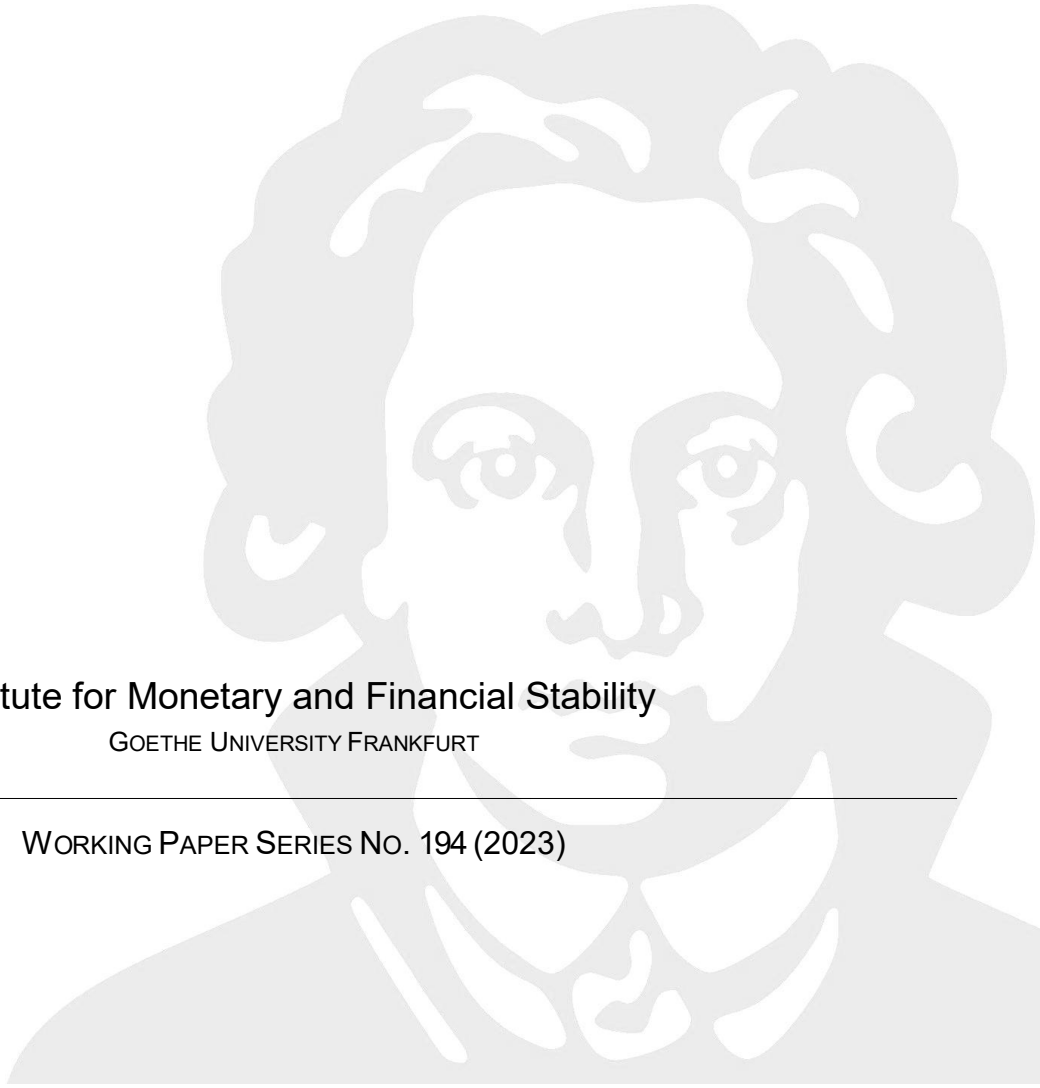
MARTIN BAUMGÄRTNER, JOHANNES ZAHNER

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



Institute for Monetary and Financial Stability
GOETHE UNIVERSITY FRANKFURT

WORKING PAPER SERIES NO. 194 (2023)

Institute for Monetary and Financial Stability
Goethe University Frankfurt
House of Finance
Theodor-W.-Adorno-Platz 3
D-60629 Frankfurt am Main
www.imfs-frankfurt.de | info@imfs-frankfurt.de

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

By Martin Baumgärtner† and Johannes Zahner‡

This draft: October 23, 2024
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 gather a text corpus that is unparalleled in size and diversity in the central bank communication literature, as well as introduce a novel approach to text quantification from computational linguistics. The combination of both 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 affect market expectations and impact monetary policy actions, substantially reducing the inflation response parameter in an estimated Taylor Rule.
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, Michael McMahon, Lawrence Christiano, Isaiah Hull, Ricardo Correa, Davide Romelli, Juri Marcucci, Andreas Joseph, 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 ECONDAT 2021, 54th Annual Conference of the MMF, RES & SES Annual Conference 2023, DNB: the Central bankers go data driven, 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.

† 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.

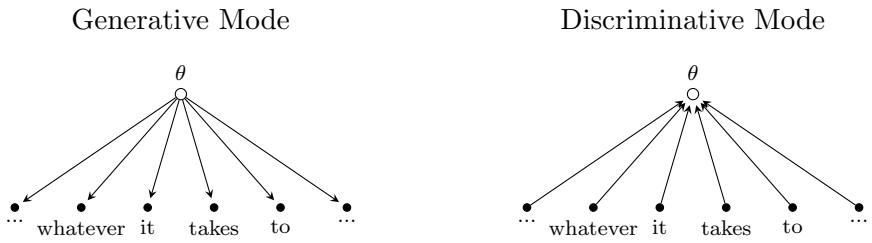
1. Introduction

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 Reserve (Fed), 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. 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 to retrieve the underlying messages from the observed variation in central bank communication ($W \rightarrow \theta$).

Figure 1 : Communication Model



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

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) to count 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. Recent research has shown such indices to be dominated by noise, rather than signal (Hayo and Zahner, 2023).

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. In this space, conceptually similar terms are located near each other. Models such as Mikolov, Yih, et al. (2013) and Pennington et al. (2014) (or more recently *ChatGPT*), leverage large corpora from a variety of sources, such as Twitter or Google articles. While performing well in contexts where the language is similar to their training data, such models are not trained for the technical language used by monetary policymakers.

In this paper, we develop a language model trained explicitly for monetary policy. Our focus is 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 of central bank communication 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.

Fourthly, we present a practical example that demonstrates how researchers may utilise our publicly available language model. We construct a leading index based on inter-meeting communications of members of the Fed, which tracks the strength of the Fed's inflation-targeting regime. In two empirical exercises, we demonstrate that this index causes shifts in market expectations and leads to a moderation of the inflation response parameter within the Fed's rule-based

monetary policy.
The remainder of this paper is structured as follows. Section 2 provides a literature 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. We utilize the best-performing language model in Section 5, to measure the Fed's inflation regime. The final section concludes this paper.

2. Related literature

There are several methods available to researchers for quantifying qualitative information for econometric models. One widely used approach in the field of central bank communication research is to bypass the explicit analysis of the qualitative textual content by instead relying on high frequency (financial) market reactions during periods when a document is published. This particular strand of literature effectively reduces the dimensionality of the documents under consideration by focusing solely on the market's interpretation of the information as captured by their responses to it.[1]
An alternative approach to working with textual data is manual classification, whereby researchers categorize sentences, paragraphs, or sections to quantify qualitative information–a method often referred to as the "narrative" approach. While this process is labor-intensive and prone to misclassification, it allows for the capture of highly specific patterns. For instance, Friedman and Schwartz (1963) analyzed internal Fed deliberations and debates to identify money supply contractions that contributed to the Great Depression, Romer and Romer (2004b) use similar narrative records to create true monetary policy shocks, Ehrmann and Fratzscher (2007) employ manual classification to compare different types of central bank communication, and Tillmann (2021) classify responses in ECB press conference's Q&A sessions to estimate a disagreement index.[2]
In the past decade, NLP methods have become prominent in economics, particularly in the analysis of central bank communication. Gentzkow, Kelly, et al. (2019), Chakraborty and Joseph (2017), and Ash and Hansen (2023) provide excellent surveys of the use of text data, with a focus on economics and monetary policy. Most applications in this field focus on so-called dictionaries that sign categories to specific terms thereby quantifying the qualitative information into few dimensions. Dictionary-based methods inherently assume that the text

[1] Among important contribution to this strand of the literature are Gürkaynak et al. (2005), Brand et al. (2010), Bernanke and Kuttner (2005), Cieslak and Schrimpf (2019), Jarociński and Karadi (2020), Swanson (2021), and Jarociński (2022), who all utilize intraday data around the reading of press-conference statements to measure the effect of monetary policy decisions.

[2] A key shortcoming of this approach is that it deciphers the message with respect to only few dimensions. Another limitation, shared with much central bank communication literature, is its focus on the supply of information, while neglecting potential demand effects. However, Tillmann (2023) recently showed that market participants typically react to communication surprises in predictable ways.

corpus provides limited information about the concept of interest and that the researcher holds strong priors regarding the specific language used to describe that concept. In certain cases, these assumptions hold, making dictionary-based approaches the most appropriate tool for analysis. Notable examples include the construction of an economic uncertainty indices through term frequency counts in news articles (e.g. Baker et al., 2016; Ferrari and Le Mezo, 2021), stock market predictions using a psychosocial dictionary on a Wall Street Journal column (Tetlock, 2007), or measuring media slant in American news-outlets from phrase frequencies in Congressional Records (Gentzkow and Shapiro, 2010).

Dictionaries have also been widely used in the context of central bank communication, some being explicitly designed to financial and monetary policy language (e.g. Loughran and McDonald, 2011; Apel and Grimaldi, 2014; Bennani and Neuenkirch, 2017; Correa et al., 2021). The necessity of central bank specific dictionaries arises from the language employed in this field. Take, for instance, the term *"liability"*, which carries a negative connotation in common parlance while it is a purely technical term within the realms of finance and monetary policy. 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, dictionaries also present significant limitations, including oversimplification of language, omission of potentially relevant information, and a lack of objectivity. First, in dictionaries the relationship between latent concepts and words characterized by a prior assumption that most words are irrelevant, resulting in the exclusion of a substantial portion of text. As Harris (1954, p. 156) points out, *"language is not merely a bag of words [dictionary] but a tool with particular properties which have been fashioned in the course of its use"*.

Moreover, dictionary construction is inherently subjective. Researchers curate a subset of a language's vocabulary based on their interpretation of each term's meaning, introducing biases that can distort the results. Finally, terms defined in dictionary are typically classified in binary terms (positive/negative, for example), thereby failing to capture the context-dependent nuances of language. As a result, dictionary approaches quantify complex concepts in low-dimensional representations, prone to measurement error. For instance, the phrase *great recession* would be classifies using Loughran and McDonald's (2011) sentiment dictionary, even though the term *great* is not meant to be positive in this context. Conversely, expanding dictionaries to include more sophisticated terms or bigrams (e.g., *great_recession*) may improve specificity but at the cost of reduced sensitivity. The trade-off illustrates how dictionaries struggle to account for the interaction between words and their meanings.

Consequently, dictionary-based methods have been found to be extremely noisy.

For instance, when Hayo and Zahner (2023) examined how much variation in sentiment-based indicators of central bank communication could be attributed to changes in macroeconomic, financial, and monetary variables, they found that these factors explained only a small fraction of the underlying variation–typically less than 5%. The authors conclude that their *"findings cast some doubt on the reliability of conclusions [...] that are based on variations in dictionary-based sentiment indicators"* (Hayo and Zahner, 2023, p. 5).

Recent research has acknowledged the the dictionary limitations, suggesting augmenting indices or combining different dictionaries. An example for the former approach is Tadle (2021) who use two dictionaries (hawkish/dovish and positive/negative), discarding a sentence's classification as hawkish if it contains more negative than positive terms. They shows how this augmented sentiment index helps explain movements in high-frequency variables during the Fed press conference. An example for the latter approach, studies such as Azqueta-Gavaldon et al. (2019), Kalamara et al. (2020), Shapiro, Sudhof, et al. (2020), Gorodnichenko et al. (2021), and Kanelis and Siklos (2022) combine multiple sentiment (and other) indices in regression models. They find that different dictionaries capture distinct aspects of an underlying corpus and complement each other. We replicate this approach in Section 4, observing moderate improvements in predictive power when combining up to four dictionaries.

In addition to these augmentations, alternatives to dictionary approaches are becoming more popular, with two prominent examples being the concepts of *similarity* and *complexity*. Similarity measures the distance between two documents' vocabulary (more on that in Section 4). Acosta and Meade (2015), Amaya and Filbien (2015), and Ehrmann and Talmi (2020) find that introductory statements of major central banks became more similar over time. Meanwhile complexity is commonly approximated using the Kincaid et al. (1975) grade level, measuring how many years of formal education are necessary to comprehend a text. 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 communication, a trend predicted by Gentzkow, Kelly, et al. (2019) quote. Word embeddings are multidimensional word representations that have been applied in various context. For instance, Azqueta-Gavaldon et al. (2019) and Cieslak, Hansen, et al. (2021) improve the aforementioned uncertainty indices, Jha et al. (2020) improve central bank sentiment indices, Hu and Sun (2021) decompose central bank vague talk, Handlan (2020) and Hansen and Kazinnik (2023) generate monetary policy shocks from Fed press statements, Bertsch et al. (2022) measures the Fed's interpretation of its financial stability mandate, Campiglio

et al. (2023) measure the "greenness" of central bank talk and Apel, Grimaldi, and Hull (2019) measure central banker disagreement. There are also interactions between embeddings (mainly from the subdomain of topic-modelling) and sentiment analysis, such as Hansen and McMahon (2016) and Fraccaroli et al. (2020).

Most existing studies rely on language models that are trained on broad, general purpose corpora, such as Wikipedia articles. Shapiro, Sudhof, et al. (2020), for instance, 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, as we argue in section 3, the lack of predictive power may not reflect an inherent flaw in the use of embeddings, but rather stem from the non-specific nature of the training corpus. Many general language models lack relevant monetary policy specific terms, such as *hicp*.

Notable exceptions, and thus methodologically the closest research to our paper, are Apel, Grimaldi, and Hull (2019) and Bertsch et al. (2022). Apel, Grimaldi, and Hull (2019) 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. Similarly, Bertsch et al. (2022) train a transformer-based models on Fed speeches from the 1960s to 2020. While their transformer embeddings are usefull for the analysis of central bank mandates, they are less general purpose compared to our model being trained on speeches by over 100 central banks. Moreover, the transformer structure makes these embeddings less straightforward to incorporate into downstream economic empirical analysis.

To the best of our knowledge, we are the first to train embeddings on a specific central bank communication corpus. 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; Bertsch et al., 2022), and on the other hand, the need to fine-tune these representations for their respective use (Loughran and McDonald, 2011).

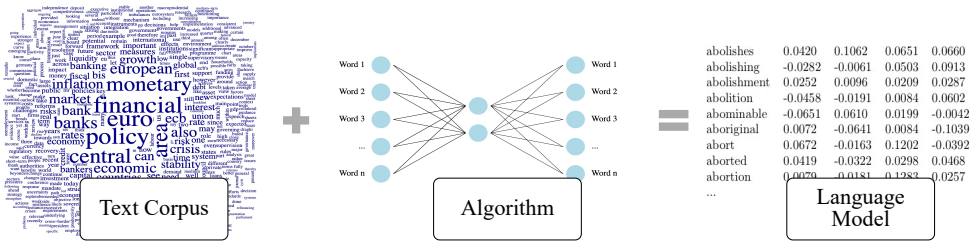
3. Methodology

> *"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 where the training of the model is independent from the application, 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.

Figure 2 : How to retrieve a language model



Note: This graph outlines the paper's structure. In Section 3.1, we introduce a novel text corpus exclusively focusing on central bank communication. Section 3.2 details the various machine learning algorithms employed. Training these algorithms on the text corpus yields the language models, which we evaluate in Section 4 and subsequently utilize in Section 5.

3.1. Text Corpus

Our text corpus reflects our paper's primary focus on central bank communication. To make the corpus as broad as possible, we acquire all English central bank speeches published by the Bank for International Settlements (BIS).[3] In addition, we collect reports, minutes, forecasts, press conferences and economic reviews gathered from central bank websites.[4] However, as Figure 3 shows, the predominant form of communication we have collected (75%) remain central bank speeches. An overview of the sources used for building the corpus can be found in Table A1.

Next, we enrich the corpus with meta-information, such as the title, speaker, role of the speaker, event where the speech was delivered, and other relevant data. In total, we collect 21,916 documents, comprising 112 million words, from 128 central banks. Table 1 provides a breakdown of the contributions by each central bank.

[3]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.

[4]We exclude media interviews to the extent possible for two main reasons. First, they are typically not systematically published on the BIS or central bank websites. Second, they do not represent strictly central bank communication, as it would be difficult omit the contributions of the interviewer on such a large scale.

As expected, the most prominent sources are the Fed, Nippon Ginkō (the Bank of Japan), and the ECB, which collectively account for around 40% of the corpus. Emerging markets are also well-represented, with the Reserve Bank of India as the fourth-largest contributor, providing over 800 speeches (approximately 4% of the corpus), and Bank Negara Malaysia (the Central Bank of Malaysia) contributing over 450 speeches (around 2%). Our extensive coverage is illustrated in Figure 3. We represent 83% of the global population and 89% of global GDP based on 2007 data.

Table 1: Corpus Summary

	Central Bank	Number of Speeches	Fraction of the Corpus
1	US Federal Reserve	3706	16.9%
2	Bank of Japan	2779	12.7%
3	European Central Bank	2481	11.3%
4	Reserve Bank of India	814	3.7%
5	Sveriges Riksbank	800	3.7%
6	Deutsche Bundesbank	703	3.2%
7	Bank of England	659	3.0%
8	Reserve Bank of Australia	644	2.9%
9	Bank of Canada	505	2.3%
10	Central Bank of Malaysia	467	2.1%
	...		
128	Banco de Guatemala	1	0.005%

Note: Own calculations; based on text-corpus as described in Section 3.

In terms of the temporal distribution, the majority of the documents stem from the past two decades. As shown in Figure 3, around 93% of the text stems from the post-2000 period, with the majority (55%) originating in the 2010s.
In contrast to the previous NLP central bank communication literature (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. To map these relationships in the embeddings, it is advantageous to identify related words and combine them as a token, for example, *federal_reserve*. To do this efficiently in our large corpus,

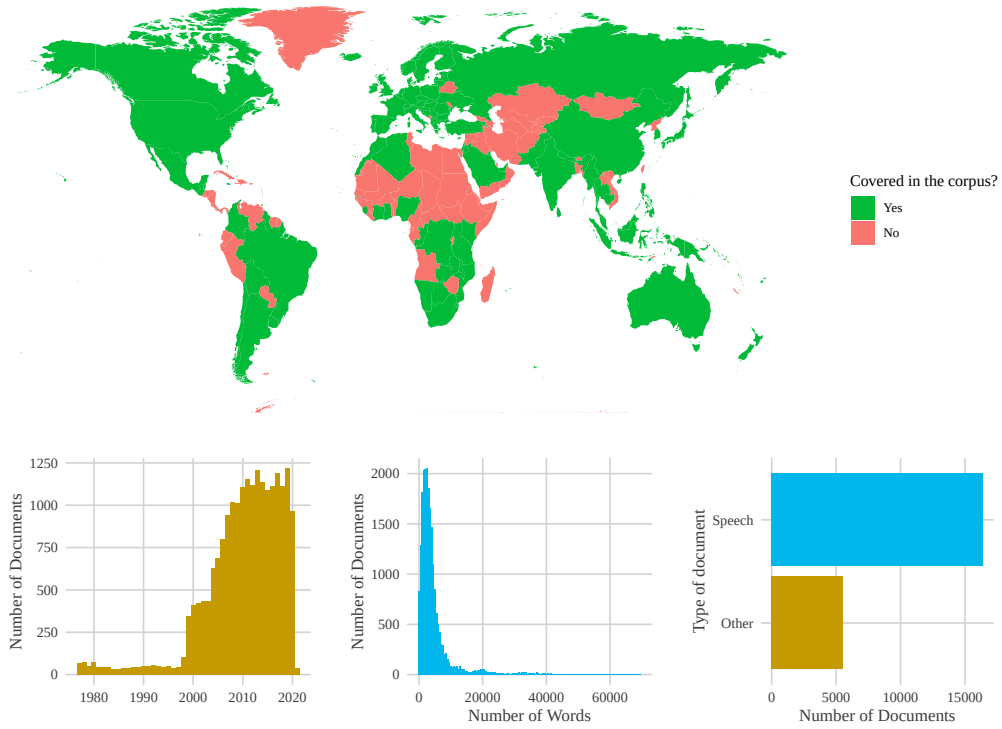


Figure 3 : Properties of the Text corpus

Note: Based on the authors' own calculations from the statistical summary of the text corpus, as introduced in Section 3. Data available upon request. The world map reflects the central bank's headquarters as a proxy for the jurisdiction of the corresponding currency.

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.

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.

3.2. Algorithm

Modern language models largely follow the proposition of linguistic Zellig Harris's (1954) distributional hypothesis, that the meaning of a word can be approximated by examining the distribution of the contexts in which it occurs. Specifically, if two words consistently appear in similar contexts, they are likely to represent similar concepts. Conversely, differences in their contextual environments signal differences in meaning. To illustrate this, consider the term *forecast* as used by then-Chairman Ben Bernanke in 2005:

> *"[...] for example, the blue chip consensus* ***forecast*** *released yesterday looks for real growth 3.6 percent this year [...]*
>
> — Ben Bernanke at the Finance Committee Luncheon, 8 March 2005

The adjacent (underlined) words surrounding *forecast* form the word's 6-word "context window". According to Harris (1954), words that appear frequently in similar contexts tend to have related meanings. For instance, in a 2011 speech, then-Vice Chair Janet Yellen used the term *outlook* in a comparable context:

> *"[...] for example, the blue chip consensus* ***outlook*** *for real gdp growth has edged down only modestly [...]*
>
> — Janet Yellen at the Economic Club of New York, 11 April 2011

This recurring pattern is not coincidental. In fact, the phrase *"the blue chip consensus"* is followed by the term *"outlook"* 32 times and *"forecast"* 10 times in our corpus, emphasizing their semantic similarity based on shared contexts. Harris's (1954) distributional hypothesis is directly incorporated into modern language models. Each word has as a numerical vector representation, known as a word embedding. For instance, the word *forecast* is represented by the following vector:

$$\mathrm{v}_{forecast} = [-0.063, -0.026, 0.0007, ...]$$

Similarly, outlook is represented by its own vector:

$$\mathrm{v}_{outlook} = [-0.051, -0.060, -0.015, ...]$$

Such vector-representation allow us to capture word similarities in a continuous, high-dimensional space. Words with similar meanings–such as *forecast* and *outlook*–will have vectors that are close to each other in the word-embedding-space. This is archieved by training a neural network to predict the target word, given the surrounding context, i.e. *P(outlook | for, example, the, blue chip, ...)*. For instance, predicting the missing word in the following sentence:

> FOR EXAMPLE THE BLUE CHIP CONSENSUS _____ RELEASED YESTERDAY LOOKS FOR REAL GROWTH

To predict well on average, given this context, the neural network must assign high probabilities to both *outlook* and *forecast*. As a consequence of the prediction

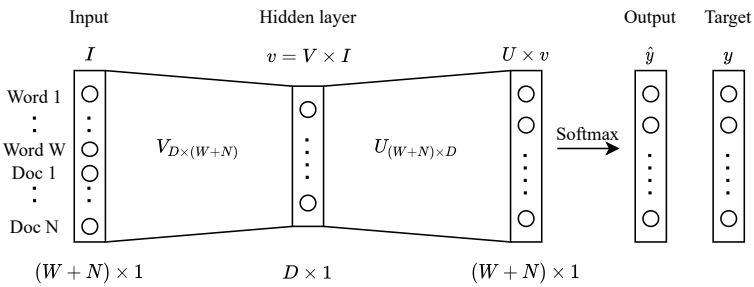
task, the algorithm places these words close to each other in the word-embedding space, ultimately capturing the semantic meaning as a byproduct. In contrast, a term like *GDP*, which is less likely to occur in this particular context, would be positioned further away in the embedding space. The distance between these three terms then accurately reflects the fact that *GDP* is a distinct concept from *outlook* or *forecast*.
The above described algorithm is called *Word2Vec*, a popular prediction-based model which employs neural networks to make these predictions from context (Mikolov, Yih, et al., 2013; Mikolov, Chen, et al., 2013; 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 the adjacent context words. We provide the mathematical summary of Word2Vec in Appendix A.A2.

While word embeddings capture the meaning of individual words, document embeddings extend this concept to entire documents (in our case speeches), retaining the same properties. For instance, the document embeddings for the two earlier speeches are the following two vectors:

$$\mathrm{v}_{Bernanke\ speech} = [-0.034, 0.028, 0.043, ...]$$
$$\mathrm{v}_{Yellen\ speech} = [-0.038, -0.042, 0.020, ...]$$

Training document embeddings alongside word embeddings provides a significant advantage when capturing broader themes. Take the quotes from Ben Bernanke and Janet Yellen above. While the context words (*blue, chip, consensus, ...*) are similar, the speeches differ in the broader context: Bernanke's quote discusses real growth above 3%, while Yellen's discusses a slowdown. Such a subtle distinction may not be fully captured by simply aggregating word embeddings. Document embeddings, on the other hand, are designed to capture the unique thematic content of the entire document. This is analogous to including a document-specific fixed effect in a traditional regression model. In this paper, we use Doc2Vec by Le and Mikolov (2014) to train document embeddings directly. Doc2Vec assigns each document its own vector, trained simultaneously with the word embeddings from the corpus. An illustration of the Doc2Vec model is provided in Figure 4.
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. A comprehensive introduction into Word2Vec, Doc2Vec, LDA, and GloVe can be found in Appendix A.A2.
Finally, an alternative to training embeddings from scratch is the use of pre-trained general language models (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

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 A1 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.

corpora. Since pre-trained language models are methodology-independent, one can find both pre-trained Word2Vec models and pre-trained GloVe models. We compare our embeddings to the two most prominent word-embedding models: Glove6B and Word2Vec Google News. GloVe6B (Pennington et al., 2014) has been 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 is trained on Google News articles.

3.3. Rhetoric Stability

If context defines meaning, than the stability of that context becomes a necessary condition for inference to consistently consistently map input words to feature values. If the context for a given word differs significantly between the training and application phases, the resulting analysis may yield biased estimates. We refer to the condition where the training and application share a stable context *"rhetorical stability"*.[5] In the following, we outline a test for assessing rhetorical stability.

We propose the following test. Retrieve the context words (i.e. the surrounding terms) for key terms, in our case monetary policy terms, such as the word "inflation". If the context terms occur in similar frequency in training and application, the word "inflation" has, per Harris (1954), the same meaning in both corpera.

In the following, we apply our procedure to assess rhetorical stability across central banks in our corpus. We use Wikipedia articles as a control group, simulating the corpus used for training general language models. This comparison also allows us to evaluate how well general models perform compared to our specialized corpus.

[5]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.

We start by obtaining two representative text samples from Wikipedia. First, we collect the whole text of 1,000 random Wikipedia pages using the *getwiki* package in *R*, referring to this dataset as "Wikipedia Random". Second, to explore the potential merits of a specialized corpus, we collect the top twenty Wikipedia pages for a subset of monetary policy terms (see next paragraph). This dataset is referred to as "Wikipedia Monetary Policy" and expected to be more similar to our corpus. From our corpus, we extract eight subsets, in particular all speeches of the most prevalant central banks: the Fed, ECB, BoE, BoJ, Bundesbank, Riksbank, BoC, and RBI.

Next, we select target words relevant to monetary policy, using Apel and Grimaldi (2014)'s Monetary Policy Dictionary.[6] We keep all words occurring within a six-word context of each target word, and standardize the frequencies for comparison across datasets. For example, the dominant term in the context of *"price"* in Wikipedia articles is *"share"*, accounting for about 0.5% of all context words. This contrasts sharply with the Fed, where *"share"* accounts for only about 0.005% of all price-related context words. Other terms such as *"stability"* emerge as significantly more relevant context words for Fed accounting for about 0.2% of all words (for the ECB is is 0.9%). In contrast, *"stability"* as a context word is virtually non-existent in the Wikipedia corpus. To formally test rhetorical stability, we estimate the following linear regression for each target word i the relative frequency f of the context word j between the Fed and k =(Wiki, Wiki MP, ...):[7]

$$f_{i,j,Fed} = \beta_0 + \beta_1 f_{i,k} + \beta_2 i + \epsilon \tag{1}$$

The results are presented in Table 2, highlighting three key findings. First, column one reveals only a small coefficient for Wikipedia Random, indicating a weak correlation between the context words used by the Fed and Wikipedia, and, with 2%, little explanatory power.

Second, the results in column two suggest that monetary policy-focused Wikipedia articles perform notably better than the random sample. Both the correlation and model fit improve three- to fourfold, suggesting that thematic specificity in the corpus significantly enhances context stability, supporting our hypothesis that domain-specific language is more consistent and and should therefore be used in training language models.

Finally, when correlations across central banks consistently exceed 0.45, a fivefold increase over the Wikipedia Random sample, and the explained variance jumps from 2–8% to 35–45%. This confirms that context stability for key monetary policy terms is much stronger in central bank speeches than in Wikipedia, even for RBI, the only developing country central bank in the sample. For example, the language in an RBI governor's speech is seven times more similar to a Fed

[6]Specifically, we use the following nine terms: inflat*, price, wage, cyclical, growth, employ, unemplo*, recover*, and cost, with the asterisk (*) serving as a placeholder.

[7]The results do not depend on the control for the target word, as shown in Table A4

Table 2: Rhetoric Stability

	Dependent variable:							
	'US Federal Reserve'							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
'Wikipedia Random'	0.10***							
	(0.004)							
'Wikipedia Monetary Policy'		0.26***						
		(0.003)						
'European Central Bank'			0.65***					
			(0.002)					
'Bank of England'				0.71***				
				(0.002)				
'Bank of Japan'					0.42***			
					(0.002)			
'Bank of Canada'						0.46***		
						(0.002)		
'Sveriges Riksbank'							0.50***	
							(0.002)	
'Reserve Bank of India'								0.57***
								(0.002)
Keyword Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	104,267	105,109	116,645	110,231	110,314	107,280	120,128	109,352
R^2	0.02	0.09	0.45	0.45	0.33	0.33	0.39	0.34

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 dependent and independent variable are relative term frequencies as defined in Section 3.3.3. Table A4 shows regression results without controls.

speech than to general Wikipedia content, and 2.3 times more similar compared to monetary policy-focused Wikipedia pages.

In summary, our findings strongly suggest that the language used by monetary policymakers differs substantially from that found in publicly available sources such as Wikipedia, but is rather stable across central banks. This in turn implies that the meaning of terms such as "inflation" differs in Wikiepdia articles compared to our corpus, reinforcing our thesis that central bank communication is best captured by specialized language models tailored to a corpus of central bank communication only.

4. Evaluation of language models

In this section, we train the algorithms introduced Section 3 on our central bank communication corpus and evaluate the corresponding language models. The aim is to determine the most effective central bank communication 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 jointly whenever possible. Since there exists no benchmark for evaluating language models in economics yet, we turn to the fields of computational linguistics, where evaluation tasks can be

broadly distinguished as intrinsic or extrinsic. Extrinsic tasks involve evaluating the embeddings against other, externally known contexts, i.e., assessing the embeddings' ability to solve specific tasks. Intrinsic procedures examine whether the embeddings reflect an assumed relationship between words.
We proceed by presenting first two extrinsic evaluations, once with a focus on the linguistic part, the first with a focus on linguistic performance and the second targeting monetary policy relevance, to identify the best performing model across both domains. Following this, we present two intrinsic evaluations on the selected model.

4.1. Extrinsic evaluation

Common extrinsic evaluation methods in computational linguistics involve tasks such as classification or named entity recognition. However, the datasets used for these tasks often evaluate embeddings in a broad context, whereas we're interested in their domain specificity. Due to a lack of established external evaluation methods in economics, we evaluate the embeddings using a two-step approach. First, we assess how well the trained models can predict words, essentially replicating the task they have been trained for. Second, following Le and Mikolov, 2014, we evaluate the models according to their predictive performance, by testing their ability to predict monetary policy shocks from the Fed and the ECB.
To provide a point of reference, we include two wildefly used and publicly available word embeddings in the evaluation tasks– Google's *Word2Vec Google News* and Princeton's *GloVe6B*.

Extrinsic Evaluation 1: Word Predictions. In the absence of an established procedure, we use an unsupervised approach that takes advantage of Harris's (1954) distributional theory, whereas a well-trained language model should be able to predict a missing word based on its surrounding context. To illustrate this, let's revisit the example from earlier:

> FOR EXAMPLE THE BLUE CHIP CONSENSUS _____ RELEASED YESTERDAY LOOKS FOR REAL GROWTH

The missing word here is "forecast". Only if the model predicts "forecast" correctly is it considered an accurate prediction. Each model is trained using a simple neural network architecture as described in the previous chapter, trained on 90% of the corpus and then evaluated for performance on the remaining 10%. The evaluation is based on the metric of "accuracy", which simply divides the number of correctly predicted words divided by the total number of guesses. To account for potential variability in this metric, we repeat this process multiple times using 10-fold cross-validation. This provides a more robust estimate of the model's performance across different data splits. The results of the evaluation are presented in Table 3.

Table 3: Extrinsic Evaluation 1: Word Prediction

Algorithm	Accuracy	Standard deviation
Third Party Word Embeddings		
Word2Vec GoogleNews	0.55	0.016
GloVe 6B	0.65	0.008
(Our) Word Embeddings		
LDA	0.06	0.014
Word2Vec Bow	0.50	0.009
Word2Vec Skipgram	0.68	0.007
Doc2Vec PVDM	0.80	0.009
Doc2Vec PVDM Pre	0.80	0.017
GloVe	0.83	0.008
Doc2Vec Bow	0.84	0.007
Doc2Vec Bow Pre	0.84	0.009

Note: Table shows the evaluation results from Section 4. Accuracy is calculates as Number of correct predictions / Total number of predictions. Standard deviation is calculated using 10-fold cross-validation. Specifications: Bow = (Distributed) Bag Of Words; PVDM = Paragraph Vector Distributed Memory; Pre = pretrained embeddings were used as more efficient starting points.

First, the overall prediction accuracy is high. All models, except LDA, achieved word prediction accuracies above 50%, meaning they get more than 50% of he predictions right. Notably, our own trained models outperform publicly available pre-trained models (Word2Vec GoogleNews and GloVe6B) by up to 20 percentage points. This finding highlights the importance of training language models on domain-specific datasets such as monetary policy documents.
Second, in terms of ranking by prediction accuracy, LDA models are outperformed by Word2Vec models, which are outperformed by both GloVe and Doc2Vec models. The similar performance of an algorithm based on word prediction (Doc2Vec) and one based on corpus-wide statistics (GloVe) is interesting and highlights the value of our agnostic approach to embedding algorithms. Third, we observed a slight (but statistically insignificant) improvement in the performance of pre-trained Doc2Vec models compared to non-pre-trained models. Finally, in terms of predictive power, the Doc2Vec Bow algorithm achieved the highest accuracy, correctly identifying almost 85% of all terms. This represents an improvement of 5 percentage points over the Doc2Vec PVDM algorithm.

Extrinsic Evaluation 2: Interest Rate Predictions. Our second evaluation task addresses a standard empirical approach in the field of central bank communication: the assessment of an interest rate rule that incorporates textual information. This approach, as explored in studies such as Aruoba and Drechsel (2022), aims to improve on the traditional Taylor rule by exploiting the informational content of central bank communication. Specifically, we isolate the

explanatory power of textual information beyond the traditional macroeconomic variables included in a Taylor rule. We achieve this by first estimating a standard Taylor rule:

$$i_t = \alpha + \beta X_t + \epsilon_t \tag{2}$$

where the short-run rate (i, Wu and Xia (2016)-shadow rate) is regressed on a set of macroeconomic variables (X_t), such as growth in consumer prices, GDP growth rate, and the current unemployment rate. The regressions are conducted on a monthly basis for both the US and the Euro Area separately (see Table A8 for Summary Statistics). The error term (ϵ) captures the unexplained variation in the interest rate after accounting for the included macroeconomic variables. We then test which language model is best at explaining that remaining (unexplained) variation.

We evaluate our embeddings against the previously used embeddings by Google and Princton, and against established dictionaries from the fields of computational linguistics (e.g. Hu and Liu, 2004; Mohammad and Turney, 2013), finance (e.g. Loughran and McDonald, 2011), and monetary economics (e.g. Bennani and Neuenkirch, 2017). In light of the potential complementarity of these indices, we also also employ a specification that incorporates them together.

For each speech i delivered at time t, we retrieve a corresponding speech representation, $d_{i,t}$. For the document embeddings, the speech representation $d_{i,t}$ directly corresponds to the pre-trained 300-dimensional vector representing the whole speech (document). For the word embeddings, we construct the speech representation $d_{i,t}$ by calculating the dot product of the word embedding matrix and the corresponding term frequency matrix for each speech, which creates a 300-dimensional vector that captures the semantic content of the speech by considering the contribution of each word, weighted by its frequency of occurrence. For the dictionaries, we incorporate sentiment information by including the calculated frequency of each category in the respective dictionary (e.g. positive, negative, anger, fear, anticipation). In addition, where applicable, we calculate a sentiment (hawkish-dovish) index taking difference between positive (dovish) and negative (hawkish) terms is divided by their sum.

We leverage the unexplained variance from the estimated Taylor rule (Equation (2)) and the speech representations, $d_{i,t}$'s, to assess the marginal explanatory power of textual information. This is achieved through the following regression:

$$\epsilon_{i,t} = h(d_{i,t}) + \nu_{i,t} \tag{3}$$

Here, $\epsilon_{i,t}$ denotes the residual component of the interest rate, $d_{i,t}$ the corresponding high-dimensional speech representation, $h(.)$ a nonlinear transformation function, and $\nu_{i,t}$ the error term. We rely on the transformation due to the high dimensionality of the speech representations (up to 300 dimensions). The high dimensionality leads to over-fitting and hinders the generalisability of the models. To address this challenge, follow Aruoba and Drechsel (2022) and use an

Elastic Net regularisation (e.g. Zou and Hastie, 2005), which identifies and removes potentially irrelevant textual features from the model (see Chakraborty and Joseph, 2017, for further details). As before, we train each model on 80% of the observations and evaluate its performance out-of-sample on the remaining 20%.

The results presented in Table 4 show the relative improvement in explaining the unexplained variance of the interest rate (i.e. ν/ϵ), hence lower values indicating better performance. For example, Loughran and McDonald's (2011) sentiment dictionary reduces the unexplained variance by around 10% (to 90%). The rows of the table are organised according to the method used (dictionary, word embedding, document embedding) and the columns represent the target interest rate (Fed or ECB).

While dictionary-based approaches capture some information, their impact is limited. At most, they reduce the unexplained variance by 13% (US) and 18% (Euro area). Even when combining all dictionaries (resulting in a 24-dimensional representation), Word embedding models perform better, reducing the unexplained variance by an additional 15-20 percentage points. This demonstrates the advantage of incorporating semantic relationships between words. Consistent with Extrinsic Evaluation 1, the word embeddings from the Doc2Vec models outperform the others. Finally, document embeddings provide an additional reduction in unexplained variance, reducing the unexplained variance by a further 10-20 percentage points. The pre-trained Doc2Vec bag-of-words algorithm achieves the highest overall reduction, down to 32% unexplained variance in the Euro area.

Due to its superior performance in both word prediction and interest rate prediction, we select the Doc2Vec bag-of-words variant with pre-trained word embeddings (bolded in all tables) as our primary language model.[8] For readability, we will henceforth refer to this model simply as "Doc2Vec".

[8]The upcoming results are robust across all Doc2Vec variants. Results are available upon request.

Table 4: Extrinsic Evaluation 2: Interest Rate Predictions

	3-month-interbank-rate	
	United States	Euro Area
No model	1.00	1.00
Dictionaries		
Loughran-McDonald Sentiment	0.90	0.87
Hu Lui Sentiment	0.89	0.83
Hawk-Dove Dictionary	0.88	0.86
NRC Word-Emotion Lexicon	0.87	0.82
All Dictionaries combines	0.85	0.74
Word Embeddings		
GloVe6B	0.76	0.62
Word2Vec GoogleNews	0.76	0.62
LDA	0.76	0.64
Word2Vec Bow	0.75	0.61
GloVe	0.72	0.58
Word2Vec Skipgram	0.70	0.60
Doc2Vec Bow	0.70	0.57
Doc2Vec Bow Pre	0.74	0.55
Doc2Vec PVDM	0.73	0.53
Doc2Vec PVDM Pre	0.68	0.52
Document Embeddings		
LDA	0.57	0.42
Doc2Vec PVDM	0.56	0.36
Doc2Vec PVDM Pre	0.56	0.36
Doc2Vec Bow	0.50	0.32
Doc2Vec Bow Pre	0.54	0.32

Note: The table shows the evaluation results across the different algorithms introduced in the previous section. 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 dictionaries are based on Loughran and McDonald (2011), Hu and Liu (2004), Bennani and Neuenkirch (2017), and Mohammad and Turney (2013) respectively, where we include both, the absolute number of identified terms, as well as the relative number (for instance the *sentiment*).

4.2. Intrinsic evaluation

Intrinsic evaluations assess the quality of the learned representations themselves, focusing on how well these representations align with our perception of those words or concepts. Unlike extrinsic scores, there is no objective metric. Consequently, these assessments are inherently subjective and should be interpreted with caution.

We conduct three intrinsic evaluations. First, we present an evaluation of the word-representation of the Doc2Vec model. Next, we comparing the word representation of our Doc2Vec model to Word2Vec GoogleNews and GloVe6B. Third, we present an evaluation of the document representations. All three evaluations rely on the cosine similarity between (word) vectors, which takes the cosine distance between two (word) embeddings. The similarity score between two vectors a and b is calculated follows:

$$S_{a,b} = \frac{a \cdot b}{||a|| \times ||b||} \tag{4}$$

Words with high similarity are considered to be semantically more similar as they are close in the vector space.

Intrinsic Evaluation 1: Semantic Similarity. First, we assess the semantic relationships between key macroeconomic terms (INFLATION, UNEMPLOYMENT and OUTPUT) learned by our Doc2Vec model. Table A5 presents the ten most similar terms in the vocabulary for each term.[9] The results in Table A5 suggest that our Doc2Vec model is capable of grouping words with similar economic meaning. For example, Doc2Vec relates terms such as CORE_INFLATION and INFLATION_EXPECTATIONS to INFLATION, which aligns with their economic meaning. The same is true for the terms of UNEMPLOYMENT and OUTPUT. Furthermore, the model appears to capture relationships between broader economic concepts, as evidenced by the high similarity between UNEMPLOYMENT and ECONOMIC_SLACK.

Intrinsic Evaluation 2: Homonyms. A major challenge for language models arises from homonyms–words that have multiple meanings depending on the context. Because each word occupies a single point in a high-dimensional embedding space, homonyms introduce noise into those models. Take the term BASEL, which may refer to either to the Swiss city or the Basel accords.[10] We expect general language models such as GloVe6B and GoogleNews to primarily associate BASEL

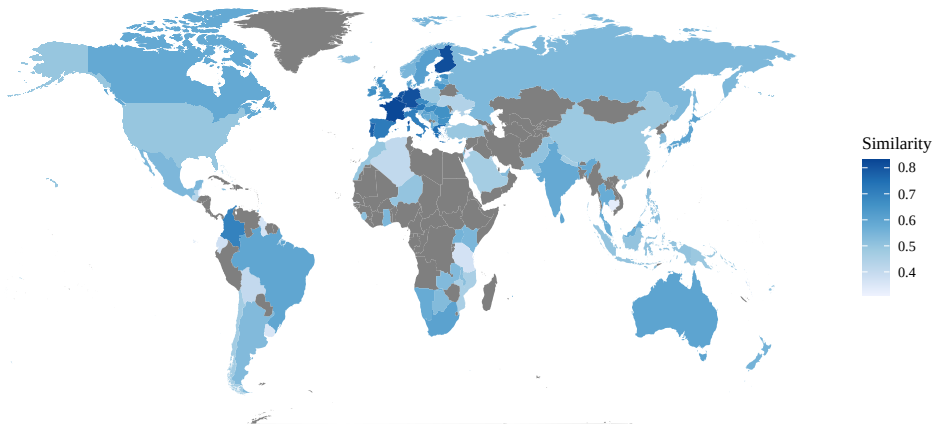
[9] On our website (https://sites.google.com/view/whatever-it-takes-bz2021) we provide an interactive tool that allows users to make the same assessment for any word in the entire vocabulary.

[10] Another example (by Loughran and McDonald, 2011) discussed earlier is the term LIABILITY. The meaning differs when used in every-day context to a financial context. In the Appendix, we provide additional examples for the interested reader.

with the city, potentially introducing noise into monetary policy research that relies on these models to quantify central bank communication. To test this hypothesis, we examining the similarity of the term BASEL across ur Doc2Vec model, GloVe6B and Word2Vec GoogleNews in Table A6. Indeed, while the latter two associate BASEL primarily with the city, whrereas our Doc2Vec associates it with banking regulation, even correctly identifying abbreviations such as Basel Committee on Banking Supervision (BCBS). Whilst in a non-monetary context the city may be a better fit in general parlour, for monetary policy communication, banking regulation is clearly the more relevant concept.

Intrinsic Evaluation 3: Central Bank Similarities. Finally, we conduct an intrinsic evaluation of the document embeddings. Our starting point is a simple conjecture: central banks in western economies with similar objectives are likely to exhibit more similar communication compared to those in other regions. We operationalise this idea by averaging the document embeddings for each central bank and then estimating their similarity to the ECB. By choosing the ECB as the reference central bank, we can also illustrate the national euro area central banks. The results are shown in Figure 5, where darker colours indicate greater similarity. Consistent with our hypothesis, central banks in Europe and North America appear to be more similar to the ECB, which is in line with our intuition. We use this observation as a starting point for our analysis of monetary policy regimes in the next chapter.

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 (4).

In summary, this section has focused on the quantification of words and documents using the previously presented algorithms. We evaluated all algorithms based on their out-of-sample prediction performance, subsequently selecting Doc2Vec due to its superior results. To further assess the quality of the Doc2Vec embeddings (both at the word- and document-level), we performed three intrinsic evaluations. These evaluations examined whether the embeddings capture relationships that correspond to our perceptions. Our results provide evidence that the Doc2Vec embeddings do indeed contain meaningful information at both word and document level.

5. Monetary policy regime classification

In the following, we want to highlight the versatility of language models to venture beyond capabilities of word-count indices, specifically, demonstrate how our our Doc2Vec language model can be used to retrieve latent messages, i.e. identifying avenues for $W \rightarrow \theta$. To date, our embeddings have been utilized in various papers: to create an indicator of the ECB's commitment to act as a lender of last resort, to measure the scientification process of the Bank of England (BoE) or to measure the development of gender biases in global central banker communication (Zahner and Baumgärtner, 2023; Goutsmedt et al., 2023; Zahner, 2024). However, in this section, we demonstrate how the Doc2Vec language model may be used to develop a leading indicator for the strength of the Fed's inflation targeting regime.[11]
According to Mishkin (1999, p. 591) a key component of inflation targeting is an *"increased transparency of the monetary policy strategy through communication with the public and the markets about the plans and objectives"* However, traditional literature predominantly identifies monetary policy regime changes by detecting structural breaks in macroeconomic time series data (e.g. Bae et al., 2012; Benati and Goodhart, 2010; Bikbov and Chernov, 2013). These methods effectively pinpoint discrete regime changes, but they rely on past observable action making their identification inherently reactive. As such, predictions in policy actions are traditionally difficult (e.g. Sims and Zha, 2006) and gradual regime changes that evolve continuously are particularly difficult to measure.
To address these limitations, we adopt an alternative methodology, more closer to the narrative approach (e.g. Romer and Romer, 2002; Romer and Romer, 2004a). We identify regime changes within the Fed by focusing on their inter-meeting communication to capture the central bank's prevailing inflation targeting stance. As such, our methodology emphasizes observable intention, rather than action, making it forward-looking and predictive given the accessibility of modern central bank communication.
We proceed as follows: First, we demonstrate that central bank communication

[11]The source code for this application can be found online at https://sites.google.com/view/whatever-it-takes-bz2021. 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.

varies systematically across monetary policy regimes, with a particular focus on inflation-targeting frameworks. Next, we leverage these communication differences to construct a novel, communication-based index that effectively tracks shifts in the Fed's inflation regime. Our index closely follows regime shifts identified in traditional, with the advantage of being continuous and leading. We further demonstrate that the index captures language patterns typically associated with the respective inflation regime. We then show that changes in this index lead to adjustments in both inflation and interest rate expectations. Finally, we find that shifts in our index correspond to systematic deviations in rule-based monetary policy actions.

5.1. Does communication differ across monetary policy regimes?

We begin the empirical analysis by examining the relationship between central bank communication and monetary policy regimes more generally. Establishing this link is central to understanding whether changes in communication reflect changes in the underlying policy regimes. To conduct this analysis, we combine two datasets.

The first dataset quantifies all speeches delivered by each central bank on an annual basis. For this, we rely on the embeddings generated by our Doc2Vec language model specified in the previous section. Specifically, we calculate the average document embedding, resulting in a 300-dimensional representation corresponding to a central bank in a given year. We then compute the cosine distance between each central bank and a reference bank within each year, which is the panel equivalent to the cross-sectional approach described in **??**. In fact, the average cosine distance over time for each central bank relative to the ECB is what we presented in Figure 5. For the purposes of this analysis, we also include the Reserve Bank of New Zealand (RBNZ) and the Fed alongside the ECB, collectively referring to them as the *reference central banks*. The choice of these three institutions is due to their adoption of different inflation targeting regimes over the past two decades (more on this below). Thus, the first dataset allows us to assess how the communication patterns of different central banks align with or diverge from those of the reference central banks on an annual basis.

The second dataset provides a measure for the monetary policy regimes. Here we rely on the annual classification by Cobham (2021).[12] Cobham's classification captures the diversity of policy regimes through ten key target variables (e.g., inflation rate, money supply, ...), that are further subdivided into 32 mutually

[12]We use monetary policy frameworks and regimes somewhat interchangeably throughout the paper. While frameworks are defined as 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), a monetary policy regime is typically defined more narrowly focusing on monetary policy objective, target and instrument (e.g. Bae et al., 2012). Since a monetary policy regime is inherently embedded within a broader policy framework, we argue that any change in the framework would necessarily entail a shift in the regime. The monetary policy classifications are based on the IMF's Article IV Consultation Reports and made available https://monetaryframeworks.org. For members of a currency union, we assign the classification of the union's lead central bank.

distinct categories, ranging from loosely structured discretionary targets to fully converging inflation targets. For the purpose of our analysis, two of these 32 distinct targeting regimes are of particular interest to us:

FIT (Fixed Inflation Targeting) is characterised by its well-defined and consistently enforced numerical inflation target. The regime is associated with transparency and predictability of monetary policy, which helps to anchor inflation expectations more effectively (e.g. Benati and Goodhart, 2010).

LIT (Loose Inflation Targeting) is characterized by a more flexible stance on its inflation target, allowing central banks to respond to a broader set of economic indicators and conditions. As a result, inflation expectations are less anchored around the inflation target (e.g. Benati and Goodhart, 2010).

With respect to our reference banks: All three are all categorized as inflation-targeting regimes throughout the sample period, but only the RBNZ has maintained a *FIT* regime. In contrast, the ECB has been classified under the *LIT* regime for the entire period, while the Fed has transitioned from the *LIT* regime to the *FIT* regime.

Merging both dataset results in over 950 bank-year observations for 88 individual central banks. We complement this dataset with corresponding macroeconomic variables –specifically, the inflation rate, unemployment rate, and log(GDP)– which allows us to control for the macroeconomic environment when analyzing communication similarity. The data sources and the descriptive statistics of all variables are documented in Table A10 and Table A11 respectively.

The empirical identification strategy is straightforward. We regress the cosine distance ($S_{i,j}$) of bank j and a reference central bank i in year t on dummy variables representing the respective monetary policy regime of bank j ($Regime$). We account for macroeconomic conditions by including the difference in our macroeconomic indicators between central bank j and the reference central bank i ($X_{i,j}$) as well as for time-specific variation by including year fixed effects:

$$S_{i,j,t} = \beta_1 Regime_{j,t} + \beta_2 X_{i,j,t} + \epsilon_{i,j,t} \tag{5}$$

The coefficients of interest are the β_1's, specifically those on inflation targeting regimes, which we report in Table 5. The full regression table can be found in Table A12.

We find the following results. First, a simple comparison of the communication similarity between inflation-targeting central banks and those with other monetary policy targets (column 1) finds a consistently positive and significant coefficient for the RBNZ, Fed, and ECB regressions. This suggests that central banks with inflation-targeting regimes tend to communicate more similarly to our reference central banks compared to banks with other regimes. In terms of economic relevance, being classified as an inflation-targeting central bank increases the similarity from 34-40% to 45-52%, or more than one standard deviation (see

Table 5: Regression results: Monetary Policy Regime classification

	Dependent Variable: Similarity towards i								
$i =$	RBNZ			Fed			ECB		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Inflation Target	0.11***	0.09***		0.12***	0.09***		0.15***	0.15***	
	(0.02)	(0.02)		(0.02)	(0.02)		(0.02)	(0.03)	
– FIT			0.14***			0.12***			0.10***
			(0.02)			(0.02)			(0.02)
– LIT			0.09***			0.10***			0.17***
			(0.02)			(0.02)			(0.02)
– FCIT			0.04			0.10**			0.15***
			(0.04)			(0.04)			(0.04)
– LCIT			0.05*			0.03			0.08***
			(0.03)			(0.03)			(0.03)
[...]									
Constant	0.34***	0.35***	0.34***	0.40***	0.42***	0.41***	0.37***	0.36***	0.37***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Rem. MPF Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Macro-controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	957	957	957	957	957	957	957	957	957
R^2	0.22	0.24	0.28	0.18	0.19	0.22	0.26	0.32	0.38
Adjusted R^2	0.20	0.21	0.26	0.16	0.17	0.19	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): 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; NoNat = no national framework. *Rem. MPF Controls* indicates controls for all monetary policy frameworks not shown in the table.

Table A11). Importantly, the result persists when controlling for macroeconomic conditions (column 2).

Next, we leverage the full range of classifications from Cobham's (2021) in column 3 by further subdividing inflation-targeting regimes into full inflation targeting (FIT) and loose inflation targeting (LIT), as well as the "converging" categories ($LCIT$, $FCIT$), which represent non-constant targeting over time. The results are interesting both within and across the three targeting central banks. For the RBNZ, which has been classifies as a FIT regime throughout our sample period, we find that communication similarity is largely driven by other FIT regime banks. In contrast, the ECB, which has maintained a $LCIT$ regime over the entire sample period, shows greater similarity in communication with other $LCIT$ regime banks. For the Fed, which transitioned from $LCIT$ to FIT in 2011, communication similarity is evenly distributed between the two regimes.

These results make us confident that one of the key factors driving similarity

in central bank communication is the adoption of a common monetary policy regimes.

5.2. The evolution of the Fed's inflation targeting regime

In the next step, we utilize the the distinction between LIT and FIT regimes to quantify the extent to which Fed communication at time t aligns with that of either regime. Specifically, for each Fed speech, we calculate its Euclidean distance to the average non-Fed FIT and LIT speech.[13] We term the resulting index IT, for "inflation targeting". Positive values indicate closer alignment with FIT regimes (stronger inflation targeting), while negative values correspond to LIT regimes (looser inflation targeting). For interpretation purposes, we standardise IT, so that a $IT = 1$ would corresponds to a one standard deviation from the mean towards the FIT regime.

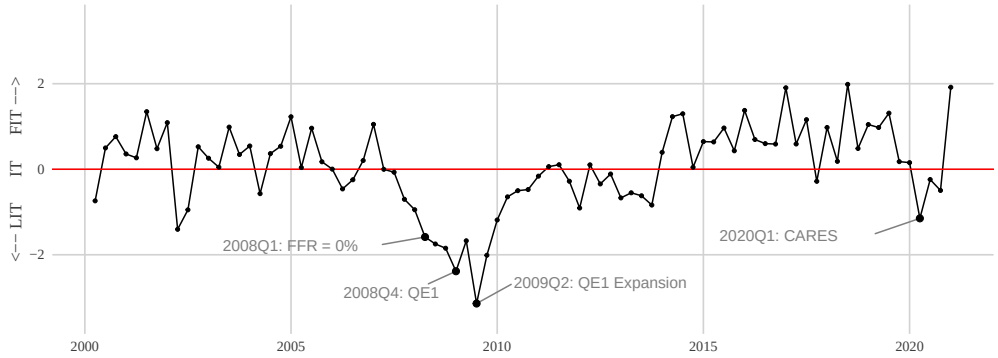
Figure 6 presents the average quarterly IT. Notably, we observe extreme negative deviations in IT that coincide with key policy events where other objectives, such as financial stability, took precedence. These include periods such as the Global Financial Crisis, when the Fed hit the zero lower bound (December 2008), the initiation (November 2008) and expansion (March 2009) of QE1, and the introduction of the CARES Act in response to the COVID-19 pandemic.

Fluctuations in our IT regime index closely track regime shifts in other studies of the Fed, such as Bae et al. (2012)–which use a Markov switching regime– and Bikbov and Chernov (2013)–which rely on a term structure model. For instance, Bae et al. (2012, Figure 5) documents a heightened inflation response–indicative of an FIT regime–during the Greenspan era, which closely mirrors the FIT stance observed in our IT index, up until the end of 2006. Similarly, Bikbov and Chernov (2013, Figure 1), who classify monetary policy into active and passive regimes, identify shifts that generally coincide with transitions in the IT index. For example, the transition to a passive regime after 2006 corresponds to a shift towards a LIT regime in our index. These similarities strengthen our confidence that the Fed's communication patterns may systematically reflect changes in policy stance, suggesting that the Fed's communication may credibly signal regime shifts.

[13]The *relative norm distance (RND)* was proposed by Garg et al. (2018) for comparison with categorical classification such as the one present. For each of the 4219 Fed speeches (s), we measure the Euclidean distance to the average of all $N = 3954$ FIT speeches (v_{FIT}) and $K = 3868$ LIT speeches (v_{LIT}) in our corpus:

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

Figure 6 : FED's stance on inflation targeting



Note: Standardized time series of the IT index, as defined in Footnote 13. Higher values indicate stronger alignment with the FIT regime, while lower values correspond to greater alignment with LIT regimes as described in Section 5. Index is based on the authors' calculations; data available upon request.

5.3. Communication patterns of FIT and LIT periods

Before evaluating the impact of our IT index on expectations and policy actions, it is important to assess whether speeches with high IT values emphasise topics consistent with FIT regimes–such as explicit mentions of inflation targets–and vice versa. We begin with two anecdotal examples where the IT index strongly suggested one regime over the other.

One of the strongest examples of a FIT speech is Donald Kohn's address at the Conference on Finance and Macroeconomics in February 2003.[14] In this speech, Kohn emphasizes the Fed's firm commitment to its inflation target:

> *"Economic contractions have frequently been led by weakness in the household sector, which often has responded to higher interest rates as the Federal Reserve acts to reverse inflation pressures"*

In addition, throughout the speech, Kohn provides a transparent outlook on the Fed's assessment of the economy and its strategy to achieve the inflation target:

> *"With production currently well below potential and inflation and inflation expectations low, it is doubtful that the temporary misalignment of rates would result in the development of any perceptible inflation pressures before the Federal Reserve would have time to take countervailing steps."*

> *"Judging from this analysis [...] it seems likely that as the economy strengthens [...] interest rates rise in response".*

[14] https://www.federalreserve.gov/boarddocs/speeches/2003/20030228/default.html

Kohn further emphasizes the importance of clarity and transparency in conveying the Fed's inflation-targeting strategy and the actions it will undertake to achieve its objective:

> *"Among other things, markets could get it wrong–for example, they could anticipate greater strength in underlying demand than is actually occurring. [...] We would set rates lower than the markets have built in, and in our various statements we would attempt to make clear our assessment of economic prospects."*

By addressing elements such as a strong commitment to the inflation target, guidance on the use of policy instruments, and the necessity for transparent communication, this speech lays out the characteristics of a FIT regime. As might be expected, with a rating that is 3 standard deviations above the mean, this speech is among the most FIT-aligned in our dataset.

Contrast this with William Dudley's opening remarks at the Transatlantic Economic Interdependence and Policy Challenges Conference in April 2013. he speech is devoid of any references to inflation, instead prioritizing discussions on non-inflationary targets such as financial stability and fiscal policy:

> *"On the regulatory side, there is considerable good news worth highlighting. In particular, substantial progress has been made in strengthening the global capital and liquidity standards for internationally active banks."*

> *"Nevertheless, the United States could be doing better. The U.S. fiscal policy program, for example, does not appear well-calibrated to the current set of economic circumstances. We have too much fiscal restraint in the short term, and too little consolidation in the long term."*

In addition, Dudley provides only vague guidance on future policy actions, offering primarily a retrospective account of the Fed's ongoing unconventional policy measures:

> *"To provide the appropriate degree of accommodation, the Federal Reserve has recently moved to an outcome-based approach in which the use of our tools is explicitly tied to developments in the economy and economic outlook. Currently, as part of this strategy, we are purchasing $85 billion of longer-term Treasuries and agency mortgage-backed securities each month."*

Dudley's speech is evidently more consistent with a LIT regime, a conclusion that is corroborated by our IT index, which ranks it among the strongest LIT-aligned speeches (approximately 3 standard deviations below the mean).

Both speeches further illustrate the imprecise measurement of traditional hawkish/dovish indices or sentiment dictionaries. Donald Kohn's speech employed

dovish language (using terms like *weakness*) and would therefore be classified as dovish with a negative tone using conventional dictionaries (e.g., Bennani and Neuenkirch, 2017; Loughran and McDonald, 2011). However, interest rates showed minimal movement in the subsequent 12 months, with the Wu and Xia (2016) rate declining by around 30 basis points, while the broader economic environment remained relatively stable. Conversely, William Dudley's speech was followed by expansionary monetary policy actions (around 150 basis points), but would be considered neutral by both traditional measures. The discrepancy between the dictionaries measures and policy actions is consistent with the findings of Hayo and Zahner (2023), who suggest that the noise in dictionary-based indices often overwhelms the intended signal. In contrast, our IT index provides an accurate representation of the underlying policy regime. There was a focus on inflation control in 2003 (as evidenced by the inflation rate and the interest rate moving together) and a more flexible approach in 2013 (reflected by the inflation rate and the interest rate moving in opposite directions).

Having established the thematic distinctions between FIT regime and LIT regime speeches based on the two anecdotal examples, the next step is to systematically quantify the differences using a topic model. Specifically, we use LDAs, an unsupervised machine learning technique among the ones we used Section 4. The advantage of LDA is that it models each speech as a mixture of latent topics. By contrasting differences in the prominence of topics across the IT regimes, it is possible to identify the varying thematic emphases placed under each regime in a systematic manner.[15]

In Figure 7 we present a word cloud representation of the most salient topics associated with the two regimes. Terms colored in red are predominantly associated with FIT regimes, while blue terms are linked primarily to LIT regimes. The thematic contrast is evident: FIT regime speeches tend to focus on objectives and targets, using terms such as *inflation*, *unemployment*, and *expectations*, reflecting a rule-based framework to inflation targeting. In contrast, LIT regime speeches cover a broader mandate, frequently addressing topics related to *banking*, *supervision*, and other financial sectors such as the *credit market*. The word cloud bolsters our confidence in the ability of IT index to effectively capture communication patterns of different inflation-targeting regimes.

[15]For a detailed description of the methodology, see Appendix A.A6.

Figure 7 : FIT/LIT specific wordcloud

Note: The word cloud illustrates the topic-word distributions (ϕ-distribution) from an LDA analysis based on Fed speeches. Red terms are associated primarily with FIT speeches, while blue terms are associated primarily with to LIT speeches. For further details, see Appendix A.A6.

5.4. Does IT communication affect expectations?

Next, we test whether regime changes are able to drive changes in market expectations. The following two hypothesis guide our analysis.

First, as Donald Kohn highights above, the Fed's communication may deliberately be strategically designed to address deviations in market expectations. Supporting this notion, Bikbov and Chernov (2013) demonstrate that monetary policy shocks have a stronger effect on short-term inflation rates in FIT (or "active") regimes. Alternatively, as suggested by Romer and Romer (2002) and Romer and Romer (2004a), changes in market expectations could also stem from shifts in policymakers' underlying beliefs about the economy's structure and the potential effectiveness of monetary policy, thereby driving adjustments in their policy stance. Regardless of the interpretation, we expect movements in IT to move inflation expectations. Since the effectiveness of such communication depends on whether the prevailing inflation rate is above or below target, we expect the impact to be moderated by the current inflation environment. Specifically we expect the effect of variation in IT to be stronger during periods when the Fed deviates further its inflation target.

Second, we expect that expectations of the Fed's policy instrument, the real interest rate, to respond in accordance with its communication signals. Specifically, we expect short-term interest rate, which are more sensitive to conventional mon-

etary policy communication shocks (e.g. Cieslak and Schrimpf, 2019), to exhibit a more strongly reaction relative to long-term rates, thereby leading to differential effects along the yield curve (e.g. Bikbov and Chernov, 2013).
To operationalize our hypotheses, we employ the empirical framework adapted from Bae et al. (2012), as follows:

$$(6)\ \ \Delta E_t[Y_{t+n}] = \alpha + \beta_1(\pi_t - 2\%) + \beta_2 IT_{t-1} + \beta_3\,(\pi_t - 2\%) \times IT_{t-1} + \beta_4 X_t + \epsilon_t$$

$E_t[Y_{t+1}]$ represents inflation- and interest rate expectations at different horizons, $\pi - 2\%$ is the deviation of the annual inflation rate from the 2% target, and IT is in the inflation regime targeting index. We use IT in lags to establish temporal precedence, which allows us to determine whether shifts in the regime *cause* subsequent changes in expectations. In order to account for variation in communication unrelated to our IT index, we control for uncertain language (e.g. Baker et al., 2016), positive and negative sentiment (e.g. Loughran and McDonald, 2011) and hawkish versus dovish language (e.g. Bennani and Neuenkirch, 2017) in our speeches. To control for press-conference forward guidance, we include Swanson's (2021) forward guidance shocks. Finally, we control for changes in the inflation risk premium and the real risk premium. The data sources and the descriptive statistics of the all covariates can be found in Table A10 and Table A11.

Table 6: Regression Results: Expectations

	Dependent variable (in Δ):							
	$E_t[\pi_{1y}]$	$E_t[\pi_{2y}]$	$E_t[\pi_{10y}]$	$E_t[r_{1m}]$	$E_t[r_{1y}]$	$E_t[r_{10y}]$	$E_t[\pi_{1y}^{Mich}]$	$E_t[\pi_{5y5y}]$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$(\pi - 2\%)$	0.01	0.01	0.002	−0.03	−0.03	−0.001	0.02	0.003
	(0.02)	(0.01)	(0.005)	(0.08)	(0.04)	(0.01)	(0.02)	(0.01)
IT_{t-1}	−0.02	−0.01	−0.002	0.08	0.04	0.003	0.01	−0.01
	(0.03)	(0.01)	(0.01)	(0.09)	(0.04)	(0.01)	(0.02)	(0.01)
$(\pi - 2\%) \times IT_{t-1}$	0.04**	0.02**	0.01**	−0.11*	−0.06**	0.001	0.02	0.01
	(0.02)	(0.01)	(0.004)	(0.06)	(0.03)	(0.005)	(0.02)	(0.01)
Constant	0.07	0.03	−0.002	−0.37	−0.25*	−0.04*	−0.01	−0.01
	(0.08)	(0.05)	(0.02)	(0.29)	(0.14)	(0.02)	(0.08)	(0.04)
Observations	240	240	240	240	240	240	240	204
R^2	0.07	0.11	0.51	0.08	0.39	0.74	0.03	0.13
Adjusted R^2	0.03	0.08	0.49	0.04	0.36	0.73	−0.003	0.09

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. IT is defined in **??** and measured in standard-deviations from its historical mean. $(\pi - 2\%)$ constitutes the annual CPI inflation rate deviation from 2%. All regressions include controls for forward guidance shocks (Swanson, 2021), uncertainty terms (e.g. Baker et al., 2016), sentiment index (e.g. Loughran and McDonald, 2011), hawkish/dovish Language (e.g. Bennani and Neuenkirch, 2017), the inflation risk premium, and Real Risk Premium. Full regression table can be found in **??**

The results are presented in Table A13. Let's first focus on financial market inflation expectations in columns 1-3. First, in accordance with our first hypothesis,

we find that the effect of changes in the IT index on expectations depends on the prevailing level of inflation. Specifically, at levels close to the target ($\pi = 2\%$), the effect is insignificant, whereas it becomes highly significant at elevated levels. The interaction term indicates that an increase in IT brings expectations closer to target. Specifically, when inflation is below target, an increase in IT (reflecting communication that closer to a FIT regime) leads to heightened inflation expectations, whereas it lowers expectations when inflation exceeds the target. The magnitude of this effect is noteworthy: at an inflation rate of 4%, a one standard deviation increase in IT lowers inflation expectations by more around one-fifth of a standard deviation (see Table A11). Notably, this effect appears to be specific to financial market participants. In column 7, where we regress inflation expectations from the Michigan survey on the IT index, we find no significant effect.

Second, we find that interest rate expectations react accordingly. In periods of low inflation, a an increase in IT lowers real interest rate expectations, while the opposite is true in periods of high inflation. Again, the magnitude is noteworthy: under the same conditions as above, a one standard deviation increase in IT increases interest rate expectations by more than 20 basis points or around one-fourth of a standard deviation. Such a response magnitude is notable, particularly given that we control for contemporaneous forward guidance shocks within the same time period. This suggests that even when market participants have access to comprehensive forward guidance, nuanced changes in communication frameworks in inter-meeting communication can still elicit marked revisions in expectations.

Finally, in support of our second hypothesis and in line with the literature (e.g. Cieslak and Schrimpf, 2019; Bikbov and Chernov, 2013), we observe that expectations at the shorter end of the spectrum are more affected than those at the longer end for both inflation and interest rate expectations. To further test this, we include the 5-year-5-year forward inflation expectation rate in column 8. In line with the previous findings, there is no significant effect on the forward inflation expectation rate, which reinforces the robustness of our results.

In order to test whether our results in general are robust to the use of alternative measurements, we substitute the IT index with the Bennani and Neuenkirch's (2017) Hawish-Dovish index and Loughran and McDonald's (2011) Sentiment index. The results can be found in Table A14, reveal insignificant effects across the board.

In summary, our findings corroborate those of previous literature, which emphasise the role of central bank communication in managing market expectations beyond traditional policy rate adjustments. (e.g. Bae et al., 2012; Cieslak and Schrimpf, 2019; Bikbov and Chernov, 2013). Specifically, we find that inter-meeting communication which signals a commitment to inflation targeting causes financial market participants to realign their expectations towards the inflation target. Our results indicate that this effect is highly context-specific, with larger

deviations from the target generating stronger responses. Additionally, we show that the impact of our communication index is most pronounced for expectations over the short-term.

5.5. Inflation targeting communication and the Taylor Rule

Finally, we test whether variations in inflation regimes, captured through our IT regime index, serve as leading indicators for deviations from conventional rule-based monetary policy. Specifically, we assess whether shifts in the Fed's communication predicts future policy responses, as implied by deviations from a standard Taylor rule. Our hypothesis builds on the previous finding, suggesting that Fed communication which aligns more closely with that of FIT regimes, markets interpret it as signaling a stronger emphasis on stabilizing inflation around the target. Conversely, periods with language that resembles LIT regimes is perceived as a signal of increased policy flexibility. Consequently, we test whether the IT index moderates the inflation response parameter in the Fed's rule based monetary policy.

The empirical identification strategy to assess our hypothesis, closely follows the previous specification. We estimate the following augmented Taylor rule:

$$i_t = \alpha + \beta_1(\pi_t - 2\%) + \beta_2 IT_t + \beta_3 (\pi_t - 2\%) \times IT_t + \epsilon_t \tag{7}$$

where i represents the Wu and Xia (2016) shadow rate, $\pi - 2\%$ is the deviation of the inflation rate from the 2% target, and IT is our lagged index. Our hypothesis (stronger inflation targeting regime amplifies the sensitivity of the policy rate to inflation deviations) is captured by the interaction term $(\pi - 2) \times IT$. Thus, we expect the coefficient of interest $\beta_3 > 0$, implying that an increase in IT (closer to FIT regimes) leads to a stronger rule-based response to inflation deviations.

We control for the Fed's dual mandate by including the unemployment rate and the output gap. The data sources, as well as information on the transformation and the descriptive statistics of the all covariates can be found in Table A10 and Table A11. The results of the regression analysis are presented in Table 7.

We find the following: While the Taylor principle cannot be rejected across specifications, once we include our IT index in column 2, the inflation response coefficient is highly dependent on the inflation regime, i.e. the IT index. When the IT index is at its historical mean (i.e., $\bar{IT} \approx 0$), such as during the Greenspan era, the inflation response is approximately 1.3–consistent with findings from Bae et al. (2012)[16].

However, a one standard deviation increase in the IT index raises the inflation response by 0.45. Conversely, a one standard deviation decrease in IT lowers the inflation response below one, indicating a potential violation of the Taylor principle. During the financial crisis ($IT \approx -2.9$) and the early COVID-19 period

[16] Bae et al. (2012) estimate a parameter of 1.32 for the period up to 1997-2005, which would respond to the average estimate during the same time period in our case.

Table 7: IT Taylor Rule Regression Table

	Dependent variable:			
	Wu-Xia Shadow Rate			
	(1)	(2)	(3)	(4)
$(\pi - 2\%)$	1.07^{***}	1.33^{***}	0.98^{***}	0.84^{***}
	(0.20)	(0.22)	(0.15)	(0.16)
Unemp. Rate			-1.09^{***}	-1.02^{***}
			(0.10)	(0.12)
Output Gap				0.37^{*}
				(0.19)
IT_{t-1}		0.13	-0.84^{***}	0.40
		(0.23)	(0.18)	(0.76)
$IT_{t-1} \times (\pi - 2\%)$		0.46^{**}	0.62^{***}	0.81^{***}
		(0.19)	(0.12)	(0.17)
$IT_{t-1} \times$ *Unemp. Rate*				-0.22^{*}
				(0.13)
$IT_{t-1} \times$ *Output Gap*				-0.41
				(0.26)
Constant	1.24^{***}	1.10^{***}	7.43^{***}	6.97^{***}
	(0.23)	(0.23)	(0.61)	(0.74)
Observations	85	84	84	84
R^2	0.25	0.32	0.72	0.74
Adjusted R^2	0.24	0.29	0.71	0.72

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. Wu-Xia Shadow Rate is from Wu and Xia (2016). *IT* is defined in **??** and measured in standard-deviations from its historical mean. $(\pi - 2\%)$ constitutes the annual CPI inflation rate deviation from 2%. Output Gap is measured as the Cyclical Component of HP Filtered Real GDP Series.

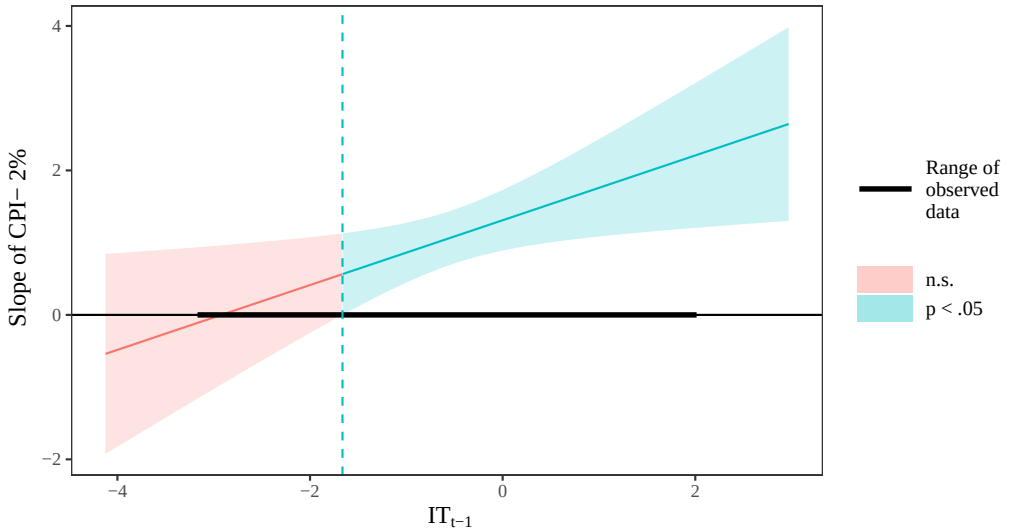
($IT \approx -2.2$), the inflation response parameter fell to near-zero, highlighting the discretionary nature of policy in these periods. The variance in the inflation parameter across regimes is consistent with findings of Benati and Goodhart (2010, Figure 14), who find estimates ranging from zero to over three.

Figure 8 illustrates the dynamic relationship between IT and the inflation response alongside Johnson-Neyman 95% level confidence. The graph confirms that high levels of IT are associated with a considerable and significant inflation responses, while low levels render the response insignificant.

To validate our results, we conduct a series of robustness checks. First, as shown in columns 3 and 4, we find that our findings are robust to the inclusion of additional business cycle indicators such as the unemployment rate and output gap. Notably, the interaction terms between IT and these variables are not significant, reinforcing the centrality of inflation-targeting communication in driving the results.

Next, we address potential concerns about omitted variables, measurement error

Figure 8 : Interaction Effect of Inflation Response and *IT*



Note: Graph is based on the second regression in Table 10. The Johnson-Neyman confidence interval are at the 95% significance level. The distribution of uncertainty is shown by the thin black line.

and variable selection. We test alternative specifications which we report in Table A15. Our results are robust against using the shadow short rate by Krippner (2020), controlling for speaker-specific variation (Hayo and Zahner, 2023)[17], and controlling for dictionary approaches as well as the Forward Guidance Shocks discussed in the previous subsection. Additionally, we re-estimate the *IT* index using an expanding window approach to ensure that results are not driven by future information affecting *IT* classifications.[18] Our results are quantitatively, and qualitatively the same.

To summarize our results provide compelling evidence that shifts in the Fed's inflation regime have significant implications for future policy actions and that our communication based index is a leading indicator for such regime shifts. Specifically, communication that more closely aligns with an inflation-targeting regime strengthens the inflation response coefficient, indicating a tighter adherence to rule-based policy during such periods. These findings underscore the causal role of central bank communication not just shaping market expectations, but guiding

[17] In order to control for speaker fixed effects, we recompute the IT index for each speech, regress the resulting IT index on speaker fixed dummies, and then use the residuals of that regression as our IT index.

[18] We use an expanding window to eliminate concerns about data leakage from future periods affecting our *IT* index. For instance, when classifying the aforementioned Ben Bernanke's 2005 speech at the Finance Committee Luncheon, we restrict the information set to speeches from 2004 and earlier. This ensures that the *IT* index for each period is based solely on data available up to that point.

subsequent 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 –, the provision of central bank specific embeddings and the development of an *inflation-targeting* regime indicator for the Fed.

First, we collect a text-corpus that is unparalleled in size (more than 20.000 speeches) and diversity (more than 100 central banks) within this literature, as both is necessary to train such a language model sufficiently. We show that our corpus offers crucial advantages over conventional corpora used in the existing literature. Second, we introduce embeddings, a class of novel machine learning algorithms from computational linguistics to quantify text. These language models generate meaningfully multidimensional vector representations for words and documents (speeches). Third, we provide high quality text-representations for central bank communication by training and evaluating multiple of these algorithms on our central bank communication corpus. The algorithm with the highest predictive power is able to generate both multidimensional representation for each word and each speech. We show that these embeddings outperform existing language models and conventional dictionaries in predicting monetary policy shocks. Finally, using the best performing embeddings, we demonstrate the broad applicability of embeddings by approximating the Fed's inflation targeting regime from its communication. Specifically, we develop a leading index that tracks deviations in the Fed's communication towards inflation targeting. Our findings indicate these deviations shift market expectations and impact monetary policy actions, leading to a substantially reduction the inflation response parameter in an estimated Taylor-Rule.

Central bank specific embeddings have important implications for policymakers and central bankers, allowing for more nuanced evaluations of their communication. We view this paper to be just a first step towards answering many exciting questions, including developing superior measures for sentiment and uncertainty, modeling institutional differences, and enhancing 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.

Aruoba, S. B., & Drechsel, T. (2022). Identifying monetary policy shocks: A natural language approach.

Ash, E., & Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics*, *15*(1), 659–688.

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*.

Bae, J., Kim, C.-J., & Kim, D. H. (2012). The evolution of the monetary policy regimes in the us. *Empirical Economics*, *43*, 617–649.

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

Benati, L., & Goodhart, C. (2010). Monetary policy regimes and economic performance: The historical record, 1979–2008. *Handbook of monetary economics* (pp. 1159–1236). Elsevier.

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

Bennani, H., & Neuenkirch, M. (2017). The (home) bias of european central bankers: New evidence based on speeches. *Applied Economics*, *49*(11), 1114–1131. https://doi.org/10.1080/00036846.2016.1210782

Bernanke, B. S., & Kuttner, K. N. (2005). What explains the stock market's reaction to federal reserve policy? *The Journal of finance*, *60*(3), 1221–1257.

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

Bikbov, R., & Chernov, M. (2013). Monetary policy regimes and the term structure of interest rates. *Journal of Econometrics*, *174*(1), 27–43.

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

Campiglio, E., Deyris, J., Romelli, D., & Scalisi, G. (2023). Warning words in a warming world: Central bank communication and climate change. *Global Research Alliance for Sustainable Finance and Investment*.

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.

Friedman, M., & Schwartz, A. J. (1963). A monetary history of the united states, 1867-1960.

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

Goutsmedt, A., Sergi, F., Claveau, F., & Fontan, C. (2023). The different paths of central bank scientization: The case of the bank of england.

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. (2017). 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., & 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

Hayo, B., & Zahner, J. (2023). What is that noise? analysing sentiment-based variation in central bank communication. *Economics Letters*, *222*, 110962.

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

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews, 168–177.

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

Kanelis, D., & Siklos, P. L. (2022). Emotion in euro area monetary policy communication and bond yields: The draghi era. *Available at SSRN 4301590*.

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.

Krippner, L. (2020). A note of caution on shadow rate estimates. *Journal of Money, Credit and Banking*, *52*(4), 951–962.

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.

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.

Mishkin, F. S. (1999). International experiences with different monetary policy regimes). any views expressed in this paper are those of the author only and not those of columbia university or the national bureau of economic research. *Journal of monetary economics*, *43*(3), 579–605.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, *29*(3), 436–465.

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

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

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

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

Romer, C. D., & Romer, D. H. (2002). The evolution of economic understanding and postwar stabilization policy.

Romer, C. D., & Romer, D. H. (2004a). Choosing the federal reserve chair: Lessons from history. *Journal of Economic Perspectives*, *18*(1), 129–162.

Romer, C. D., & Romer, D. H. (2004b). A new measure of monetary shocks: Derivation and implications. *American economic review*, *94*(4), 1055–1084.

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

Sims, C. A., & Zha, T. (2006). Were there regime switches in u.s. monetary policy? *American Economic Review*, *96*(1), 54–81. https://doi.org/10.1257/000282806776157678

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. (2021). Financial Markets and Dissent in the ECB's Governing Council. *European Economic Review*, *139*, 103848.

Tillmann, P. (2023). 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. (2024). 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*.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320.

Appendix

A1. Corpus Statistics

Table A1: Sources for the Text Corpus

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

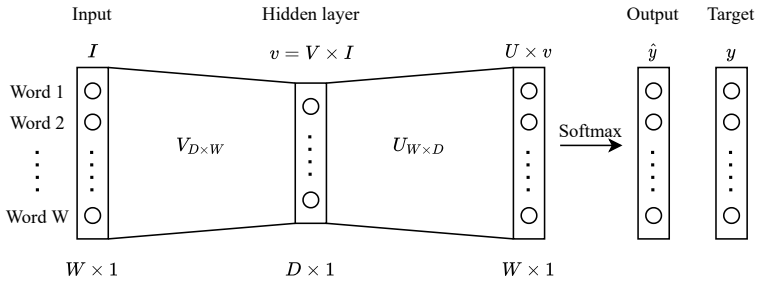
Note: The table summarizes the number of documents (n) by sources in the our text corpus.

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 A1.

Figure A1 : 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.

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:

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

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 A2, 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.A2.

Table A2: 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).

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).[19] The hyperparameters are summarized in the following table:

Table A3: 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

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

A3. Rhetorical Stability

Table A4: Robustness: Rhetoric Stability

	Dependent variable:							
	'US Federal Reserve'							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
'Wikipedia Random'	0.09***							
	(0.004)							
'Wikipedia Monetary Policy'		0.26***						
		(0.003)						
'European Central Bank'			0.66***					
			(0.002)					
'Bank of England'				0.71***				
				(0.002)				
'Bank of Japan'					0.42***			
					(0.002)			
'Bank of Canada'						0.46***		
						(0.002)		
'Sveriges Riksbank'							0.50***	
							(0.002)	
'Reserve Bank of India'								0.57***
								(0.002)
Keyword control	No	No	No	No	No	No	No	No
Observations	104,267	105,109	116,645	110,231	110,314	107,280	120,128	109,352
R^2	0.01	0.07	0.45	0.45	0.32	0.33	0.38	0.34

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 dependent and independent variable are relative term frequencies as defined in 3.3.

A4. Intrinsic Evaluation

Table A5: Intrinsic Evaluation 1: Similarity of key monetary policy terms.

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

Note: The table shows the most similar terms to the words *inflation*, *unemployment* and *output* according to the cosine distance of the underlying word embeddings as defined by Equation (4). The langauge model is the Doc2Vec model chosen in Section 4. The underscore is used to highlight collocations as described in Section 3.

Table A6: 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 (4). The underscore is used to highlight collocations as described in Section 3.

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 A7: Additional Intrinsic Evaluation: Homonym across language models.

Doc2Vec	GloVe6B	Word2Vec GoogleNews
ngfs	afforestation	greener
climate-related	forestation	sustainability
green_finance	beautification	greened
climate_change	reforestation	green
paris_agreement	canker	Greening
climate-	jagielka	greenest
greener	citrus	composting
frank_elderson	punxsutawney	revitalization
greenhouse	gartside	Greenest
climate_change	colonizing	Greener

Note: The table shows for the Doc2Vec and the two genereal corpus models the ten most similar words to the word *"greening"* according to the cosine distance of the underlying word embeddings as defined by Equation (4). The underscore is used to highlight collocations as described in Section 3.

A5. Extrinsic Evaluation II - Summary Statistics

Table A8: Summary Statistics Evaluation

Statistic	N	Mean	St. Dev.	Pctl(25)	Median	Pctl(75)
Fed						
Shadow Rate (Wu-Xia, 2016)	315	1.97	2.66	−0.19	1.55	4.92
Inflation Rate	315	2.17	1.14	1.53	2.14	2.88
Unemployment Rate	315	5.79	1.84	4.50	5.30	6.20
Production Growth	315	0.40	2.02	0.01	0.66	1.17
ECB						
Shadow Rate (Wu-Xia, 2016)	291	0.25	3.77	−2.29	1.26	3.41
Inflation Rate	291	1.63	0.92	1.00	1.80	2.30
Unemployment Rate	291	9.43	1.32	8.37	9.23	10.40
Production Growth	291	0.29	2.86	−0.38	0.47	1.26

A6. Monetary Policy Regimes - LDA analysis

We implement our LDAs model following a broad strand in the literatyre on central bank communication (e.g. Hansen and McMahon, 2016; Hansen, McMahon, and Prat, 2017; Wischnewsky et al., 2021; Schmeling and Wagner, 2019).
First, we preprocess the speeches by removing standard stopwords as is standard practice in the literature. We then apply the LDA model using the TOPICMODELS package in R to construct a 25-topic model, maintaining default hyperparameters to ensure simplicity and comparability across studies. The only critical parameter in this context is the number of topics, which determines the level of topic granularity. We choose 25 topics to strike a balance between granularity and interpretability, drawing from an average value used in previous studies, which show significant variation: Hansen, McMahon, and Prat (2017) employ 50 topics, Hansen and McMahon (2016) utilize 15 topics, and Schmeling and Wagner (2019) use a range of 2 to 8 topics. In unreported robustness checks, we confirm that varying the number of topics does not materially alter our findings.
Each speech is then assigned to the topic with the highest probability. For example, consider William Dudley's opening remarks at the Transatlantic Economic Interdependence and Policy Challenges Conference from Section 5. The LDA model assigns this speech a probability of 39% for Topic 8 and 16% for Topic 16, with all remaining topics having probabilities below 2%. Consequently, the speech is classified under Topic 8.
To identify which topics are predominantly featured in LIT and FIT speeches, we estimate a simple regression of our IT index on dummies for each topic, where each dummy equals 1 if the speech is classified under that topic, and 0 otherwise:

$$IT_i = \beta_1 \, Topic \; 1_i + \beta_2 \, Topic \; 2_i + ... + \beta_{25} \, Topic \; 25_i$$

We run two separate regressions: first, with IT as a continuous variable and second, using a binary variable that equals 1 if $IT > 0$ (indicating an FIT regime) and 0 otherwise. The regression results are presented in Table A9. Based on the coefficients, we identify Topics 1, 4, and 25 as LIT topics, and Topics 2, 7, and 15 as FIT topics. The former group consistently shows the lowest coefficients, while the latter exhibits the highest coefficients across both regressions.
Finally, for each of the six topics, we extract the most probable terms from the topic-word distributions (which represents the probability of a word appearing in a given topic – in LDA terminology: ϕ-distribution). The word cloud visualizations (Figure 7) reflect these topic-specific term probabilities, providing a depiction of the distinct topic structures in FIT and LIT communication.

Table A9: Regression Results: LDA topics

	Dependent variable:	
	IT	$IT > \bar{IT}$
	OLS	*logistic*
	(1)	(2)
Topic 1	0.01***	−1.47***
	(0.002)	(0.45)
Topic 2	0.03***	3.67***
	(0.002)	(0.56)
Topic 3	0.02***	2.56***
	(0.003)	(0.56)
Topic 4	−0.002	−0.40
	(0.002)	(0.56)
Topic 5	0.02***	2.64***
	(0.003)	(0.56)
Topic 6	0.004	0.54
	(0.003)	(0.54)
Topic 7	0.02***	3.20***
	(0.003)	(0.63)
Topic 8	0.01***	1.75***
	(0.002)	(0.51)
Topic 9	0.01**	0.94*
	(0.003)	(0.55)
Topic 10	0.01**	0.74
	(0.003)	(0.54)
Topic 11	0.001	0.25
	(0.002)	(0.50)
Topic 12	0.01***	0.87*
	(0.002)	(0.49)
Topic 13	0.02***	2.53***
	(0.003)	(0.53)
Topic 14	0.01**	1.22**
	(0.003)	(0.52)
Topic 15	0.03***	4.13***
	(0.003)	(0.75)
Topic 16	0.003	0.59
	(0.002)	(0.49)
Topic 17	0.004	0.77
	(0.003)	(0.53)
Topic 18	0.002	−0.22
	(0.002)	(0.55)
Topic 19	0.02***	2.55***
	(0.002)	(0.52)
Topic 20	0.02***	2.94***
	(0.002)	(0.54)
Topic 21	0.02***	1.78***
	(0.003)	(0.54)
Topic 22	0.01***	1.61***
	(0.003)	(0.53)
Topic 23	0.01***	1.91***
	(0.003)	(0.54)
Topic 24	0.02***	2.51***
	(0.002)	(0.48)
Topic 25	−0.002	−0.06
	(0.002)	(0.52)
Observations	1,850	1,850

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.

A7. Monetary Policy Regimes - Summary Statistics

Table A10: MP Frameworks – Data Sources

	Source	Transformation
	Main Regression 1: MP Targets	
MP Frameworks	monetaryframeworks.org	–
Unemployment Rate	(1) World Bank (SL.UEM.TOTL.ZS) (2) FRED (LRHUTTTTEZA156S)	Total Unemployment Rate in %
Inflation Rate	(1) World Bank (FP.CPI.TOTL.ZG) (2) FRED (FPCPITOTLZGEMU)	CPI in annual % change
GDP	(1) World Bank (NY.GDP.PCAP.CD) (2) FRED (NYGDPPCAPCDEMU)	Natural logarithm of GDP per capita in current US$
	Main Regression 3: Taylor Rule	
Wu-Xia Shadow Rate	Federal Reserve of Atlanta	Quarterly average.
Inflation Rate	FRED (CPIAUCSL)	CPI in annual % change; Quarterly average.
Output Gap	FRED (GDPC1)	Cyclical Component of HP Filtered Series ($\lambda = 1600$)
Unemployment Rate	FRED (UNRATE)	Quarterly average of Total Unemployment Rate.
Shadow Short Rate	Reserve Bank of New Zealand	Quarterly average.

Table A11: MP Frameworks – Summary Statistics

	N	Mean	St. Dev.	Pctl(25)	Median	Pctl(75)
Main Regression 1: MP Targets						
Cobham's (2021) MP targets						
ITs	975	0.62	0.49	0	1	1
– FIT	975	0.23	0.42	0	0	0
– LIT	975	0.36	0.48	0	0	1
– FCIT	975	0.01	0.08	0	0	0
– LCIT	975	0.03	0.17	0	0	0
LSD	975	0.24	0.43	0	0	0
ERTs	975	0.10	0.31	0	0	0
WSD	975	0.01	0.11	0	0	0
ERfix	975	0.02	0.12	0	0	0
NoNat	975	0.001	0.03	0	0	0
MixedTs	975	0.01	0.08	0	0	0
RBNZ Controls						
Similarity to RBNZ	957	0.39	0.10	0.32	0.38	0.44
Inflation Δ to NZL	957	1.48	4.25	−0.95	0.65	2.76
log(GDP) Δ to NZL	957	−0.74	1.47	−1.90	−0.25	0.42
Unemployment Rate Δ to NZ	957	1.40	5.18	−1.91	0.41	3.69
Fed Controls						
Similarity to Fed	957	0.48	0.09	0.41	0.49	0.55
Inflation Δ to US	957	1.19	4.25	−1.28	0.35	2.73
log(GDP) Δ to US	957	−1.28	1.42	−2.37	−0.83	−0.14
Unemployment Rate Δ to US	957	1.30	5.03	−1.96	0.32	3.35
ECB Controls						
Similarity to ECB	957	0.52	0.11	0.44	0.51	0.58
Inflation Δ to EA	957	−0.88	5.45	−3.67	−0.91	1.62
log(GDP) Δ to EA	957	0.22	1.90	−1.02	0.15	1.79
Unemployment Rate Δ to EA	957	−0.80	4.90	−3.99	−1.68	1.14
Main Regression 2: Expectations						
RND	241	0.01	1.01	−0.71	0.03	0.70
Inflation Rate - 2%	241	0.17	1.24	−0.46	0.13	1.01
$\Delta E_t[\pi_{1y}]$	241	−0.01	0.39	−0.21	0.03	0.20
$\Delta E_t[\pi_{2y}]$	241	−0.01	0.22	−0.14	0.01	0.12
$\Delta E_t[\pi_{10y}]$	241	−0.01	0.11	−0.08	−0.0001	0.06
$\Delta E_t[\pi^M_{1y}ich]$	241	−0.002	0.34	−0.10	0.00	0.10
$\Delta E_t[\pi_{5y5y}]$	204	−0.002	0.19	−0.07	0.01	0.07
$\Delta E_t[r_{1m}]$	241	−0.01	1.33	−0.65	−0.07	0.70
$\Delta E_t[r_{1y}]$	241	−0.01	0.82	−0.42	−0.02	0.33
$\Delta E_t[r_{10y}]$	241	−0.01	0.19	−0.12	−0.01	0.10
Δ Infl. Risk Premium	241	−0.0002	0.05	−0.02	0.003	0.02
Δ Real Risk Premium	241	−0.0004	0.05	−0.03	−0.01	0.02
Sentiment (LM, 2016)	241	0.17	0.15	0.08	0.17	0.27
Uncertainty (BBD, 2016)	241	1.61	1.47	0.69	1.23	2.00
Hawk-Dove (BN, 2017)	241	0.44	0.25	0.30	0.47	0.61
FG Shocks (Swanson, 2021)	241	−3.85	80.15	−17.62	0.00	14.78
Main Regression 3: Taylor Rule						
RND	85	−0.0004	0.99	−0.55	0.11	0.60
RND (Speaker Fixed Effects)	85	0.04	0.97	−0.63	0.18	0.68
RND (Expanding Window)	85	−0.01	1.02	−0.80	0.15	0.56
Shadow Rate (Wu, Xia, 2016)	85	1.42	2.46	−0.50	1.13	2.83
Shadow Short Rate (Krippner, 2021)	85	1.23	2.67	−0.17	1.01	2.98
Inflation Rate - 2%	85	0.17	1.16	−0.40	0.11	0.92
Unemployment Rate	85	5.78	1.81	4.40	5.30	6.67
Output Gap	85	0.16	1.15	−0.48	0.16	0.95
Uncertainty (BBD, 2016)	85	1.52	0.81	0.95	1.33	1.85
Hawk-Dove (BN, 2017)	85	0.46	0.15	0.38	0.46	0.55
Sentiment (LM, 2016)	85	0.16	0.10	0.10	0.16	0.22
FG Shocks (Swanson, 2021)	82	−2.40	63.46	−40.07	3.01	35.31

A8. Monetary Policy Regimes - Extended Regression Tables

Table A12: Regression results: Monetary Policy Framework classification

	Dependent Variable: Similarity towards i								
$i =$	RBNZ			Fed			ECB		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
ITs	0.11***	0.09***		0.12***	0.09***		0.15***	0.15***	
	(0.02)	(0.02)		(0.02)	(0.02)		(0.02)	(0.03)	
– FIT			0.14***			0.12***			0.10***
			(0.02)			(0.02)			(0.02)
– LIT			0.09***			0.10***			0.17***
			(0.02)			(0.02)			(0.02)
– FCIT			0.04			0.10**			0.15***
			(0.04)			(0.04)			(0.04)
– LCIT			0.05*			0.03			0.08***
			(0.03)			(0.03)			(0.03)
NoNat	−0.10	−0.10	−0.09	−0.12	−0.13	−0.13	−0.03	−0.03	−0.03
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.10)	(0.09)	0.09)
ERTs	0.05*	0.04	0.05*	0.04*	0.02	0.03	0.07***	0.07***	0.06**
	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
LSD	0.05**	0.06**	0.07***	0.05**	0.05**	0.05**	0.04	0.06**	0.05**
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)
MixedTs	0.02	0.001	0.01	0.04	0.01	0.02	0.12***	0.10**	0.10**
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)
WSD	0.13***	0.12***	0.14***	0.14***	0.12***	0.13***	0.14***	0.14***	0.14***
	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)
Constant	0.34***	0.35***	0.34***	0.40***	0.42***	0.41***	0.37***	0.36***	0.37***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Macro-controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	957	957	957	957	957	957	957	957	957
R^2	0.22	0.24	0.28	0.18	0.19	0.22	0.26	0.32	0.38
Adjusted R^2	0.20	0.21	0.26	0.16	0.17	0.19	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; NoNat = no national framework.

Table A13: Extended Regression Results: Expectations

	Dependent variable (in Δ):							
	$E_t[\pi_{1y}]$	$E_t[\pi_{2y}]$	$E_t[\pi_{10y}]$	$E_t[r_{1m}]$	$E_t[r_{1y}]$	$E_t[r_{10y}]$	$E_t[\pi_{1y}^{Mich}]$	$E_t[\pi_{5y5y}]$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$(\pi - 2\%)$	0.01	0.01	0.002	−0.03	−0.03	−0.001	0.02	0.003
	(0.02)	(0.01)	(0.005)	(0.08)	(0.04)	(0.01)	(0.02)	(0.01)
IT_{t-1}	−0.02	−0.01	−0.002	0.08	0.04	0.003	0.01	−0.01
	(0.03)	(0.01)	(0.01)	(0.09)	(0.04)	(0.01)	(0.02)	(0.01)
$(\pi - 2\%) \times IT_{t-1}$	0.04**	0.02**	0.01**	−0.11*	−0.06**	0.001	0.02	0.01
	(0.02)	(0.01)	(0.004)	(0.06)	(0.03)	(0.005)	(0.02)	(0.01)
Uncertainty	0.01	0.01	0.002	−0.05	−0.02	−0.001	0.02	0.01
	(0.02)	(0.01)	(0.004)	(0.06)	(0.03)	(0.004)	(0.02)	(0.01)
Sent	−0.30	−0.16	−0.04	0.96	0.55	0.02	−0.01	−0.03
	(0.20)	(0.11)	(0.04)	(0.68)	(0.35)	(0.05)	(0.18)	(0.10)
Hawk	−0.10	−0.04	−0.002	0.64	0.41**	0.06**	−0.06	−0.01
	(0.12)	(0.06)	(0.02)	(0.40)	(0.20)	(0.03)	(0.11)	(0.06)
FG Shocks	−0.01	0.01	0.02**	0.08	0.04	0.03***	0.03	0.05***
	(0.03)	(0.02)	(0.01)	(0.11)	(0.06)	(0.01)	(0.03)	(0.02)
Infl.Risk.Premium	1.75***	0.81**	0.04	−6.63***	−11.72***	−0.53***	0.47	−0.47
	(0.60)	(0.33)	(0.12)	(2.03)	(1.03)	(0.15)	(0.54)	(0.28)
Real.Risk.Premium	−0.67	0.57	1.60***	5.28**	8.01***	3.81***	0.30	1.25***
	(0.71)	(0.39)	(0.14)	(2.40)	(1.21)	(0.18)	(0.63)	(0.35)
Constant	0.07	0.03	−0.002	−0.37	−0.25*	−0.04*	−0.01	−0.01
	(0.08)	(0.05)	(0.02)	(0.29)	(0.14)	(0.02)	(0.08)	(0.04)
Observations	240	240	240	240	240	240	240	204
R^2	0.07	0.11	0.51	0.08	0.39	0.74	0.03	0.13
Adjusted R^2	0.03	0.08	0.49	0.04	0.36	0.73	−0.003	0.09

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. *IT* is defined in **??** and measured in standard-deviations from its historical mean. $(\pi - 2\%)$ constitutes the annual CPI inflation rate deviation from 2%. *Hawk*, *Sent* and *Uncertainty* measure sentiment (e.g. Loughran and McDonald, 2011), hawkish/dovish Language (e.g. Bennani and Neuenkirch, 2017) and uncertainty terms (e.g. Baker et al., 2016). FG Shocks are by Swanson (2021). The inflation risk premium and the real risk premium are based on treasury yields, inflation data, inflation swaps, and survey-based measures of inflation expectations (Source: Federal Reserve Bank of Cleveland).

A9. Monetary Policy Regimes - Robustness Test

Table A14: Robustness Test: Expectations

	Dependent variable (in Δ):								
	$E_t[\pi_{1y}]$	$E_t[\pi_{2y}]$	$E_t[\pi_{10y}]$	$E_t[\pi_{1y}]$	$E_t[\pi_{2y}]$	$E_t[\pi_{10y}]$	$E_t[\pi_{1y}]$	$E_t[\pi_{2y}]$	$E_t[\pi_{10y}]$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$(\pi - 2\%) \times IT_{t-1}$	0.04**	0.02**	0.01**						
	(0.02)	(0.01)	(0.004)						
$(\pi - 2\%) \times Hawk_{t-1}$				0.02	0.01	0.01			
				(0.02)	(0.01)	(0.004)			
$(\pi - 2\%) \times Sent_{t-1}$							−0.04*	−0.02*	−0.01
							(0.02)	(0.01)	(0.004)
Observations	240	240	240	241	241	241	241	241	241
R^2	0.07	0.11	0.51	0.05	0.10	0.50	0.05	0.10	0.50
Adjusted R^2	0.03	0.08	0.49	0.01	0.06	0.48	0.02	0.06	0.48

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. IT is defined in Section 5.5.2 and measured in standard-deviations from its historical mean. $Hawk$ and $Sent$ measure sentiment (e.g. Loughran and McDonald, 2011) and hawkish/dovish Language (e.g. Bennani and Neuenkirch, 2017). $(\pi - 2\%)$ constitutes the annual CPI inflation rate deviation from 2%. All regressions include controls for $Hawk$, $Sent$, IT, forward guidance shocks (Swanson, 2021), uncertainty terms (e.g. Baker et al., 2016), the inflation risk premium, and Real Risk Premium. Regressions in column 1-3 to those in Table 9 in the main text.

Table A15: Robustness Test: IT Taylor Rule Regression Table

	Dependent variable:			
	Shadow Short Rate	Wu-Xia Shadow Rate		
Rob. Test:	*(Shadow rate)*	*(Speaker FE)*	*(FG + Dict.)*	*(Rolling Window)*
$(\pi - 2\%)$	0.64^{***}	1.03^{***}	0.84^{***}	1.09^{***}
	(0.17)	(0.17)	(0.17)	(0.18)
Unemp. Rate	-1.26^{***}	-0.80^{***}	-1.01^{***}	-1.02^{***}
	(0.14)	(0.13)	(0.14)	(0.19)
Output Gap	0.37^{*}	0.30	0.44^{**}	0.16
	(0.20)	(0.22)	(0.21)	(0.22)
IT_{t-1}	0.66	-0.40	0.51	1.15
	(0.84)	(0.84)	(0.79)	(1.01)
$IT_{t-1} \times (\pi - 2\%)$	0.69^{***}	0.64^{***}	0.82^{***}	0.71^{***}
	(0.18)	(0.20)	(0.18)	(0.18)
$IT_{t-1} \times Unemp.\ Rate$	-0.26^{*}	-0.01	-0.24^{*}	-0.29
	(0.14)	(0.14)	(0.14)	(0.18)
$IT_{t-1} \times Output\ Gap$	-0.34	0.08	-0.40	-0.12
	(0.28)	(0.25)	(0.28)	(0.23)
Hawk			1.35	
			(1.40)	
Sent			0.29	
			(2.05)	
Uncertainty			-0.01	
			(0.20)	
FG Shocks			-0.001	
			(0.002)	
Constant	8.11^{***}	5.66^{***}	6.22^{***}	6.63^{***}
	(0.81)	(0.79)	(1.31)	(1.04)
Observations	84	84	81	84
R^2	0.74	0.66	0.76	0.70
Adjusted R^2	0.71	0.63	0.72	0.68

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. Wu-Xia Shadow Rate is from Wu and Xia (2016), Shadow Short Rate is from Krippner (2020), IT is defined in **??** and measured in standard-deviations from its historical mean. $(\pi - 2\%)$ constitutes the annual CPI inflation rate deviation from 2%. Output Gap is measured as the Cyclical Component of HP Filtered Real GDP Series. *Hawk*, *Sent* and *Uncertainty* measure sentiment (e.g. Loughran and McDonald, 2011), hawkish/dovish Language (e.g. Bennani and Neuenkirch, 2017) and uncertainty terms (e.g. Baker et al., 2016). FG Shocks are by Swanson (2021).

IMFS WORKING PAPER SERIES

Recent Issues

193 / 2023	Alexander Meyer-Gohde	Numerical Stability Analysis of Linear DSGE Models - Backward Errors, Forward Errors and Condition Numbers
192 / 2023	Otmar Issing	On the importance of Central Bank Watchers
191 / 2023	Anh H. Le	Climate Change and Carbon Policy: A Story of Optimal Green Macroprudential and Capital Flow Management
190 / 2023	Athanasios Orphanides	The Forward Guidance Trap
189 / 2023	Alexander Meyer-Gohde Mary Tzaawa-Krenzler	Sticky information and the Taylor principle
188 / 2023	Daniel Stempel Johannes Zahner	Whose Inflation Rates Matter Most? A DSGE Model and Machine Learning Approach to Monetary Policy in the Euro Area
187 / 2023	Alexander Dück Anh H. Le	Transition Risk Uncertainty and Robust Optimal Monetary Policy
186 / 2023	Gerhard Rösl Franz Seitz	Uncertainty, Politics, and Crises: The Case for Cash
185 / 2023	Andrea Gubitz Karl-Heinz Tödter Gerhard Ziebarth	Zum Problem inflationsbedingter Liquiditätsrestriktionen bei der Immobilienfinanzierung
184 / 2023	Moritz Grebe Sinem Kandemir Peter Tillmann	Uncertainty about the War in Ukraine: Measurement and Effects on the German Business Cycle
183 / 2023	Balint Tatar	Has the Reaction Function of the European Central Bank Changed Over Time?
182 / 2023	Alexander Meyer-Gohde	Solving Linear DSGE Models with Bernoulli Iterations
181 / 2023	Brian Fabo Martina Jančoková Elisabeth Kempf Luboš Pástor	Fifty Shades of QE: Robust Evidence
180 / 2023	Alexander Dück Fabio Verona	Monetary policy rules: model uncertainty meets design limits

179 / 2023	Josefine Quast Maik Wolters	The Federal Reserve's Output Gap: The Unreliability of Real-Time Reliability Tests
178 / 2023	David Finck Peter Tillmann	The Macroeconomic Effects of Global Supply Chain Disruptions
177 / 2022	Gregor Boehl	Ensemble MCMC Sampling for Robust Bayesian Inference
176 / 2022	Michael D. Bauer Carolin Pflueger Adi Sunderam	Perceptions about Monetary Policy
175 / 2022	Alexander Meyer-Gohde Ekaterina Shabalina	Estimation and Forecasting Using Mixed-Frequency DSGE Models
174 / 2022	Alexander Meyer-Gohde Johanna Saecker	Solving linear DSGE models with Newton methods
173 /2022	Helmut Siekmann	Zur Verfassungsmäßigkeit der Veranschlagung Globaler Minderausgaben
172 / 2022	Helmut Siekmann	Inflation, price stability, and monetary policy – on the legality of inflation targeting by the Eurosystem
171 / 2022	Veronika Grimm Lukas Nöh Volker Wieland	Government bond rates and interest expenditures of large euro area member states: A scenario analysis
170 / 2022	Jens Weidmann	A new age of uncertainty? Implications for monetary policy
169 / 2022	Moritz Grebe Peter Tillmann	Household Expectations and Dissent Among Policymakers
168 / 2022	Lena Dräger Michael J. Lamla Damjan Pfajfar	How to Limit the Spillover from an Inflation Surge to Inflation Expectations?
167 / 2022	Gerhard Rösl Franz Seitz	On the Stabilizing Role of Cash for Societies
166 / 2022	Eva Berger Sylwia Bialek Niklas Garnadt Veronika Grimm Lars Other Leonard Salzmann Monika Schnitzer Achim Truger Volker Wieland	A potential sudden stop of energy imports from Russia: Effects on energy security and economic output in Germany and the EU
165 / 2022	Michael D. Bauer Eric T. Swansson	A Reassessment of Monetary Policy Surprises and High-Frequency Identification