MACIE PAPER SERIES

Marburg Centre for Institutional Economics

MACIE

Nr. 2021/01

Whatever it takes to understand a central banker - Embedding their words using neural networks.

Johannes Zahner
MACIE, Philipps-Universität Marburg

Martin Baumgärtner
THM Business School, JLU Gießen

Marburg Centre for Institutional Economics • Coordination: Prof. Dr. Elisabeth Schulte
c/o Research Group Institutional Economics • Barfuessertor 2 • D-35037 Marburg

Phone: +49 (0) 6421-28-23196 • Fax: +49 (0) 6421-28-24858 •
www.uni-marburg.de/fb02/MACIE • macie@wiwi.uni-marburg.de

Philipps Universität Marburg

Whatever it takes to understand a central banker - Embedding their words using neural networks.*

By Martin Baumgärtner† and Johannes Zahner‡

Draft: August 2, 2021

Dictionary approaches are at the forefront of current techniques for quantifying central bank communication. This paper proposes embeddings – a language model trained using machine learning techniques – to locate words and documents in a multidimensional vector space. To accomplish this, we gather a text corpus that is unparalleled in size and diversity in the central bank communication literature, as well as introduce a novel approach to text quantification from computational linguistics. Utilizing this novel text corpus of over 23,000 documents from over 130 central banks we are able to provide high quality text-representations –embeddings– for central banks. Finally, we demonstrate the applicability of embeddings in this paper by several examples in the fields of monetary policy surprises, financial uncertainty, and gender bias.

JEL: C45, C53, E52, Z13
Keywords: Word Embedding, Neural Network, Central Bank Communication, Natural Language Processing, Transfer Learning

* We are grateful to helpful comments from Matthias Neuenkirch, Bernd Hayo, Jens Klose, Peter Tillmann, Elisabeth Schulte, Ania Zaleska and other participants of the 7th International Young Finance Scholars' Conference and the MAGKS Doctoral Colloquium 2021. Furthermore, this paper immensely benefited from the support of Matthias Neuenkirch and Hamza Bennani during the collection of its text-corpus. Finally, we want to thank Peter Winker for discussing the relevant text-mining methods.

† THM Business School, JLU Gießen, Germany, martin.baumgaertner@posteo.de

‡ Corresponding author. School of Business and Economics, Institutional Economics Research Group, Philipps-Universität Marburg, Germany, johannes.zahner@wiwi.uni-marburg.de.

I. Introduction

Over the last few decades, there has been an increase in the analysis and interpretation of central bank communication (Blinder et al., 2008). This development was accelerated by the zero lower bound and the emergence of forward guidance, wherein central bankers recognized the possibility to complement actions with well-placed language to steer market participants towards a desired equilibrium path. As a result, central banks increased their communication substantially. The Federal Open Market Committee (FOMC), for example, started publishing press conferences since 2011, and the European Central Bank (ECB) began disclosing monetary policy meeting minutes in 2015.

Consequently, a substantial body of literature has emerged that developed methods for quantifying different aspects of such communication. The literature is still primarily driven by dictionary approaches, in which pre-defined dictionaries, such as Loughran and McDonald (2011), Apel and Grimaldi (2014), and Picault and Renault (2017), are used to count terms (for example, positive and negative) to extract a single dimension (for example, sentiment) from a text corpus. However, these methods fall short of incorporating the language's richness, its multidimensionality, and its context-dependence. Moreover, dictionary approaches are inherently subjective, as discussed in Gentzkow, Kelly, et al. (2019, p. 554)'s survey on text mining in economics, where the authors emphasize that *"dictionary-based methods heavily weigh prior information."*

To address these weaknesses, we turn to linguistic and computer science, using machine learning tools to develop a novel *language model*. Such a model can be estimated from a set of texts – the *corpus* –, and an *algorithm* that locates words a multidimensional vector space. In this vector space, conceptually similar terms are mapped in close proximity, reflecting meaningful relationships.

Following recent advances in computational linguistics, we propose (word) embeddings as a novel language model for quantifying central bank communication. Therefore, the main objective of this paper is to bring computational linguistics research into the economic sphere. By developing a language model trained explicitly for monetary policy, our focus is essentially twofold. On the one hand, we sharpen the previously broad focus of embeddings, while, on the other hand, we enhance content extraction compared to the simplicity of dictionary approachesWe see this paper as an essential step in the endeavor of modern text quantification, initialized by Gentzkow, Kelly, et al. (2019, p. 553) who state that *"approaches [...] which use embeddings as the basis for mathematical analyses of text, can play a role in the next generation of text-as-data applications in social science"*.

This paper contributes to the current literature on several fronts. First, we collect a novel text-corpus unparalleled in size and diversity. The corpus, which contains approximately 23.000 speeches by 130 central banks, is considerably larger than any previously used in the central bank communication literature. Second, this paper introduces novel machine learning algorithms for text quantifying. We com-

pare a multitude of different algorithms according to objective criteria. Third, by training the novel algorithm on the novel text corpus, we introduce a language model previously unseen in monetary policy (and likely economics). We demonstrate how this language model can be used in various applications throughout this paper, such as examining the effect of central bank speeches during the Euro Area crisis, predicting monetary policy surprises, comparing central bank objectives, and measuring gender bias. Finally, by making the language model publicly available, this paper's most important contribution is to make this new string of research accessible to other researchers, allowing them to incorporate embeddings into their own research.

The remainder of this paper is structured as follows. Section II provides a literature overview of the current state of natural language processing (NLP) in monetary economics. In Section III we introduce both the text corpus and the algorithms, combining both elements into language models. We then evaluate the quality of the resulting embeddings in the central bank context in Section IV before applying the best performing language model in Section V.The final section concludes this paper.

II. Related literature

NLP has established itself in the central banking literature with an abundance of high-quality research. There are several methods available to researchers for quantifying qualitative information; Gentzkow, Kelly, et al. (2019) provides an excellent survey on the use of text data with a focus on economics.

Rather than the explicit analysis of text, tracking market reactions during periods when a text is published is a frequent dimensionality reduction method. This strand of literature disregards the qualitative data provided and instead entirely focuses on the market's interpretation of the text. Among successful implementations are Gürkaynak et al. (2005), Brand et al. (2010), Jarociński and Karadi (2020), and Swanson (2021) who utilize intraday data around the reading of press-conference statements to measure the effect of monetary policy decisions.

When working with text data, a different approach is to manually classify them, whereby humans categorize sentences, paragraphs or even sections and thus quantify the qualitative information themselves. Although the process is labour-intensive and prone to misclassification, it allows the researcher to capture highly specific patterns. Ehrmann and Fratzscher (2007) use manual classification to compare different types of communication between central banks, and Tillmann (2020) classifies answers during the ECB press conference's Q&A to estimate a disagreement index.

However, most applications today concentrate on rule-based classification utilizing computers. Precisely, the majority of NLP in economics focuses on so-called dictionary methods, whereby a predefined dictionary classifies certain words, thereby quantifying the qualitative information into few dimensions. Famous examples in economics include the calculation of an uncertainty and recession index by

counting respective terms in news-articles (e.g. Baker et al., 2016; Ferrari and Le Mezo, 2021), stock market predictions using a psychosocial dictionary on a Wall Street Journal column (Tetlock, 2007), or measuring media slant in American news-outlets from phrase frequencies in Congressional Records (Gentzkow and Shapiro, 2010). There are also numerous applications utilizing dictionaries in the context of central bank communication. In fact, dictionaries have been explicitly designed for the use in financial and central bank context (e.g. Loughran and McDonald, 2011; Apel and Grimaldi, 2014; Picault and Renault, 2017; Correa et al., 2021). The peculiarity of the terminology spoken in the central bank context necessitates the usage of such central bank-specific dictionaries. These dictionaries have been applied in numerous ways, for example, to measure implied inflation targets (Shapiro and Wilson, 2019; Zahner, 2020), home biases of central bankers (Hayo and Neuenkirch, 2013) or financial stability objectives (Peek et al., 2016; Wischnewsky et al., 2021).

The benefit of dictionary-based methods is their ease of understanding and evaluation through their straightforward and transparent quantification of an underlying corpus. However, at the same time, they lack objectivity and omit relevant information. By definition, dictionaries are subjective, as researchers define a subset of a language's vocabulary based on their own assessment of the underlying true meaning of the respective word. Furthermore due to the low dimensionality, dictionaries are incapable of capturing nuance as well as interactions between terms. For example, the phrase *great recession* is classified as neutral in Loughran and McDonald's (2011) sentiment dictionary, even though the term *great* is not meant to be positive in this context. Finally, a substantial portion of text is omitted when relying on a dictionary, an argument made before by Harris (1954, p. 156), who state that *"language is not merely a bag of words but a tool with particular properties which have been fashioned in the course of its use"*.

Recent research recognizes and highlights the dictionary approach's disregarding element, suggesting either augmenting such an index or combining different dictionaries to improve predictive power. Tadle (2021), for instance, uses the former approach utilizing two dictionaries (one for hawkish/dovish and the other for positive/negative), rejecting a sentence's classification as hawkish or dovish if it contained more negative than positive terms. The author shows how this augmented sentiment index helps explain movements in high-frequency variables during the FOMC press conference. Another famous example is the interaction of topic-modelling and sentiment analysis by Hansen and McMahon (2016) and Fraccaroli et al. (2020). A different approach is applied by Azqueta-Gavaldon et al. (2019), Kalamara et al. (2020), Shapiro, Sudhof, et al. (2020), and Gorodnichenko et al. (2021), who combine different sentiment indices in a regression model at the same time. They find that different dictionaries capture various aspects of an underlying corpus and can thus complement each other.

In addition to these augmentations, alternatives to dictionary approaches are becoming more popular. One example is the concept of *similarity*, which is oper-

ationalized using the distance between two documents' vocabulary. This metric gained popularity through Acosta and Meade (2015), Amaya and Filbien (2015), and Ehrmann and Talmi (2020), who find that introductory statements became more similar over time. Another example is the measurement of verbal complexity, which is commonly approximated with the Flesch-Kincaid grade level by Kincaid et al. (1975). Smales and Apergis (2017) and Hayo, Henseler, et al. (2020) illustrate that markets react strongly concerning the complexity of the information communicated in press statements. As helpful as these new approaches are, some of the corpus' relevant underlying information remains neglected. For example, exchanging the term *inflation* with *deflation* does not change the level of complexity but substantially alters the message.

In the last years, embeddings have entered the realm of monetary policy, following a trend predicted by Gentzkow, Kelly, et al.'s (2019) quote in the introduction. Word embeddings are multidimensional word representations that are used to measure similarity in Twitter tweets (Masciandaro et al., 2020), in the development of a real-time economic sentiment index (Aguilar et al., 2021), for the improvement of the Euro Area uncertainty index (Azqueta-Gavaldon et al., 2019), for the decomposition of central bank vague talk (Hu and Sun, 2021), and to measure central banker disagreement (Apel, Grimaldi, and Hull, 2019). Generally, economic research relies on general language models trained on a general text corpus such as Wikipedia. Shapiro, Sudhof, et al. (2020), for example, use such embeddings in their analysis of news articles. The authors are unconvinced by the results and resort to the modified dictionary approach mentioned earlier. However, the lack of predictive power is most likely the result of the limited sample size and may possibly be due to the absence of specificity in the training corpus. For example, some general language models lack relevant monetary policy specific terms, such as *hicp*. One notable exception, and thus methodologically the closest research to our paper, is Apel, Grimaldi, and Hull (2019), who employ a recurrent neural network to develop their disagreement metric, thereby training word embeddings as a byproduct. However, the authors neither disclose information about their embeddings, nor use them outside this specific context.

To the best of our knowledge, we are the first to train embeddings on a specific text corpus and apply the language model to a variety of applications. Thereby, this paper touches two different literature strings. On the one hand, in the development of novel text-representation (Apel, Grimaldi, and Hull, 2019), and on the other hand, in the need to fine-tune these representations for their respective use (Loughran and McDonald, 2011).

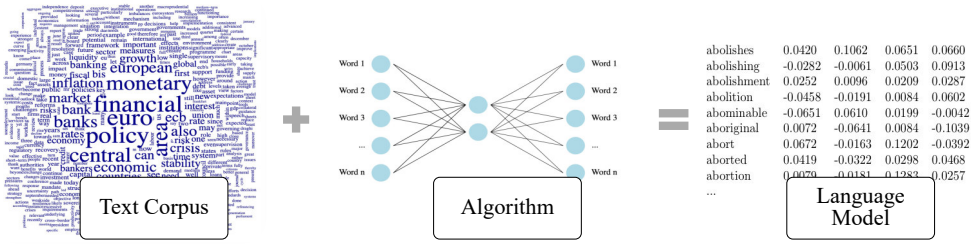
III. Methodology

> *"The meaning of words lies in their use"*
>
> — Wittgenstein (1958, p. 80)

A language model maps a text corpus into an n-dimensional space, whereby the

model itself can be arbitrarily simple. Take, for instance, dictionary approaches in sentiment analysis that classify terms as positive, negative and neutral, thereby mapping a corpus' vocabulary into a single dimension. This paper's proposed language model is a multidimensional representation called embedding, received by training an algorithm on a text corpus. A stylised overview of the procedure - and an overview of the structure of this section - is provided in Figure 1.

Figure 1 : How to get a language model



A. Text Corpus

Our text corpus reflects our paper's primary objective on monetary policy. To make the corpus as broad as possible, we acquire all English central bank speeches published by the Bank for International Settlements (BIS).[1] We complement the corpus with as much meta-information as possible, collecting title, speaker, role of speaker, event at which the speech was delivered, and further information. In the next step, we enrich the corpus with documents gathered from central bank websites. Among them are reports, minutes, forecasts, press conferences and economic reviews. To keep our corpus as homogeneous as possible, we exclude all presentations and scientific papers. The former usually contain little coherent text; the latter are primarily oriented towards the academic literature in their jargon and are thus not official central bank communication. The use of information on the respective central bank allows us to create features for the country, the currency area and each central banker. We provide a set of descriptive illustrations in the appendix.

Compared to the previous NLP application in monetary policy (e.g. Amaya and Filbien, 2015; Hansen and McMahon, 2016; Ehrmann and Talmi, 2020), we apply a minimum of pre-processing on the text corpus. This is generally done in the embeddings literature (e.g. Mikolov, Yih, et al., 2013) since similar words should be close in the vector space, which eliminates the need for standardisation through stemming, lemmatisation or removal of stopwords. As a result, we limit the pre-processing to improve the expressiveness of the word tokens. First, we identify

[1] We determine the language of the individual text using Google's Compact Language Detector 3.

Table 1: Corpus Summary

Source	Type	n
BIS	Speech	16,627
FED	Minute, Press Conference, Transcript, Agenda, Blue-, Green-, Teal-, Beige- and Red-Book	2,238
BOJ	Minute, Economic Report, Release, Outlook Report	2,187
ECB	Minute, Press Conference, Economic Outlook, Blog	343
Riksbank	Minute, Economic Review, Monetary Policy Report	330
Australia	Minute	159
Poland	Minute	156
Iceland	Minute	101

Note: The table summarizes the number of documents (n) by sources in the our text corpus.

so-called collocations, that is, words with specific meaning when used together. It is important to notice the distinctive features of collocation and context were already highlighted by Firth (1957), whereas *"collocation is not to be interpreted as context, by which the whole conceptual meaning is implied"* but as *"mere word accompaniment"*. One example is the words *federal* and *reserve*, which have one specific meaning when used together. Another example is the word *quantitative*, which in itself means expressible in terms of quantity. In contrast, *quantitative easing* represents a specific measure of central banks that cannot be concluded from its individual parts. To map these relationships in the embeddings, it is advantageous to identify related words and combine them as a token, for example, *federal_reserve* or *quantitative_easing*. To do this efficiently in our large corpus, we use the algorithm introduced by Blaheta and Johnson (2001) to obtain a basic set of collocations. Furthermore, we form collocations from all speakers of the BIS corpus. For example, *ben* and *bernanke* becomes *ben_bernanke*.

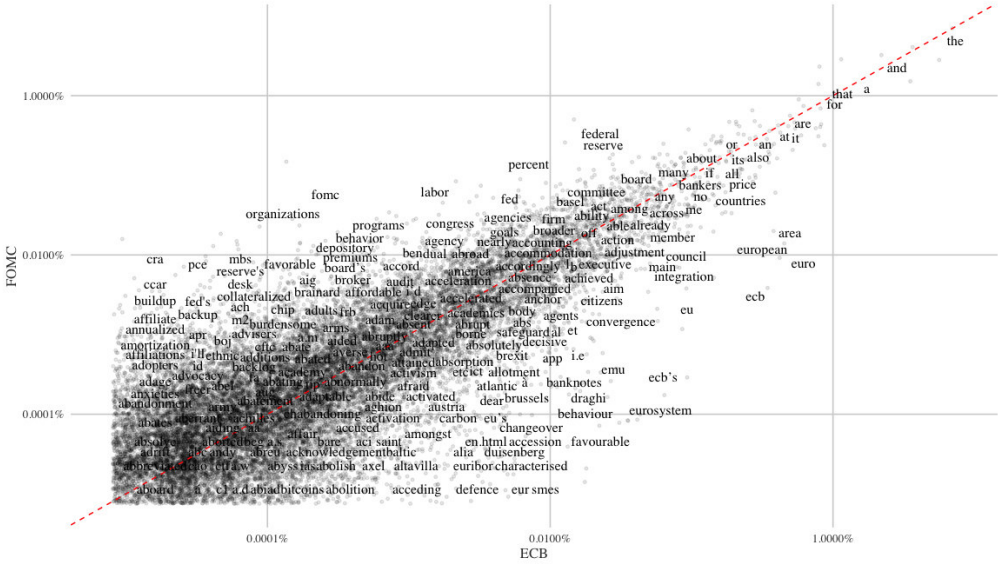
Second, to keep the embeddings as uniform as possible, we replace several unique entities with placeholder tokens. Therefore, all email addresses are encoded as [email], URLs by [url], Unicode tokens by [unicode] and decimal numbers by [decimal]. Furthermore, we remove all apostrophes and quotation marks. In a final step, we convert the entire text to lower case.

Our final corpus includes over 23.000 documents, more than 100 million individual word-tokens, more than 130 central banks worldwide and over 1,000 individual speakers. As a result, on the one hand, we have a text corpus that is unprecedented in quantity and diversity in the monetary policy literature, and on the other, containing highly specific central bank vocabulary.

The corpus' homogeneity is what we address next. To compare the central bank's jargon, we estimate and compare the relative word frequency for the seven most frequent central banks in our sample. An illustration for the ECB and the FOMC is provided in Figure 2. Formally testing the homogeneity, we discover that neither of the six central banks has a correlation below 98 percent in their relative word

use when compared to the ECB, implying that jargon is very homogeneous across central banks.[2] We conclude from this that the institutions do not differ in any relevant way concerning their jargon. We conduct the same test for the underlying years and derive qualitatively and quantitatively the same result.

Figure 2 : Illustration of frequency of used terms between FOMC and ECB



Notes: This graph depicts the wording of the FOMC and its European counterpart. The relative frequency of each word is measured for both central banks and presented in this jitter plot. To make it easier to read, numbers and terms that appeared only in the texts of one central bank (mainly names) are removed. In addition both we scaled both axes by the logarithm and added noise, so the correlation is even stronger than shown here.

B. Algorithm

Our emphasis in this subsection is on the introduction of algorithms that provide numerical representations for text documents. Due to the evident shortcomings of traditional dictionary approaches mentioned in the previous section, we turn to linguistics and computer science for our language model.
Modern language models follow the proposition of leading linguistic academic Zellig S. Harris in their pursuit of superior text representation. According to Harris (1954, p.151), *"meaning is not a unique property of language, but a general characteristic of human activity"*, implying that the distinction between meaning and

[2]The precise values are: Federal Reserve (Fed): 98.7% (t = 884.67), Riksbank: 98.0% (t = 585), Bank of England (BoE): 98.9% (t = 966), Bank of Japan (BoJ): 98.4% (t = 668), Bundesbank: 99.4% (t = 1257), and Central Bank of India: 98.5% (t = 783). The results are also illustrated in the Appendix in Figure A2.

the quantifiable properties of language is not always evident. His distributional theory builds on this observation and approximates the meaning of words using the distribution over the environments (context) a word occurs. If a word (for example, *outlook*) can be found repeatedly in the same environments as another word (for example, *forecast*), these words represent a similar concept, whereas the difference in environments corresponds to the difference in meaning.

In the following, we will introduce four algorithms building on the distributional hypothesis that we will subsequently apply to obtain embeddings. These algorithms can be broadly divided into two categories: prediction-based methods and count-based methods. The former use surrounding words to make predictions, whereas the latter uses corpus-wide statistical properties such as word co-occurrence – that is, how often words appear together. The following section introduces both categories and their most prevalent techniques.

Prediction-based algorithms

Prior to formally introducing the algorithms, we provide a simplistic example to facilitate comprehension between the concepts of terms, target words and context. Following Harris (1954)'s distributional theory, a word's meaning is based on the environment in which it appears. The context of a word, the set of its surrounding words, operationalizes this environment. Given a context window of one, the context of the word *brighter* (called the target word) in the following sentence would be *this* and *outlook*:

> *"[...] this* **brighter** *outlook remains subject to considerable uncertainty, also regarding the path of the pandemic [...]"*
>
> — Christine Lagarde, IMF Spring Meetings, 8 April 2021

The prediction-based algorithms are generally tasked to predict the target word given the context words, i.e. *p(brighter | this, outlook)*. They then proceeds with the next target word, i.e. to predict *p(outlook | brighter, remains)*, then *p(remains | outlook, subject)* and so on.[3] By optimizing some objective function, the algorithm improves its ability to predict target words based on their context. Note how the approach directly incorporates the previously stated linguistic premise by Harris (1954) whereas similar words occur in the same context. It also becomes evident why the context is key. Assume the model is given the (slightly larger) context *"this brighter outlook remains subject to considerable..."* from the preceding sentence and is tasked with predicting the next word. To perform well on this task on average, it must not only assign a high probability to the word *uncertainty*, but also to semantically similar words that frequently occur in the

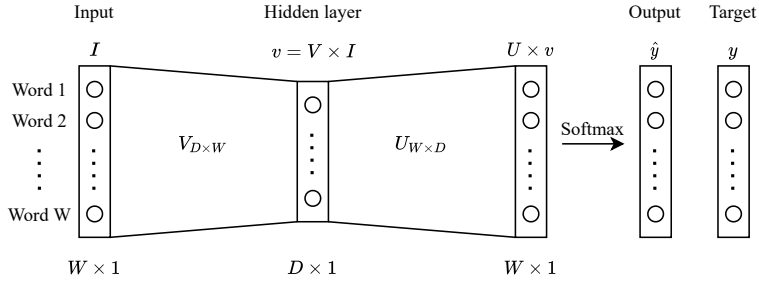
[3]The demonstrated example is called *continuous bag of words* model. In addition, there is a reverse approach, i.e. the algorithm is tasked to predict the context from the target word. This method is called *skip-gram*.

same context, such as *risk*. As a consequence of the prediction task, the algorithm places these words close to each other in the word-embedding space, ultimately capture the semantic meaning as a byproduct.

Word2Vec

The Word2Vec model of Mikolov, Yih, et al. (2013), Mikolov, Chen, et al. (2013), and Mikolov, Sutskever, et al. (2013) is based on the above principle. Building on the work of Bengio et al. (2003), Collobert and Weston (2008), and Turian et al. (2010), the authors propose a neural network capable of predicting words from their context. In doing so, the algorithm is both accurate and efficient.
Mathematically, Word2Vec, and similar prediction-based models, are single-layer log-linear models based on the inner product between two word vectors. The hidden layer's size determines the dimensionality of the word-embedding's representation. An illustration of such a model is provided in Figure 3.

Figure 3 : Graphical illustration of Mikolov, Yih, et al. (2013)'s Word2Vec model.



Notes: This figure illustrates the model architecture of a feed-forward neural network with three layers. The first layer is called the input layer, the second hidden layer, and the third output layer. The connections between the layer (particularly the nodes) are called weights and adjusted during the training process. The ensuing word-embedding matrix is, therefore, the projection of the input layer into the hidden layer. A second weight matrix maps the hidden layer into the output layer.

Formally, the target of the neural network underlying the Word2Vec approach is to predict a single word w_t – the target word – based on its surrounding words w_c – its context – for a vocabulary size W. The objective of the network is to maximize the log-likelihood:

$$L = \frac{1}{T}\sum_{t=1}^{T} \log p(w_t|w_c). \tag{1}$$

The probability of word w_t, given the words w_c is estimated using the following softmax function:

$$p(w_t|w_c) = \frac{exp(v_{w_t}^T v_{w_c})}{\sum_{w=1}^{W} exp(v_w^T v_{w_c})} \tag{2}$$

where v_i is the embedding vector. In other words, the models' functional structure represents a single linear hidden layer linked to a softmax output layer, where the exponential function prevents negative numbers and could be omitted without loss of generality. The objective is maximized using an iterative optimization algorithm (stochastic gradient descent, see, e.g. Chakraborty and Joseph, 2017; Athey, 2019) to identify a local – in best case global – maximum. Ultimately, we are only interested in the vector representations for the target words $\hat{V}$, as those are the corresponding embeddings.

There are several interesting points to note from this approach. First, the hidden layer's size is equivalent to the dimensionality D of the embeddings by design. This size has traditionally been set to 300 (e.g. Mikolov, Yih, et al., 2013), but different sized representations are entirely feasible. Second, it is apparent that the window size m (the context) significantly impacts the embedding. Since each word in the context has equal weight on the target prediction, a broad word context may not capture important semantic meaning. In contrast, a very narrow context may miss relevant details. The initial calibrations of Word2Vec and Doc2Vec (the following algorithm) used single-digit window sizes, namely five (Mikolov, Sutskever, et al., 2013) and eight (Le and Mikolov, 2014).

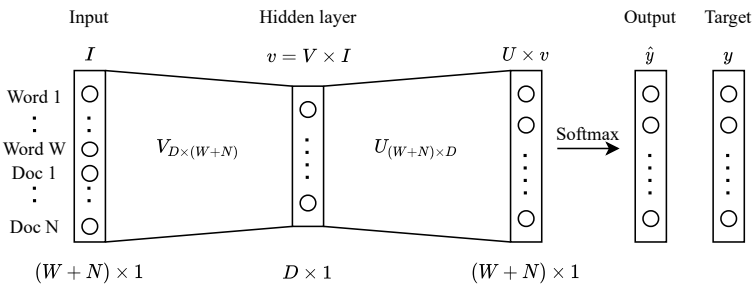
Third, due to the unsupervised nature of this machine learning model, there is no necessity to provide labelled data. In other words, no manual input is required to obtain the desired word embeddings, which is a substantial advantage since training such models necessitates a large training corpus. Furthermore, if the underlying text is sufficiently homogeneous, researchers can use a much larger text-corpus during the training phase of the language model compared to its final application. We utilise this advantage by training the central bank specific language model on texts from numerous central banks.

Doc2Vec

There are several extensions to the original Word2Vec model. The Doc2Vec approach by Le and Mikolov (2014), which proposes the inclusion of document specific information in the input layer, is one notable example. In its simplest form, Doc2Vec incorporates an ID for each document into the neural network's input layer, resulting in an embedding vector for each document. This representation is referred to as document embedding in the remainder of this paper. An illustration of the Doc2Vec model is provided in Figure 4.

This approach is intuitively similar to controlling for specific characteristics in traditional economic regressions, such as country-dummies in a panel regression. The main advantage of Doc2Vec over Word2Vec is that the document embedding can be used as a summary of the document in subsequent regressions. In Section IV and Section V, we will demonstrate how similarity in document embeddings may be used in a regression model. However, it should be noted that, unlike word embeddings, document embeddings cannot be easily transferred to new corpora.

Figure 4 : Graphical illustration of Le and Mikolov (2014)'s Doc2Vec model.



Notes: This figure is intended to provide an illustration of the Doc2Vec model architecture. It is inspired by Le and Mikolov (2014)'s depiction. The only difference to Figure 3 is the additional document ID being fed into the neural network. The ensuing word-embedding and document-embedding is the projection of the input layer into the hidden layer.

Count-based algorithm

An alternative to obtaining embeddings through neural networks is leveraging corpus-wide statistics to obtain word representations. Our analysis focuses on two approaches: one designed for topic modelling and the other developed explicitly as a substitute for the previously introduced prediction-based algorithms.

LDA

The most famous example of a count-based model in economics is unquestionably the Latent Dirichelet Allocation (LDA) algorithm. Since its introduction by Blei, Ng, et al. (2003), it has been used in monetary policy numerous times (e.g. Tobback et al., 2017; Hansen, McMahon, and Tong, 2019; Wischnewsky et al., 2021). We will not formally introduce the concept of LDA here owing to its popularity in economics and central banking. Interested readers are directed to Bholat et al. (2015) for an introduction to LDA in monetary policy text-mining applications. The premise of LDA is that documents contain a combination of latent topics, which themselves are based on a distribution over words in the underlying corpus. The generative probabilistic model is used in most economic applications to uncover latent topics in a corpus. As a byproduct, LDA generates topic distributions over the vocabulary as well, a concept closely related to the embedding matrices of prediction-based approaches, which is why we incorporate LDA into our analysis.

However, there are several distinctions between our application and previous ones in economics. First, to the best of our knowledge, these "topic"-embeddings have never been used in an economic context. Second, the number of topics – an important hyperparameter in LDA– varies widely across applications, ranging from two (Schmeling and Wagner (2019)) to 70 (Hansen, McMahon, and Prat (2018)), although in general, the number of topics does not exceed 50 in the

economic literature. As our objective is to maximise predictive power and to keep LDA comparable to others algorithms, we cover a much larger number of topics, namely 300. Finally, in economic applications, the identification and analysis of latent topics are generally the main priority. We refrain from interpreting (or even selecting) topics in the same fashion as we do for all other algorithms.

GloVe

The most famous count-based algorithm in NLP is the global factorization method, called GloVe. Following the success of Word2Vec, Pennington et al. (2014) propose GloVe, which trains a language model on global word co-occurrences. The approach is based on the notion that the global relative probability of terms, co-occurring in the same context, captures the relevant semantic information. Formally, the following least squared regression model is proposed:

$$L = \sum_{t,c=1}^{W} f(X_{t,c})(w_t^T w_c + b_c + b_t - \log X_{t,c})^2. \tag{3}$$

In Equation (3) w_t is the word-embedding vector for word t, $f(.)$ is a concave weighing function, b_c and b_t are bias expressions, and $X_{t,c}$ the co-occurrence counts for the context and target word within a defined window. Equation (3) is then iteratively optimized given the scale of the regression. The authors find substantial improvements over Word2Vec using the same corpus, vocabulary, and window size.

General corpus models

As mentioned in the introduction, no attempts have been made to train embeddings specifically for the economic context to the best of our knowledge. This may be due to the computational burden, the necessary amount of text, or other factors. An alternative to training embeddings from scratch is the use of pre-trained general language models called *transfer learning*(e.g. Binette and Tchebotarev, 2019; Doh et al., 2020; Istrefi et al., 2020; Shapiro, Sudhof, et al., 2020; Hu and Sun, 2021). These are open-source language models that have been trained on large general corpora. Since pre-trained language models are methodology-independent, one can find both pre-trained GloVe models and pre-trained Word2Vec models. We compare our results to two such general models as a benchmark: Glove6B and Word2Vec Google News.[4]

In Table 2, we provide an overview of all algorithms and corpora applied in this paper to train the language models. Since many algorithms can be computed in

[4]GloVe6B (Pennington et al., 2014) is trained on 6 billion tokens from Wikipedia text and News articles with a vocabulary of 0.4 million tokens. Word2Vec News Articles (Le and Mikolov, 2014) results from the original paper and is trained on Google News articles.

different configurations, we test different specifications here. The hyperparameters we used for each model can be found in the Appendix A.A2.

Table 2: Model Overview

Model	Word embedding	Document embedding	Corpus
Word2Vec	x		CB corpus
Word2Vec GoogleNews	x		Google News
GloVe	x		CB corpus
GloVe6B	x		Wikipedia/Gigaword
Doc2Vec	x	x	CB corpus
LDA	x	x	CB corpus

Note: The columns 'Word embedding' and 'Document embedding' refer to the model language model's ability to generate the respective embeddings. 'CB' is used as an abbreviation for 'Central Bank'. Word2Vec GoogleNews refers to the Le and Mikolov (2014) language model and GloVe6B refers to Pennington et al. (2014).

IV. Evaluation of language models

In this section, we apply the algorithms introduced in the previous section to our corpus and evaluate the results. In this way, we expect to answer the question of which algorithm summarizes the content of the text corpus best and thus provides the most convincing language model. Due to the algorithm's heterogeneity – Doc2Vec and LDA estimate document embeddings in addition to word embeddings – we proceed by estimating a word representation and a document representation whenever possible.
Since there exists no benchmark for evaluating language models in economics yet, we turn to the fields of traditional linguistics. There, evaluation tasks can be broadly distinguished as intrinsic or extrinsic. Intrinsic procedures examine whether the embeddings reflect an assumed relationship between words. One typical task would be to determine whether the vectors indicate associations similar to humans' perceptions. Another task would be the ability to find word analogies that resemble real analogies. We present several intrinsic evaluations at the second part of this section.

A. Extrinsic evaluation

Extrinsic tasks involve evaluating the vectors against other, externally known contexts, i.e., assessing the embeddings' ability to solve specific tasks. Typical methods would be classification tasks or named-entity recognition. However, the datasets on which these tasks generally rely on are designed to evaluate embeddings in a broad context, when we are interested in the opposite, their domain specificity. Due to a lack of external evaluation methods, we follow Le and Mikolov (2014) and evaluate each model's predictive performance in a classification task.

Our evaluation task concerns the current interest rate level of the ECB and FOMC, which we forecast using the respective central bank's speeches.[5] Since we are primarily concerned with the correct level, we divide the corresponding 3-month interbank rates into quintiles to derive our evaluation target.[6] Finally, as we are interested in the best possible performance, we employ a neural network to predict the respective interest rate levels with our embeddings.[7] This algorithm allows for complex non-linear relationships between the individual dimensions, which may be relevant. Each language model is trained on 75% of our data (the training sample), with the remaining observations serving as the test set for out-of-sample evaluation.

Table 3 summarizes the accuracy of the evaluation results split by Document- and Word Embedding and task. Since there exists several variants in the Word2Vec and Doc2Vec algorithms and we aim for a broad comparison, we estimate all variants. The name in column one starts with the algorithm followed by the variant's abbreviations.[8]

Our evaluation yields some interesting results. First, the federal funds rate level appears to be more challenging to predict across models. Second, we find a consistent difference in the level of accuracy between document embeddings and word embeddings. While the former are consistently above 40% accurate, only a few word embedding models achieve this level. Finally, the Doc2Vec algorithm appears to be most suitable for our context, outperforming the others on both the document and word levels.

As a result, we decide to concentrate on Doc2Vec as our primary algorithm. The Bag-of-words variant with pre-trained word-embeddings is explicitly chosen because of its high performance on the document level (being close to par for the ECB task and best at the FOMC task) and consistently good performance on the word embeddings task.[9]

[5]Note that we evaluate our language model on a sub-sample of the available embeddings. However, Appendix A.A3 demonstrates that the presented results are robust on a more general task.

[6]It is not uncommon in machine learning and monetary policy to convert a regression analysis into a classification one. The previously discussed Apel, Grimaldi, and Hull (2019) are one noteworthy example.

[7]With a few exceptions, the network structure closely resembles the representation in the previous section. We employ a single hidden layer neural network with 64 units, dropout regularization, and a Relu activation. Softmax activation is used once more in the output layer. The Adam optimizer is used to train the model on a categorical cross-entropy loss function. We tested various specifications, but the performance does not change substantially. The exact parameterization are available upon request.

[8]For further details, see Mikolov, Sutskever, et al. (2013) and Le and Mikolov (2014).

[9]Please note that the upcoming results are robust for the individual Doc2Vec variants. Results are available upon request. To ease comprehension, we will refer in the following to the language model "Doc2Vec Bow Pre" only as "Doc2Vec".

Table 3: Evaluation results of algorithms.

Algorithm	3-month Euribor	3-month FFR
Document Embeddings		
Doc2Vec Bow Pre	0.74	0.61
Doc2Vec Bow	0.75	0.59
Doc2Vec PVDM	0.70	0.48
Doc2Vec PVDM Pre	0.67	0.52
LDA	0.55	0.42
Word Embeddings		
Doc2Vec PVDM Pre	0.41	0.35
Doc2Vec Bow Pre	0.40	0.28
Doc2Vec PVDM	0.44	0.22
GloVe	0.38	0.22
Word2Vec GoogleNews	0.36	0.31
GloVe 6B	0.34	0.19
LDA	0.25	0.22
Doc2Vec Bow	0.21	0.25
Word2Vec Bow	0.20	0.21
Word2Vec Skipgram	0.19	0.21

Note: The table shows the evaluation results across the different algorithms introduced in the previous section. The accuracy was evaluated on a classification task with five categories + one outside option if the model was unsure. Therefore the expected performance would be $1/6 \approx 0.17$. With regards to the specifications: Bow = (Distributed) Bag Of Words; PVDM = Paragraph Vector Distributed Memory; Pre = pretrained embeddings were used as more efficient starting points.

B. Intrinsic evaluation

Following the extrinsic evaluation, we turn to an intrinsic assessment of our Doc2Vec model. As stated at the outset of this section, these assessments are inherently subjective and should therefore be viewed cautiously. The presented intrinsic evaluations are based on the cosine distance in the embeddings space, which is a measure of similarity between two-word vectors a and b of length n, and defined as follows:

$$similarity_{a,b} = \frac{a \cdot b}{||a|| \times ||b||} = \frac{\sum_{i=1}^{n} a_i \times b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \times \sqrt{\sum_{i=1}^{n} b_i^2}} \tag{4}$$

In the first evaluation, we compare the embeddings based on their assessment of which terms are most similar to a given word. We then assess the embeddings' ability to handle homonyms. Finally, we determine the central bank's similarity score and evaluate whether the relationship show meaningful results.
We present the first intrinsic evaluation of our embedding in Table 4, which lists

the most similar words based on the cosine distance to the words *inflation*, *unemployment*, and *output*. For example, the words *unemployment* and *joblessness* are relatively close to each other in our embedding space.

Table 4: Intrinsic Evaluation: Similarity in selected word embeddings.

inflation	unemployment	output
core_inflation	unemployment_rate	nonfarm_business
inflation_expectations	natural_rate	sector
economic_slack	joblessness	per_hour
underlying_inflation	jobless	output_growth
inflation_outlook	labor_force	producers
price_inflation	unemployed	manufacturing_output
actual_inflation	labor_market	factory
disinflationary	economic_slack	hourly_compensation
inflation_rate	unemployment_rates	business_equipment
disinflation	participation_rate	labor_costs

Note: The table shows the most similar terms to the words *inflation*, *unemployment* and *output* according to the cosine distance of the underlying word embeddings as defined by Equation (4). The underscore is used to highlight collocations as described in Section III.A.

It is evident that our language model is capable of grouping words with semantically similar meanings. For example, it is reassuring that through the our training process, several terms containing the word *inflation*, such as *core_inflation* and *inflation_expectations*, are grouped together. The same is true for the terms *unemployment* and *output*. Furthermore, it appears that the language model captures the relationships between economic concepts such as *unemployment* and *labor market*.

Next, we turn to an evaluation of homonyms. Some homonyms arise because their meanings differ in different contexts. Since our language model is very context-specific, the issue with certain homonyms should be less prevalent than in language models trained on a more general context. In the following, we illustrate this by estimating the similarity to the term *basel* and comparing our results to the general language model GloVe6b and GoogleNews. The results can be found in Table 5, where we can see that *basel* is associated with the city in GloVe6b and some abbreviations in Word2Vec GoogleNews, but it is only associated with banking regulation vocabulary in our language model. Remarkably, it even correctly matches abbreviations such as the Basel Committee on Banking Supervision (BCBS).[10]

[10]In the Appendix, we provide additional examples for the interested reader.

Table 5: Intrinsic Evaluation: Homonym across language models.

doc2vec	GloVe6B	Word2Vec GoogleNews
basel_committee	zurich	abbr
basle	basle	Tst
capital_accord	zürich	iva
basel_accord	bern	tHe
bcbs	switzerland	Neurol
basle_committee	stuttgart	BASLE
basel_ii	hamburg	PARAGRAPH
basel_iii	cologne	tellus
consultative	lausanne	Def.
minimum_capital	schaffhausen	Complementarity

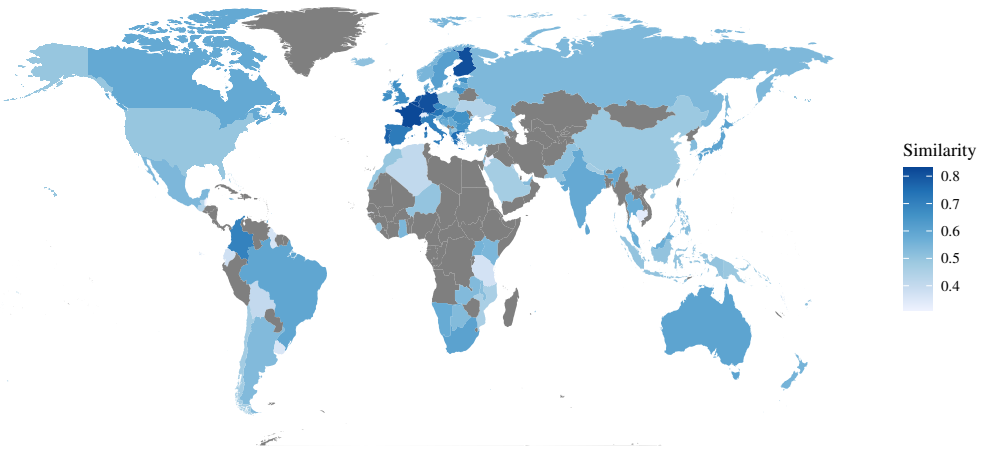
Note: The table shows for the Doc2Vec and the two genereal corpus models the ten most similar words to the word *basel* according to the cosine distance of the underlying word embeddings as defined by Equation (4). The underscore is used to highlight collocations as described in Section III.A.

Finally, we turn to an intrinsic evaluation of the document embeddings. Here, we measure the similarity between central banks, assuming that central banks in western countries are more akin to one another based on similar objectives. We operationalise this idea by averaging the document embeddings for each central bank and estimating their similarity towards the ECB. The result is depicted in Figure 5 with darker colors indicating greater similarity. It appears that central banks in Europe and North America are closest to the ECB, which is consistent with our assumption.[11] This observation is investigated further in our first application in Section V.

To summarize, we used the previously introduced algorithms for quantifying words and documents in this section. We evaluated all methods using out-of-sample prediction and chose the one with the highest overall predictive power. Subsequently, we used three intrinsic assessments to determine whether previously assumed relationships are embedded in our model. We conclude that the embeddings contain meaningful information at both the word and document level.

[11] Note that we find the same results when using word embeddings.

Figure 5 : Central banks' similarity



Notes: This graph illustrates the cosine distance between the average ECB document embedding and all average central bank document embeddings in our dataset. Darker colors depict a lower distance, i.e. a higher similarity. The cosine distance is defined in Equation (4).

V. Applications

Genberg, Karagedikli, et al. (2021) suggest in their survey on machine learning in central banking that empirical approaches in the realm of monetary policy serve one of four purposes: data description, forecasting, structural analysis, and decision communication. The previous section concentrated on the description component, whereas now we apply embeddings in forecasting exercises and structural analysis. In particular we present four potential application in monetary policy, finance and sociology, using our previously chosen language model.
The first application assesses whether central banks' objectives drive the differences in similarity we found in the previous section. We find that inflation targeting central banks are more similar. Next, we use Mario Draghi's *whatever it takes* speech to create an indicator of the ECB's commitment to act as a lender of last resort. We find that in times of crisis ECB communication can calm financial markets. In our third application, we investigate prejudice and biases in the technical language of central bankers. The final application is in forecasting, where we put our embeddings to the test as a predictor for monetary policy surprises.
The applications are intended to provide case studies for the use of embeddings via transfer learning. Please note that the source code for all applications can be found online.[12] This is done for two reasons: First, we want other researchers to be able to comprehend and replicate our findings. Second, and most importantly,

[12]https://sites.google.com/view/whatever-it-takes-bz2021

it should demonstrate how conveniently embeddings can be incorporated into one's own research.

A. Inflation targeting

The first application investigates factors that influence central bank similarity, using the document similarity index introduced in the previous section as a dependent variable. We are particularly interested whether these results are influenced by similar institutional preferences. As a result, we investigate whether the relative similarity to the ECB can be explained by the adoption of inflation targeting, since this is among the most prevalent and observable institutional settings.
In a first step, we label all central banks as "inflation targeting" after their official announcement as an institution aiming for a specific inflation rate, resulting in 44 central banks being classified as such. Next, we determine the average embedding of all central banks. If an inflation target was announced between 1999 and 2020, the institution is divided based on the date of the respective announcement. As outlined in Section IV, we calculate the similarity of those average embeddings to the one representing the ECB. Finally, using these similarities as dependent variable and the inflation-objective dummy as independent variable, we run an OLS regression. The results are displayed in column one of Table 6.

Table 6: Inflation target regression

	Dependent variable:					
	Similarity to ECB					
	Document embedding			Word embedding		
	(1)	(2)	(3)	(4)	(5)	(6)
Inflation target	0.124***	0.081***	0.077***	0.003***	0.003***	0.003***
	(0.018)	(0.020)	(0.022)	(0.001)	(0.001)	(0.001)
Euro Area		0.120***			−0.001	
		(0.028)			(0.001)	
ECB member			0.091***			−0.001
			(0.026)			(0.001)
Constant	0.490***	0.489***	0.488***	0.995***	0.995***	0.995***
	(0.011)	(0.010)	(0.010)	(0.0003)	(0.0003)	(0.0003)
Observations	142	142	142	139	139	139
R^2	0.300	0.482	0.469	0.122	0.130	0.128

Note: Coefficients are estimated using an OLS regression. Standard errors are displayed in parentheses. ***, **, * indicate significance at the 1, 5, and 10 per cent level, respectively.

According to the regression results, adopting an inflation target appears to significantly increase the similarity to the ECB. However, since this effect may be explained by factors such as common currency or membership in the EU, we control for both in columns two and three. While the magnitude of the effect

decreases, it remains positive and statistically significant. Both controls enter the regression with positive and significant coefficients.
As a robustness test, we run the same regression using the similarity between word embeddings.[13] We find that the adoption of an inflation target remains a highly significant variable. This result makes us confident that one of the factors driving central bank similarity is the adoption of a mutual objective.

B. Whatever it takes

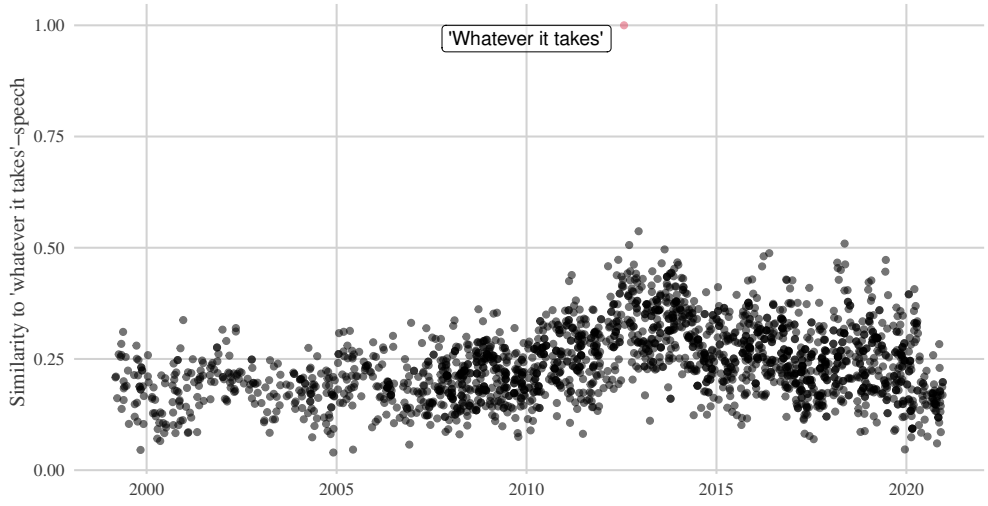
The second application focuses on the effect of central bank communication in times of heightened uncertainty, utilizing the document space. Although there is literature on this topic using word embeddings (Azqueta-Gavaldon et al., 2019), its focus is on measuring uncertainty using news articles.
We showcase a novel approach utilizing the cosine distance between the central bank document representations. The focal point is the famous speech by Mario Draghi in London on 26 July 2012, containing the iconic quote: *"Within our mandate, the ECB is ready to do whatever it takes to preserve the euro. And believe me, it will be enough."* This is widely interpreted as the ECB signaling its willingness to act as a lender of last resort if necessary.
Exploiting the particularity of this speech, we calculate the cosine distance between the ECB's remaining speeches to this event's embedding, thereby creating a time-series of an index, indicating the central bank's willingness to act as a lender of last resort. Figure 6 illustrates that, particularly during the Euro Area crisis, the embeddings of central bank speeches appear more similar to the *whatever it takes*-speech. To investigate whether the similarity to that speech can calm financial markets in times of heightened uncertainty, we run the following regression:

$$\Delta spread_{10y,t} = wit_{simil,t} + Unc_t + wit_{simil,t} \times Unc_t + X_t + \epsilon_t \tag{5}$$

where $\Delta spread_{10y}$ is the change in the spread between Greek 10-year and German 10-year government bonds and wit_{simil} is the cosine similarity of each speech to the *whatever it takes* (wit) speech.[14] We use three different specifications as uncertainty measures Unc: First, the implied volatility of the STOXX50 on the day before the speech ($VSTOXX$), second a decomposition of the VSTOXX into uncertainty (UC) and risk aversion (RA) based on Bekaert et al. (2021),[15] and finally the ECB's daily CISS index (Hollo et al., 2012). X represents a set of control variables, among them a dummy for the wit speech, Moodys agency ratings for Greek bonds, European and U.S. stock prices, monetary policy surprises based on Altavilla et al. (2019), and a dummy for the ECB's different central

[13] Recall that the jargon used by central banks is very similar as highlighted in Section III.A.
[14] Note that, due to the irregularity of speeches, we use the difference in bond prices between the day before a speech and the closing price of the day after a speech.
[15] We thank Marie Hoerova for providing the data series.

Figure 6 : Similarity of all European Central Bank speeches to the "Whatever it takes" speech.



Notes: This graph illustrates the cosine distance between a speech and the *whatever it takes* speech. The cosine distance is defined in Equation (4).

bank presidents. Since considerable risk of autocorrelation, we integrate the first lag of the bond spreads.

The results can be found in Table 7. Starting with the first specification, we find a positive and highly significant relationship between $VSTOXX$ and bond spreads, which is consistent with finance theory. Furthermore, there is a clear effect due to the actual speech of Mario Draghi that had a significant negative impact on the spread. Due to the interaction term the effect direction of wit_{simil} depends on the level of uncertainty and changes with increasing uncertainty. At low uncertainty ($VSTOXX < 20$), the coefficient is positive and then becomes negative. A possible explanation for the initial positive effect would be that a *whatever it takes* speech has exactly the opposite effect at low uncertainty. When financial markets are calm, such a speech could be interpreted as a signal of impending troubles. In this situation, the speech would become a self-fulfilling prophecy, triggering spreads to rise.

We find no major differences in the other specifications. The sign of the similarity variable remains positive and significant in both cases, but it reverses as the level of uncertainty rises. Only the configuration with the $CISS$ shows a generally lower level of significance. To control for possible other effects, we add additional variables to our model.[16] None of these variables cause the coefficients of interest

[16]The full table can be found in Appendix A.A4

Table 7: Regression results: Whatever it takes

	Dependent variable:		
	$\Delta spread_{10y}$		
$Unc_t =$	$VSTOXX_{pd,t}$	$CISS_{pd}$	UC_{pd}
wit_{simil}	1.416***	0.353**	0.485***
	(0.482)	(0.161)	(0.179)
$wit_{simil} \times Unc_t$	−0.070***	−2.911**	−0.020***
	(0.026)	(1.262)	(0.007)
Unc_t	0.016***	0.675**	0.005***
	(0.006)	(0.287)	(0.002)
RA_{pd}			−0.0001
			(0.001)
wit_{dummy}	−1.303***	−1.140***	−1.424***
	(0.317)	(0.406)	(0.278)
L(spread10y_d, 1)	0.248**	0.249**	0.249**
	(0.115)	(0.115)	(0.115)
Constant	−0.318	−0.125	−0.123
	(0.283)	(0.235)	(0.267)
Moodys Rating	Yes	Yes	Yes
MP shocks	Yes	Yes	Yes
Stock prices	Yes	Yes	Yes
President Dummy	Yes	Yes	Yes
Observations	2,028	2,028	2,028
R^2	0.116	0.113	0.116

Note: Coefficients are estimated using an OLS regression. Standard errors are displayed in parentheses. ***, **, * indicate significance at the 1, 5, and 10 per cent level, respectively. The test statistics are calculated with heteroscedasticity and autocorrelation robust (HAC) standard errors.

to change substantially.
Overall, we conclude that both Mario Draghi's speech and similar speeches can lower the spread between government bonds when tensions are high and may thus be part of a targeted forward guidance strategy.

C. Gender Bias

The next application is in an area of monetary policy that is rarely studied: the analysis of biases in central bankers' language. Biases have been found in ordinary language on numerous occasions. However, it may be informative if the very technical language of central bankers contains the same prejudices. Gender bias was chosen as an example of potential partiality in the embeddings primarily because of its contemporary relevance and to showcase how even central bankers' technical jargon might be biased.
Our analysis builds on a fast growing literature that identifies biases in publicly available embeddings (e.g. Caliskan et al., 2017; Garg et al., 2018; Manzini et al.,

2019; Sweeney and Najafian, 2019; Badilla et al., 2020), including those used as general models in the previous section. Inherent in those approaches is the idea that language reflects the latent biases of the underlying institutions. Therefore any language model derived from a biased text corpus must inherent these biases as well.

We are following Garg et al. (2018), who proposed the *relative norm distance (RND)* to represent the latent variable of a bias, a metric that measures a group's association with a neutral word. When two groups are compared, the latent bias of either group can be estimated by their distance towards the neutral term. In practice, the authors recommend gathering two lists of terms (i.e., male and female pronouns) and then averaging their embeddings. The distance between these averages and a neutral word (i.e., childcare) can then be used to calculate the prejudice of this neutral term. For instance, if the distance for the female average embedding is smaller than the distance for the male average embedding, the term is more closely associated with women, and vice versa. Formally for the word list v_a and v_b with n dimensions each and a neutral word w, the RND can be calculated by:

$$RND_{a,b} = \sqrt{\sum_{i=1}^{n}(w - v_{a,n})^2} - \sqrt{\sum_{i=1}^{n}(w - v_{b,n})^2} \tag{6}$$

To test for underlying biases in our embedding, we collect study programs and their respective gender ratios in Bachelor programs across Europe.[17]

Table 8: Academic profession association by gender

Female pronouns	Male pronouns
childcare	fashion
wildlife	physics
nursing	architecture
pre-school	mechanics
welfare	computer
education	automation

Note: The table replicates the findings of the RND measure as introduced in Garg et al. (2018). It illustrates the subset of occupations most associated with gender pronouns.

Next, we estimate the RND for each study program with respect to a set of male

[17]We use data from Eurostat on students enrolled in Bachelors Programs by sex. The dataset can be found on this link.

Table 9: Regression results - Gender Bias

	Dependent variable:
	Relative norm distance
Fraction of female students	0.039^{***}
	(0.013)
Constant	-0.030^{***}
	(0.008)
Observations	67
R^2	0.113

Note: The RND measure is used as defined in Equation (6). Higher values indicate closer association to female pronouns and lower values closer association with make pronouns. The respective pronouns can be found in Footnote 18. Coefficients are estimated using an OLS regression. Standard errors are displayed in parentheses. ***, **, * indicate significance at the 1, 5, and 10 per cent level, respectively.

and female pronouns as suggested by Garg et al. (2018).[18] The most feminine and masculine programs according to our language model can be found in Table 8. With a few exceptions, male pronouns are most closely associated with STEM fields, whereas female pronouns are most closely associated with care-taking and education.

To formally test whether this bias in association may be driven by the dominance (or lack thereof) of any gender in the respective academic profession, we run a simple OLS regression with the former as explanatory variable. The result can be found in Table 9. The regression indicates that female participation is much higher in fields closer associated with female pronouns. The effect is statistically significant and economically relevant.

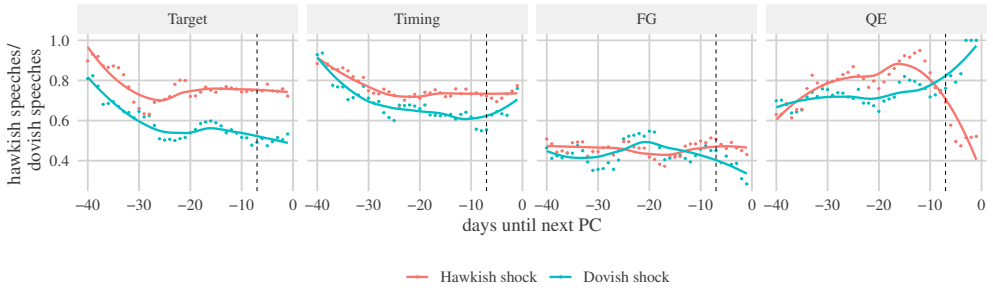
Importantly, these findings do not imply that any specific central banker or institution is communicating a gender bias on purpose. Rather, we believe that general social patterns, such as occupational gender distribution, are likely to be reflected in central bank texts as well. We hope to emphasize that any text (and thus its embeddings) are not without prejudice and should therefore be used with caution.

[18] Specifically, the following pronouns are used: Female pronouns (v_a): she, daughter, hers, her, mother, woman, girl, herself, female, sister, daughters, mothers, women, girls, females, sisters, aunt, aunts. Male pronouns (v_b): he, son, his, him, father, man, boy, himself, male, brother, sons, fathers, men, boys, males, brothers, uncle, uncles, nephew. A complete list of the academic fields is available upon request.

D. *Predicting monetary policy surprises*

In our final application, we turn towards the prediction of financial variables, specifically whether ECB speeches can accurately predict the central banks' monetary policy. To investigate whether there is predictive power in the embeddings, we turn to the monetary policy surprises by Altavilla et al. (2019). The authors construct four surprises based on different parts of the term structure using high-frequency financial data around the ECB press-conference. The researchers use a rotated factor model to calculate the change in Overnight Index Swap (OIS) rates from one month to 10 years on four latent variables. They call the relevant factors target, timing, forward guidance (FG) and quantitative easing (QE) according to their effect horizon.[19] Since these policy surprises are expected to be unpredictable (otherwise, markets would price in the change), this provides an interesting evaluation with respect to the wealth of information provided by the embeddings.

Figure 7 : Relative frequency of hawkish and dovish speeches preceding an ECB press-conference.



Notes: This graph depicts the relative frequency of hawkish to dovish speeches, as measured by the relative distance between a speech and hawkish to dovish press-conferences separated by Altavilla et al.'s (2019) policy surprises and measured in days. The upper red line represents the frequency of hawkish speeches preceding a hawkish press-conference. In contrast, the lower blue line represents the frequency of hawkish speeches preceding a dovish press-conference. The frequencies were measured with a sliding window of +-3 days.

We begin by categorizing each press-conference and surprise as *hawkish* if its surprise is positive and *dovish* if it is negative. Using the RND introduced in the previous application, we measure the relative euclidean distance for each speech with respect to all hawkish and dovish press-conferences and classify speeches as

[19] For the US Gürkaynak et al. (2005) identified a target and path factor. Due to the unique institutional setting of the ECB, the path factor can be further separated into timing, FG and QE (e.g. Brand et al., 2010; Swanson, 2021). The target surprise loads most on one-month OIS rates, timing on 6-month rates, FG on 2-year rates, and QE on 10-year rates. In general, a positive surprise corresponds to an increase in OIS rates and thus to restrictive monetary policy and vice versa.

hawkish if they are relatively closer to the hawkish press-conferences and dovish otherwise. First anecdotal evidence of these speeches' potential predictive power can be found in Figure 7, where we find a difference in the relative frequency of hawkish speeches preceding a hawkish press-conference and vice versa. While the difference increases closer to the relevant press-conference for target and timing surprises, the pattern is less clear for FG and QE. For the former, there is no clear difference, and for the latter, the gap fluctuates.[20]

Table 10: Regression results: Altavilla et. al. (2019) shocks

	Dependent variable:							
	Target		Timing		FG		QE	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Frequency	1.46***	1.47**	0.83**	1.03**	−0.05	0.06	0.48	0.03
	(0.43)	(0.65)	(0.42)	(0.51)	(0.61)	(0.63)	(1.48)	(2.00)
BBD Uncertainty		0.01*		0.002		0.001		−0.004
		(0.003)		(0.003)		(0.003)		(0.01)
EONIA		0.30		0.03		−0.04		−1.36
		(0.19)		(0.18)		(0.16)		(4.75)
GDP growth		−0.14		0.05		−0.03		0.10
		(0.11)		(0.08)		(0.08)		(0.12)
Unemployment rate		0.13		0.001		−0.01		−0.01
		(0.14)		(0.13)		(0.13)		(0.56)
Inflation		−0.16		0.21		−0.03		−0.36
		(0.21)		(0.20)		(0.19)		(0.67)
Constant	−1.57***	−3.65**	−0.31	−1.08	0.15	0.17	−0.51	0.49
	(0.32)	(1.62)	(0.31)	(1.62)	(0.31)	(1.51)	(1.16)	(7.06)
Observations	195	195	195	195	194	194	52	52
Log Likelihood	−117.82	−113.96	−131.83	−130.63	−134.10	−133.70	−35.84	−34.63
Akaike Inf. Crit.	239.64	241.92	267.65	275.26	272.19	281.40	75.68	83.27

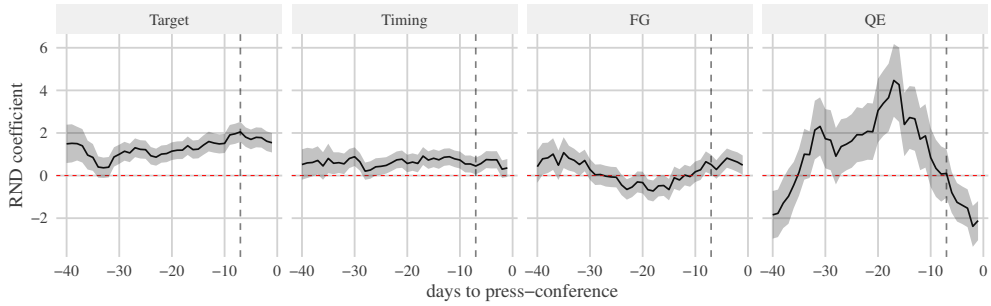
Note: The dependent variables are the monetary policy surprises by Altavilla et al. (2019). Frequency depicts the relative frequency of hawkish to dovish speeches prior an ECB press-conference. Each speech is categorized by the relative distance to the average hawkish and dovish press conference, for each policy surprises respectively. BBD Uncertainty represents the current Baker et al. (2016) uncertainty index. Coefficients are estimated using an OLS regression. Standard errors are displayed in parentheses. ***, **, * indicate significance at the 1, 5, and 10 per cent level, respectively.

We run a Probit model regression with the relative frequency as independent variable to formally test this relationship. The results can be found in Table 10. Several macroeconomic variables, such as the Euro Over Night Index Average (EONIA) rate, current unemployment, and inflation rate, are also included. We find the following: First, the frequency is statistically significant for surprises with a short horizon. Both target and timing surprises show a significant

[20]Note that QE surprises are not available before 1 October 2014, as QE surprises are not expected in the Euro Area until that date. Accordingly, the number of observations varies between surprises.

positive correlation between the relative number of hawkish speeches and the direction of the surprise. With an average of 10 speeches per press-conference, each additional hawkish speech increases the probability of observing such a hawkish target or timing policy surprise by 10-15%, depending on the shock. Second, this relationship remains even when incorporating the different control variables, so we assume that the speeches and embeddings contain more recent information.
Although we cannot provide empirical evidence as to why pre-decision speeches have predictive power for monetary policy surprises, it seems helpful to point out a possible theoretical channel from the literature. Bauer and Swanson (2020) find that Fed surprises are correlated with macroeconomic news. This news could also be reflected in the central bank's speeches and thus have predictive potential.

Figure 8 : Regression results of rolling window approach.



Notes: This graph depicts the hawk-frequency coefficient from the regression results of table 10 re-estimated using a rolling window of +-3 days. The y-axis depicts the days to the next ECB press-conference. The grey area is the standard deviation for the respective coefficient.

We find these results particular fascinating since we do not filter speeches at this point, i.e. all speeches are equally weighted, whether they occur ten days before a press-conference or 40 days before. However, it seems unlikely that all speeches carry the same weight since the executive board has a quiet period prior to press-conferences and since it seems unlikely that future monetary policy can be communicated this effectively months in advance. To investigate whether the results may be affected by either the short end (through the quiet period phase) or the long end (though monetary policy uncertainty), we run the same regression using a rolling window. Since we cannot effectively control for macroeconomic variables, we use only the frequency of hawkish to dovish speeches as a dependent variable. The resulting coefficients as well its standard error, are illustrated in Figure 8.
The findings are not uniform across surprises but can be summarized as follows. First, about one month before the press-conference, the predictions become reliable. Although the results vary between horizons, the general pattern remains

the same: the coefficient stabilizes or rises around 20-30 days before the press-conference. Second, we observe the quiet period's expected effect. The grey dotted line depicts the seven days leading up to the press-conference. There, the coefficients become insignificant and thus are no longer a reliable predictor of the monetary policy stance. We find qualitatively and quantitatively the same results for the first three surprises when we use the narrower window for our regression, but a substantial improvement in significance for the QE surprise (2.95^{**}).

Finally, we test the embeddings' out-of-sample performance to evaluate whether the language model has actual predictive power. We use an expanding horizon approach to estimate all regressions in order to formally test forecasting performance and avoid look-ahead bias (Chakraborty and Joseph, 2017). The model is parameterized based on observations prior to 2017, and the predicted policy surprise are compared to their true values between 2017 and 2020 (33 observations, i.e. $\approx 20\%$ of the sample). We specifically choose this period since we are interested in the predictive power of our model during different time periods.[21] The results can be found in Table 11. Across the different surprises, the accuracy of the predictions is remarkable, all predict higher than 50% correctly.[22] The accuracy appears to decrease with increasing horizon, which is consistent with our earlier findings. This result may provide first evidence that speeches target expectations on the shorter side.[23]

Table 11: Out of sample accuracy Altavilla surprises.

Policy surprise	Accuracy
Target	70 %
Timing	61 %
Quantitative easing	58 %
Forward guidance	52 %

Note: This table summarizes the out-of-sample prediction performance across different surprises. The models are estimated on ECB speeches before 2017-01-01 and evaluated on speeches after 2017-01-01. The accuracy displayed is constitutes the fraction of correct predictions by all predictions.

[21] Since we anticipate a shift in jargon as the COVID-19 pandemic hits the global economy in 2020, this event provides an interesting basis for evaluating the language model outside its training environment. However, we tried other time periods as well and came to the same conclusions.

[22] It is worth noting that we also tested word embeddings in this prediction task. Although they did not outperform the document embeddings, they did provide surprisingly good prediction as well: Target: 64%, Timing: 58%; FG: 58% and QE: 55%. The results are available upon request.

[23] It is important to note that, due to the small number of remaining observations for the training phase, we are not particularly confident in our QE results. We selected multiple time periods, both longer and shorter, and found that QE performed fairly consistently above 50%. However, we would caution against over-interpreting this result based on only 22 observations.

This result makes us confident that i) speeches have predictive power beyond previous findings and ii) that the embeddings can capture some of it. The findings provide many potential future research questions regarding the most relevant dimensions in the embedding space and factors affecting those. Furthermore, we employ a simple linear model, whereas recent contributions such as Kalamara et al. (2020) and Hinterlang (2020) demonstrate how machine learning (and, in particular, neural networks) could be applied to such prediction tasks.

VI. Conclusion

Understanding the communication of central banks has developed to be a substantial entity in monetary policy, with dictionary approaches at the forefront of current techniques to quantify their speeches, press-conference and reports. In this paper, we expanded this literature in three ways: the compilation of a novel text-corpus, the introduction of algorithms stemming from computational linguistic to extract embeddings – a language model – and the provision of central bank specific embeddings.

First, we collect a text-corpus that is unparalleled in size and diversity within this literature, as both is necessary to train such a language model sufficiently. Then, we introduce embeddings, a novel approach from computational linguistics to quantify texts. These language models are trained using machine learning techniques that locate words and documents in a multidimensional vector space. It has been demonstrated that these embeddings can capture meaningful real-world relationships. Finally, we are able to provide high quality text-representations for central bank communication by training and evaluating different algorithms using an objective criteria. The algorithm with the highest predictive power is able to generate both multidimensional word and document representations.

Within this paper we highlighted the broad applicability of embeddings by illustrating four prominent examples in the fields of central bank institutions, financial uncertainty, gender bias, and monetary policy shock prediction. For example, we illustrate that our language model is able to extract relevant information to forecast future monetary policy shocks from public speeches. Throughout our applications, we emphasize several techniques for extracting the abundance of information contained within embeddings. In our work with embeddings, we found that similarities — euclidean and cosine — are a suitable metric for integrating textual information into economic models or investigating them as dependent and independent variables themselves. Furthermore, we highlight how the use of embeddings in neural networks is a field to be further explored in future research.

Our approach has important implications for policymakers and central bankers, allowing for more nuanced ex-ante and ex-post evaluations of communication strategies, such as obtaining preliminary assessments of future communication. We believe this paper to be just a first step toward answering many exciting questions, for example extracting superior measures for concepts such as sentiment, or uncertainty, modelling institutional differences, and improving real-time

predictions. We hope that by making our language models publicly available, we will be able to assist in this process.

References

Acosta, M., & Meade, E. E. (2015). Hanging on every word: Semantic analysis of the fomc's postmeeting statement. *FEDS Notes*, (2015-09), 30.

Aguilar, P., Ghirelli, C., Pacce, M., & Urtasun, A. (2021). Can news help measure economic sentiment? an application in covid-19 times. *Economics Letters*, *199*, 109730.

Altavilla, C., Brugnolini, L., Gürkaynak, R. S., Motto, R., & Ragusa, G. (2019). Measuring euro area monetary policy. *Journal of Monetary Economics*, *108*, 162–179. https://doi.org/10.1016/j.jmoneco.2019.08.016

Amaya, D., & Filbien, J.-Y. (2015). The similarity of ECB's communication. *Finance Research Letters*, *13*, 234–242. https://doi.org/10.1016/j.frl.2014.12.006

Apel, M., Grimaldi, M., & Hull, I. (2019). *How much information do monetary policy committees disclose? evidence from the fomc's minutes and transcripts* (Working Paper Series No. 381). Sveriges Riksbank. Stockholm. http://hdl.handle.net/10419/215459

Apel, M., & Grimaldi, M. B. (2014). How informative are central bank minutes? *Review of Economics*, *65*(1), 53–76.

Athey, S. (2019). 21. the impact of machine learning on economics. *The economics of artificial intelligence* (pp. 507–552). University of Chicago Press.

Azqueta-Gavaldon, A., Hirschbühl, D., Onorante, L., Saiz, L. et al. (2019). Sources of economic policy uncertainty in the euro area: A machine learning approach. *ECB Economic Bulletin*, *5*.

Badilla, P., Bravo-Marquez, F., & Pérez, J. (2020). Wefe: The word embeddings fairness evaluation framework. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 430–436. https://doi.org/10.24963/ijcai.2020/60

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The quarterly journal of economics*, *131*(4), 1593–1636.

Bauer, M., & Swanson, E. T. (2020). *The Fed's Response to Economic News Explains the "Fed Information Effect"* (Working Paper).

Bekaert, G., Hoerova, M., & Xu, N. R. (2021). *Risk, monetary policy and asset prices in a global world* (SSRN Scholarly Paper ID 3599583). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.3599583

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *The Journal of Machine Learning Research*, *3*, 1137–1155.

Bholat, D. M., Hansen, S., Santos, P. M., & Schonhardt-Bailey, C. (2015). Text mining for central banks. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2624811

Binette, A., & Tchebotarev, D. (2019). *Canada's Monetary Policy Report: If Text Could Speak, What Would It Say?* (Staff Analytical Notes No. 2019-5). Bank of Canada. https://ideas.repec.org/p/bca/bocsan/19-5.html

Blaheta, D., & Johnson, M. (2001). Unsupervised learning of multi-word verbs. *Association for Computational Linguistics Workshop on Collocation (2001)*, 54–60.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. *Text mining* (pp. 101–124). Chapman; Hall/CRC.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Blinder, A. S., Ehrmann, M., Fratzscher, M., De Haan, J., & Jansen, D.-J. (2008). Central bank communication and monetary policy: A survey of theory and evidence. *Journal of Economic Literature*, *46*(4), 910–45.

Brand, C., Buncic, D., & Turunen, J. (2010). The Impact of ECB Monetary Policy Decisions and Communication on the Yield Curve. *Journal of the European Economic Association*, *8*(6), 1266–1298. https://doi.org/10.1111/j.1542-4774.2010.tb00555.x

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Chakraborty, C., & Joseph, A. (2017). *Machine learning at central banks* (tech. rep.). Bank of England. Working Paper.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25th international conference on Machine learning*, 160–167. https://doi.org/10.1145/1390156.1390177

Correa, R., Garud, K., Londono, J. M., & Mislang, N. (2021). Sentiment in central banks' financial stability reports. *Review of Finance*, *25*(1), 85–120.

Doh, T., Song, D., Yang, S.-K. et al. (2020). Deciphering federal reserve communication via text analysis of alternative fomc statements.

Ehrmann, M., & Fratzscher, M. (2007). Communication by Central Bank Committee Members: Different Strategies, Same Effectiveness? *Journal of Money, Credit and Banking*, *39*(2-3), 509–541. https://doi.org/10.1111/j.0022-2879.2007.00034.x

Ehrmann, M., & Talmi, J. (2020). Starting from a blank page? semantic similarity in central bank communication and market volatility. *Journal of Monetary Economics*, *111*, 48–62.

Ferrari, M., & Le Mezo, H. (2021). *Text-based recession probabilities* (Working Paper Series No. 2516). European Central Bank. https://ideas.repec.org/p/ecb/ecbwps/20212516.html

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*.

Fraccaroli, N., Giovannini, A., & Jamet, J.-F. (2020). *Central banks in parliaments: A text analysis of the parliamentary hearings of the bank of england, the european central bank and the federal reserve*. ECB Working Paper.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Genberg, H., Karagedikli, Ö. et al. (2021). *Machine learning and central banks: Ready for prime time?* (Tech. rep.). The South East Asian Central Banks Research and Training Centre.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–574. https://doi.org/10.1257/jel.20181020

Gentzkow, M., & Shapiro, J. M. (2010). What drives media slant? evidence from us daily newspapers. *Econometrica*, *78*(1), 35–71.

Gorodnichenko, Y., Pham, T., & Talavera, O. (2021). *The voice of monetary policy* (Working Paper No. 28592). National Bureau of Economic Research. https://doi.org/10.3386/w28592

Gürkaynak, R. S., Sack, B., & Swanson, E. T. (2005). Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements. *International Journal of Central Banking*, *1*(1), 39.

Hansen, S., & McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, *99*, S114–S133.

Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the FOMC: A computational linguistics approach. *The Quarterly Journal of Economics*, *133*(2), 801–870. https://doi.org/10.1093/qje/qjx045

Hansen, S., McMahon, M., & Tong, M. (2019). The long-run information effect of central bank communication. *Journal of Monetary Economics*, *108*, 185–202. https://doi.org/10.1016/j.jmoneco.2019.09.002

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

Hayo, B., Henseler, K., Rapp, M. S., & Zahner, J. (2020). *Complexity of ECB Communication and Financial Market Trading* (MAGKS Joint Discussion Paper Series in Economics No. 201919). Philipps-University Marburg. https://ideas.repec.org/p/mar/magkse/201919.html

Hayo, B., & Neuenkirch, M. (2013). Do Federal Reserve Presidents communicate with a Regional bias? *Journal of Macroeconomics*, *35*, 62–72. https://doi.org/10.1016/j.jmacro.2012.10.002

Hinterlang, N. (2020). *Predicting monetary policy using artificial neural networks* (tech. rep.). Deutsche Bundesbank. Discussion Paper.

Hollo, D., Kremer, M., & Lo Duca, M. (2012). *Ciss-a composite indicator of systemic stress in the financial system* (tech. rep.). ECB.

Hornik, K., & Grün, B. (2011). Topicmodels: An r package for fitting topic models. *Journal of Statistical Software*, *40*(13), 1–30.

Hu, N., & Sun, Z. (2021). *Uncertain talking at central bank's press conference: News or noise?* (SSRN Working Paper Series). HKIMR Working Paper.

Istrefi, K., Odendahl, F., & Sestieri, G. (2020). Fed communication on financial stability concerns and monetary policy decisions: Revelations from speeches.

Jarociński, M., & Karadi, P. (2020). Deconstructing Monetary Policy Surprises—The Role of Information Shocks. *American Economic Journal: Macroeconomics*, *12*(2), 1–43. https://doi.org/10.1257/mac.20180090

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., & Kapadia, S. (2020). *Making text count: economic forecasting using newspaper text* (Working Papers No. 865). Bank of England. https://doi.org/10.2139/ssrn.3610770

Kincaid, J. P., Fishburne Jr., R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel* (tech. rep.). Naval Technical Training Command Millington TN Research Branch.

Lau, J. H., & Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. *Proceedings of the 1st Workshop on Representation Learning for NLP*, 78–86.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *International conference on machine learning*, 1188–1196.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings [arXiv: 1904.04047]. *arXiv:1904.04047 [cs, stat]*. http://arxiv.org/abs/1904.04047

Masciandaro, D., Romelli, D., & Rubera, G. (2020). *Does it fit? tweeting on monetary policy and central bank communication* (tech. rep.). SUERF.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. http://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 746–751.

Peek, J., Rosengren, E. S., & Tootell, G. (2016). *Does fed policy reveal a ternary mandate?* (Working Papers). Federal Reserve Bank of Boston.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Picault, M., & Renault, T. (2017). Words are not all created equal: A new measure of ECB communication. *Journal of International Money and Finance*, *79*, 136–156. https://doi.org/10.1016/j.jimonfin.2017.09.005

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2).

Schmeling, M., & Wagner, C. (2019). *Does central bank tone move asset prices?* (CEPR Discussion Papers No. 13490). C.E.P.R. https://EconPapers.repec.org/RePEc:cpr:ceprdp:13490

Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2020). Measuring news sentiment. *Journal of Econometrics*.

Shapiro, A. H., & Wilson, D. J. (2019). *Taking the fed at its word: A new approach to estimating central bank objectives using text analysis* (Working Paper Series). Federal Reserve Bank of San Francisco. https://doi.org/10.24148/wp2019-02

Smales, L., & Apergis, N. (2017). Does more complex language in FOMC decisions impact financial markets? *Journal of International Financial Markets, Institutions and Money*, *51*, 171–189. https://doi.org/10.1016/j.intfin.2017.08.003

Swanson, E. T. (2021). Measuring the effects of federal reserve forward guidance and asset purchases on financial markets. *Journal of Monetary Economics*, *118*, 32–53. https://doi.org/10.1016/j.jmoneco.2020.09.003

Sweeney, C., & Najafian, M. (2019). A transparent framework for evaluating unintended demographic bias in word embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1662–1667. https://doi.org/10.18653/v1/P19-1162

Tadle, R. C. (2021). Fomc minutes sentiments and their impact on financial markets. *Journal of Economics and Business*, 106021. https://doi.org/https://doi.org/10.1016/j.jeconbus.2021.106021

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*(3), 1139–1168.

Tillmann, P. (2020). *Financial Markets and Dissent in the ECB's Governing Council* (Working Paper No. 48-2020). MAGKS Joint Discussion Paper Series in Economics.

Tobback, E., Nardelli, S., & Martens, D. (2017). *Between hawks and doves: Measuring central bank communication* (Working Paper Series). ECB.

Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 384–394.

Wischnewsky, A., Jansen, D.-J., & Neuenkirch, M. (2021). Financial stability and the fed: Evidence from congressional hearings. *Economic Inquiry*, *59*(3), 1192–1214. https://doi.org/10.1111/ecin.12977

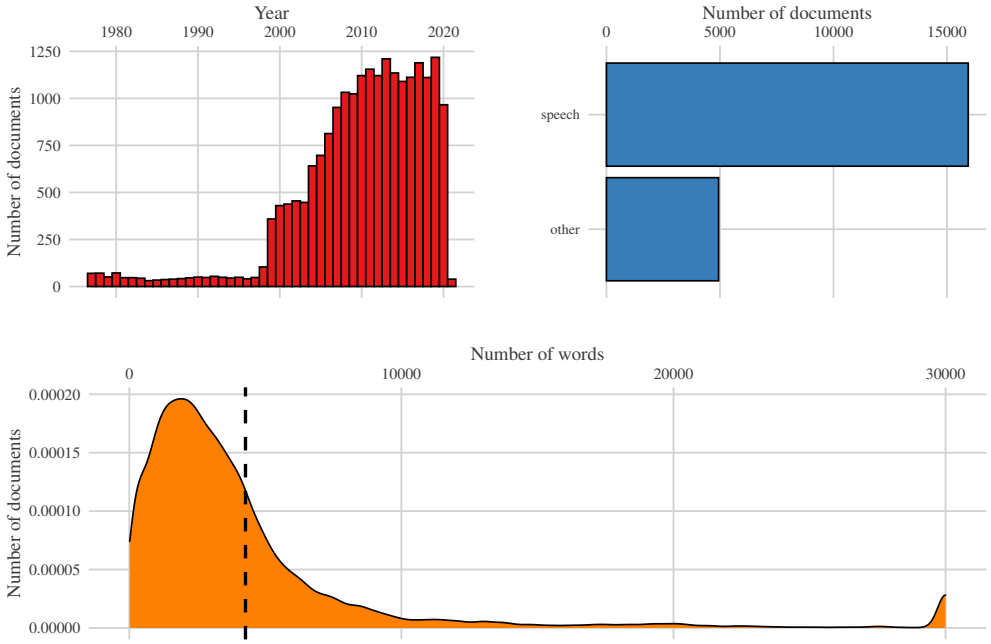
Wittgenstein, L. (1958). *Philosophische untersuchungen (zweite auflage). english edition*.

Zahner, J. (2020). *Above, but close to two percent. evidence on the ecb's inflation target using text mining* (MAGKS Joint Discussion Paper Series in Economics). Philipps-University Marburg.

Appendix

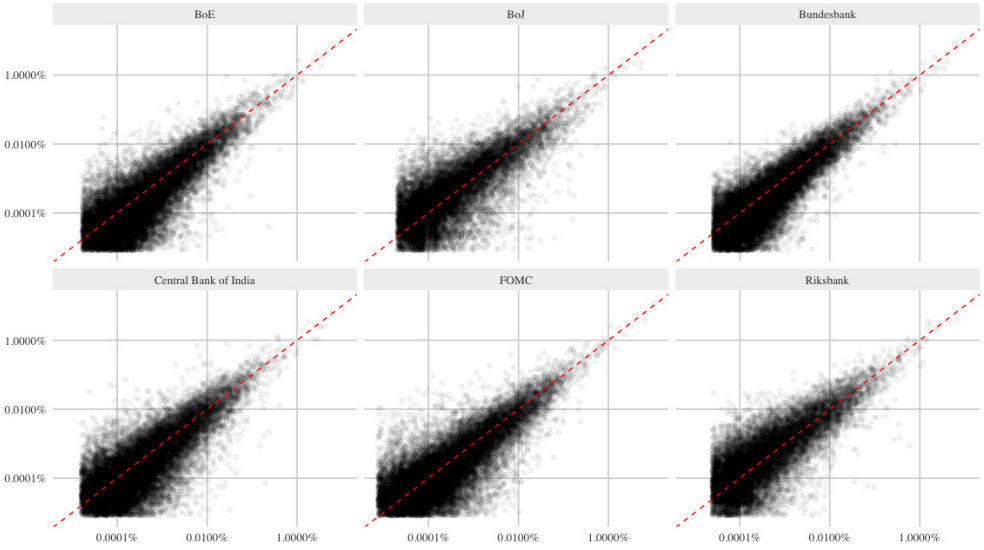
A1. Graphical illustrations of text corpus

Figure A1 : Descriptive summary of the corpus



Notes: This figure shows the basic properties of our central bank corpus, broken down by year, type, and word length Documents with more than 30,000 words grouped in the *other* category.

Figure A2 : Illustration of frequency of used terms between ECB other central banks.



A2. Language Model specifications

The following are the hyperparameters we use. For the Word2Vec model we refer to Mikolov, Yih, et al. (2013) and Rehurek and Sojka (2011) and for the GloVe model we use Pennington et al.'s (2014) specification. The parameters of the Doc2Vec model are based on Lau and Baldwin (2016). For the LDA we use the findings of Blei and Lafferty (2009) as well as few modifications by Hornik and Grün (2011).[24] The hyperparameters are summarized in the following table:

[24]For the Gibbs sampling draws we chose a burnin rate of 1000, sampled 2000 iterations and returned every fifth iteration.

Table A1: Hyperparameter Settings for Evaluation

Method	Dim	Window Size	Sub-Sampling	Negative Sample	Itera-tions	learning-rate	alpha	delta
Doc2Vec-DBOW	300	15	0.0001	5	20	0.05	-	-
Doc2Vec-DM	300	5	0.0001	5	20	0.05	-	-
Word2Vec	300	5	0.0001	5	10	0.05	-	-
GloVe	300	-	-	10 20	0.1	0.75	-	
LDA	300	-	-	-	-	-	0.166	0.01

A3. Additional evaluation

External evaluation

In addition to our economic evaluation task we test our whole embeddings in a more general setting. This should serve as a robustness test with a different task, different empirical methodologies, and far more central bank participation. We select classification tasks that are uninteresting in and of themselves to reduce the risk of spurious correlation between the embeddings and potential application outcome variables (Athey, 2019). In particular, the classification task used here is to predict each speech's central bank and publication year, assuming that higher performance implies a language model's relative superiority.

Following current research like Chakraborty and Joseph (2017), the assessment is carried out using out-of-sample testing via cross-validation. In particular, we use five-fold cross-validation, where each model is trained on four-fifths of the dataset and evaluated on the remaining fifth. This process is repeated five times, with the evaluation's accuracy estimated on each fold. We use the following two machine learning techniques for the classification task: K-Nearest-Neighbor (KNN) and random forest.[25]

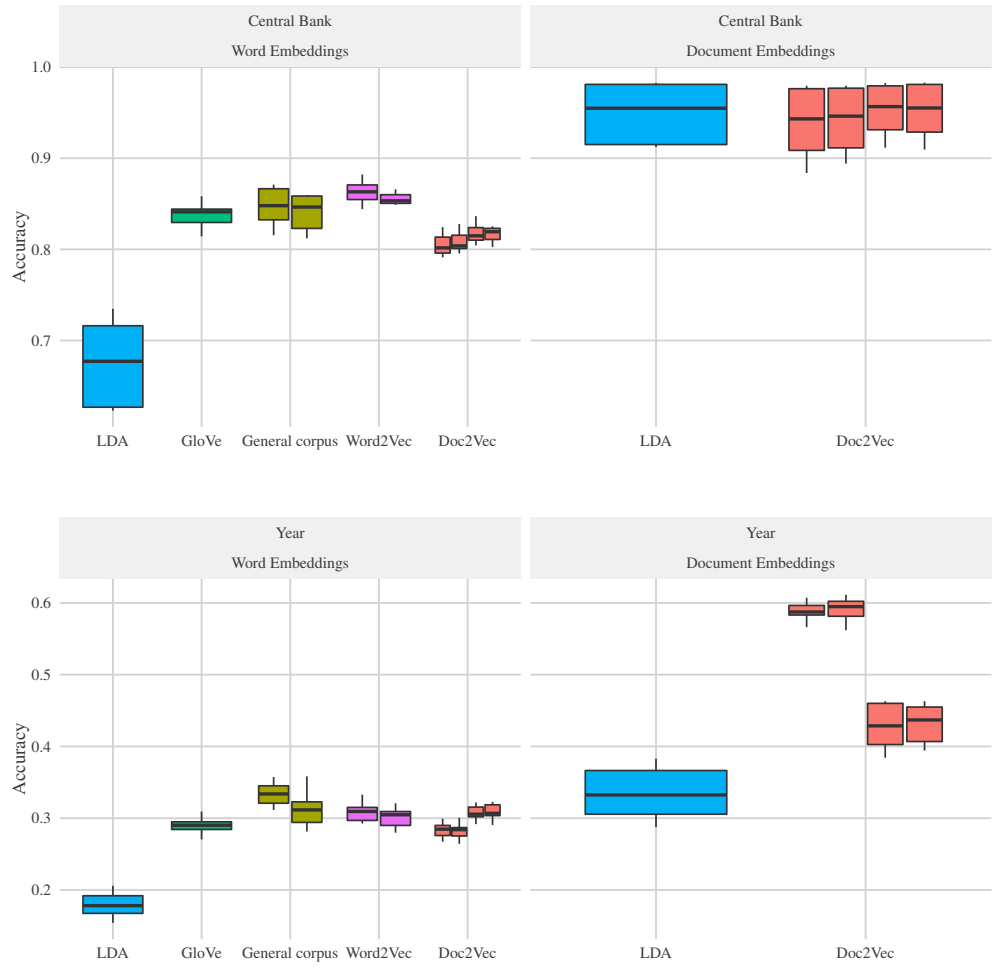
The word embedding results are illustrated in Figure A3, with one algorithm per row and one prediction task per column. The expected accuracy from guessing would be 0.25 for the central bank prediction and 0.06 for the year prediction.

The result is similar to the results from the main text. Document embeddings seem to be better suited for summarizing text. For word embeddings, only minor differences are found between the algorithms. Thus, it seems that in these more general tasks, unlike in the economics-related tasks, our word embeddings do not have a clear corpus advantage over the general language models. However, they are not worse either. This again emphasizes the potential of our embeddings in the analysis of central banks. Interestingly, there appears no clear trend between

[25] A great introduction into both non-parametric methods as well as the performance metric is provided by Chakraborty and Joseph (2017).

KNN and Random Forest with regard to performance, which is – concerning the latter ones' complexity – remarkable. KNN appears to be better in predicting the central banks, whereas random forest is slightly superior in the year predictions.

Figure A3 : Evaluation of Embeddings



Notes: This graph depicts the evaluation of different algorithms as discussed in this chapter. The measurement on the y-axis is accuracy of the underlying task, which is measured as $(true\ positive + true\ negative)/(number\ of\ observation)$.

Similar to our *basel* example, we find problems with potentially distorting contexts in general language models if we look at the term *greening*: While Word2Vec GoogleNews associates the colour with this term and Glove6B climate change, our language model associates this topic with terms from the area of climate policy regarding green finance.

Table A2: Additional Intrinsic Evaluation: Homonym across language models.

doc2vec	GloVe6B	Word2Vec GoogleNews
ngfs	afforestation	greener
climate-related	forestation	sustainability
green_finance	beautification	greened
climate_change	reforestation	green
paris_agreement	canker	Greening
climate-	jagielka	greenest
greener	citrus	composting
frank_elderson	punxsutawney	revitalization
greenhouse	gartside	Greenest
climate_change	colonizing	Greener

Note: The table shows for the Doc2Vec and the two genereal corpus models the ten most similar words to the word *"greening"* according to the cosine distance of the underlying word embeddings as defined by Equation (4). The underscore is used to highlight collocations as described in Section III.A.

A4. Applications - Robustness checks

Table A3: Application 2: Whatever it takes - Full table

	Dependent variable:		
	$\Delta\text{spread}_{10y}$		
	(5)	(6)	(7)
$\text{wit}_{\text{simil}}$	1.416*** (0.482)	0.353** (0.161)	0.485*** (0.179)
$\text{wit}_{\text{simil}} \times \text{VSTOXX}_{\text{pd}}$	−0.070*** (0.026)		
$\text{wit}_{\text{simil}} \times \text{ciss}_{\text{pd}}$		−2.911** (1.262)	
$\text{wit}_{\text{simil}} \times \text{UC}_{\text{pd}}$			−0.020*** (0.007)
$\text{VSTOXX}_{\text{pd}}$	0.016*** (0.006)		
ciss_{pd}		0.675** (0.287)	
UC_{pd}			0.005*** (0.002)
RA_{pd}			−0.0001 (0.001)
$\text{wit}_{\text{dummy}}$	−1.303*** (0.317)	−1.140*** (0.406)	−1.424*** (0.278)
altavilla.Target	−0.034 (0.038)	−0.031 (0.038)	−0.034 (0.038)
altavilla.Timing	0.001 (0.008)	0.002 (0.008)	0.001 (0.008)
altavilla.FG	0.005 (0.007)	0.005 (0.007)	0.005 (0.007)
altavilla.QE	−0.024 (0.019)	−0.025 (0.018)	−0.024 (0.019)
lag_asset.sp500	−0.0001 (0.0001)	−0.0001 (0.0001)	−0.0001 (0.0001)
lag_asset.stoxx	0.0001* (0.00004)	0.0001 (0.00004)	0.0001* (0.00004)
MoodysA2	−0.049 (0.067)	−0.045 (0.067)	−0.046 (0.067)
MoodysA3	0.386** (0.168)	0.393** (0.170)	0.379** (0.166)
MoodysBa1	0.063 (0.042)	0.075* (0.044)	0.058 (0.041)
MoodysBa3	0.194 (0.120)	0.192 (0.121)	0.191 (0.117)
MoodysB1	0.154* (0.089)	0.148 (0.090)	0.146* (0.088)
MoodysB3	0.159* (0.089)	0.157* (0.089)	0.156* (0.088)
MoodysCaa1	0.106 (0.106)	0.109 (0.104)	0.102 (0.106)
MoodysCaa2	0.186* (0.108)	0.185* (0.108)	0.181* (0.107)
MoodysCaa3	0.083 (0.107)	0.090 (0.104)	0.080 (0.106)
MoodysCa	0.109 (0.207)	0.130 (0.206)	0.103 (0.205)
MoodysC	−0.060 (0.139)	−0.047 (0.131)	−0.060 (0.139)
lag(spread10y_d, 1)	0.248** (0.115)	0.249** (0.115)	0.249** (0.115)
presidentDuisenberg	−0.091 (0.207)	0.027 (0.195)	−0.073 (0.204)
presidentLagarde	0.087** (0.042)	0.074* (0.044)	0.084** (0.041)
presidentTrichet	−0.044 (0.197)	−0.016 (0.192)	−0.036 (0.196)
Constant	−0.318 (0.283)	−0.125 (0.235)	−0.123 (0.267)
Observations	2,028	2,028	2,028
R^2	0.116	0.113	0.116
F Statistic	10.529***	10.153***	10.101***

Note: Coefficients are estimated using an OLS regression. Standard errors are displayed in parentheses. ***, **, * indicate significance at the 1, 5, and 10 per cent level, respectively. The test statistics are calculated with heteroscedasticity and autocorrelation robust (HAC) standard errors.

As a robustness test we replicate the job example of Garg et. al (2018) using female and male names. We use occupation data from Eurostat and match all descriptions with Garg et. al's (2018) pronouns. The following are the results:

Table A4: Regression results - Gender Bias

	Dependent variable:
	Relative norm distance
Fraction of female students	0.0003*
	(0.0001)
Constant	−0.004
	(0.009)
Observations	32
R^2	0.092

Note: The RND measure is used as defined in Equation (6). Higher values indicate closer association to female pronouns and lower values closer association with make pronouns. The respective pronouns can be found in Footnote 18. Coefficients are estimated using an OLS regression. Standard errors are displayed in parentheses. ***, **, * indicate significance at the 1, 5, and 10 per cent level, respectively.