Joint Discussion Paper Series in Economics

by the Universities of

Aachen · Gießen · Göttingen
Kassel · Marburg · Siegen

ISSN 1867-3678

No. 30-2021

Martin Baumgärtner and Johannes Zahner

Whatever it Takes to Understand a Central Banker – Embedding their Words Using Neural Networks.

This paper can be downloaded from

https://www.uni-marburg.de/en/fb02/research-groups/economics/macroeconomics/research/magks-joint-discussion-papers-in-economics

Coordination: Bernd Hayo • Philipps-University Marburg
School of Business and Economics • Universitätsstraße 24, D-35032 Marburg
Tel: +49-6421-2823091, Fax: +49-6421-2823088, e-mail: hayo@wiwi.uni-marburg.de

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‡

This draft: August 14, 2023
First 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 utilize 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. This allows us to provide high-quality central bank-specific textual representations and demonstrate their applicability by developing an index that tracks deviations in the Fed's communication towards inflation targeting. Our findings indicate that these deviations in communication significantly impact monetary policy actions, substantially reducing the reaction towards inflation deviation in the US.

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 Bernd Hayo, Jens Klose, Peter Tillmann, Michalis Haliassos, Juri Marcucci, Andreas Joseph, Michael McMahon, Isaiah Hull, Ricardo Correa, Davide Romelli, Matthias Neuenkirch, Robin Lumsdaine, Ulrich Fritsche, Christian Conrad, Elisabeth Schulte, Ania Zaleska, Linda Shuku, Christoph Pfeufer, Jeffrey Ziegler, Annick van Ool, and participants at the Advanced analytics: new methods and applications for macroeconomic policy, Conference on Non-traditional Data, the Machine Learning and Natural Language Processing in Macroeconomics, 54th Annual Conference of the MMF, RES & SES Annual Conference 2023, the Central bankers go data driven: Applications of AI and ML for policy and prudential supervision, the VfS Annual Conference 2022: Big Data in Economics, the Banca d'Italia Research Seminar, the 4th International and Interdisciplinary Conference on the Quantitative and Computational Analysis of Textual Data, the 6th International Conference on Applied Theory, Macro and Empirical Finance, the Workshop "Digital Methods in History and Economics", the RGS Doctoral Conference, and the 7th International Young Finance Scholars' Conference for their feedbacks.
Declarations of interest: none (applies to all authors).

† THM Business School, JLU Gießen, Germany, martin.baumgaertner@posteo.de
‡ Corresponding author. Chair of Macroeconomics and Finance, Goethe University Frankfurt, Germany, zahner@econ.uni-frankfurt.de.

1. Introduction

What did European Central Bank (ECB) president Mario Draghi mean on July 26, 2012, when he stated that *"within [its] mandate, the ECB is ready to do whatever it takes to preserve the euro"*? According to the current literature on central bank communication quantification, this is a neutral sentence. However, the message contained in the statement was nothing short of extraordinary for financial market participants and monetary policy experts; in fact, it marked a turning point in the ongoing euro crisis. We propose a novel language model in this paper that is able to capture such subtleties.

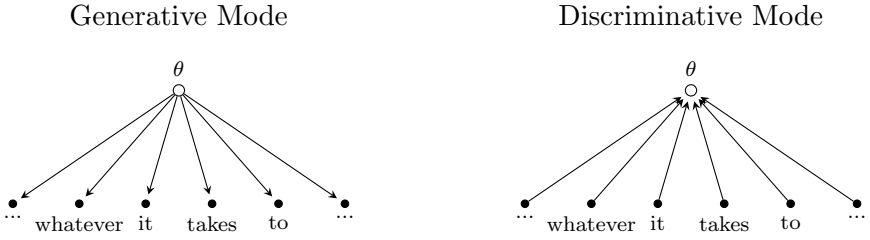
Over the last few decades, there has been an increase in the use of unstructured big data in monetary policy, in particular in the analysis and interpretation of central bank communication (Blinder et al., 2008). This development was certainly accelerated by the zero lower bound and the emergence of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

The analysis of central bank communication is based on the presumption that it contains latent messages (θ) by the monetary policymakers, which are worth extracting. These messages can be discrete, such as a bank's stance in a policy debate, or continuous, such as signaling policy direction or communicating the bank's preferences. While not observable directly, the θ's generate variations in the communication, and hence the words used (W), a process depicted on the left-hand side of Figure 1. Since only the outcome of this sampling process can be directly observed, it is the receivers' job to infer the underlying message from the variation in W, as illustrated by the right-hand side. This paper aims to provide a representation for words that allows simple models to retrieve the underlying messages from the observed variation in central bank communication ($W \rightarrow \theta$).

Decoding of messages is most effective when the language used is stable, homogeneous, and represented in its richness. The current string in the central bank communication literature uses pre-defined dictionaries, such as Loughran and McDonald (2011), Apel and Grimaldi (2014), and Picault and Renault (2017) for counting terms (for example, positive and negative words) to extract a single dimension (for example, sentiment) from a document. Such a practice equates to an extreme prior of the informativeness of the vast majority of communicated terms, which may only suffice for simple messages, thereby falling short of capturing the domain-specific richness of the representation.

To address these shortcomings, modern linguistics and computer science has turned to machine learning to develop novel *language models*. Such models are estimated from a set of text – the *corpus* –, and an *algorithm* that locates words in a multidimensional vector space. Conceptually similar terms are mapped in

Figure 1 : Communication Model



Note: The illustration is adapted from Lowe (2021, p. 10).

close proximity, Meanwhile models such as Mikolov, Yih, et al. (2013) and Pennington et al. (2014), leverage large corpora from a variety of sources, such as Twitter or Google searches, and thus allow for little inference about the technical language used by monetary policymakers, violating the condition of stable and homogeneous communication.

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 approaches. We 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 one previously used in the central bank communication literature. Second, this paper introduces novel machine learning algorithms for text quantifying. We compare a multitude of different algorithms according to objective criteria. Doc2Vec, an algorithm that leverages the word and document space, outperforms the others in our evaluation. Third, by training the novel algorithm on the novel text corpus, we introduce a language model previously unseen in monetary policy.

Our language model demonstrates its usefulness by effectively quantifying the central bank objectives of different central banks. We then develop a time series that monitors deviations in the Fed's inflation targeting communication. Results based on a reaction function highlight the substantial impact of communication deviations on monetary policy actions. Our findings indicate that the response to inflation deviations in the U.S. is substantially reduced, if the communication shifts away from inflation targeting.

The remainder of this paper is structured as follows. Section 2 provides a lit-

erature overview of the current state of natural language processing (NLP) in monetary economics. In Section 3 we introduce both the text corpus and the algorithms, combining both elements into language models used to represent W. We then evaluate the quality of the resulting embeddings in the central bank context in Section 4 before applying the best-performing language model in Section 5, essentially providing possibilities of inference ($W \rightarrow \theta$). The final section concludes this paper.

2. Related literature

Natural language processing (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 as captured by (the aggregate consequences of) their responses to it. Among successful implementations are Gürkaynak et al. (2005), Brand et al. (2010), Cieslak and Schrimpf (2019), Jarociński and Karadi (2020), Swanson (2021), and Jarociński (2022) 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.[1]

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

[1] One notable shortcoming the quantification literature (and this paper), is the focus on the supply of provided information, omitting potential demand effects. However, Tillmann (forthcoming) shows that market participants react to surprises in the expected manner.

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 omit relevant information. In terms of Figure 1, the $\theta \rightarrow W$ relationship is characterized by a prior of zero for the majority of the modeled words. The issue of excluding a substantial portion of text has been articulated before by Harris (1954, p. 156), stating 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"*.

In addition, dictionaries are inherently 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 and the coarseness of the interpretation of the message that comes along with it, 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.

Recent research recognizes and highlights the dictionary approach's limitations to capture the messages' meanings, 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 contains 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 as captured by its measure 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, p. 533) quote. Word embeddings are multidimensional word representations that are used to measure similarity in Twitter tweets (Masciandaro et al., 2020), for the improvement of uncertainty indices (Azqueta-Gavaldon et al., 2019; Cieslak, Hansen, et al., 2021), for the decomposition of central bank vague talk (Hu and Sun, 2021), for the analysis of FOMC introductory statements (Handlan, 2020), and for measuring central banker disagreement (Apel, Grimaldi, and Hull, 2019).[2] Economic research in this field relies on general language models trained on a general text corpus such as Wikipedia. Shapiro, Sudhof, et al. (2020), for example, use Pennington et al.'s (2014) 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 in Shapiro, Sudhof, et al. (2020) and may 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. Their embeddings, however, are not a focal part of the paper and are thus not suitable for general-purpose quantifying central bank communication.[3]

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 contributes to two current desiderata in this literature. On the one hand, the development of novel text-representation (Apel, Grimaldi, and Hull, 2019), and on the other hand, the need to fine-tune these representations for their respective use (Loughran and McDonald, 2011).

[2]There are some finance applications relying on embeddings, such as Araci (2019), Jha et al. (2020), and Rahimikia et al. (2021).

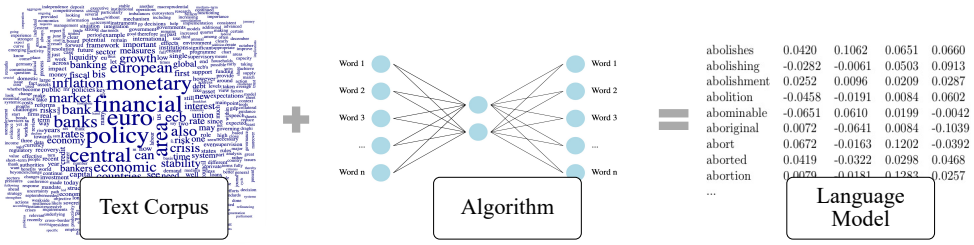
[3]Following the publication of our working paper, a number of authors have adapted our approach of training embeddings specific to central bank communication, such as Bertsch et al. (2022) and Hansen and Kazinnik (2023)

3. Methodology

> *"The meaning of words lies in their use. [...] One cannot guess how a word functions. One has to look at its use, and learn from that."*
>
> — 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, derived from training an algorithm on a text corpus. Embeddings, thereby, provide a nuanced representation of the words (W). Our paper proposes a method for text classification that is detached from causal inference, called *transfer learning*. Transfer learning describes a process in which specialized knowledge is gained by working on one task and is subsequently applied to a different, but related, task. As a result, we avoid potential conflicts that arise when dimension reduction and the application of dimension-reduced variables are performed simultaneously (e.g. Egami et al., 2018).
Figure 2 provides a stylized overview of the procedure how to retrieve a language model. The figure also reflects the structure of the remainder of this section.

Figure 2 : How to retrieve a language model



3.1. Text Corpus

Our text corpus reflects our paper's primary focus 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).[4] 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.

[4]https://www.bis.org/cbspeeches/index.html. We determine the language of the individual documents using Google's Compact Language Detector 3 and clean the corpus accordingly.

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.[5] 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 institutions 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.

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.

In contrast to the previous NLP applications 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 in near proximity 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 so-called collocations, that is, words with specific meaning when used together. The distinctive features of collocation and context were already highlighted by Firth (1957, p. 11), 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 instrument 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

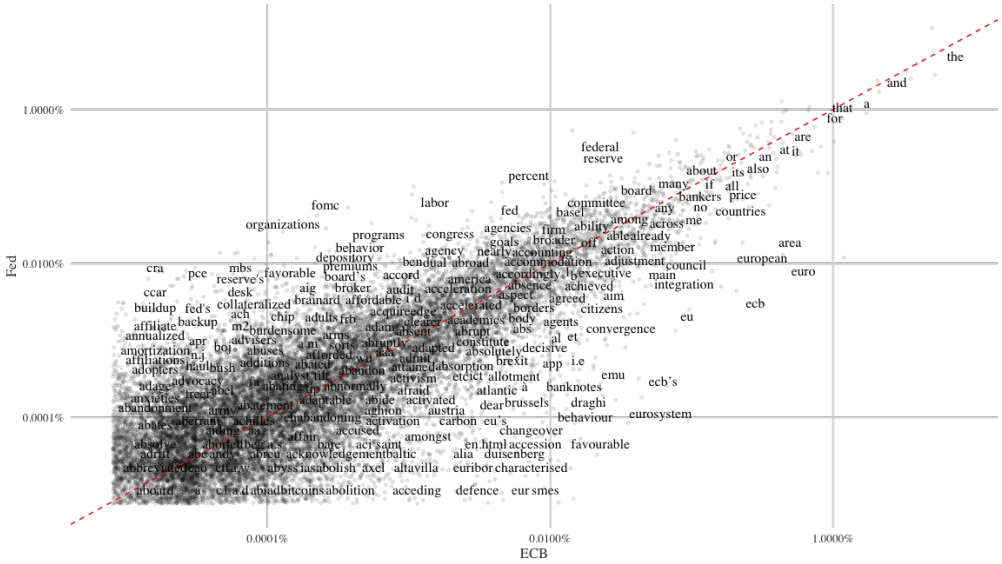
[5]Media interviews have not been collected; they are typically not systematically posted on the BIS or central bank websites. The corpus is screened to ensure that each document is represented only once.

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 over a period from the 1980s to early 2021.[6] The result is a corpus of text that, on the one hand, is unprecedented in quantity and diversity in the monetary communication literature, and, on the other hand, contains exclusively highly specific central bank vocabulary.

Figure 3 : Illustration of frequency of used terms between Fed and ECB



Note: This graph depicts the wording of the Fed and ECB. 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 we scaled both axes by the logarithm and added noise, so the correlation is even stronger than shown here.

[6]The distribution of documents is heavily skewed to the left, as central banks have only increasingly published documents over the Internet since the 2000s. An overview is provided in Figure A1 in the Appendix.

The corpus' stability with respect to word use is what we address next. We test whether the usage of language can be assumed rhetorically stable across institutions and time, a necessary condition to allow for inference.[7] We use the central bank's jargon, the relative word frequency, for the seven most frequent central banks in our sample as an approximation. An illustration of the relative word frequencies for the ECB and the FOMC is provided in Figure 3. Formally testing 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.[8] We conclude from these observations that the institutions do not differ in any relevant way concerning their jargon. In order to test for temporal stability, we compare the jargon of the central bankers across time. The results remain quantitatively the same, illustrating that the usage of language did not change markedly.[9]

3.2. Algorithm

Modern language models follow the proposition of linguistic Zellig S. Harris in their pursuit of superior text representation. Harris (1954) 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. The context of a word, the set of its adjacent 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

Prediction-based algorithms embody this concept. Their operational principle revolves around predicting a target word using the surrounding context words, predict the target word given the context words, i.e. *P(brighter | this, outlook)*. Note how the approach directly incorporates the previously stated distributional semantics by Harris (1954) whereas similar words occur in the same context.

[7] An example of rhetorical instability is the Google Flu Trends Project (Lazer et al., 2014), which used flu-related Google searches to predict medical appointments. The project was discontinued in 2015 due to severe misjudgment by the algorithm caused by changes in search behavior.

[8] The Pearson correlation coefficients of the relative word-frequency of the ECB towards the respective central bank are: Federal Reserve (Fed): 98% (t = 884), Riksbank: 98% (t = 585), Bank of England (BoE): 98% (t = 966), Bank of Japan (BoJ): 98% (t = 668), Bundesbank: 99% (t = 1257), and Central Bank of India: 98% (t = 783). The results are also illustrated in the Appendix in Figure A2.

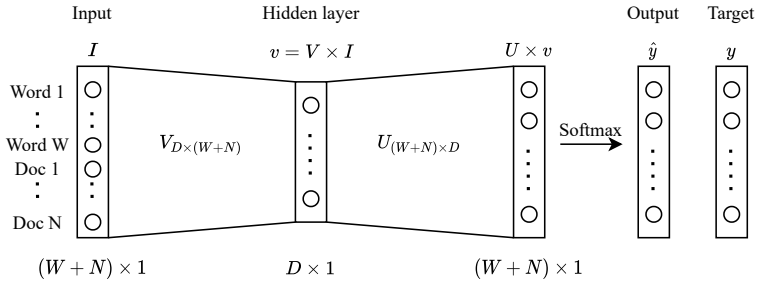
[9] Formally, we estimate the Pearson correlation coefficients between the relative word-frequency of all central banks in a given year and the relative word-frequency of all central banks in every other year. The coefficients are > 0.98 and highly significant for all possible combinations. This does not imply that the language has not changed. Rather, it indicates that most terms that were frequently used in 2001, for example, are still frequently used in 2019. Some terms may have changed considerably over time.

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 ____"* 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 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 capturing the semantic meaning as a byproduct.

Word2Vec, a popular prediction-based model, employs neural networks to make these predictions from context Mikolov, Yih, et al. (2013), Mikolov, Chen, et al. (2013), and Mikolov, Sutskever, et al. (2013). At its core, Word2Vec features a single linear hidden layer connected to a softmax output layer. Its primary objective is to forecast the target word given its neighboring context words.

Building upon the foundational idea of Word2Vec, Doc2Vec was introduced by Le and Mikolov (2014). What sets Doc2Vec apart is its ability to embed entire documents. By incorporating document-specific information into the neural network's input layer, every document receives a unique ID. This results in a distinct embedding vector for each document, capturing its overarching semantic essence. 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.

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



Note: 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 A3 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.

An alternative to obtaining embeddings through neural networks is leveraging corpus-wide statistics to obtain word representations, such as Latent Dirichelet Allocation (LDA) or GloVe (e.g., Blei, Ng, et al., 2003; Pennington et al., 2014). We will demonstrate, however, that prediction based methods, outperform corpus-wide methods.[10]

[10] A comprehensive introduction into Word2Vec, Doc2Vec, LDA, and GloVe can be found in Appendix A.A2.

Finally, as mentioned in the introduction, to the best of our knowledge, so far no attempts have been made to train embeddings specifically for the central bank context. This may be due to the computational burden or the necessary amount of text. An alternative to training embeddings from scratch is the use of pre-trained general language models using 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 all our embeddings to two such general models as a benchmark: Glove6B and Word2Vec Google News.[11]

4. Evaluation of language models

In this section, we apply the algorithms introduced in the previous section to our corpus and evaluate the corresponding language models. We aim to determine the algorithm that best summarizes the content of our central bank corpus 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.[12]
Since there exists no benchmark for evaluating language models in economics yet, we turn to the fields of computational 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 embeddings 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 in the second part of this section.

4.1. Extrinsic evaluation

Extrinsic tasks involve evaluating the embeddings 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 are designed to evaluate embeddings in a broad context, while we are interested in the opposite, their domain specificity. Due to a lack of external evaluation methods, we benchmark the embeddings in the following two steps. First, we test how well the models can predict words and then assess the predictive performance of each model in a

[11] 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.

[12] Whenever we evaluate the word embeddings on document level, we average over all word vectors of a document.

monetary policy classification task (Le and Mikolov, 2014). We demonstrate in Appendix A.A4 that the presented results are robust to more general tasks.
In the absence of an established procedure, we use an unsupervised approach that takes advantage of Harris's (1954) notion that context defines the meaning. Good language models should be able to predict terms using their adjacent words. Thus, each model is presented a task such as the following: predict the word *substantial* given the bag of words [*outlook, remains, subject, to, uncertainty, also, regarding, the*].[13]

Table 2: Evaluation results of word prediction task.

Algorithm	Accuracy	Standard deviation
Word Embeddings		
Doc2Vec Bow	0.846	0.007
Doc2Vec Bow Pre	0.844	0.009
GloVe	0.831	0.008
Doc2Vec PVDM	0.803	0.009
Doc2Vec PVDM Pre	0.800	0.017
Word2Vec Skipgram	0.678	0.007
GloVe 6B	0.646	0.008
Word2Vec GoogleNews	0.546	0.016
Word2Vec Bow	0.502	0.009
LDA	0.064	0.014

Note: The table shows the evaluation results across the different algorithms introduced in the previous section. The accuracy was evaluated as the Number of correct predictions / Total number of predictions. 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.

The results are depicted in Table 2. There are four noteworthy results. First, the Doc2Vec and GloVe models perform best, correctly predicting more than 80% of the words. Second, the Bag-of-Words models outperform the others in this group, by an additional 5% points higher accuracy. Third, each model's performance does not vary much across folds. In this context, it should be noted that there is no statistically significant difference between the two top models. Finally, the general language models do not fare as well in terms of relative performance, which emphasizes the importance to train on monetary policy documents.
Our second evaluation task concerns the current interest rate level of the ECB and Fed, which we forecast using the respective central bank's texts.[14] Since we are

[13]We train a neural network with a hidden layer. The results presented are simulated out-of-sample predictions with 10-fold cross-validation. More information is provided in Appendix A.A4.

[14]Our evaluation contrasts with the empirical literature on central bank communication, which tends to identify changes rather than levels of policy rates. However, such an exercise would have limited

primarily concerned with the correct level, we divide the corresponding 3-month interbank rates into quintiles to derive our evaluation target.[15] We are interested in the best performance, therefore, we employ a neural network to predict the respective interest rate levels with our embeddings.[16] 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 prediction.

Table 3: Evaluation results of algorithms.

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

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 uninformed 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.

Table 3 summarizes the accuracy of the predictions split by Document- and Word Embedding as well as task. Since there exist several variants in the Word2Vec

our analysis to three categories (increase, no change, decrease), with very unequal distribution. Thus, distinguishing between the algorithms becomes fairly difficult, which is why we decided to use levels instead.

[15] 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.

[16] We employ a single hidden layer neural network with 64 units and dropout regularization. We tested various specifications, but the performance does not change substantially. The exact parameterization is available upon request.

and Doc2Vec algorithms and we aim for a broad comparison, we estimate them all. The name in column one starts with the algorithm followed by the variant's abbreviations.
Our evaluation yields some interesting results. First, the federal funds rate level appears to be more challenging to predict across all 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, confirming previous results by 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 (bold in all tables) is chosen because it performs best in word prediction and also performs consistently well in monetary policy classification.[17]

4.2. 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 interpreted cautiously. The presented intrinsic evaluations are based on the cosine distance in the embeddings space, which is a measure of similarity ($S_{a,b}$) between two-word vectors a and b of length n, and defined as follows:

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

In the first evaluation, we select economic concepts in the word embedding space and assess the most similar words to these concepts. The results are presented in Table 4, for the words *inflation*, *unemployment*, and *output*.[18]
It is evident that our language model is capable of grouping words with semantically similar meaning. For example, it is reassuring that 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. Homonyms arise because their meaning differs in different context. 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

[17]The upcoming results are robust across all Doc2Vec variants. Results are available upon request. To ease readability, we will refer in the following to the language model "Doc2Vec Bow Pre" only as "Doc2Vec".

[18]On our website (https://sites.google.com/view/whatever-it-takes-bz2021) we provide an interactive tool that allows users to make the same assessment for any word in the entire vocabulary.

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 (1). The underscore is used to highlight collocations as described in Section 3.3.1.

this property by estimating the similarity to the term *basel* and compare 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).[19]

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 intuition.[20] This observation is the starting point of our investigation into monetary policy frameworks in Section 5.

To summarize, we used the previously introduced algorithms for quantifying words and documents in this section. We evaluated all methods using out-of-sample predictions and selected the Doc2Vec on the basis of this evaluation. Subsequently, we used three intrinsic assessments to determine whether previ-

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

[20] The same chart with a different central bank as a comparison group is available upon request.

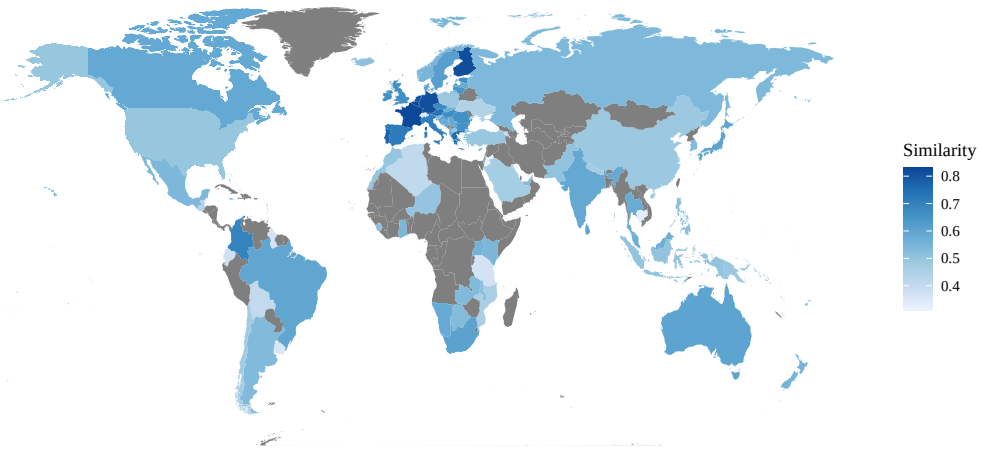
Table 5: Intrinsic Evaluation: Similarity to Basel across language models

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 (1). The underscore is used to highlight collocations as described in Section 3.3.1.

ously assumed relationships are embedded in our model. We conclude that the embeddings contain meaningful information at both the word and document level.

Figure 5 : Central banks' similarity



Note: 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 (1).

5. Monetary policy framework classification

In this section, we will demonstrate how our our Doc2Vec language model can be used to retrieve latent messages, i.e. identifying avenues for $W \rightarrow \theta$. While our focus lies on the identification of monetary policy framework, the application is intended to provide case studies for the use of embeddings via transfer learning. The source code can be found online.[21] This is done for two reasons: First, we want other researchers to be able to comprehend and replicate our findings. Second, and most importantly, it should demonstrate how conveniently embeddings can be incorporated into one's own research.

The first application assesses whether central banks' objectives drive the differences in textual similarity we reported in the previous section. We find that inflation targeting central banks are indeed more similar.

In two companion papers, we demonstrate how our embeddings may be used to create an indicator of the ECB's commitment to act as a lender of last resort (Zahner and Baumgärtner, 2023) and to investigate prejudices and biases in the technical language of central bankers across the globe (Zahner, 2023).

5.1. Measuring Monetary policy frameworks

We investigate whether banks' institutional settings can explain the differences in the similarity of communication. Institutional classifications are inherently mul-

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

tidimensional; we only address aspects that are considered relevant for monetary policy.

Our analysis relies extensively on Cobham (2021), who uses the IMF's Article IV Consultation Reports to classify de jure monetary policy frameworks, on an annual basis following the end of the Bretton Woods system. A monetary policy framework refers to the *"objectives pursued by the monetary authorities, but also the set of constraints and conventions within which their monetary policy decisions are taken."* (Cobham, 2021, p. 1). Cobham identifies ten target variables (inflation, money supply, and others) that can be further subdivided into 32 mutually distinct categories, ranging from *loosely structured discretionary targets* to *fully converging inflation targets*. The classification, which covers approximately 150 central banks, is available online. Merging the monetary policy framework with our corpus yields more than 80 central bank classifications and more than 800 country-year observations.[22]

In the first step, we use all texts of a year to calculate an average annual embedding. To do this, we compute the average of all vector elements over all documents of a year, element by element.

Next, we calculate the cosine distance, as explained in Section 4, between a central bank's average annual embedding and a specified institution's embedding.

The question as to which particular monetary policy institution is used for comparison is ultimately left to the researcher's discretion. Since we focus on monetary policy objectives and inflation targeting is prevalent, we select three institutions that have different histories with respect to this objective. Specifically, we selected the first inflation-targeting central bank (the Reserve Bank of New Zealand (RBNZ)) and two prominent ones in our corpus (the ECB and the Fed) because they provide interesting variations given their different institutional settings and objectives, e.g., the Fed has a dual mandate, while the ECB has a primary and a secondary objective. In the following, we will refer to those three as benchmark central bank.

Econometrically, we run an OLS regression of the similarities ($S_{i,j,t}$) between a central bank j and the benchmark central bank $i \in \{RBNZ, FED, ECB\}$ on the central bank target ($Target_{j,t}$) defined by Cobham (2021) at time t. In practice, this means that central bank communication can change over time, as we can only compare, for instance, the Fed and the Bank of England in the same year. To control for macroeconomic conditions, we take the difference of three macroeconomic indicators (inflation, unemployment, log(GDP)) towards the benchmark central bank, i.e., $\Delta X_{i,j,t} = X_{j,t} - X_{i,t}$. Finally, we control for euro area members ($EA_{j,t}$).[23].

$$S_{i,j,t} = Target_{j,t} + \Delta X_{i,j,t} + EA_{j,t} + \epsilon_i \tag{2}$$

[22] Members of a currency area are assigned the classifications of the currency area's lead central bank, as opposed to omitting these observations.

[23] Neither the choice of macroeconomic variables nor the dummy seems to affect the results. Results are available upon request

In a first step, we examine the general differences between institutions labelled *inflation targeting* and those otherwise using a dummy variable (ITs) that takes the value of one for inflation targeting central banks. Results are reported in specification (1) in Table 6. We find a consistently positive, significant, and economically relevant coefficient in all three benchmarks, suggesting that central banks with inflation targeting communicate more similarly relative to the RBNZ, the Fed, and ECB. When accounting for macroeconomic differences and euro area banks, the results persist.

Table 6: Regression results: Monetary Policy Framework classification

	Dependent Variable: Similarity towards bank i								
$i =$	RBNZ			Federal Reserve			European Central Bank		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
ITs	0.07***	0.11***		0.06***	0.09***		0.09***	0.16***	
	(0.02)	(0.03)		(0.02)	(0.02)		(0.02)	(0.03)	
- FIT			0.17***			0.12***			0.12***
			(0.03)			(0.02)			(0.03)
- LIT			0.11***			0.10***			0.18***
			(0.03)			(0.02)			(0.02)
- FCIT			0.05			0.10***			0.14***
			(0.04)			(0.04)			(0.04)
- LCIT			0.06*			0.02			0.09***
			(0.03)			(0.03)			(0.03)
Constant	0.32***	0.31***	0.29***	0.32***	0.43***	0.42***	0.40***	0.41***	0.41***
	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)
Rem. MPF Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Macro. Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	84	821	821	84	825	825	83	821	821
R^2	0.15	0.18	0.24	0.20	0.19	0.22	0.28	0.31	0.37
Adjusted R^2	0.11	0.17	0.23	0.16	0.18	0.21	0.24	0.30	0.36

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. We adapt the notations directly from Cobham (2021): ITs = inflation targets; LIT = loose inflation targeting; LCIT = loose converging inflation targeting; FIT = full inflation targeting; FCIT = full converging inflation targeting; WSD = well structured discretion; LSD = loose structured discretion; ERTs = exchange rate targets; MixedTs = mixed targets; NNF = no national framework. "Rem. MPF Controls" indicates controls for all monetary policy frameworks not shown in the table.

In a second step, we examine the similarities on an annual basis and, moreover, include all regimes. Thus we now exploit the full range of Cobham's (2021) classifications (the footnote of Table 6 provides an overview). The results are shown in model (2). For all three benchmarks, the coefficients on the inflation target increase significantly.

In a final step, we are interested which inflation targeting characteristics influence

our results. Therefore, we partition the inflation targeting category further into loose inflation targeting LIT (e.g., euro area, US until 2011, South Africa) and full inflation targeting FIT (e.g., New Zealand, US since 2011, Poland), as well as a converging category for each, representing non-constant targeting over time. The results are interesting both within and across benchmarks, due to the different relative weights depending on the institution being compared. The similarity towards the RBNZ (always FIT) is significantly higher for inflation-targeting institutions only. The Fed, which transitioned from LIT to FIT, has a nearly equal weight between both, whereas the ECB (which has always been LIT) is closer to LIT institutions. [24]

This result makes us confident that one of the factors driving the similarity among central banks embeddings is the adoption of a mutual objective and framework, which implies that researchers may use public communication when abstracting central banks monetary policy framework.

5.2. The Federal Reserve's inflation targeting

In the following analysis, we examine whether differences in central bank communication also lead to differences in policy actions. The Fed's transition between FIT and LIT provides a unique opportunity to study the impact of communication shifts on policy actions. Our aim is to determine whether there is empirical evidence that communication influences policy actions, i.e., we seek to shed light on the potential impact of central bank communication on actual policy decisions. To formulate our hypothesis, we test whether deviations in communication have led to deviations in the form of rule-based monetary policy and its associated parameters. Our focus will be on the Taylor rule, which serves as a widely-used guide for monetary policymakers.

One concern with measuring the Fed's stance is that of identification, since both FIT and LIT are closely related concepts, making it difficult to introduce substantial variance into the measurement of their effects. We therefore propose the following approach, based on the *relative norm distance (RND)* proposed by Garg et al.'s (2018): For each Fed speech s_t, we measure the Euclidean distance to the average of all N FIT speeches (v_{FIT}) and K LIT speeches (v_{LIT}) in our corpus, excluding Fed speeches:

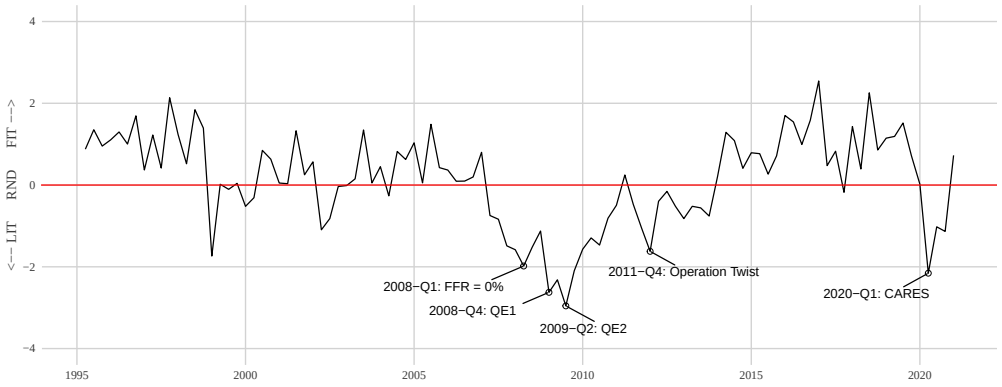
$$RND_t = \sqrt{\left(s_t - \frac{1}{N}\sum_{n=1}^{N}(v_{FIT,n})\right)^2} - \sqrt{\left(s_t - \frac{1}{K}\sum_{k=1}^{K}(v_{LIT,k})\right)^2} \tag{3}$$

One interpretation of the relationship between the deviation from rule-based monetary policy and the RND measure is that it becomes more relevant during crises when other goals, such as financial stability, seem to matter more.

[24] As a robustness check, we conduct the same regression using the similarity between the word embeddings.[25] We find that the adoption of an inflation target remains a highly significant variable.

After standardizing the resulting time series index, we present the quarterly time series in Figure 6, where positive deviations indicate communication more aligned with full inflation targeting and vice versa. In particular, we observe significant deviations from full inflation targeting during major economic events, such as the financial crisis and the outbreak of the COVID-19 pandemic, which may indicate a shift in policy objectives to address prevailing economic challenges. For instance, during the Financial Crisis, the Fed lowered the Federal Funds Rate to a range of zero to 0.25 percent in December 2008. Additionally, QE1 was announced in November 2008, expanded in March 2009, and complemented by the revival of "Operation Twist" at the end of 2011. Furthermore, as a response to the COVID-19 pandemic, in the first quarter of 2020, the Fed implemented various measures, including the CARES Act, as part of its efforts to support the economy during the crisis. Overall, the alignment of extreme negative deviations in the index with crucial monetary policy events seems to underscore the complex relationship between central bank actions and our index.

Figure 6 : FED's stance on inflation targeting



Note: .

To test the contemporaneous effect of our index on actual policy decisions, we formulate the following basic Taylor rule:

$$i_t = \alpha + \Delta\pi_t + RND_t + (\Delta\pi_t \times RND_t) + \epsilon_t \tag{4}$$

where i_t represents the Wu and Xia (2016) shadow rate, $\Delta\pi_t$ denotes the inflation deviation from the target, RND_t is our index. Given that our index (RND) aims to measure the Fed's stance on inflation targeting, our hypothesis is that it will impact the inflation reaction coefficient, i.e. the interaction term between inflation

deviation and RND. The results of this regression analysis are presented in Table 7.

Table 7: RND Taylor Rule Regression Table

	Dependent variable:					
	Interest Rate					
	(1)	(2)	(3)	(4)	(5)	(6)
$\Delta\pi$	1.28***	1.43***	1.42***	1.45***	1.39***	1.13***
	(0.16)	(0.16)	(0.15)	(0.15)	(0.22)	(0.17)
Unemp. Rate			−0.99***	−1.00***		−0.83***
			(0.15)	(0.16)		(0.12)
Output Gap			−0.28**	−0.16		0.07
			(0.14)	(0.16)		(0.13)
RND		0.005***	−0.003*	−0.002	0.0001	−0.01*
		(0.002)	(0.002)	(0.01)	(0.002)	(0.01)
RND × $\Delta\pi$		0.41***	0.53***	0.44***	0.45***	0.47***
		(0.12)	(0.11)	(0.12)	(0.15)	(0.13)
RND × Unemp. Rate				−0.01		0.14
				(0.14)		(0.13)
RND × Output Gap				0.14		0.24**
				(0.12)		(0.10)
Constant	0.02***	0.02***	0.08***	0.08***	0.01***	0.06***
	(0.002)	(0.002)	(0.01)	(0.01)	(0.002)	(0.01)
Speaker FE	No	No	No	No	Yes	Yes
Observations	125	125	125	125	89	89
R^2	0.34	0.43	0.59	0.60	0.33	0.67
Adjusted R^2	0.33	0.41	0.57	0.57	0.31	0.64
Note:						*p<0.1; **p<0.05; ***p<0.01

We find the following: First, the Taylor principle is fulfilled across all specifications. Second, once we include our RND measure, we observe a significant and substantial effect on the inflation response coefficient. In the baseline specification (2), a one standard deviation increase in our index (indicating a shift towards greater inflation targeting) leads to a 0.41 increase in the inflation response. Conversely, a one standard deviation decrease in the RND measure lowers the inflation response to a level almost below one, indicating no compliance with the Taylor principle. To be more specific: Our findings suggest that during the financial crisis, as well as at the beginning of COVID-19, the response parameter for inflation fell essentially close to zero. In addition, we find that our index increases the explained variance by about one-third.

We then conduct a series of robustness tests. First, we demonstrate that our results remain robust to the inclusion of business cycle indicators such as the output gap (adjusted with an HP filter) and the unemployment rate in columns three and four. We also find no significant coefficient when interacting our index with these indicators, a promising result considering that our intention was to

measure inflation targeting. Interestingly, the interaction coefficient with inflation increases (with no change in the inflation coefficient itself), suggesting that our earlier result can be viewed as a lower bound.
Second, in light of recent evidence suggesting that central bank communication is highly speaker-specific (Hayo et al., 2023), we include a specification in which we control for speaker fixed effects in columns five and six. To do this, we recompute the RND index for each speech, regress the resulting RND index on speaker fixed effects, and then use the residuals as our RND index. The results from this approach remain both quantitatively and qualitatively consistent with our previous findings.
To summarize, we examine whether differences in central bank communication, particularly with respect to inflation targeting, lead to differences in policy actions. Using a Taylor rule, we show that communication that is more closely aligned with the inflation target has an impact on the inflation response. Specifically, we find that inflation targeting communication increases the inflation response coefficient, providing evidence of the link between central bank communication and policy decisions.

6. 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-conferences and reports. In this paper, we expanded the research frontier in four 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. Finally, we show how these approach may be used to evaluate past monetary policy decisions by the Fed.
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 text. 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. Third, 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. Finally, we have demonstrated the broad applicability of embeddings by illustrating that our language model effectively approximates central bank objectives. Specifically, we were able to create an index that tracks deviations in the Fed's communication towards inflation targeting. Our findings indicate that these deviations in communication significantly impact monetary policy actions, substantially reducing the reaction towards inflation deviation in the US.

Throughout our applications, we emphasize several techniques for extracting the abundance of information contained within embeddings. We found that similarities — euclidean and cosine — are a suitable metric for integrating textual information into economic models, investigating them as dependent and independent variables. 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.

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

Angelico, C., Marcucci, J., Miccoli, M., & Quarta, F. (2022). Can we measure inflation expectations using twitter? *Journal of Econometrics*, *228*(2), 259–277.

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.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

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*.

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

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

Bertsch, C., Hull, I., Lumsdaine, R. L., & Zhang, X. (2022). Central bank mandates and monetary policy stances: Through the lens of federal reserve speeches. *Available at SSRN 4255978*.

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

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

Cieslak, A., Hansen, S., McMahon, M., & Xiao, S. (2021). Policymakers' uncertainty. *Available at SSRN 3936999*.

Cieslak, A., & Schrimpf, A. (2019). Non-monetary news in central bank communication. *Journal of International Economics*, *118*, 293–315.

Cobham, D. (2021). A comprehensive classification of monetary policy frameworks in advanced and emerging economies. *Oxford Economic Papers*, *73*(1), 2–26.

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.

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., & Stewart, B. M. (2018). How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*.

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

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.

Handlan, A. (2020). Text shocks and monetary surprises: Text analysis of fomc statements with machine learning. *Published Manuscript*.

Hansen, A. L., & Kazinnik, S. (2023). Can chatgpt decipher fedspeak? *Available at SSRN*.

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

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. (2022). Central bank information effects and transatlantic spillovers. *Journal of International Economics*, *139*, 103683.

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

Jha, M., Liu, H., & Manela, A. (2020). Does finance benefit society? a language embedding approach. *A Language Embedding Approach (July 18, 2020)*.

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.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of google flu: Traps in big data analysis. *science*, *343*(6176), 1203–1205.

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

Lowe, W. (2021). Text as data: Summer school.

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

Rahimikia, E., Zohren, S., & Poon, S.-H. (2021). Realised volatility forecasting: Machine learning via financial word embedding. *arXiv preprint arXiv:2108.00480*.

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

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.

Tillmann, P. (Forthcoming). Macroeconomic surprises and the demand for information about monetary policy. *International Journal of Central Banking*.

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*.

Wu, J. C., & Xia, F. D. (2016). Measuring the macroeconomic impact of monetary policy at the zero lower bound. *Journal of Money, Credit and Banking*, *48*(2-3), 253–291.

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.

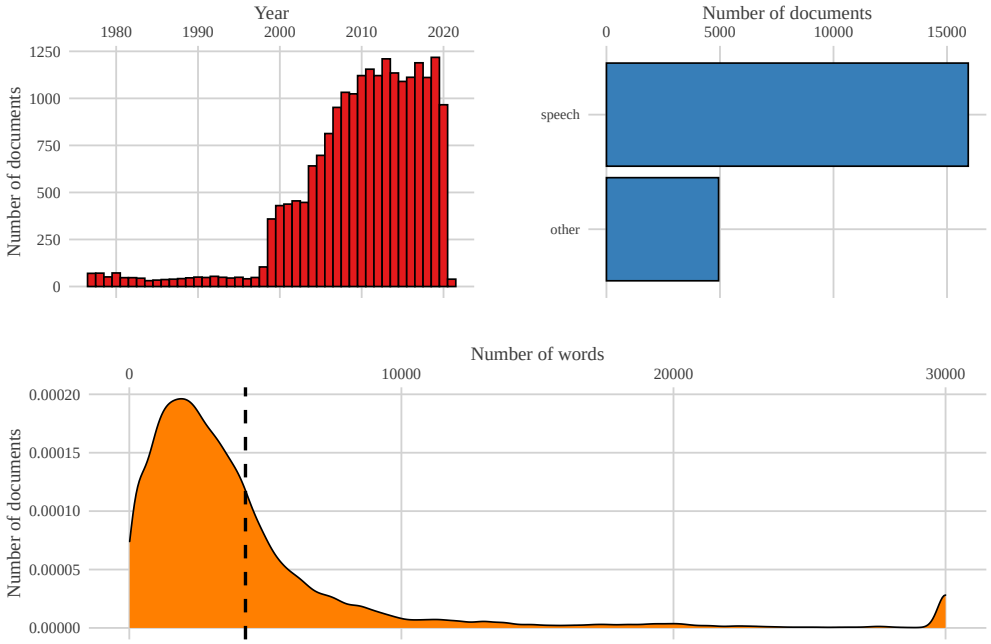
Zahner, J. (2023). Mind the gap: Assessing gender bias in central bank communication. *Unpublished Manuscript*.

Zahner, J., & Baumgärtner, M. (2023). Mastering market sentiments: Decoding the 'whatever it takes' effect. *Unpublished Manuscript*.

Appendix

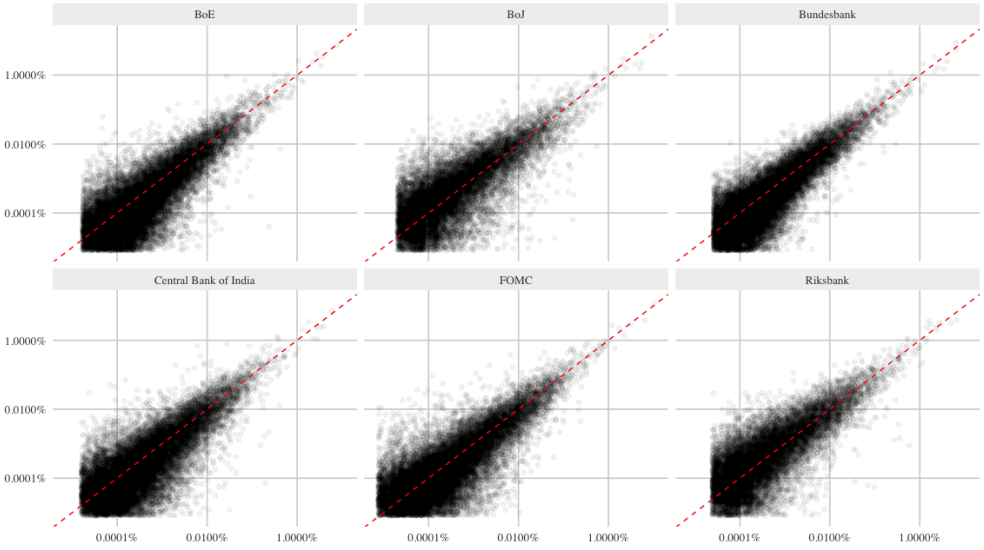
A1. Graphical illustrations of text corpus

Figure A1 : Descriptive summary of the corpus



Note: 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. Overview: Language Models

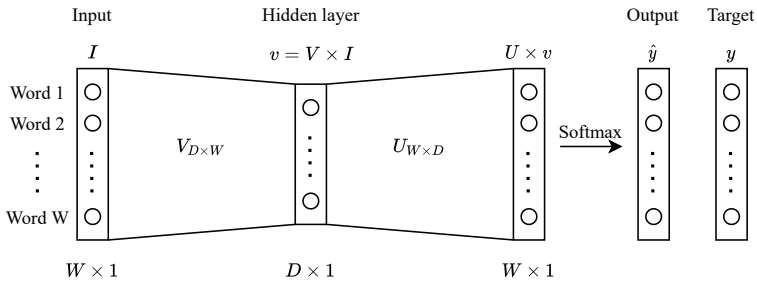
Word2Vec

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 algorithms, 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 A3.

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 over all T observations:

$$L = \frac{1}{T} \sum_{t=1}^{T} \log P(w_t | w_c). \tag{A1}$$

Figure A3 : Graphical illustration of the model of Mikolov, Yih, et al. (2013).



Note: 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.

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

$$P(w_t|w_c) = \frac{exp(u_{w_t}^T v_{w_c})}{\sum_{w=1}^{W} exp(u_w^T v_{w_c})} \tag{A2}$$

where v_{w_c} 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, 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 (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.

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. For example, in Section 4 and Section 5, we demonstrate how similarity in document embeddings may approximate in institutional differences by central banks. However, it should be noted that, unlike word embeddings, document embeddings cannot be easily transferred to new corpora.

LDA

The most famous example of a count-based model in economics is unquestionably the LDA algorithm. Since its introduction by Blei, Ng, et al. (2003), it has been used in monetary policy numerous times (e.g. Hansen and McMahon, 2016; Tobback et al., 2017; Hansen, McMahon, and Tong, 2019; Wischnewsky et al., 2021; Angelico et al., 2022). 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 NLP 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 GloVe, a global factorization method. Following the success of Word2Vec, Pennington et al. (2014) propose GloVe, which trains a language model on 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:

$$L = \sum_{t,c=1}^{W} f(X_{t,c})(w_t^T w_c + b_c + b_t - \log X_{t,c})^2. \tag{A3}$$

In Equation (A3) 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 (A3) 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.

In Table A1, 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 different configurations, we test also different specifications. The hyperparameters we use for each model can be found in Appendix A.A3.

Table A1: 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).

A3. Language Model specifications

We use the hyperparameters for our models. 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).[26] The hyperparameters are summarized in the following table:

Table A2: 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

A4. Additional evaluation

External evaluation I

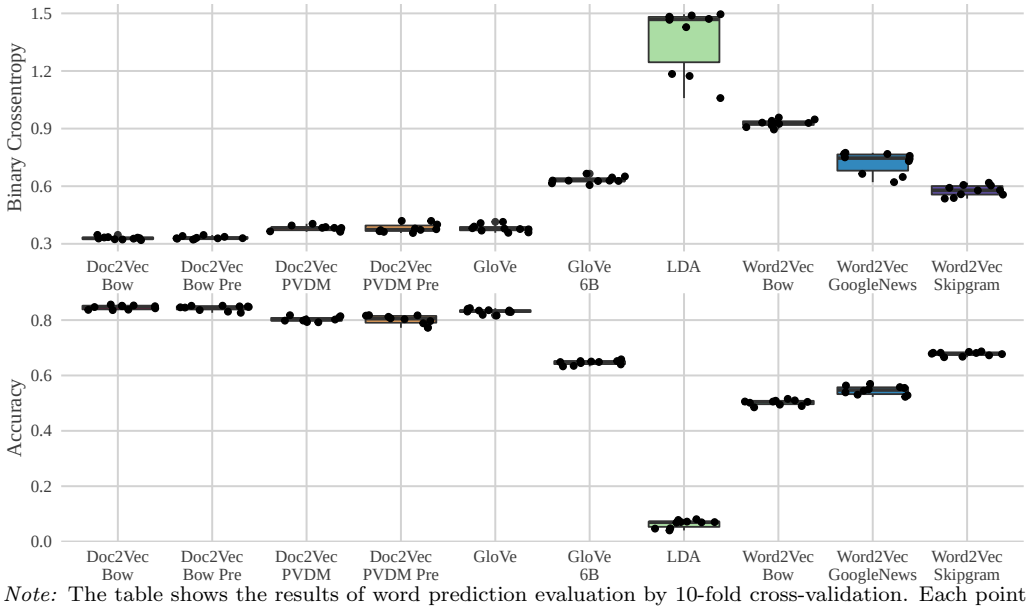
The neural network is based on the Word2Vec skip-gram algorithm. Starting from a central word, the model is to predict the context, the surrounding words. We use a neural network with two embedding matrices. The first is the (word) embedding matrix of the language models mentioned in Figure A1. The second matrix, which represents the context, is first randomly initalized. Both elements are combined using the dot product and a sigmoid layer. We predict which other word is most likely to be nearby for each word. The critical difference to the Word2Vec skip-gram structure is that the first matrix is kept constant throughout training. This ensures that the word embeddings are evaluated even after training rather than an adapted version. We simulate out-of-sample prediction using 10-fold cross validation to ensure a fair comparison between embeddings. Each model is, in each fold, first trained on 90% of the observations.[27] Then, the performance is checked using the remaining 10%. The average overall ten out-of-sample predictions are used as a benchmark for evaluating our embeddings. Figure (A4) shows the

[26]For the Gibbs sampling draws we chose a burn-in rate of 1000, sampled 2000 iterations and returned every fifth iteration.

[27]We train each model with a window size of 1 and 10 negative examples. During backpropagation, the weights of the target matrix are not adjusted. In total, each model is initially trained for 20 epochs.

detailed results for the individual folds per model. The mean values of the folds correspond to the values in Table 2.

Figure A4 : Evaluation Results Word Prediction



Note: The table shows the results of word prediction evaluation by 10-fold cross-validation. Each point corresponds to one test result. The boxplots summarize these results per evaluated model. The measurement on the y-axis is binary cross-entropy and accuracy. For the former, low values indicate good performance and for the latter, high values.

External evaluation II

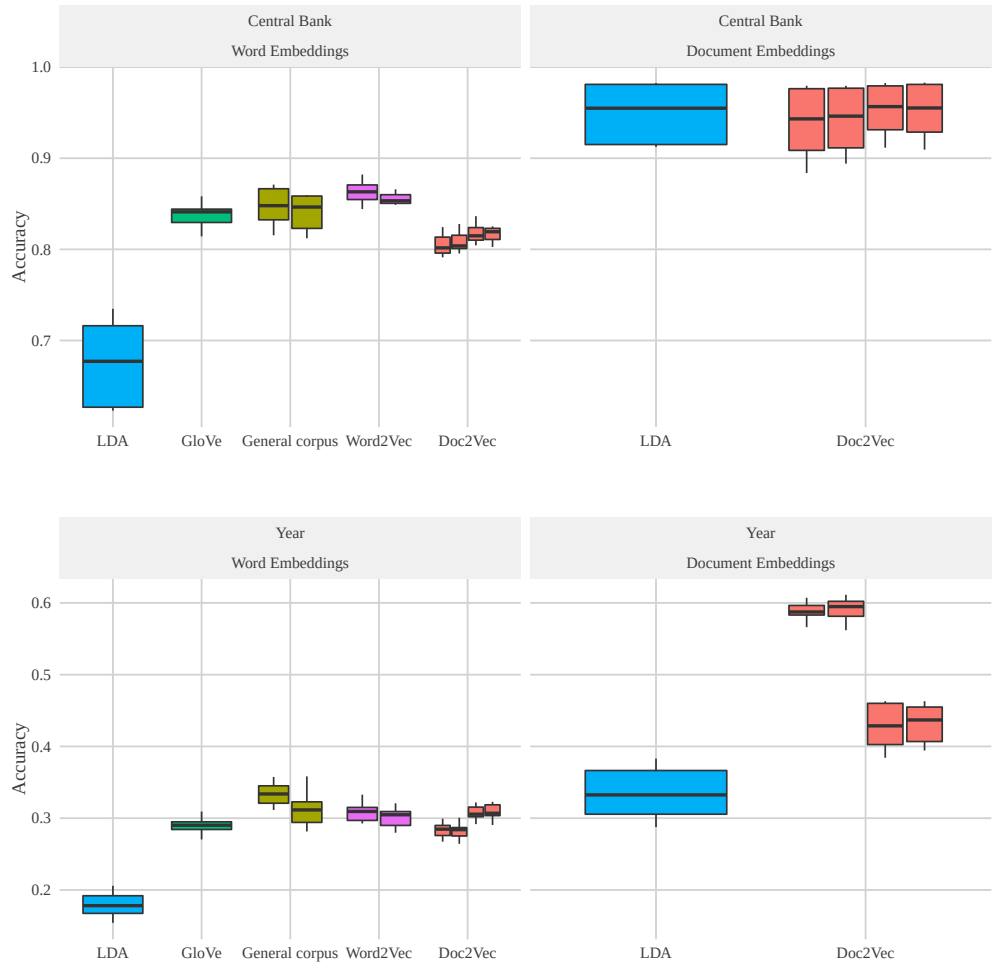
In addition to our economic evaluation task we test our whole embeddings in a more general setting. This should serve as a robustness test with a different task, different empirical methodologies, and far more central bank participation. We select classification tasks that are uninteresting in and of 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.[28]
The word embedding results are illustrated in Figure A5, 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 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.

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

Figure A5 : Evaluation of Embeddings



Note: 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 A3: 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 (1). The underscore is used to highlight collocations as described in Section 3.3.1.

A5. Applications - Robustness checks

Table A4: 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 A5: 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 (3). Higher values indicate closer association to female pronouns and lower values closer association with make pronouns. The respective pronouns can be found in **??**. 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.

Table A6: Regression results - gender focus

	Dependent variable:	
	gender_focus	
	(1)	(2)
2010 - ECB diversity strategy	0.01***	0.01**
	(0.004)	(0.01)
2013 - ECB employment	0.03***	0.04***
	(0.004)	(0.01)
2019 - ECB women scholarship	0.02***	0.03**
	(0.01)	(0.01)
Year		−0.000
		(0.001)
Constant	0.1***	0.1***
	(0.003)	(0.01)
Observations	2,183	2,183
R^2	0.04	0.04
Adjusted R^2	0.04	0.04
Note:	*p<0.1; **p<0.05; ***p<0.01	