term based approaches do not render any consistent results.

6 The Effect of Journalist Characteristics

In order to investigate the effect of journalist characteristics, models have been constructed who explain the several textual distance measures by means of a series of independent variables which consist of both article and journalist characteristics. Due to the fact that for many of the articles in the database more than one journalist has been found, journalist characteristics are in practice article dependent. The different distance measures that have been constructed can be investigated using regression analysis. For each of the measure series I have constructed a model by means of a top-down approach in which I have omitted the variable with the least significant coefficient[15] one by one iteratively. This procedure has been done using both the `banking` and `centralb` dummies on the one hand and the `C or B` dummy on the other. Models have been selected based on the adjusted R^2-value.

[15] An exception on this rule has been the dummy for the `WSJ`. If this dummy were to be excluded, the `FT` dummy would loose (part) of its interpretation.

6.1 Dictionary Based Approach

Of the four constructed models, the one explaining the differences in proportion of negative words is best performing (measured by the adjusted R^2-value). Speeches are covered using a smaller excess of negative words than statements. Whereas the Wall Street Journal and Reuters write with a similar sentiment difference, the Financial times is found to report using a larger proportion of negative words. Articles appearing in printed media also exhibit larger proportions of negative words than similar articles which are only published online. As the authors of an article have more writing experience, they are, on average, also less inclined to use more negative words than the central bank communication that is being covered. Experience within banks or central banks have different effects: one of the authors having banking experience leads to a smaller proportion of excess negative words, while the reverse hold for central banking experience. Male and domestic authors tend to report using larger proportions of negative words compared to their source, while authors who have won an award or have a postgraduate degree tend to do the reverse. Having an undergraduate degree from one of the top 100 universities in the Shanghai ranking or having done a 'relevant' undergraduate major will lead to smaller proportions of negative words.

Table 8: Models for several differences in proportions based on the Loughran and McDonald (2011) dictionary. Estimated coefficients are displayed together with their p-values.

	d_prop_neg		d_prop_pos		d_tone		d_prop_unc	
Constant	0.0111	0.000	-0.0017	0.000	-0.0130	0.000	-0.0034	0.000
Speech	-0.0065	0.000	-0.0040	0.000	0.0026	0.000	0.0008	0.000
FT	0.0032	0.000	0.0010	0.000	-0.0022	0.000	0.0014	0.000
WSJ	0.0001	0.829	1.85E-6	0.993	-0.0001	0.905	2.99E-6	0.994
Print	0.0007	0.004			-0.0006	0.028	0.0002	0.163
Experience	-4.63E-5	0.032			0.0001	0.016	0.0002	0.000
Banking	-0.0019	0.000			0.0021	0.000	-0.0006	0.025
Centralb	0.0063	0.000			-0.0067	0.000	0.0019	0.000
B or C			-0.0003	0.036				
Male	0.0032	0.000	0.0005	0.008	-0.0028	0.000	0.0007	0.003
Advanced	-0.0013	0.000	0.0003	0.054	0.0016	0.000	-0.0007	0.002
Award	-0.0021	0.001			0.0019	0.006	0.0007	0.118
Domestic	0.0007	0.005	0.0006	0.000	-0.0001	0.728	-0.0004	0.019
Shanghai100	-0.0044	0.000			0.0043	0.000		
Undergraduate	-0.0016	0.000			0.0016	0.000	-0.0012	0.000
Adjusted R^2	0.208336		0.106369		0.103603		0.033768	
Log likelihood	30015.87		36871.53		28874.28		33525.37	
F-statistic	197.3792		171.0434		87.2468		29.25278	
Observations	9702		100001		9702		9702	

The model explaining the differences in proportion of positive words has around half of the explanatory power of the one on negative words. Much less article and journalist characteristics are of explanatory value in this model. Whether an article is published in print or online media, the working experience of a journalist, the undergraduate major and university of the journalist and whether the journalist has ever won an award are no predictors of the difference in proportion of positive words. The Financial Times covers using significantly more positive words than Reuters and the Wall Street Journal. Speeches are also

covered using more positive words than statements. Domestic journalists, male journalists and journalists with an advanced degree report with a smaller shortage of positive words, on average. Journalists with (central) banking experience tend to use proportionally less of these words. The difference in overall tone is being determined in roughly the same way as the difference in proportion of negative words with one notable exception: the `domestic` variable. Whether an author is domestic does not influence the overall tone of the media; while it does lead to both a higher proportion of positive and negative words. Ahern and Sosyura (2015) also find journalists with more experience and more relevant education to report more accurately. Contrary to their results, however, the journalist's gender and quality of education are found to be relevant.

The model which aims to explain differences in the proportion of uncertainty related words has less explanatory power than the ones previously discussed. Speeches are covered using a smaller shortage of these words than statements. The same holds for articles appearing in the Financial Times compared to those appearing in the Wall Street Journal or Reuters. Journalists with more experience, central banking experience or those who are male are, on average, using proportionally more uncertainty related words. The opposite holds true for journalists who have banking experience, post graduate education, a relevant undergraduate degree or journalists who are domestic.

Please refer to Appendix D for the full analysis of the differences in proportion using the Laver and Garry (2000) dictionary. Overall, media coverage on speeches contains higher proportions of economic policy related words than coverage of statements. The Wall Street Journal uses least of these, followed by the Financial Times; Reuters uses most, but still less than its source. Experienced journalists are likely to report using less policy related words than their younger colleagues. Journalists with banking experience on average also cover central bank communication using less policy related words. Those journalists with advanced education or those covering domestic affairs are less inclined to use policy related words as well. Journalists who have won an award, or journalists who have completed an undergraduate major in a relevant field or at a university from the top 100 of the Shanghai 2015 university ranking cover with relatively more policy related words.

6.2 Term Based Approach

Contrary to the dictionary approach, the measures obtained by the term based approach are not independent of document length. The aforementioned analysis has been performed once using the previously used variables, and once while incorporating the length of the central bank communication[16] `c_token`. The analysis has been done for both `c_token` and its logarithm. Including source length in the model quadruples the adjusted R^2 values of the models.

As expected, the Dice and Jacquard distance measures perform very similarly. Although the coefficients do not correspond in value, their signs are identical and their significance levels are of similar order. When correcting for length, speeches are found to be covered

[16]Regression analysis has shown that the length of the central bank communication as measured by the number of tokens is related to the term based measures most. This is due to the relatively large variation in this variable compared to the other measures of text length.

Table 9: Final models for several measures of lexical distance using a term based approach. Estimated coefficients are displayed together with their p-values.

	Dice		Jacquard		Cosine		JS		KL	
Constant	0.1693	0.000	0.0935	0.000	-0.0172	0.036	0.0160	0.000	5.5889	0.000
Speech	0.0100	0.000	0.0063	0.000	-0.0067	0.009	-0.0004	0.000	-0.0280	0.039
FT	-0.0204	0.000	-0.0117	0.000	-0.0216	0.000	0.0012	0.000	0.1850	0.000
WSJ	-0.0090	0.017	-0.0043	0.069	-0.0218	0.000	0.0003	0.001	0.0698	0.007
Print							-0.0001	0.000		
Experience	0.0013	0.000	0.0008	0.000	0.0011	0.000	-6.34E-6	0.030	0.0020	0.033
Banking	-0.0031	0.209	-0.0013	0.382	-0.0057	0.067	0.0007	0.000	-0.0514	0.002
Centralb	0.0296	0.000	0.0177	0.000	0.0274	0.000	-0.0004	0.000	-0.2030	0.000
Gender	0.0109	0.000	0.0064	0.000	0.0188	0.000	-0.0003	0.000	-0.0539	0.000
Advanced	-0.0081	0.000	-0.0051	0.000	-0.0100	0.000	-0.0003	0.000		
Award	0.0170	0.000	0.0097	0.000	0.0284	0.000	-0.0004	0.000	-0.1810	0.000
Domestic	0.0019	0.265	0.0011	0.301	0.0099	0.000	0.0004	0.000	-0.0182	0.117
Shanghai100	0.0110	0.000	0.0068	0.000	0.0224	0.000	-0.0003	0.000	-0.0155	0.134
Undergraduate	-0.0112	0.000	-0.0066	0.000	-0.0139	0.000	0.0002	0.000		
c_token	-1.15E-5	0.000	-6.78E-6	0.000						
log(c_token)					0.0313	0.000	-0.0013	0.000	-0.2233	0.000
Adjusted R^2	0.1755		0.1614		0.1864		0.6224		0.2898	
Log likelihood	11773.01		16263.32		9477.046		49470.46		-7140.639	
F-statistic	159.843		144.617		171.964		1143.311		372.181	
Observations	9702		9702		9702		9702		10006	

more accurately by the media than statements. Reuters is most accurate, followed by the WSJ and FT respectively. Male journalists cover less accurately than female journalists. Journalists with more years of writing experience, central banking experience, who have won an award, are domestic or those who have done an undergraduate degree at one of the top 100 universities of the Shanghai 2015 university ranking, on average, report more accurately. Journalists with advanced education, or those who have a relevant undergraduate degree report less accurately. The larger the central banking communication is, the smaller the Dice and Jacquard measures tend to be. This uncovers one of the systematic flaws in the term based approach: missing terms are punished rather harshly due to the fact that the inner product does not count any similarity for these, while it does contribute to the vector length.

The cosine measure gives rather similar results as the Dice and Jacquard. Two pronounced differences are being observed. First of all, the cosine measure finds speeches to be covered less accurately than statements. A second difference is observed at the `c_token` variable. The logarithm performs better here, but does have a positive sign. Note that due to the fact that speeches are much longer than statements, these results are very dependent on the way in which document length is included in the model. Due to the measures' dependence on document length it is difficult to draw reliable conclusions.

The model aiming to explain the JS divergence measure has a high R^2 value of 0.623. This is also the only of the term based approach measures where it matters for reporting accuracy whether the article appeared in printed media or online. Articles in printed media tend to be more accurate than articles of similar nature published online. Speeches are covered more accurately than statements, and the Financial Times covers with smallest accuracy, followed by the Wall Street Journal and then Reuters. The variables `domestic`

and `advanced` are the only ones with differing signs compared to the cosine measure. For JS, domestic journalists report less accurately, while people with advanced education tend to report more accurately.

The KL divergence measure was insensitive to more variables than any of the other measures. All significant journalist characteristics but writing experience have a positive effect on reporting accuracy (please refer to Table 9 for more details). Longer documents tend to be covered more accurately, while speeches are more accurately covered than statements of similar characteristics. The news agency performance observed for the other measures is observed for KL as well.

7 Conclusion

On average, the media use higher proportions of negative words and smaller proportions of positive words than releases by the central bank contain. This makes media coverage of central bank communication to have a more negative overall tone than its source. This confirms the findings by Ahern and Sosyura (2015) that media tend to sensationalize. Moreover, the media use smaller proportions of uncertainty related words than new central bank releases contain, finding support for the claim that central bank communication might be more nuanced than media coverage. The Laver and Garry (2000) dictionary is too specific and not specifically designed for a central banking context. The difference in the overall usage of policy related words, however, does reveal some insights. When inspecting the different countries in the dataset, media coverage on all countries but the Eurozone contains less policy words than the central bank communication. Moreover, experienced journalists cover, on average, with less policy related words than their younger colleagues.

The Dice and Jacquard measures perform very similarly, while the cosine measure behaves slightly differently. The KL divergence is stronger correlated to the cosine measure than the JS divergence. The scaled versions of the inner product turn out not to be invariant to the length of the central banking communication they cover. This makes especially the difference in reporting accuracy on statements and speeches hard to determine; since speeches are much longer than statements. The different measures do not give a univocal view on the difference. Together with the more difficult interpretation, this leaves the dictionary approach to render more meaningful results overall.

Speeches are covered using smaller proportions of negative and smaller proportions of positive words than statements. The overall tone with which speeches are covered is more positive than that with which statements are covered. Moreover, the cosine measure and language model divergence measures find speech coverage to be more accurate than statement coverage. Although all three agencies use more negative words and less positive words than central banks, the Financial Times uses proportionally more positive and negative words than the Wall Street Journal and Reuters. With an overall most negative tone, the FT tends to sensationalize most. The term based measures, however, find Reuters to be significantly less accurate than the other agencies. In all countries the tone of the media is more negative than that of the central bank; an effect which is strongest for Canada and least strong for New Zealand. Media tend to cover speeches less negatively when they are

given by the president of a central bank than when the vice president has given the speech. The overall tone of printed articles is more negative than the tone of online articles: printed media have the tendency to sensationalize more than online media. The term based measures all find printed articles to be more accurate than online articles, however. In 1996 and the years after the crisis, higher proportions of uncertainty related words are used by the media. These proportions are still smaller than those used by the central banks.

As the authors of an article have more writing experience, they tend to use less negative words. Male and domestic authors use more negative words, while those who have won an award, or score better in one of the undergraduate degree variables use proportionally less of these. The difference in overall tone is being determined in roughly the same way as the difference in proportion of negative words. However, whether an author is domestic does not influence the overall tone of the media; while it does lead to both a higher proportion of positive and negative words. Journalists with more writing experience or central banking experience are on average using more uncertainty related words while those with banking experience or those who are domestic use proportionally less of these. All term based measures find Reuters to report most accurately. Moreover, all of them find articles written by female journalists to have a smaller lexical distance than their male colleagues. Journalists with central banking experience, who have won an award or those who have obtained an undergraduate degree at one of the top 100 universities of the Shanghai 2015 academic ranking report more accurately than those without. All measures but the KL divergence measure find more experienced journalists to report more accurately. Most divergence measures also find domestic journalists to be more accurate than non-domestic authors. Those with a relevant undergraduate degree report, on average, less accurately.

8 Limitations and Suggestions for Further Research

This paper is centered around quantitative textual analysis. Due to the limited time frame, the measures by which similarity measures have been constructed are not state-of-the-art ones. Ultimately, counting the frequency of words is the corner stone of the similarity measures which have been utilized. Extensions which make use of bigrams or trigrams might capture similarity in a better way. Moreover, current measures penalize similarity measures when certain words are only present in one of the texts. Corrections for synonyms (or even antonyms) could present a more reliable framework for computing textual distance. More advanced machine learning algorithms would form a very welcome addition to the research design. In further research, using more advanced distance measures would improve the quality of the results.

Another limitation is the nature of the data. The data set used in this research is relatively old and covers only a very short time period after the financial crisis. Since the crisis has dramatically changed the outlines of the financial system, it would be of added value to take more recent years into consideration as well. Additionally, the journalist finding algorithm could be improved. Especially for Reuters and the Wall Street Journal progress can be made by finding more journalist names. Access to databases containing university diplomas obtained and journalist employment might further improve the quality of the journalist characteristics.

For the computation of the KL divergences, a λ value of 0.9 has been used, as suggested by Fernández (2007). Since this value must be tuned for optimal results (Verheij et al., 2012), it might very well be that within this specific context the value of 0.9 is not optimal. Further research could look into this optimization process deeper. For doing this, more efficient software should be made or better hardware could be employed to limit computation times.

References

Ahern, K. R. and Sosyura, D. (2015). Rumor has it: Sensationalism in financial media. *Review of Financial Studies*, page hhv006.

Apel, M. and Grimaldi, M. (2012). The information content of central bank minutes. *Riksbank Research Paper Series*, (92).

Bigi, B. (2003). Using kullback-leibler distance for text categorization. In *European Conference on Information Retrieval*, pages 305–319. Springer.

Blinder, A. S., Ehrmann, M., Fratzscher, M., De Haan, J., and Jansen, D.-J. (2008). Central bank communication and monetary policy: A survey of theory and evidence. Technical report, National Bureau of Economic Research.

Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526.

Fernández, R. T. (2007). The effect of smoothing in language models for novelty detection. In *Proceedings of the 1st BCS IRSG conference on Future Directions in Information Access*, pages 28–29.

Grosse, I., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., Oliver, J., and Stanley, H. E. (2002). Analysis of symbolic sequences using the jensen-shannon divergence. *Physical Review E*, 65(4):041905.

Gurun, U. G. and Butler, A. W. (2012). Don't believe the hype: Local media slant, local advertising, and firm value. *The Journal of Finance*, 67(2):561–598.

Hiemstra, D. (2009). Language models. In *Encyclopedia of Database Systems*, pages 1591–1594. Springer.

Laver, M. and Garry, J. (2000). Estimating policy positions from political texts. *American Journal of Political Science*, pages 619–634.

Li, F. (2006). Do stock market investors understand the risk sentiment of corporate annual reports? *Available at SSRN 898181*.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Loughran, T. and McDonald, B. (2014). Measuring readability in financial disclosures. *The Journal of Finance*, 69(4):1643–1671.

Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467.

Thada, V. and Jaglan, V. (2013). Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm. *International Journal of Innovations in Engineering and Technology*.

Verheij, A., Kleijn, A., Frasincar, F., and Hogenboom, F. (2012). A comparison study for novelty control mechanisms applied to web news stories. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on*, volume 1, pages 431–436.

A Appendix - Table

Table 10: Number of media articles per month.

Month	Articles	Month	Articles
Jan	6819	July	5497
Feb	6870	August	5293
Mar	8519	September	7681
Apr	5744	October	8036
May	5693	November	7591
June	8534	December	6506

Table 11: Different textual similarities in relation to speeches or statements (reference category). Estimated coefficients including their significance are displayed.

	d_prop_less	d_prop_eq	d_prop_pro	d_prop_neg	d_prop_pos	d_prop_unc	d_tone	Inner	Dice	Jacquard	Cosine	JS	KL
constant	-0.004474	-0.007218	-0.000391	0.006263	-4.36E-04	-0.003132	-0.006699	304.9754	0.143341	0.079217	0.192046	0.009224	4.332341
st. e.	4.61E-05	8.07E-05	1.36E-05	5.48E-05	3.04E-05	3.78E-05	6.08E-05	2.72E+00	3.61E-04	2.28E-04	4.58E-04	1.59E-05	2.93E-03
t-stat.	-9.70E+01	-89.41879	-2.88E+01	114.3068	-14.35336	-82.9534	-110.226	111.9227	397.0562	347.0032	419.0286	581.0995	1480.513
prob.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
speech	0.003105	-0.000569	-0.001065	-0.005108	-4.51E-03	5.23E-04	0.000597	332.7848	-0.026006	-0.014693	0.026875	-0.003074	-0.357771
st. e.	7.12E-05	0.000125	2.10E-05	8.46E-05	4.69E-05	5.83E-05	9.38E-05	4.21E+00	5.57E-04	3.52E-04	7.07E-04	2.45E-05	4.52E-03
t-stat.	4.36E+01	-4.568493	-50.80829	-60.40863	-96.12647	8.973681	6.360269	79.13702	-46.67952	-41.70659	37.9965	-125.4932	-79.22287
prob.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 12: Different textual similarities in relation to the news agencies (Reuters is the reference category). Estimated coefficients together with their significance are displayed.

	d_prop_less	d_prop_eq	d_prop_pro	d_prop_neg	d_prop_pos	d_prop_unc	d_tone	Inner	Dice	Jacquard	Cosine	JS	KL
constant	-0.003534	-0.008963	-0.000979	0.002139	-0.002652	-0.003741	-0.004792	413.9843	0.127041	0.069776	0.196806	0.008312	4.217759
st. e.	4.17E-05	7.16E-05	1.23E-05	4.83E-05	2.87E-05	3.34E-05	5.33E-05	2.525761	0.000326	0.000206	0.000412	1.53E-05	2.71E-03
t-stat.	-84.65446	-125.1157	-79.34223	44.28441	-92.54862	-112.0112	-89.8646	163.9048	389.6297	339.1502	477.4465	543.4529	1555.845
prob.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
FT	0.0017	0.006316	0.000497	0.008255	0.001424	0.003349	-0.006832	96.32571	1.90E-02	1.16E-02	2.63E-02	-1.08E-03	-1.47E-01
st. e.	8.88E-05	1.52E-04	2.62E-05	1.03E-04	6.09E-05	7.10E-05	1.13E-04	5.369849	0.000693	0.000437	0.000876	3.25E-05	5.76E-03
t-stat.	19.15489	41.46773	18.96732	80.37937	23.3633	47.1608	-60.2656	17.93825	27.41962	26.44387	30.04937	-33.13364	-25.44835
prob.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
WSJ	0.000203	0.002877	0.000527	0.003897	0.000405	0.001939	-0.003491	148.9942	2.01E-02	1.22E-02	1.53E-02	-2.14E-03	-7.57E-02
st. e.	0.000137	0.000235	4.05E-05	0.000159	9.41E-05	1.10E-04	1.75E-04	8.295912	0.001071	0.000676	0.001354	5.02E-05	8.90E-03
t-stat.	1.480673	12.2261	13.0082	24.55959	4.307791	17.67144	-19.93541	17.95995	18.81439	18.09446	11.29938	-42.66118	-8.497085
prob.	0.1387	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 13: Different textual similarities for the different countries in the sample (the U.S. is the reference category here). Estimated coefficients together with their significance are displayed.

	d_prop_less	d_prop_eq	d_prop_pro	d_prop_neg	d_prop_pos	d_prop_unc	d_tone	Inner	Dice	Jacquard	Cosine	JS	KL
constant	-0.003088	-0.008015	-0.00088	0.0027	-0.002923	-0.005039	-0.005623	337.8346	0.123752	0.067646	0.179958	0.007606	4.276307
st. e.	0.0000461	7.22E-05	0.0000135	0.0000552	0.0000316	0.0000359	0.000061	2.601672	0.00035	0.000222	0.000444	0.0000165	0.002968
t-stat.	-66.98447	-111.0413	-65.37113	48.92086	-92.59078	-140.1789	-92.23138	129.8529	3.54E+02	304.7407	405.3293	459.8814	1.44E+03
prob.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Australia	-0.000438	0.001804	-0.00054	-0.000542	0.001247	0.004431	0.001789	18.57574	0.045376	0.027612	0.051616	0.0000978	-0.316452
st. e.	0.000224	0.000351	0.0000655	0.000268	0.000154	0.000175	0.000297	12.65381	0.001702	0.00108	0.002159	0.0000804	0.014437
t-stat.	-1.953423	5.13964	-8.250574	-2.019033	8.121027	25.34116	6.032992	1.467995	26.66507	25.57562	23.90295	1.215939	-21.91958
prob.	0.0508	0.0000	0.0000	0.0435	0.0000	0.0000	0.0000	0.1421	0.0000	0.0000	0.0000	0.2240	0.0000
Canada	-0.002109	-0.009873	0.000265	0.004528	0.000676	0.005831	-0.003852	141.6965	0.082573	5.17E-02	0.104626	0.000698	-4.06E-01
st. e.	0.000265	0.000415	0.0000773	0.000317	1.81E-04	0.000207	0.00035	14.94977	2.01E-03	1.28E-03	0.002551	0.000095	1.71E-02
t-stat.	-7.962786	-23.80469	3.429444	14.27796	3.73E+00	28.22854	-10.99503	9.478179	4.11E+01	4.05E+01	41.01059	7.343172	-2.38E+01
prob.	0.0000	0.0000	0.0006	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
EMU	0.002499	0.014354	0.000548	0.000956	-6.73E-04	0.003206	-0.001628	626.1823	-0.018879	-0.010337	0.07146	-0.001082	-0.387552
st. e.	0.0000928	0.000145	0.0000271	0.000111	6.36E-05	0.0000724	0.000123	5.239088	0.000705	0.000447	0.000894	0.0000333	0.005977
t-stat.	26.91966	98.75981	20.22173	8.598324	-1.06E+01	44.29449	-13.26235	119.5213	-26.79499	-23.12474	79.92776	-32.48534	-64.8362
prob.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
England	0.000602	-0.014829	-0.001966	0.006864	0.003513	0.006198	-0.003351	-110.3466	0.048868	0.028747	0.028054	0.002788	-0.008042
st. e.	0.000126	0.000197	0.0000367	0.00015	0.000086	0.000098	0.000166	7.089944	0.000953	0.000605	0.00121	0.0000451	0.008089
t-stat.	4.793222	-75.39139	-53.59892	45.63568	40.83205	63.27083	-20.16798	-15.56382	51.25322	47.52236	23.18708	61.86554	-0.994242
prob.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3201
Japan	-0.00523	-0.007848	0.001364	0.006518	0.003955	0.00763	-0.002563	-39.94422	0.044464	0.026668	0.034827	0.002414	-0.015878
st. e.	0.000126	0.000197	0.0000367	0.00015	0.000086	0.000098	0.000166	7.089944	0.000953	0.000605	0.00121	0.0000451	0.008089
t-stat.	-41.6392	-39.89868	37.18746	43.33883	45.97488	77.88621	-15.42567	-5.633926	46.63418	44.08468	28.78512	53.55592	-1.962905
prob.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0497
Swits.	-0.005586	0.008779	-0.000454	0.000979	0.000613	0.005074	-0.000366	-82.22029	0.042278	0.025832	0.028769	0.003258	-0.081358
st. e.	0.000317	0.000497	0.0000926	0.00038	0.000217	0.000247	0.00042	17.9079	0.002408	0.001528	0.003056	0.000114	0.020431
t-stat.	-17.60731	17.67098	-4.9008	2.57637	2.819939	20.50605	-0.872032	-4.591286	17.55537	16.90648	9.414013	28.6189	-3.981973
prob.	0.0000	0.0000	0.0000	0.0100	0.0048	0.0000	0.3832	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
New Zea.	-0.002336	-0.009452	0.000792	-0.003041	0.000385	0.002297	0.003426	133.066	0.083319	0.051392	0.098709	0.00133	-0.370874
st. e.	0.000365	0.000572	0.000107	0.000437	0.00025	0.000285	0.000483	20.61368	0.002772	0.001759	0.003518	0.000131	0.023519
t-stat.	-6.39591	-16.52767	7.425517	-6.954701	1.539978	8.066036	7.093192	6.455229	30.05579	29.22039	28.06017	10.15085	-15.76946
prob.	0.0000	0.0000	0.0000	0.0000	0.1236	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 14: Different textual similarities between media coverage and central bank speeches in relation to whether the speech is given by the president or vice president (reference category) of a central bank. Estimated coefficients including their significance are displayed.

	d_prop_less	d_prop_eq	d_prop_pro	d_prop_neg	d_prop_pos	d_prop_unc	d_tone	Inner	Dice	Jacquard	Cosine	JS	KL
constant	-0.003333	-0.005239	-0.000944	0.001632	-4.97E-03	-0.002295	-0.006602	850.9348	0.125996	0.069653	0.262493	0.006536	3.785488
st. e.	2.06E-04	3.11E-04	6.66E-05	2.43E-04	1.38E-04	1.55E-04	2.81E-04	1.43E+01	1.68E-03	1.08E-03	2.29E-03	4.89E-05	1.37E-02
t-stat.	-1.62E+01	-16.85543	-1.42E+01	6.725912	-36.02781	-14.83202	-23.46481	59.69414	74.9096	64.56133	114.7162	133.5608	276.7057
prob.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
president	0.002104	-0.002731	-0.000549	-0.000511	2.40E-05	-3.36E-04	0.000535	-228.5052	-0.009284	-0.005499	-0.046706	-0.000414	0.20268
st. e.	2.14E-04	0.000322	6.90E-05	2.51E-04	1.43E-04	1.60E-04	2.91E-04	1.48E+01	1.74E-03	1.12E-03	2.37E-03	5.07E-05	1.42E-02
t-stat.	9.85E+00	-8.485911	-7.958807	-2.034336	0.167993	-2.097011	1.836525	-15.48286	-5.331431	-4.923064	-19.71526	-8.17398	14.30957
prob.	0.0000	0.0000	0.0000	0.0419	0.8666	0.0360	0.0663	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 15: Different textual similarities in for printed and online (reference category) articles. Estimated coefficients together with their significance are displayed.

	d_prop_less	d_prop_eq	d_prop_pro	d_prop_neg	d_prop_pos	d_prop_unc	d_tone	Inner	Dice	Jacquard	Cosine	JS	KL
constant	-0.003385	-0.008294	-0.000912	0.003188	-2.49E-03	-0.003305	-0.005679	428.3025	0.130135	0.07171	0.200317	0.008134	4.19491
st. e.	3.88E-05	6.69E-05	1.15E-05	4.60E-05	2.67E-05	3.13E-05	5.02E-05	2.35E+00	3.04E-04	1.92E-04	3.84E-04	1.43E-05	2.53E-03
t-stat.	-8.72E+01	-124.0246	-7.95E+01	69.36481	-93.38546	-105.6402	-113.0879	182.2331	427.9746	373.8011	521.2721	567.8615	1660.459
prob.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
print	0.001306	0.005098	0.00045	0.005666	9.78E-04	2.39E-03	-0.004688	99.87326	0.013921	0.008139	0.018346	-0.001224	-0.077848
st. e.	9.58E-05	0.000165	2.83E-05	1.13E-04	6.58E-05	7.72E-05	1.24E-04	5.80E+00	7.50E-04	4.73E-04	9.48E-04	3.53E-05	6.23E-03
t-stat.	1.36E+01	30.89454	15.89025	49.96365	14.85123	30.99654	-37.83395	17.2211	18.55401	17.19456	19.34785	-34.64051	-12.48801
prob.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 16: Different textual similarities over the different days of the week (Monday is the reference category). Estimated coefficients together with their t-statistics are displayed.

	d_prop_less	d_prop_eq	d_prop_pro	d_prop_neg	d_prop_pos	d_prop_unc	d_tone	Inner	Dice	Jacquard	Cosine	JS	KL
constant	-0.002814	-7.60E-03	-0.001083	0.004388	-0.003898	-0.001759	-0.008286	670.2266	0.128421	0.070968	0.227347	0.006944	3.978508
t-stat.	-20.10943	-31.26196	-26.08191	25.9882	-40.43633	-15.62556	-45.22382	80.55147	116.6539	102.1552	165.0207	133.616	441.195
Tues.	-0.001366	0.0000501	0.000426	-0.000052	0.002206	-0.002576	0.002258	-353.7038	0.007991	0.004243	-0.040154	0.001505	0.320257
t-stat.	-8.843337	0.186851	9.304083	-0.279266	20.73446	-20.72746	11.16614	-38.52204	6.578045	5.534688	-26.41154	26.23424	32.18305
Wednes.	-0.001234	-0.001516	0.000409	-0.000317	0.000962	-0.001987	0.001279	-342.9453	0.00985	0.005404	-0.037046	0.000877	0.298832
t-stat.	-7.859239	-5.552726	8.776572	-1.671895	8.895996	-15.72008	6.221088	-36.72386	7.97242	6.931018	-23.95871	15.03461	29.52632
Thurs.	1.22E-03	0.001298	-0.0000693	-7.19E-04	0.001742	-0.0000142	0.002461	-82.46474	-0.003749	-0.002365	-0.007986	0.000877	0.090064
t-stat.	7.997729	4.886186	-1.52807	-3.900803	16.54276	-0.115194	12.29825	-9.075213	-3.1186	-3.117378	-5.307559	15.45061	9.145315
Fri.	-0.001006	0.000694	0.000616	0.001174	0.001783	0.001347	0.000609	-174.1849	0.008563	0.005134	-0.009075	0.000766	0.108358
t-stat.	-5.077346	2.02E+00	10.49108	4.913143	13.06538	8.450656	2.34621	-14.78884	5.494968	5.220887	-4.653172	10.41427	8.487942
Satur.	1.28E-03	3.45E-03	0.000931	-0.000331	0.000464	0.001552	0.000795	35.01434	0.00152	0.001445	0.030284	-0.000779	-0.070463
t-stat.	1.836983	2.842013	4.490889	-0.392405	0.963869	2.759772	0.868731	0.842409	0.27644	0.416492	4.400352	-3.001734	-1.564216
Sun.	0.000666	0.008632	0.000411	-0.005421	-0.001324	0.000965	0.004097	63.45525	-0.024593	-0.014359	-0.026499	-0.000969	0.121738
t-stat.	0.905738	6.754608	1.885338	-6.10902	-2.613017	1.630026	4.255034	1.451235	-4.250989	-3.933146	-3.66017	-3.548601	2.568954

Table 17: Different textual similarities over the different months of the year (January is the reference category). Estimated coefficients together with their t-statistics are displayed.

	d_prop_less	d_prop_eq	d_prop_pro	d_prop_neg	d_prop_pos	d_prop_unc	d_tone	Inner	Dice	Jacquard	Cosine	JS	KL
constant	-0.003999	-7.17E-03	-0.000745	0.006601	-0.00179	-0.003115	-0.008391	401.8184	0.132382	0.072956	0.196287	0.007962	4.237967
t-stat.	-32.46922	-33.54036	-20.40528	44.65668	-21.20169	-31.17149	-52.16085	53.9496	136.6085	119.3282	160.8214	174.2581	529.1531
Febr.	0.001538	-0.001584	-0.000334	-0.004202	-0.001501	-0.0000295	0.002701	147.3249	-0.004635	-0.002411	0.023651	-0.000744	-0.162626
t-stat.	8.846025	-5.247295	-6.480701	-20.13768	-12.59411	-0.208788	11.89584	14.01287	-3.388456	-2.79359	13.72781	-11.53194	-14.38489
March	0.00113	-0.000104	-0.000225	-0.003593	-0.0000692	0.000628	0.003524	13.41785	0.002956	0.001765	0.00549	-0.000165	-0.037236
t-stat.	6.839359	-0.36317	-4.599293	-18.11489	-0.61111	4.686297	16.32518	1.342614	2.27327	2.151285	3.352482	-2.686247	-3.464978
April	4.54E-04	-0.001237	-0.0002	-3.43E-03	-0.000407	0.001387	0.003024	97.46968	0.005312	0.003402	0.019484	-0.000323	-0.145332
t-stat.	2.493899	-3.913511	-3.697807	-15.69294	-3.259252	9.382433	12.71003	8.848882	3.706672	3.762236	10.79425	-4.777744	-12.27005
May	0.001129	0.001292	-0.0000719	-0.003707	-0.001291	-0.000862	0.002416	44.52882	-0.003283	-0.001953	0.004592	0.000237	-0.059963
t-stat.	6.18555	4.08E+00	-1.32886	-16.91721	-10.31616	-5.814806	10.13189	4.032801	-2.284925	-2.154255	2.537571	3.505044	-5.050297
June	1.10E-03	5.75E-05	-0.000281	-0.001765	-0.001437	0.000558	0.000328	24.80178	-0.002082	-0.001202	0.000792	0.0000482	-0.038785
t-stat.	6.65843	0.200437	-5.747511	-8.902071	-12.6925	4.16107	1.519757	2.48268	-1.601988	-1.465634	0.483738	0.786132	-3.610457
July	0.002626	-0.002678	-0.000705	-0.004438	-0.002286	0.001059	0.002152	204.41	-0.012784	-0.007203	0.027296	-0.00091	-0.198652
t-stat.	14.24345	-8.369203	-12.90529	-20.05905	-18.09611	7.076647	8.936408	18.33532	-8.813672	-7.871282	14.94097	-13.30537	-16.57085
August	-1.83E-03	1.37E-03	0.000212	-0.001251	0.001639	0.0000352	0.00289	-125.7007	0.006707	0.003731	-0.012625	0.000728	0.074752
t-stat.	-9.801393	4.22E+00	3.839	-5.592961	12.83926	0.23307	11.87681	-11.15679	4.575399	4.033784	-6.83768	10.53766	6.17004
Sept.	-0.0000951	-0.000463	0.0000528	-0.001862	-0.000136	0.0008	0.001726	-9.455214	0.002513	0.001472	-0.002105	0.000222	0.014685
t-stat.	-0.562316	-1.576657	1.052408	-9.166867	-1.173861	5.827819	7.807493	-0.923963	1.887525	1.751844	-1.255316	3.536077	1.334406
Oct.	0.001218	-0.001392	0.000224	-0.002144	-0.000473	0.0000348	0.001671	41.46689	0.002031	0.001372	0.006244	-0.0000677	-0.071303
t-stat.	7.27354	-4.78897	4.515253	-10.66792	-4.120135	0.256243	7.640755	4.094901	1.541598	1.650836	3.762755	-1.08938	-6.548128
Nov.	0.002236	0.001295	0.0000797	-0.001792	-0.000113	-0.000497	0.001679	73.82916	0.00139	0.000817	0.01242	0.000289	-0.072514
t-stat.	13.17718	4.394694	1.584893	-8.800319	-0.972863	-3.610316	7.576144	7.194552	1.041436	0.969918	7.385802	4.589927	-6.571488
Dec.	-1.37E-04	-0.0000323	0.0000592	-0.002157	-0.00036	-0.000741	0.001797	18.36846	0.000681	0.000348	0.002675	0.00031	0.003977
t-stat.	-0.777836	-0.105471	1.132703	-10.19658	-2.980554	-5.182525	7.805634	1.723274	0.490778	0.397514	1.531671	4.736825	0.346989

Table 18: Different textual similarities over the different years the sample consists of. Estimated coefficients together with their t-statistics are displayed.

	d_prop_less	d_prop_eq	d_prop_pro	d_prop_neg	d_prop_pos	d_prop_unc	d_tone	Inner	Dice	Jacquard	Cosine	JS	KL
constant	-0.004882	-9.04E-03	-0.000448	0.006198	-0.001183	-0.000923	-0.007381	517.4179	0.144373	0.080146	0.214323	0.007441	4.144192
t-stat.	-50.03381	-54.86485	-15.7274	53.31573	-17.67845	-11.90141	-58.07899	87.47535	189.5463	166.6263	221.5473	208.4715	651.3289
y1996	0.003363	0.004576	-0.000448	0.00133	-0.001813	0.002553	-0.003143	-8.97734	-0.010161	-0.005652	0.018289	0.002191	-0.022068
t-stat.	6.732466	5.42642	-3.076022	2.235218	-5.293186	6.430217	-4.831327	-0.296443	-2.605617	-2.295217	3.692579	11.98725	-0.677432
y1997	0.003785	-0.001444	-0.001482	-0.005688	-0.002586	-0.001701	0.003103	-87.68639	-0.035071	-0.020503	-0.025033	-0.000281	0.011127
t-stat.	16.54595	-3.739829	-22.20718	-20.8733	-16.48605	-9.355335	10.41543	-6.323765	-19.64183	-18.1837	-11.0386	-3.362359	0.746015
y1998	-0.000678	-0.00401	-0.000584	-0.002825	-0.001213	-0.000405	0.001612	-164.5127	-0.029234	-0.017028	-0.03076	0.001873	0.179459
t-stat.	-3.954833	-13.86701	-11.68279	-13.83793	-10.32263	-2.972945	7.224334	-15.83953	-21.85818	-20.16153	-18.10826	29.89106	16.06289
y1999	1.21E-03	-0.001793	-0.000302	-4.76E-03	-0.00138	-0.002355	0.00338	-112.608	-0.027831	-0.016658	-0.023291	0.001143	0.064698
t-stat.	8.414246	-7.394909	-7.210857	-27.81212	-14.01075	-20.62896	18.06605	-12.93134	-24.8189	-23.52478	-16.35393	21.74309	6.906882
y2000	0.00219	0.0000789	-0.000155	-0.005646	-0.001651	-0.005127	0.003995	-128.1802	-0.024253	-0.014589	-0.021896	0.001278	0.110475
t-stat.	14.68714	3.13E-01	-3.556219	-31.78232	-16.15114	-43.2652	20.5711	-14.18205	-20.83864	-19.84983	-14.81265	23.43029	11.36314
y2001	2.57E-03	8.78E-03	0.000708	0.000609	-0.000692	-0.005282	-0.001301	-170.409	0.007358	0.004485	-0.000425	0.001718	0.103713
t-stat.	14.97361	30.26146	14.11029	2.974733	-5.873597	-38.67627	-5.813377	-16.36007	5.486055	5.29498	-0.249368	27.33174	9.256362
y2002	0.002221	0.008136	-0.000227	-0.00054	-0.000612	-0.00221	-0.0000718	-121.4856	-0.012866	-0.007684	-0.017392	-0.0000549	0.054503
t-stat.	14.32319	31.07813	-5.010995	-2.921537	-5.751937	-17.92555	-0.355536	-12.92113	-10.62719	-10.05046	-11.31045	-0.968228	5.389032
y2003	1.42E-03	8.11E-03	-0.000883	-0.000659	-0.000849	-0.003438	-0.00019	-137.4641	-0.010093	-0.005766	-0.012461	0.000726	0.104715
t-stat.	7.36044	2.48E+01	-15.6508	-2.860085	-6.405601	-22.38521	-0.755872	-11.73423	-6.691012	-6.052862	-6.504062	10.2654	8.309805
y2004	0.003043	0.006012	-0.001951	-0.002167	-0.002415	-0.004147	-0.000248	-129.6722	-0.011924	-0.007434	-0.016091	0.000311	0.003706
t-stat.	17.96059	21.02237	-39.45439	-10.73403	-20.78536	-30.79657	-1.123151	-12.62489	-9.015766	-8.901047	-9.578887	5.018599	0.3354
y2005	0.003048	0.007544	-0.001272	-0.001193	-0.002274	-0.001754	-0.001081	-61.02084	-0.008972	-0.005421	-0.009592	-0.0000659	-0.001482
t-stat.	16.53933	24.24713	-23.64756	-5.431924	-17.99028	-11.97101	-4.501948	-5.461085	-6.235339	-5.966023	-5.248844	-0.977235	-0.123263
y2006	0.003137	0.001713	-0.000465	-0.001349	-0.000601	-0.000581	0.000748	31.10195	-0.008918	-0.005199	0.003183	0.00018	-0.092824
t-stat.	17.675	5.716621	-8.976418	-6.379157	-4.938093	-4.118697	3.235875	2.89094	-6.437382	-5.943287	1.808842	2.771152	-8.019826
y2007	3.09E-03	0.000443	0.0000649	0.000175	0.000781	-0.00134	0.000605	-29.80244	-0.015037	-0.009184	-0.014354	0.000637	0.022764
t-stat.	18.98615	1.610817	1.365443	0.904041	6.98991	-10.34984	2.852832	-3.018281	-11.82653	-11.43878	-8.888785	10.68333	2.143206
y2009	0.000545	-0.003498	-0.000155	-0.003102	-0.002042	-0.0000263	0.00106	47.19849	0.004425	0.002893	0.008432	-0.0007	-0.058196
t-stat.	3.629276	-13.8013	-3.542823	-17.34189	-19.83358	-0.220202	5.422657	5.185877	3.775761	3.908423	5.664428	-12.74229	-5.944304
y2010	0.001877	-0.004487	-0.000573	-0.002063	-0.003221	0.0000229	-0.001158	-3.06369	-0.001749	-0.001001	-0.000479	-0.000802	0.054901
t-stat.	8.212063	-11.63088	-8.583765	-7.577098	-20.55368	0.126197	-3.889081	-0.221135	-0.980612	-0.888917	-0.211534	-9.588725	3.683888

B Appendix - Programming code

B.1 Code Used for the Loughran and McDonald (2011) Dictionary

```
#Code used to run the Loughran and McDonald 2011 dictionary and retrieve document information
#Code can be used for CB statements, CB speeches and media coverage of speeches and statements.

#Install and import the quanteda package
install.packages("quanteda")
library(quanteda)

#Import text files, generate corpus
textfileFS <- textfile("C:/Users/Gebruiker/Documents/Data_Scriptie/Statements/*.txt",
docvarsfrom="filenames", sep="_", docvarnames=c("Country","Date"))
corpusFS <- corpus(textfileFS)

#Set language to English, tokenize and compute document feature matrices
metadoc(corpusFS, "language") <- "english"
tokenize(corpusFS, removeNumbers = TRUE, removePunct = TRUE, removeSeparators = TRUE)
mydfm <- dfm(corpusFS)

#Get tokens, types and sentences and export them
statementsentence <- nsentence(corpusFS)
statementtoken <- ntoken(corpusFS)
statementtype <- ntype(corpusFS)
dfstatementsentence <- as.data.frame(statementsentence)
dfstatementtoken <- as.data.frame(statementtoken)
dfstatementtype <- as.data.frame(statementtype)
write.csv(dfstatementsentence, "F:/CB_statements_sentence.csv")
write.csv(dfstatementtype, "F:/CB_statements_type.csv")
write.csv(dfstatementtoken, "F:/CB_statements_token.csv")

#Get info on documents and export it
dfinfo <- as.data.frame(corpusFS$documents)
dfinfo$texts<-NULL
write.csv(dfinfo, "F:/CB_statements_info.csv")

#import and run the Loughran and McDonald 2011 dictionary, and export results
findict <- dictionary(file = "C:/Users/Gebruiker/Dropbox/Thesis/Dictionaries/Loughran_&_McDonald
2014.cat", format = "wordstat")
FinAn <- dfm(corpusFS, dictionary=findict)
dataframefinan <- as.data.frame(FinAn)
write.csv(dataframefinan, "F:/CB_statements_finan.csv")
```

B.2 Code Used for the Laver and Garry (2000) Dictionary

```
#Code used to run the Laver and Garry 2000 dictionary
#Code can be used for CB statements, CB speeches and media coverage of speeches and statements.

#Install and import the quanteda package
install.packages("quanteda")
library(quanteda)

#Import all relevant text files, retrieve their document names and put them into a corpus
textfileFS <- textfile("C:/Users/Gebruiker/Documents/Data_Scriptie/Statements/*.txt",
docvarsfrom="filenames", sep="_", docvarnames=c("Country","Date"))
corpusFS <- corpus(textfileFS)

#Assign the English language and make sure words are seperated
metadoc(corpusFS, "language") <- "english"
tokenize(corpusFS, removeNumbers = TRUE, removePunct = TRUE, removeSeparators = TRUE)

#Import the Laver and Garry 2000 dictionary
lgdict <- dictionary(file = "http://www.kenbenoit.net/courses/essex2014qta/LaverGarry.cat",
format = "wordstat")

#Generate document feature matrices, run dictionary and write output.
Analysis <- dfm(corpusFS, dictionary=lgdict)
dataframelgdict <- as.data.frame(Analysis)
write.csv(dataframelgdict, "F:/CB_statements_lgdict.csv")
```

B.3 Code Used for to Compute Distances with the Vector Model

```
#Code used to compute the jacquard, dice and cosine distances.
#Code can be used for media coverage of speeches and statements separately.

#Install and import the quanteda package
install.packages("quanteda")
library(quanteda)

#import the cb communication and media coverage pairs
INPUT <- read.table("F:/Example/ex.txt")
NumberArticles <- dim(INPUT)[1]

MediaName <- NULL
CBName<-NULL
InnerProduct <- 0
NormDFMMedia <- 0
NormDFMCB <- 0
CosDist <- 0
JacDist <- 0
DicDist <- 0

#initialize empty output data frame
OUTPUT <- data.frame(NameMedia= numeric(NumberArticles), NameCB=numeric(NumberArticles),
Inner= numeric(NumberArticles), Cos = numeric(NumberArticles), Jac = numeric(NumberArticles),
Dic = numeric(NumberArticles))

#compute the relevant variables for all pairs of documents
for(i in 1:NumberArticles){

#empty everything before the start of a new iteration
MediaName <- NULL
CBName<-NULL
InnerProduct <- 0
NormDFMMedia <- 0
NormDFMCB <- 0
CosDist <- 0
JacDist <- 0
DicDist <- 0

MediaName <- INPUT[i,1]
CBName <- INPUT[i,2]

#Folder names with slash at the end
DefaultLinkMedia <- "C:/Users/Gebruiker/Documents/Data_Scriptie/Data_relevant_and_irrelevant/
Speeches/Combined/"
DefaultLinkCB <- "C:/Users/Gebruiker/Documents/Data_Scriptie/Speeches/"

#Generate paths to relevant names
MediaLink <- paste(DefaultLinkMedia, MediaName, sep="")
CBLink <- paste(DefaultLinkCB, CBName, sep="")

#import relevant documents and perform necessary formalities
TextMedia <- textfile(MediaLink)
TextCB <- textfile(CBLink)
CorpusMedia <- corpus(TextMedia)
CorpusCB <- corpus(TextCB)
metadoc(CorpusMedia, "language") <- "english"
metadoc(CorpusCB, "language") <- "english"
tokenize(CorpusMedia, removeNumbers = TRUE, removePunct = TRUE, removeSeparators = TRUE)
tokenize(CorpusCB, removeNumbers = TRUE, removePunct = TRUE, removeSeparators = TRUE)

#generate dfm vectors and allow for mathematical operations
mydfmMedia <- dfm(CorpusMedia, ignoredFeatures = c(stopwords("english")))
mydfmCB <- dfm(CorpusCB, ignoredFeatures = c(stopwords("english")))
mydfmCombined <- rbind(mydfmMedia, mydfmCB)
DFM<-as.data.frame(mydfmCombined)
DFMMedia <- data.matrix(DFM[1,], rownames.force = NA)
```

```
DFMCB <- data.matrix(DFM[2,], rownames.force = NA)

#compute distances
InnerProduct <- DFMMedia %*% t(DFMCB)
NormDFMMedia <- sqrt(sum(DFMMedia^2))
NormDFMCB <- sqrt(sum(DFMCB^2))
CosDist <- InnerProduct/(NormDFMCB*NormDFMMedia)
JacDist <- InnerProduct/(NormDFMCB^2+NormDFMMedia^2-InnerProduct)
DicDist <- 2*InnerProduct/(NormDFMCB^2 + NormDFMMedia^2)

#assign distances to output matrix
OUTPUT[i,1] <- toString(MediaName)
OUTPUT[i,2] <- toString(CBName)
OUTPUT[i,3] <- toString(InnerProduct)
OUTPUT[i,4] <- toString(CosDist)
OUTPUT[i,5] <- toString(JacDist)
OUTPUT[i,6] <- toString(DicDist)

#used to keep track of the time
print(i)

}

write.csv(OUTPUT, "F:/Cosine_analysis_speeches.csv")
```

B.4 Code Used for to Compute Jensen-Shannon Divergence

```
#Code used to compute JS divergence
#Code can be used for media coverage of speeches and statements separately

#Install and import the quanteda package
install.packages("quanteda")
library(quanteda)

#import the cb communication and media coverage pairs
INPUT <- read.table("F:/OLD/SERIOUS/statements.txt")
NumberArticles <- dim(INPUT)[1]

MediaName <- NULL
CBName<-NULL
Theta1<-NULL
Theta2<-NULL
M<-NULL
temp1<-NULL
temp2<-NULL
KL1<-NULL
KL2<-NULL
JS<-NULL

#initialize empty output data frame
OUTPUT <- data.frame(NameMedia= numeric(NumberArticles), NameCB=numeric(NumberArticles),
 J_S= numeric(NumberArticles))

#compute the relevant variables for all pairs of documents
for(i in 1:NumberArticles){

#empty everything before the start of a new iteration
MediaName <- NULL
CBName<-NULL
Theta1<-NULL
Theta2<-NULL
M<-NULL
temp1<-NULL
temp2<-NULL
KL1<-NULL
KL2<-NULL
JS<-NULL

MediaName <- INPUT[i,1]
CBName <- INPUT[i,2]
#Folder names with slash at the end
DefaultLinkMedia <- "C:/Users/Gebruiker/Documents/Data_Scriptie/Data_relevant_and_irrelevant/
Statements/Combined/"
DefaultLinkCB <- "C:/Users/Gebruiker/Documents/Data_Scriptie/Statements/"

#Generate paths to relevant names
MediaLink <- paste(DefaultLinkMedia, MediaName, sep="")
CBLink <- paste(DefaultLinkCB, CBName, sep="")

#import relevant documents and perform necessary formalities
TextMedia <- textfile(MediaLink)
TextCB <- textfile(CBLink)
CorpusMedia <- corpus(TextMedia)
CorpusCB <- corpus(TextCB)
metadoc(CorpusMedia, "language") <- "english"
metadoc(CorpusCB, "language") <- "english"
tokenize(CorpusMedia, removeNumbers = TRUE, removePunct = TRUE, removeSeparators = TRUE)
tokenize(CorpusCB, removeNumbers = TRUE, removePunct = TRUE, removeSeparators = TRUE)

#generate dfm vectors and allow for mathematical operations
mydfmMedia <- dfm(CorpusMedia, ignoredFeatures = c(stopwords("english")))
mydfmCB <- dfm(CorpusCB, ignoredFeatures = c(stopwords("english")))
mydfmCombined <- rbind(mydfmMedia, mydfmCB)
```

```
DFM<-as.data.frame(mydfmCombined)
DFMMedia <- data.matrix(DFM[1,], rownames.force = NA)
DFMCB <- data.matrix(DFM[2,], rownames.force = NA)

Theta1 <- DFMMedia/sum(DFMMedia)
Theta2 <- DFMCB/sum(DFMCB)
M <- (Theta1+Theta2)/2

#compute JS divergence
temp1<-log(Theta1/M)
temp2<-log(Theta2/M)
KL1 <- Theta1 %*% t(replace(Theta1, Theta2==-Inf,0))
KL2 <- Theta2 %*% t(replace(Theta2, Theta2==-Inf,0))
JS <- (KL1 + KL2)/2

#assign JS values to output matrix

OUTPUT[i,1] <- toString(MediaName)
OUTPUT[i,2] <- toString(CBName)
OUTPUT[i,3] <- toString(JS)

#used to keep track of the time
print(i)

}

write.csv(OUTPUT, "F:/JS_analysis_statements.csv")
```

B.5 Code Used for to Compute Kullback-Leibler Divergence

```
#Code used to compute KL divergence
#Code can be used for media coverage of speeches and statements separately

#Install and import the quanteda package
install.packages("quanteda")
library(quanteda)

#Create dfm for all CB communication, has to be done in parts due to memory limitations
textfileStatements1 <- textfile("C:/Users/Gebruiker/Documents/Data_Scriptie/
Data_relevant_and_irrelevant/Statements/Combined/part_1/*.txt")
corpusStatements1 <- corpus(textfileStatements1)
textfileStatements1 <- NULL
metadoc(corpusStatements1, "language") <- "english"
tokenize(corpusStatements1, removeNumbers = TRUE, removePunct = TRUE, removeSeparators = TRUE)
mydfmStatements1 <- dfm(corpusStatements1, ignoredFeatures = c(stopwords("english")))
corpusStatements1 <- NULL

textfileStatements2 <- textfile("C:/Users/Gebruiker/Documents/Data_Scriptie/
Data_relevant_and_irrelevant/Statements/Combined/part_2/*.txt")
corpusStatements2 <- corpus(textfileStatements2)
textfileStatements2 <- NULL
metadoc(corpusStatements2, "language") <- "english"
tokenize(corpusStatements2, removeNumbers = TRUE, removePunct = TRUE, removeSeparators = TRUE)
mydfmStatements2 <- dfm(corpusStatements2, ignoredFeatures = c(stopwords("english")))
corpusStatements2 <- NULL

textfileStatements3 <- textfile("C:/Users/Gebruiker/Documents/Data_Scriptie/
Data_relevant_and_irrelevant/Statements/Combined/part_3/*.txt")
corpusStatements3 <- corpus(textfileStatements3)
textfileStatements3 <- NULL
metadoc(corpusStatements3, "language") <- "english"
tokenize(corpusStatements3, removeNumbers = TRUE, removePunct = TRUE, removeSeparators = TRUE)
mydfmStatements3 <- dfm(corpusStatements3, ignoredFeatures = c(stopwords("english")))
corpusStatements3 <- NULL

COMBINED <- rbind(mydfmStatements1, mydfmStatements2, mydfmStatements3)
SUM <- sum(COMBINED)

mydfmStatements1 <- NULL
mydfmStatements2 <- NULL
mydfmStatements3 <- NULL

#Take the media-central bank couples as input
INPUT <- read.table("F:/OLD/SERIOUS/statements.txt")
NumberArticles <- dim(INPUT)[1]

#Initialize a series of relevant variables
MediaName <- NULL
CBName<-NULL
Theta1<-NULL
Theta2<-NULL
temp<-NULL
temp1<-NULL
lambda <- 0.9

#Initialize output matrix
OUTPUT <- data.frame(NameMedia= numeric(NumberArticles),
NameCB=numeric(NumberArticles), K_L= numeric(NumberArticles))

#Go over all different pairs of media and central bank communication
for(i in 1:NumberArticles){

MediaName <- NULL
```

```
CBName<-NULL
Theta1<-NULL
Theta2<-NULL
KL12<-0

#retrieve article names
MediaName <- INPUT[i,1]
CBName <- INPUT[i,2]

#assign folder links to name
DefaultLinkMedia <- "C:/Users/Gebruiker/Documents/Data_Scriptie/Data_relevant_and_irrelevant/
Statements/Combined/"
DefaultLinkCB <- "C:/Users/Gebruiker/Documents/Data_Scriptie/Statements/"

#generate paths to files
MediaLink <- paste(DefaultLinkMedia, MediaName, sep="")
CBLink <- paste(DefaultLinkCB, CBName, sep="")

#import documents and perform formalities
TextMedia <- textfile(MediaLink)
TextCB <- textfile(CBLink)
CorpusMedia <- corpus(TextMedia)
CorpusCB <- corpus(TextCB)
metadoc(CorpusMedia, "language") <- "english"
metadoc(CorpusCB, "language") <- "english"
tokenize(CorpusMedia, removeNumbers = TRUE, removePunct = TRUE, removeSeparators = TRUE)
tokenize(CorpusCB, removeNumbers = TRUE, removePunct = TRUE, removeSeparators = TRUE)

#generate document feature matrices and make them operational
mydfmMedia <- dfm(CorpusMedia, ignoredFeatures = c(stopwords("english")))
mydfmCB <- dfm(CorpusCB, ignoredFeatures = c(stopwords("english")))
mydfmCombined <- rbind(mydfmMedia, mydfmCB)
DFM<-as.data.frame(mydfmCombined)
DFMMedia <- data.matrix(DFM[1,], rownames.force = NA)
DFMCB <- data.matrix(DFM[2,], rownames.force = NA)

#compute theta vectors
Theta1 <- DFMMedia/sum(DFMMedia)
Theta2 <- DFMCB/sum(DFMCB)

#do the same procedure for all words in the vector
for (j in 1:length(DFMMedia)){
LOG<-NULL
temp1<-NULL
temp<-NULL
if (Theta1[j]==0){
} else if (Theta2[j]==0){
temp1 = colnames(DFMCB)[j]
temp = (1-lambda)*sum(COMBINED[,temp1])/SUM
KL12 <- KL12 + Theta1[j]*log(Theta1[j]/temp)
} else {
LOG <- log(Theta1[j]/Theta2[j])
LOG <- replace(LOG,LOG==-Inf,0)
LOG <- replace(LOG, is.nan(LOG),0)
KL12 <- KL12 + Theta1[j]*LOG
}
}

#assign values to the output frame
OUTPUT[i,1] <- toString(MediaName)
OUTPUT[i,2] <- toString(CBName)
OUTPUT[i,3] <- toString(KL12)

print(i)

}

write.csv(OUTPUT, "F:/KL_analysis_statements_laptop.csv")
```

C Appendix - Vector Space Model Distances

Define the following texts:

- A = "I have"
- B = "I have have"
- C = "I have no"
- D = "I have have no"
- E = "I have no money"
- F = "I have have have"

Table 19: Several lexical distance measures for the vector space model from text A to the others.

	Cosine	Dice	Jacquard
B	0.95	0.86	0.75
C	0.82	0.80	0.67
D	0.87	0.75	0.60
E	0.71	0.67	0.50
F	0.89	0.67	0.50

From Table 19 one can observe that the Dice and Jacquard measure perform similarly. The cosine measure, however, deviates from the other two in two cases: for text C and D and for text E and F. In both cases the cosine measure prefers the text with repetitive elements over those without repetition instead of newly used words. For comparative purposes, consider the generalized version of B and F, say G which consists of the term "I" together with n times the term "have". For G, the cosine distance to A monotonically decreases to 0.707 since $\lim_{n\to\infty} cos(A, G) = \frac{1}{\sqrt{2}}$. The other two measures, however, decrease to 0 when n approaches infinity.

D Appendix - Laver and Garry (2000) Proportions

The differences in proportions obtained by using the Laver and Garry (2000) dictionary can be investigated by using regression analysis. For each of the difference series I have constructed a model by means of a top-down approach in which I have omitted the variable with the least significant coefficient[17] one by one iteratively. This procedure has been done using both the `Banking` and `Centralb` dummies on the one hand and the `C or B` dummy which takes value one if any of the authors has either banking or central banking specific experience. Models have been selected based on the value of the adjusted R^2-value. The four final models are presented in Table 20 below. One can see that all four models have limited explanatory power.

Table 20: Models for several differences in proportions based on the Laver and Garry (2000) dictionary. Estimated coefficients are displayed together with their p-values.

	d_prop_less		d_prop_eq		d_prop_pro		d_total	
Constant	-0.0045	0.000	-0.0021	0.004	-0.0002	0.228	-0.0069	0.000
Speech	0.0030	0.000			-0.0008	0.000	0.0020	0.000
FT	0.0008	0.017	-0.0020	0.000	-0.0004	0.000	-0.0017	0.014
WSJ	-0.0003	0.607	-0.0080	0.000	-0.0011	0.000	-0.0096	0.000
Print	0.0010	0.000	-0.0006	0.116	0.0002	0.025	0.0008	0.074
Experience	-0.0001	0.000	-0.0001	0.019	1.12E-05	0.072	-0.0002	0.000
Banking	0.0025	0.000	-0.0039	0.000	-0.0008	0.000	-0.0023	0.001
Centralb	-0.0020	0.000	0.0026	0.000	0.0008	0.000	0.0012	0.141
Gender	-0.0010	0.000	0.0012	0.017	0.0004	0.000	0.0008	0.209
Advanced	0.0005	0.042	-0.0022	0.000	-0.0003	0.001	-0.0019	0.002
Award	-0.0010	0.052	0.0084	0.000	0.0007	0.000	0.0084	0.000
Domestic	0.0001	0.766	-0.0030	0.000	-0.0007	0.000	-0.0034	0.000
Shanghai100	0.0002	0.261	0.0058	0.000	0.0004	0.000	0.0063	0.000
Undergraduate	0.0011	0.000			0.0002	0.053	0.0015	0.007
Adjusted R^2	0.0430		0.0465		0.0418		0.0394	
Log likelihood	31848.22		26593.26		42089.32		24001.26	
F-statistic	34.50		45.34		33.54		31.6206	
Observations	9702		9995		9702		9702	

In the model for `d_prop_less` it turns out that speeches have more policy words related to limited government intervention than statements. Articles by the FT also contain more policy words related to limited government intervention compared to articles by Reuters and the WSJ. The same holds for printed articles, which also have a larger proportion of these words compared to online articles. Journalists with much experience, central banking experience, advanced education or those who are male use smaller proportions of policy words related to limited government intervention. Articles written by journalists with banking experience, who have won an award or have a relevant undergraduate degree exhibit relatively higher proportions of policy words related to a small government.

Both the FT and the WSJ use proportionally less words from the category 'equal state' than Reuters. Articles written by journalists with more experience, banking experience,

[17] An exception on this rule has been the dummy for the `WSJ`. If I were to exclude this one, the `FT` dummy would have lost (part) of its interpretation.

advanced education or who are domestic exhibit smaller proportions of words related to an 'equal state'. On the other hand, articles written by journalists with central banking experience, who are male, have won an award or have finished an undergraduate degree at one of the universities in top 100 of the Shanghai academic ranking will be inclined to use relatively more of these words. The conclusions for `d_prop_pro` are of similar nature. Only printed media have relatively higher proportions of words related to intervention compared to online media, while speeches are covered with less of these compared to statements.

Overall, media coverage on speeches contains more economic policy related words than that of statements. The WSJ uses least of these, followed by the FT; Reuters uses most, but still less than their source. Experienced journalists are likely to cover with less policy related words compared to their younger colleagues. Journalists with banking experience on average also cover central bank communication using less policy related words. Gender is not of any relevance in this context. Those journalists with advanced education or those covering domestic affairs are less inclined to use policy related words as well. Those journalists who have won an award, or journalists who have completed an undergraduate major in a relevant field or at a university from the first 100 on the Shanghai 2015 academic ranking cover with more relatively more policy related words.