Ambiguity-Robust Dictionaries using Contextualized Embeddings

Jamil Civitarese and Sylvan Zheng*

July 28, 2026

Abstract

Keyword dictionaries remain widely used in text analysis for their simplicity, transparency, and replicability. However, sensitivity to keyword selection and inability to account for context introduce measurement error that attenuates estimates and obscures real relationships. We introduce ambiguity-robust dictionaries, a text-as-data method that leverages contextual word embeddings to produce more precise measures while preserving the key advantages of dictionary approaches. Our method generates contextualized embeddings for all keyword instances, then applies constrained fuzzy clustering to determine the degree of each instance's membership in a target cluster defined by a small set of researcher-identified anchor words. This approach allows the construction of ambiguity-robust document level word counts, increasing precision and reducing sensitivity to keyword selection. We demonstrate the method using three applications: UN environmental discourse, ethnic bias in Kenyan judicial politics, and moral language in US Congressional speech. In each application, our method yields substantively more robust and precise estimates than conventional approaches.

Word count: 9,205

*Wilf Family Department of Politics, New York University. Preliminary draft.

1 Introduction

The use of keyword dictionaries to measure latent concepts in text is a foundational technique in computational social science. Despite the rise of more complex models, dictionary-based approaches remain widely used; in recent years, researchers have used keyword dictionaries to measure quantities such as the sentiment of magazine articles about judicial rulings (Bailey et al., 2025), the degree of religiosity in military group communications (Ying, 2026), and the degree of intuition versus evidence-based rhetoric in U.S. congressional speeches (Aroyehun et al., 2025). Dictionary models have important advantages: in addition to their low cost and ease of use, document measures based on simple word counts are transparent, interpretable, and easily replicable. These properties allow empirical research that leverages dictionary measures to uphold values central to rigorous scientific inquiry.

However, the parsimony of traditional dictionary methods creates an important weakness. Dictionary methods are unable to account for context: words are often polysemous, and their meaning can shift dramatically depending on the surrounding text. The analyst interested in measuring the degree to which a politician's speech is concerned with climate change may mistake references to a downtown "business environment" for references to the ecological environment. This creates a dilemma for the researcher when constructing lists of dictionary keywords: the inclusion of potentially ambiguous terms could improve but also deteriorate the measurement quality of the dictionary. Substantive analyses using measures derived from dictionary approaches risk sensitivity to specific inclusion or exclusion of *a priori* reasonable word choices, leaving the conclusions vulnerable to the "garden of forking paths" risk outlined by Gelman and Loken (2013). While supervised learning approaches and even more advanced models (e.g., large language models) resolve this weakness by using more complex statistical modeling, we argue that this introduces a tradeoff against the core advantages of dictionary approaches: cost, transparency, interpretability, and replicability.

This paper introduces ambiguity-robust dictionaries, a methodological innovation

designed to resolve this tradeoff, accounting for contextual variation in word usage while preserving the advantages of dictionary approaches. Our method leverages recent advances in embedding and clustering machine learning techniques to identify and discount keyword usages that are inconsistent with the researcher's target concept. First, we generate contextualized embeddings (Khodak et al., 2018) for each instance of a dictionary word in the corpus. Then, based on a researcher-defined subset of unambiguous "anchor" words we use a constrained fuzzy clustering algorithm to calculate the membership score in the anchor cluster for each contextualized keyword. While traditional approaches would count the number of keyword occurrences within a document, our method sums the 0-1 membership scores within the document to provide a context-regularized word count. The result is a more accurate, de-noised measure that explicitly models out-of-context word usage. This directly addresses the most significant limitation of dictionary measures while maintaining high levels of transparency, interpretability, and replicability.

We illustrate the usage and benefits of our method through three empirical applications. First, we analyze environmental rhetoric in decades of discourse in the UN General Debate. We show that our method successfully identifies and discounts instances of both out-of-context ambiguous keywords such as "environment" as well as fully off-topic keywords such as "trade" or "economy." Then, we demonstrate the substantive stakes. A large literature in international and comparative environmental politics argues that democracies tend to express stronger environmental commitments in multilateral arenas due to domestic accountability pressures, exposure to normative networks, and responsiveness to civil society demands (Neumayer, 2002; Bättig and Bernauer, 2009). We show that this relationship is highly sensitive to word choice when using conventional dictionaries; however, when using our ambiguity-robust approach, the statistical correlation of this relationship is preserved no matter the level of dictionary contamination.

Second, we show that our method protects substantive conclusions against garden

of forking paths critiques by providing dramatic precision and interpretability advantages. We replicate the findings of Choi et al. (2022) on ethnic bias in Kenyan judicial politics, who use dictionaries to find a significant relationship between trust keywords used in judicial decisions and judge-appellant coethnicity. Using a permutation test, we show that this result is vulnerable to dictionary specification when using conventional methods: across a sample of plausible trust dictionaries, we recover statistically significant results in fewer than 20% of draws. However, our ambiguity robust method dramatically reduces this vulnerability by modeling out of context usage — results using contextually aware measure of trust estimate a statistically significant effect in over 80% of dictionary draws. This equates to a precision increase roughly equivalent to increasing the sample size by a factor of six, creating an invaluable tool when data collection is limited. Our method also allows us to investigate which keywords most contribute to the target concept, providing an important layer of interpretability to our findings. We show that judges build trust in coethnic decisions by specifically appealing to evidence and respected societal figures rather than abstract legal rhetoric.

Finally, we demonstrate the ability of our method to generate new insights on large scale datasets by analyzing the relationship between campaign contributions and negative moral language in over 800,000 US congressional speeches from 1980-2022. We document a rise in the prevalence of negative moral language and offer suggestive evidence for a donor selection mechanism that is considerably sharpened by the increased precision of the ambiguity-robust dictionary measure. Legislators who use one standard deviation more negative moral language receive 20% more individual campaign contributions when measured using the ambiguity-robust method, compared to 12% using conventional dictionary counts. Moreover, this relationship has increased sharply over time: by 2020-2022, legislators who use one standard deviation more negative moral language receive 50% more contributions compared to 2000. A within-legislator analysis shows limited evidence that changes in rhetorical style are associated with changes in donations, suggesting that the relationship operates primarily through selection rather

than marginal incentives. Together, these results indicate that the rising prevalence of negative moral language in Congress coincides with a strengthening cross-sectional alignment between moralized rhetoric and donor support.

Our paper makes three contributions to the text-as-data literature. First, we make the case for continued investment in dictionary methods even as large language models and supervised classifiers become more accessible, arguing that dictionaries offer unique advantages in transparency, replicability, and the construction of continuous measures that these alternatives cannot easily match. Second, we develop a general method for constructing context-sensitive dictionary measures that nests conventional word counts as a special case. The method requires only a standard dictionary, a small set of anchor words, and pre-trained word embeddings — inputs that are already available for over 150 languages (Wirsching et al., 2025) — and runs on consumer-grade hardware even for corpora with millions of documents. We provide an R package to facilitate the ease of use of this method. Third, we demonstrate across three substantively distinct applications that the precision gains from our method are not merely theoretical: ambiguity-robust dictionaries recover relationships that conventional counts consistently underestimate or fail to detect completely.

2 Why Dictionaries?

Dictionary methods are still widely used in contemporary political science research. We review articles employing text-as-data dictionary methods published in three leading political science journals — the *American Political Science Review* (APSR), the *American Journal of Political Science* (AJPS), and the *Journal of Politics* (JOP) — between 2022 and 2026 (including articles listed as forthcoming at the time of writing).[1] Across the 33 articles we identified, 18 use dictionary methods as a primary measurement

[1] Articles were identified through a combination of journal table-of-contents browsing and keyword searches for "dictionary" on JSTOR, library catalogues, and the journal websites. The review is intended as illustrative rather than exhaustive, but provides a useful snapshot of methodological practice in top-tier American politics journals.

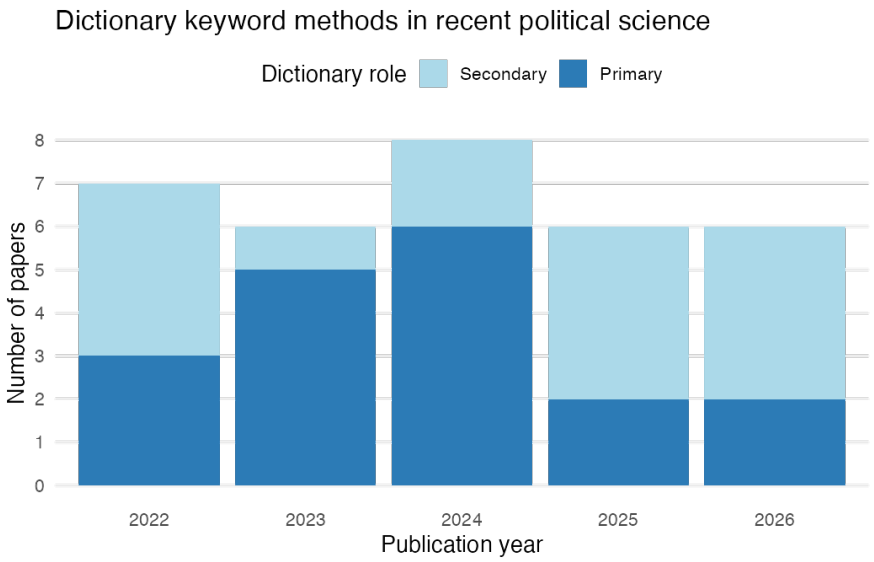


Figure 1: Count of articles published in APSR, AJPS, or JOP using dictionary measures. Articles coded as "primary" use the dictionary measure as the main empirical specification while articles coded as "secondary" use the dictionary measure as a robustness check on an alternative method.

strategy, while 15 use dictionaries in a secondary or supporting role — for instance, as a robustness check on a topic-model or embedding-based result. Figure 1 shows the breakdown of primary v. secondary dictionary use over time (the list of all articles is presented in Appendix A). The applications of dictionary methods span a broad substantive range, including sentiment and "surprise" in news and magazine reporting (Bailey et al., 2025), gender salience in corporate disclosures (Latura and Weeks, 2023) and party manifestos, framing of military intervention in presidential speeches, religiosity in elite political communication (Ying, 2026), intuition- versus evidence-based rhetoric in legislative speech (Aroyehun et al., 2025), and the prevalence of "effective" or "uncivil" language in legislator newsletters and talk-show transcripts (Naftel et al., 2025). In many of these articles, the dictionary-based measure is the central dependent or independent variable on which substantive conclusions rest.

However, the emergence of more sophisticated text-as-data methods, from supervised machine learning to large-scale transformer language models (LLMs), marks a significant development in computational social science. These techniques increasingly allow the analyst to annotate and analyze text documents with higher and higher lev-

els of predictive power. Given these developments, dictionary approaches may appear archaic: why count instances of keywords when the state of the art allows researchers to analyze text using highly performant multi-billion parameter models? We argue that dictionary-based approaches (still) provide a scientifically robust foundation for text analysis, providing a number of advantages even as more advanced modeling approaches continue to gain popularity. The simplicity of dictionaries facilitates replication, transparency, and interpretability, values central to theory driven social science inquiry. Dictionary-based methods also permit the construction of intuitive continuous measures, something that is difficult to do with more complex supervised ML or LLM models (Wu et al., 2023). Finally, dictionary-based methods are also far less costly than many alternative methods, and are more easily portable to resource constrained environments.

For a quantitative measure to be effective for theory testing and development, researchers must be able to articulate and defend the connection between a theoretical construct and its empirical operationalization (Goertz, 2006). Dictionary methods require this act of theoretical specification: the development of a dictionary represents an explicit, transparent, and falsifiable set of claims about which words are indicative of the target concept. For instance, when Baker et al. (2016) constructs a dictionary to measure economic policy uncertainty, the selection of terms like "risk," "deficit," and "regulation" provides a concrete and testable operationalization whose limitations are open to scrutiny and debate. In contrast, "black-box" systems such as large language models are composed of billions of interacting parameters and are challenging to audit directly (Turpin et al., 2023; Lyu et al., 2023). Moreover, high predictive performance does not guarantee valid measurement: a model can achieve high accuracy in label prediction while failing to capture the theoretically relevant concepts. Jankowski and Huber (2023) demonstrate this in the case of populism detection, finding that a supervised classifier learned to differentiate parties based on surface cues—party names, language choice—rather than manifesto content, even assigning high populism scores

when manifestos were replaced with texts from non-populist parties. Without the capacity to examine a model's decision rules, such failures of content validity can go undetected.

Dictionaries also offer important advantages in replicability. For scientific knowledge to be cumulative, findings should be verifiable in addition to interpretable. LLMs, while highly performant, can face reproducibility challenges. These models are stochastic, meaning the same input can produce different outputs on separate runs (Ornstein et al., 2025). Proprietary model drift is also a concern; an analysis performed with a specific version of an LLM may be difficult to replicate later if the underlying model has been altered (Barrie et al., 2024). Open-weight models running on local hardware and hyperparameter selection have both been proposed as solutions to improve the replicability concern of LLMs (Spirling, 2023); however, these choices come with tradeoffs with cost and performance that are not yet well understood. In contrast, a dictionary-based analysis is deterministic and replicable; given the same lexicon and corpus, the results will be identical each time, providing a stable foundation for cumulative science.

Another key advantage of dictionary methods is their ability to generate fine-grained, continuous measures of a concept's salience within a document. Many phenomena of interest to social scientists exist on a continuum; the goal of measurement is often not simply to classify a document but to place it on a meaningful scale. Dictionary methods accomplish this naturally; the proportion or number of dictionary keywords found in a given document yields a direct and continuous measure of a concept's prevalence. Large language models and supervised classifiers, by contrast, are designed for classification. While it is common to repurpose a model's predicted probability as a pseudo-continuous score, this value represents the model's confidence that a document belongs to a given class, not a measure of the intensity of the latent trait within the text (Wu et al., 2023). A document scored at 0.8 is not necessarily more "populist" than one scored at 0.6; the model is simply more confident in its classification of the former. The long tradition of scaling methods in political science, from Wordscores

(Laver et al., 2003) and onward, underscores the discipline's need for true continuous measures.[2]

The practical advantages of dictionaries should also not be overlooked. The computational cost of dictionary methods in the modern computing landscape is near zero, meaning that researchers can iterate freely without worrying about budgetary constraints, slow batch job timing, or unavailable service provision. In particular, these advantages are most noteworthy for junior researchers without access to large research resources and for tasks that involve millions of documents - something not unreasonable in today's big data world. Although LLM performance continues to improve, these improvements largely come with increasing computational scale of frontier models. Inference with LLMs will continue to be expensive, if not in financial terms then always in temporal terms as researchers will be limited by the network transfer times to frontier model providers if not their rate limits. In addition, there are numerous research settings where the application of heavier weight models such as LLMs is inappropriate. Increasing quantities of rich digital trace data are being made available to researchers, but these data are often held within restricted data enclaves where Internet access and compute resources are limited (Feal et al., 2024). Under such conditions, large language models may be impractical. There remains a time and place for dictionary analysis purely for practical purposes, even at today's computational scale.

Our method is part of a broader trend toward hybrid solutions that balance performance with transparency. Latent Semantic Scaling (Watanabe, 2021), for example, uses seed words and corpus-internal LSA embeddings to produce semi-supervised continuous measures, and has seen wide adoption in recent political science research (Zollinger, 2024; Baturo et al., 2025), while Häffner et al. (2023) shows how deep learning can be used to construct more accurate and interpretable domain-specific dictionaries. Ambiguity-robust dictionaries share this semi-supervised logic but use instance-level contextualized embeddings rather than type-level word scores, directly addressing

[2]Note these scaling methods place documents on a bipolar dimension—measuring relative position rather than conceptual prevalence.

Table 1: A Comparison of Approaches to Text-as-Data

Feature	**Traditional Dictionaries**	**Supervised Classifiers**	**LLMs**	**Ambiguity-Robust Dictionaries**
Interpretability	High. Explicit link between words and concept.	Low. Model learns nonlinear feature combinations.	Very Low ("Black box").	High. Explicit link between words and concept.
Replicability	High.	High.	Low. Stochastic outputs.	High.
Content Validity	Moderate. Vulnerable to polysemy and context-insensitivity.	Low. Prone to learning spurious correlations from labels.	Unknown. Opaque nature prevents systematic validation.	High. Explicitly models and discounts out-of-context word usage.
Measure Type	Continuous (e.g., word proportion).	Categorical or Discrete (Low dimensionality).	Categorical or Discrete (Low dimensionality).	Continuous.
Data Requirement	Low. Requires only a word list.	High. Requires large set of human-labeled training documents.	Low (for few-shot prompting).	Low. Requires only a word list.
Generality	Low. Susceptible to ambiguity; Dictionaries need to be tailored to context.	Moderate. Learns context from labels but depends on validation.	Very High.	Moderate to High. Learns context from anchor words.

polysemy at the token level. Table 1 summarizes how ambiguity-robust dictionaries compare to existing approaches.

3 Presentation of the Method

We propose an ambiguity-robust dictionary method for measuring conceptual attention in text corpora. Starting from a corpus $\mathcal{C}$, a dictionary $\mathcal{D}$, a subset of anchor words $\mathcal{A} \subseteq \mathcal{D}$, and a pre-trained embedding space, the procedure:

(1) maps every *instance* of a dictionary word to an *à la carte* (ALC) contextualized vector;

(2) reduces embedding dimensionality with UMAP;

(3) performs constrained fuzzy clustering, forcing anchors into the on-concept cluster; and

(4) aggregates cluster memberships to produce document-level scores.

The remainder of this section formalizes the objects involved and details each step.

3.1 Definitions

- **Corpus.** $\mathcal{C} = \{d_1, \ldots, d_D\}$ with $d_j = (t_{j1}, \ldots, t_{jN_j})$.
- **Dictionary.** $\mathcal{D} \subseteq \mathcal{V}$ (token types); $\mathcal{A} \subseteq \mathcal{D}$ are anchor words.
- **Instances of Words.** $\mathcal{I} = \{(j, n) : t_{jn} \in \mathcal{D}\}$; let $M = |\mathcal{I}|$.
- **Contextualised embeddings.** For each $i = (j, n) \in \mathcal{I}$, $e_i \in \mathbb{R}^P$ computed via ALC with base vectors $v(\cdot)$.

3.2 Steps of the Algorithm

3.2.1 Contextualized Embeddings

The first step of the algorithm is to create contextualized embeddings for each instance of words present in the dictionary. Formally, each observational unit $i = (j, n) \in \mathcal{I}$, corresponding to a token in position n of document j, is mapped to a contextualized embedding vector $e_i \in \mathbb{R}^P$, where P is the dimensionality of the embedding space (typically $P = 300$). These embeddings capture the meaning of the focal word token as used in its local context.

We use the `context` package in R, which implements the *à la carte* (ALC) embedding approach described by Rodriguez et al. (2023) and builds on the methodological framework of Khodak et al. (2018). This approach incorporates the co-occurrence structure of words surrounding each token to produce a vector representation that is sensitive to contextual usage, rather than relying solely on static word embeddings. Given the context of a word instance and a pre-trained embedding model, analysts can generate high-quality vector representations even with very few occurrences of the target term. The result of this step is a matrix $\mathbf{E} \in \mathbb{R}^{M \times P}$, where each row $e_i^\top$ corresponds to the contextualized vector for a token instance $i \in \mathcal{I}$. This matrix serves as the backbone of our analysis: each row plays the role that a dictionary-based word count would in a traditional approach.

For social science applications, Wirsching et al. (2025) provide pretrained GloVe (Pennington et al., 2014) and fastText (Bojanowski et al., 2017) embeddings for 157 languages, along with the projection matrices needed to construct à la carte embeddings. These embeddings are not particular to political science, being trained with Wikipedia data. Since our method depends solely on dictionaries and base embeddings, these resources significantly reduce the barriers to implementing ambiguity-robust dictionaries in multilingual research. In our application sections, we rely on pre-trained GloVe embeddings to ensure comparability with the paper we replicate (Choi et al., 2022).

3.2.2 Dimensionality Reduction (UMAP)

The instance-level embeddings in $\mathbf{E} \in \mathbb{R}^{M \times 300}$, typical length of word embeddings, are informative but high-dimensional, which is both redundant (many dimensions are correlated) and unwieldy for clustering or downstream modeling. To obtain a compact representation that preserves salient semantic structure, we apply Uniform Manifold Approximation and Projection (UMAP; McInnes et al., 2018). UMAP is a non-linear dimensionality reduction method that models the data as lying on a low-dimensional manifold. It constructs a weighted k-nearest-neighbors (kNN) graph in the high-dimensional space. After a process to reduce the dimensionality of data, points that are close in the original space are encouraged to remain close after projection, while the method also retains enough global structure to avoid fragmenting clusters.

We chose UMAP over alternative dimensionality reduction techniques because it strikes a balance between preserving local neighborhoods—keeping semantically similar words close in the low-dimensional space—while also maintaining the global structure of the embedding. Moreover, UMAP offers competitive runtime performance through scalable approximate kNN search; UMAP can also be trained on a subsample of $\mathbf{E}$ and then applied to new instances, which further reduces computational costs. These properties make UMAP a pragmatic default for reducing contextualized embeddings prior to clustering. In our application, we reduce the 300-dimensional matrix $\mathbf{E} \in \mathbb{R}^{M \times 300}$ to a 10-dimensional representation $\mathbf{Z} \in \mathbb{R}^{M \times 10}$.

It is important to note that, despite approximate kNN accelerations, UMAP's computational cost increases with the number of instances M, both in terms of runtime and memory required to construct the neighbor graph and perform the optimization. To manage this complexity, we limit the training sample for UMAP to 50,000 word instances, drawn from $\mathbf{E}$. In our implementation, we use simple random sampling. However, if there is concern about underrepresenting infrequent dictionary terms, stratified sampling can be employed to ensure that each dictionary entry contributes proportionally to the training set.

3.2.3 Constrained Fuzzy C-means

We use clustering as a regularization device to attenuate the influence of rare, marginal, or out-of-context uses of dictionary words. The idea is to let tokens that occur in anchor-like contexts collectively shape a reference cluster while allowing other usages to contribute only proportionally to their semantic proximity. We employ fuzzy C-means (FCM) primarily because it affords an immediate way to encode anchor constraints: anchor instances can be pinned to a designated cluster, and the standard FCM updates naturally propagate this constraint to the centroid estimates at every iteration. Alternative clustering choices are possible. In particular, one can fit a multivariate Gaussian mixture model (GMM) by maximum likelihood expectation-maximization and fix the responsibilities for anchor points; similarly, hard clustering (e.g., K-means with must-link constraints) can be used provided the algorithm permits anchor constraints that shape centroid updates.

Setup and Objective. Let $\mathbf{Z} \in \mathbb{R}^{M\times 10}$ be the UMAP-reduced matrix whose i-th row $z_i^\top$ encodes the contextualized embedding for token instance i. We seek C clusters with centroids $\{c_k\}_{k=1}^{C} \subset \mathbb{R}^{10}$ and fuzzy memberships $\mathbf{U} = (u_{ik}) \in [0,1]^{M\times C}$ satisfying $\sum_{k=1}^{C} u_{ik} = 1$ for all i. With fuzzifier $m = 2$, the FCM criterion is

$$J(\mathbf{U}, \{c_k\}) \;=\; \sum_{k=1}^{C}\sum_{i=1}^{M} u_{ik}^{m}\, \|z_i - c_k\|_2^2 \quad \text{with} \quad m = 2. \tag{1}$$

Anchor instances are constrained to a designated cluster. Let $\mathcal{I}_A \subset \{1,\dots,M\}$ be the index set of anchor tokens and $r : \mathcal{I}_A \to \{1,\dots,C\}$ the mapping that assigns each anchor to a cluster (often $r(i) \equiv 1$). The constraint is

$$u_{i\,r(i)} = 1\,, \qquad u_{ik} = 0 \;\; \forall k \neq r(i), \quad \text{for all } i \in \mathcal{I}_A. \tag{2}$$

These memberships remain fixed throughout the iterations; non-anchor rows are updated as usual.

Iterative Updates. In our implementation, we apply FCM to $\mathbf{Z}$ using the following iterative scheme. Denote by d_{ik} the (squared) Euclidean distance from point z_i to centroid c_k: $d_{ik} \equiv \|z_i - c_k\|_2^2$.

(i) **Initialization.** Randomly initialize centroids $\{c_k^{(0)}\}_{k=1}^C$ by sampling C rows of $\mathbf{Z}$ (seeded for reproducibility). Initialize $\mathbf{U}^{(0)}$ with random positive entries and row-normalize. For anchors $i \in \mathcal{I}_A$, overwrite the row to enforce $u_{i\,r(i)}^{(0)} = 1$ and $u_{ik}^{(0)} = 0$ for $k \neq r(i)$.

(ii) **Centroid update.** At iteration $t = 1, 2, \ldots,$ update each centroid via the weighted mean

$$c_k^{(t)} \;=\; \frac{\sum_{i=1}^M \left(u_{ik}^{(t-1)}\right)^m z_i}{\sum_{i=1}^M \left(u_{ik}^{(t-1)}\right)^m}, \quad \text{with } m = 2. \tag{3}$$

Effect of anchors. Since $u_{i\,r(i)}^{(t-1)} = 1$ for anchor i, anchors exert full weight on their assigned centroid $c_{r(i)}^{(t)}$, thereby pulling this centroid toward the anchor manifold. If a denominator becomes numerically tiny (a nearly empty cluster), we re-initialize $c_k^{(t)}$ to a randomly chosen point to avoid degeneracy.

(iii) **Membership update (non-anchors).** For each non-anchor $i \notin \mathcal{I}_A$, compute distances $d_{ik}^{(t)} = \|z_i - c_k^{(t)}\|_2^2$. If $d_{ik}^{(t)} = 0$ for some k, set $u_{ik}^{(t)} = 1$ and $u_{i\ell}^{(t)} = 0$ for $\ell \neq k$. Otherwise, with $m = 2$ and squared distances, the FCM membership update is

$$u_{ik}^{(t)} \;=\; \left(\sum_{\ell=1}^C \left(\frac{d_{ik}^{(t)}}{d_{i\ell}^{(t)}}\right)^{\frac{1}{m-1}}\right)^{-1} \;=\; \left(\sum_{\ell=1}^C \frac{d_{ik}^{(t)}}{d_{i\ell}^{(t)}}\right)^{-1}, \tag{4}$$

since $\frac{1}{m-1} = 1$ when $m = 2$. After updating all non-anchor rows, *re-enforce* the anchor constraints: set $u_{i\,r(i)}^{(t)} = 1$ and $u_{ik}^{(t)} = 0$ for $k \neq r(i)$ whenever $i \in \mathcal{I}_A$. This guarantees that the constraint is respected even under numerical noise.

(iv) **Convergence check.** Compute the centroid shift $\Delta^{(t)} = \sum_{k=1}^C \|c_k^{(t)} - c_k^{(t-1)}\|_2^2$. Terminate when $\Delta^{(t)} < \varepsilon$ (e.g., $\varepsilon = 10^{-5}$) or when a maximum number of iterations is reached. The returned soft memberships are $\mathbf{U}^{(t)}$; *hard* assignments may

be obtained post hoc by $\hat{k}(i) = \arg\max_k u_{ik}^{(t)}$.

How the constraint changes FCM. Relative to unconstrained FCM, the only change is that some rows of $\mathbf{U}$ are held fixed at each iteration. However, this local change has a global effect: (i) anchors contribute fully to the on-concept centroid through the u_{ik}^m weights, (ii) subsequent distance calculations reflect the anchor-shaped centroids, and (iii) decision boundaries shift to align with the anchor manifold. In practice, this stabilizes the semantics of the target cluster and regularizes away token usages that fall far from anchor contexts by assigning them smaller memberships. A constrained GMM/EM variant fixes responsibilities for anchors and updates means and covariances accordingly; hard K-means with must-link constraints is also feasible, though it lacks graded memberships. Any alternative should (a) permit anchor constraints and (b) propagate them to centroid (or component) updates so that anchors shape cluster geometry rather than serving only as a post-hoc filter.

3.2.4 Aggregation to Document-Level

Let $\mathcal{I}_j \subseteq \mathcal{I}$ denote the set of dictionary instances that occur in document j, and recall that the length (total token count) of document j is N_j, so $d_j = (t_{j1}, \ldots, t_{jN_j})$. We define document-level *attention* to the target topic as the share of all tokens in d_j that (i) match the dictionary (standard measure) or (ii) match the dictionary and are weighted by their on-concept membership (ambiguity-robust measure).

Standard dictionary attention. The formula for attention score with standard word counting is:

$$\text{Attn}_j^{\text{dict}} \;=\; \frac{1}{N_j} \sum_{n=1}^{N_j} \mathbf{1}\{t_{jn} \in \mathcal{D}\} \;=\; \frac{|\mathcal{I}_j|}{N_j}\,. \tag{5}$$

Ambiguity-robust attention. Let $u_{1i} \in [0, 1]$ be the on-concept fuzzy membership for instance $i \in \mathcal{I}_j$. The ambiguity-robust attention score is

$$\text{Attn}_j^{\text{rob}} \;=\; \frac{1}{N_j} \sum_{i \in \mathcal{I}_j} u_{1i} \,. \tag{6}$$

Both quantities are normalized by the *total* word count N_j and are therefore interpretable as the fraction of the document devoted to the topic. When the on-concept memberships are set to unity for all dictionary instances, $u_{1i} \equiv 1$, the ambiguity-robust measure reduces to the standard dictionary measure: $\text{Attn}_j^{\text{rob}} = \text{Attn}_j^{\text{dict}}$.

3.3 Recommended Implementation Practices

Our approach occupies a middle ground between fully supervised classifiers and traditional dictionary counts, combining researcher control with modern representation learning. By anchoring the "on-concept" cluster to unambiguous terms and propagating that constraint through fuzzy C-means, the method yields context-aware weights that remain transparent and interpretable. In this sense, it innovates on standard dictionaries without abandoning their virtues: it nests the classic count as a special case, yet discounts tokens whose local usage departs from the intended concept. Nonetheless, certain arbitrary decisions common to traditional dictionary-based approaches and the cluster-analysis literature persist. In this section, we outline how to mitigate them.

3.3.1 Designing Dictionaries and Anchor Words

A practical advantage of the method is that it allows researchers to extend their dictionaries to include words that are loosely or ambiguously related to the concept of interest without the risk of inflating measurement error. Because our approach uses contextual embeddings and cluster-based weighting, token instances that appear out of context will receive lower weights and contribute minimally to the final score. This reduces the incentive to omit potentially relevant but polysemous terms and ensures that documents

using conceptually aligned language—when supported by anchor-defined context—are not undercounted.

At the same time, the method introduces an additional degree of freedom through the choice of anchor words, which can substantially shape the resulting outputs. To address this, we draw on the qualitative literature on concept formation (Goertz, 2006). In this framework, a concept is structured in three layers: the first layer is the overarching concept itself; the second layer consists of its ontological constituents—factors that capture essential aspects of the concept; and the third layer consists of indicators that allow the empirical operationalization of these constituents. Within a dictionary approach, indicators correspond to words, while the ontological constituents jointly constitute the concept through an additive structure. This implies that secondary dimensions contribute linearly to the concept and are substitutable—one indicator may compensate for another. For example, references to social policy in a text are often composed of multiple subconcepts, such as education policy and healthcare policy.

To capture such multidimensional concepts, we propose building dictionaries that explicitly acknowledge this layered structure. Each subconcept may require its own set of anchor words, since these dimensions, while often correlated in the embedding space, should exhibit at least some degree of semantic separation (otherwise they would collapse into a single constituent rather than distinct entities). Incorporating at least one anchor word per subconcept strengthens the clustering algorithm's capacity to reflect the full breadth of the target concept. As a robustness check, we recommend a leave-one-out procedure within each set of subconcept-specific anchor words to evaluate the stability of the results.

3.3.2 Selecting The Number of Clusters

The decision over the number of clusters is a common problem in clustering algorithms. While there are some data-driven considerations over this, the purpose of our algorithm is to measure concepts. As such, researches must have qualitative considerations over

the number of clusters. These considerations are more feasible when the researcher herself tailored the dictionaries. However, our method is also available for off-shelf dictionaries (e.g., lexicons on emotions such as Emolex, developed by Mohammad and Turney, 2013). In that case, we recommend a data-driven process to select the number of clusters. We will discuss both approaches.

Theory-Driven Methods. As the purpose of our method is mitigating the impact of polysemy in word counts, a theory-driven way of determining the number of clusters in the algorithm is to consider carefully each word in the dictionary and think of alternative uses for it. For instance, in a toy example about environmental policy, the word environment itself is ambiguous as it can refer to "business environment", for instance, and then align with a concept of economic development policies. By considering the alternative contexts where each word might be used, the researchers can think on how related these contexts are. If they converge to an alternative concept, these words can become just a single alternative cluster. If the polysemy is not frequent as well, this word might be not considered a concept itself and agglomerated into a generic alternative cluster. Rigorous domain of the text and the concept allow this kind of theory-driven method to be applied in research.

Data-Driven Methods. A simple and effective way to choose the number of clusters in fuzzy clustering is to use the Xie–Beni (XB) index (Xie and Beni, 2002). Intuitively, XB asks whether clusters are (i) *compact*—points lie close to their assigned centroids—and (ii) *well separated*—centroids are far from one another. When clusters are tight and distinct, XB is small; when clusters are diffuse or crowded together, XB is large. Because it summarizes the core trade-off in clustering with a single interpretable quantity, XB provides a practical, data-driven criterion for selecting K that aligns closely with substantive expectations about coherent, distinct usages of a concept.

Let $\mathbf{Z} \in \mathbb{R}^{M\times 10}$ be the UMAP-reduced matrix (rows $z_i^\top$), $\{c_k\}_{k=1}^K$ the centroids, and $\mathbf{U} = (u_{ik}) \in [0,1]^{M\times K}$ the fuzzy memberships with $\sum_{k=1}^K u_{ik} = 1$. Define squared

Euclidean distances $d_{ik} = \|z_i - c_k\|_2^2$. The Xie–Beni index is

$$\mathrm{XB}(K) = \frac{\sum_{i=1}^{M}\sum_{k=1}^{K} u_{ik}^2\, d_{ik}}{M \cdot \min_{k\neq\ell} \|c_k - c_\ell\|_2^2}, \qquad \text{smaller is better.} \tag{7}$$

An advantage of the Xie-Beni index is its invariance to different clusterings for the same set of anchor words. Anchor instances are pinned to their designated cluster ($u_{ik} = 1$ for that cluster). This typically reduces within-cluster dispersion in the numerator and can enlarge centroid separation in the denominator when anchors identify a distinctive semantic region. Because the *same* anchor set is used for every K under consideration, their influence is applied consistently, so XB remains an apples-to-apples criterion for choosing K.

4 Applications

4.1 Environmental Politics in the UN General Assembly

The United Nations General Debate Corpus (UNGDC) is a comprehensive collection of statements delivered by member states at the annual UN General Debate since 1946 (Baturo et al., 2017; Jankin et al., 2017). It includes the official English texts of speeches—either as delivered or using UN translations—covering thousands of country statements from nearly all member states across the postwar period. The speakers are typically high-ranking representatives, most often heads of state or government, foreign ministers, or permanent representatives to the UN. Unlike roll-call votes in the General Assembly, which are shaped by strategic alliances and external constraints, General Debate speeches allow governments to articulate priorities in their own words and with fewer restrictions. As such, the UNGDC offers political scientists a valuable resource for tracing state preferences, issue attention, and the evolution of foreign policy agendas over time.

In this application, we use attention to environmental and climate issues in UNGDC floor speeches to illustrate how our method yields measures that are robust to the inclusion of ambiguous and even unrelated terms in a dictionary of climate and environment related keywords. We then show how these differences in measurement can affect substantive inferences about what kinds of institutions are more likely to express environmental commitments in multilateral arenas.

We begin by developing a dictionary tailored to the study of environmental politics in the international arena. The dictionary is designed to be broad enough to capture diverse aspects of environmental discourse while remaining precise and domain-relevant. It includes terms central to global debates, such as *climate*, *deforestation*, *renewable*, and *biodiversity*, but deliberately excludes words like *sustainable*, which in diplomatic contexts may refer to debt or fiscal policy rather than environmental issues. To illustrate our method, we score each speech's attention to environmental issues using this baseline dictionary and the anchor words *environmental* and *emissions* using $k = 2$ clusters. Figure 2 visualizes the results using the first two dimensions of the UMAP projection of each speech fragment's contextual embedding vector. Each fragment is colored according to its membership score, with green and yellow colors indicating higher on-concept fragments. The figure illustrates how our method separates keyword instances into on-target and off-target usages. Off-target text fragments illustrated in the figure use keywords "environment," but in contexts that reveal the word usage has little to do with environmental politics (eg, "for the creation of a climate of peace and international security"). By contrast, on-target text fragments use the same keywords in a manner that is consistent with the concept of interest (eg, "ambitious decision to combat climate change and preserve our environment").

A primary implication of our method's ability to account for context and discount off-topic word usages is that our measures will be robust to addition of ambiguous or even irrelevant words to the dictionary. By contrast, conventional word counts will be quickly overrun with measurement error as noisy terms are added. To illustrate this, we

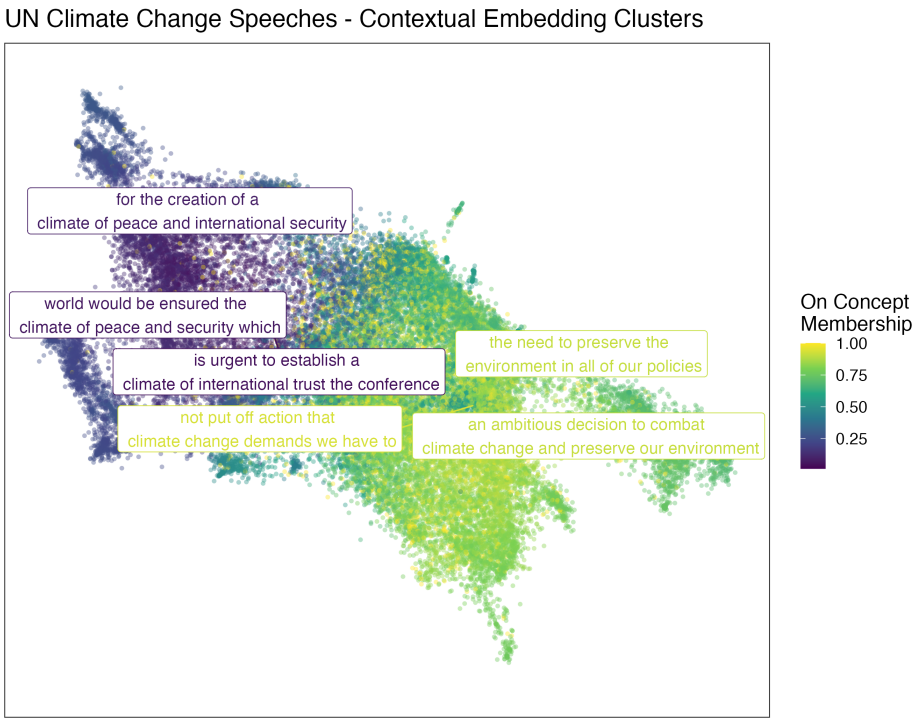


Figure 2: Visualization of climate speech fragments using first two dimensions of the UMAP projection. Navy points represent climate keyword instances with lower target cluster scores, while green and yellow points represent instances with higher target cluster scores.

develop two additional dictionaries. In the "ambiguous" dictionary, we add five additional words to the baseline dictionary such as "sustainable" and "natural" which could plausibly reflect environmental discourse but also have alternative usages which could add unwanted noise to the measure. In the "noise" dictionary, we add to the baseline dictionary five terms related to international economics.[3] We then score each speech's attention to climate politics using these three dictionaries and both conventional and ambiguity-robust methods, yielding six different measures of environmental attention. Scores are based on the proportion of dictionary keywords in each speech; adjusted scores are calculated by dividing the sum of membership scores for each contextualized keyword instance by the total number of words.

Figure 3 compares these different measures of each speech's relative attention to climate by plotting the raw proportion of words in each speech that are found in the

[3]The complete list of dictionary keywords can be found in the appendix.

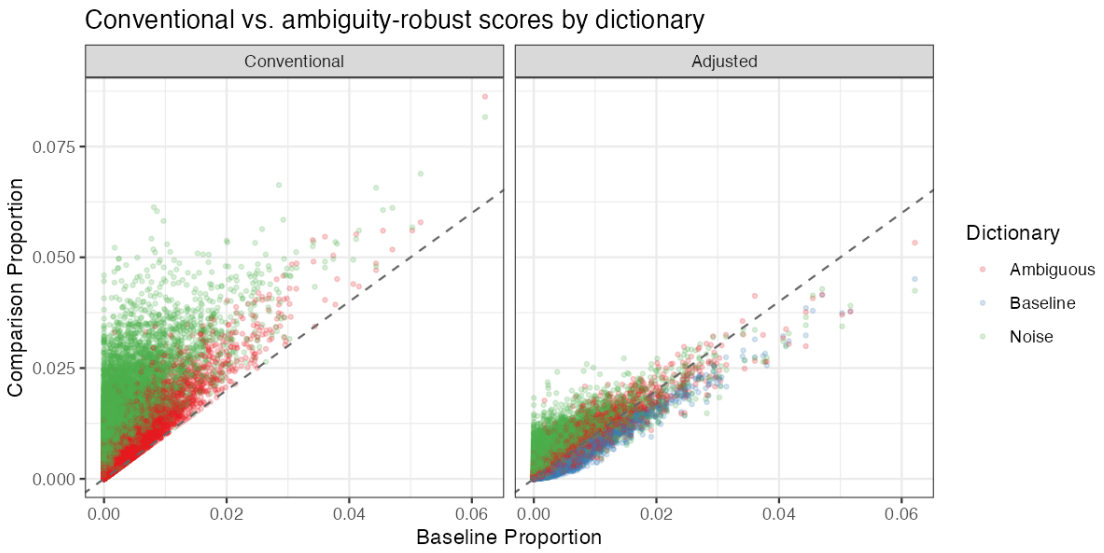


Figure 3: Environmental attention of UNGDC speeches by dictionary (baseline, ambiguous, or noisy) and method (conventional vs. ambiguity-robust). Each point represents a single speech. The conventional proportion of environmental keywords (using the baseline dictionary) of each speech is plotted on the X axis, while the Y axis plots a comparison proportion varying the dictionary (indicated by color) or method (conventional vs. adjusted, indicated by facet). Adjusted proportions are calculated by dividing the sum of membership scores for each contextualized keyword instance by the total number of tokens.

baseline dictionary on the X axis, and a comparison score using an alternative dictionary or adjusted proportions based on our ambiguity-robust method on the Y axis. The left panel of the figure illustrates how when using conventional word counts, the addition of ambiguous and noisy terms to the keyword dictionary causes substantive instability in the measure. The correlation between measures based on the baseline dictionary and the ambiguous dictionary is 0.91; this correlation drops to 0.68 when comparing the baseline against the noisy dictionary (see Table A3 for full matrix of correlation coefficients between all six measures of environmental attention). This is not necessarily surprising: including terms related to international trade means that the noisy dictionary measures attention both to environmental politics and international economics. However, we show in the right panel of Figure 3 that the application of our ambiguity-robust adjustment yields measures substantively more consistent with the baseline specification even when using the polluted, noisy dictionary. The correlation between the baseline measure and adjusted measures based on the ambiguous dictio-

nary is 0.94; this correlation only drops to 0.82 when using adjusted scores based on the noisy dictionary. This analysis provides an intuitive scaffolding for the advantages of our method. While conventional dictionary measures are sensitive to the inclusion of off-topic keywords, our method successfully captures and discards these off-topic keyword instances when constructing adjusted document level scores of environmental attention.

Next, we evaluate how these variations in measurement affect substantive inferences. A large literature in international and comparative environmental politics argues that democracies tend to express stronger environmental commitments in multilateral arenas due to domestic accountability pressures, exposure to normative networks, and responsiveness to civil society demands (e.g., Neumayer, 2002; Bättig and Bernauer, 2009). In the context of the UNGDC floor speeches, we expect that democracies express more attention to climate and environmental politics than non-democracies. To test this relationship, we estimate a set of models using a two-way fixed effects specification. Let Dem_{ct} be a democracy indicator and Env_{ct} denote environmental attention defined as the proportion of environmental keywords in a given floor speech. Then, we estimate

$$\text{Env}_{ict} = \beta \, \text{Dem}_{ct} + \alpha_c + \gamma_t + \varepsilon_{ict}, \tag{8}$$

with country and year fixed effects and standard errors clustered at the country level. This panel specification identifies the relationship between changes in democracy status within countries and changes in environmental attention while accounting for general time trends.

We compare results obtained with the baseline dictionary to those from the "noisy" dictionary. Columns 2 and 4 of Table 2 show that the relationship between democracy and environmental attention remains stable even as noise is introduced when using the ambiguity-robust method. In contrast, results obtained using traditional dictionary counts (columns 1 and 3) deteriorate sharply when using the "noisy" dictionary. This demonstrates the value of our method for empirical applications: precise measurement

that automatically discounts off-topic keyword instances allows the identification of substantive relationships.

Table 2: Regressions between dictionary-based measures of Environmental Concern and Democracy.

	Environmental Concern at the UNGA			
Democracy (BMR)	0.0005**	0.0004**	0.0007	0.0005*
	(0.0002)	(0.0002)	(0.0006)	(0.0003)
Country FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Dictionary	Baseline	Baseline	Noisy	Noisy
Number of Clusters	Count	K = 2	Count	K = 3
Observations	9,269	9,269	9,269	9,269
R^2	0.43848	0.45775	0.40960	0.43874

Clustered (Country) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Taken together, these results show that the ambiguity-robust dictionary method closely tracks conventional environmental measures, recovers the expected association between democracy and climate discourse, and exhibits markedly greater resilience to noise and lexical ambiguity. This makes ambiguity-robust dictionaries a valuable tool for analyzing thematic variation in political speech, particularly in domains where vocabulary is context-dependent or easily conflated with adjacent policy frames.

4.2 Emotions in Kenyan Judicial Politics

Choi et al. (2022) analyze nearly 10,000 criminal appeals in Kenya between 2003 and 2017 and show that judges are 3 to 5 percentage points more likely to rule in favor of coethnic appellants. To probe the mechanism behind this bias, they examine the relationship between appellant-judge coethnicity and the prevalence of trust and disgust oriented language used in the text of judicial decisions. The authors measure trust and disgust using keyword dictionaries constructed using a data driven procedure — starting from small seed dictionaries for each concept they then capture the 2,000 most closely related terms in a custom embedding space trained on their corpus of judicial texts. Armed with these keyword dictionaries, they then compute the proportion of keywords

in each ruling and estimate the effect of appellant coethnicity on the proportion of trust and disgust related keywords using linear regression including judge and courthouse-year fixed effects. They find that decisions involving coethnic appellants contained significantly more trust-related keywords (e.g., confidence, credibility, honesty), but no decrease in disgust-related keywords compared to non-coethnic appellants and interpret this pattern as pointing to in-group favoritism rather than out-group hostility as the source of bias.

While our earlier application highlighted the ability of the ambiguity-robust method to discount artificially noisy terms (eg, inserting keywords about international trade into a dictionary about climate politics), here we demonstrate the precision of our method in the context of realistic variation in keyword choice. One challenge researchers face when constructing keyword dictionaries is how to avoid Gelman and Loken (2013)'s "garden of forking paths" — given that a plausible theoretical justification exists for a large number of possible keywords and that substantive results may be highly sensitive to keyword choices, it is desirable to use a principled approach to minimize the sensitivity of substantive conclusions to any particular resolution of these dictionary choices. However, even principled approaches can pose challenges. Conventional sentiment dictionaries such as Mohammad and Turney (2013) offer a set of off-the-shelf words associated with various emotional concepts, but applying them in Choi et al. (2022)'s context may be inappropriate because affective valence is assigned to commonly used legal words such as *law* or *sentence*. To avoid this problem, the authors choose the embedding-based strategy described above to tune keywords to the linguistic idiosyncrasies of the corpus. This approach also introduces significant risks. Embeddings capture statistical context rather than semantic polarity, leading to many false positives such as direct antonyms (e.g., *honest* and *dishonest*) and words that co-occur due to syntax or domain-specific jargon rather than conceptual relevance. The keyword dictionary used by the authors for "trust" language includes several unrelated or even opposing terms like *pain*, *false*, and *incompetent*. This highlights one of the key weak-

nesses of dictionary methods: even when using principled keyword selection approaches such as externally provided, validated keyword lists or sophisticated data driven procedures, the resulting dictionary can contain words that are at best ambiguous and at worst indicating the complete opposite of the target concept.

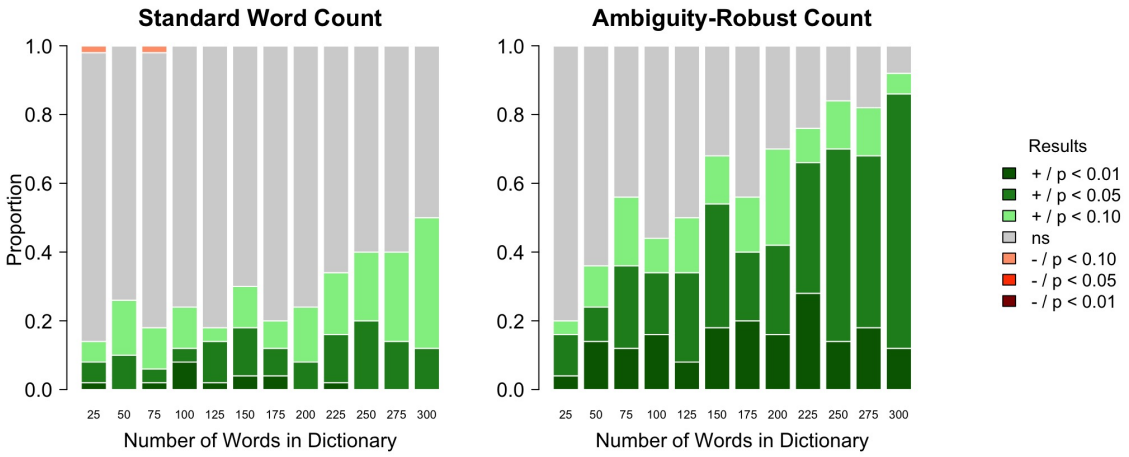


Figure 4: Results of Regressions with Dictionary Permutations.

We evaluate the sensitivity of the coethnicity-trust relationship to realistic variation in keyword choices using a permutation test. We choose a random subset of "trust" terms from the NRC (Mohammad and Turney, 2013) lexicon, repeating this process 50 times each for varying dictionary sizes (25-300 words)[4], yielding N=600 sampled dictionaries. This set of dictionaries represents a sample from the universe of "plausible" dictionaries that a researcher interested in measuring trust may choose to utilize. Then, for each of these generated dictionaries we measure trust in each judicial decision using both classical counts and our ambiguity robust method. We use a fixed number of $k = 2$ clusters for ambiguity-robust models, chosen to balance parsimony and the Xie-Beni index (see Appendix C). Finally, we re-estimate the coethnicity-trust regression model for each sampled dictionary and measurement method (standard word count versus ambiguity-robust).

Figure 4 shows the distribution of p-values from this set of tests. When using conventional word counts, the robustness of the substantive relationship between trust and

[4]The NRC lexicon contains 1,200 total trust terms

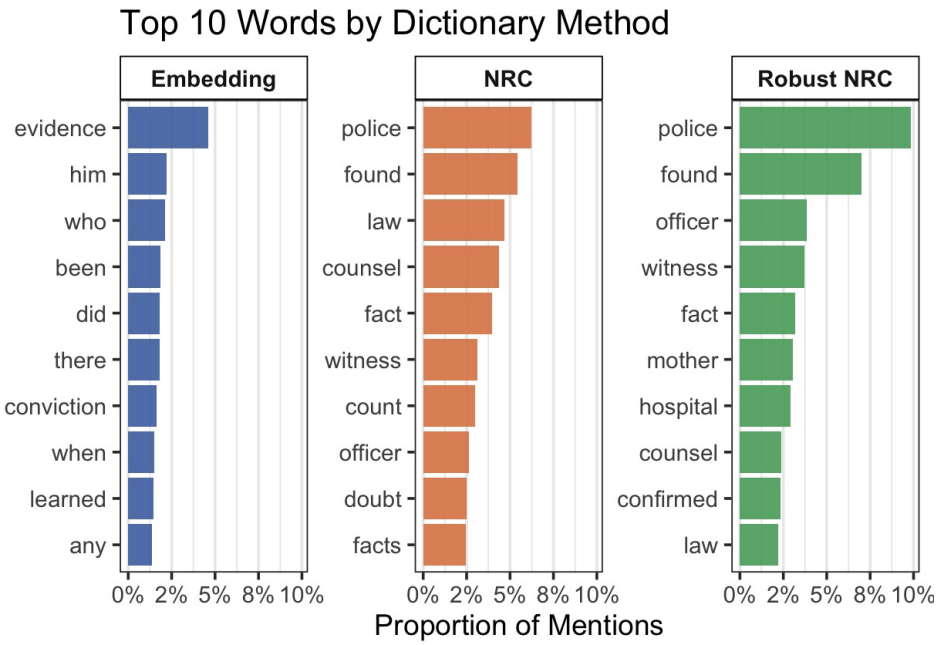


Figure 5: Top 10 Words by Dictionary Method. Comparison of term proportions across the embedding expansion, standard NRC, and ambiguity-robust methods.

co-ethnicity is highly sensitive to dictionary choice. Fewer than 20% of tests yield a p-value less than 0.05, even at the maximum dictionary size (n=300). This sensitivity is substantively concerning — it could be that the relationship between trust and coethnicity depends on unusual operationalizations of "trustful" language. However, ambiguity-robust measures tell a strikingly different story. Analysis using the ambiguity-robust measures yield $p < 0.05$ results almost 50% of the time with $n \in [100, 250]$ and over 80% of the time with $n = 300$. This suggests that the ambiguity-robust method is able to successfully and reliably eliminate the influence of words that are contextually irrelevant to the target concept. As a result, estimates are stable, converging on the same finding regardless of the specific keywords being included. The improvement in precision offered by the ambiguity-robust method is dramatic: in terms of statistical power, the difference between $\beta = 0.2$ and $\beta = 0.8$ is roughly equivalent to increasing the sample by a factor of six.

In addition to increasing precision, the use of our method also allows insight into which words are most informative in driving the relationship of interest. We measure trust language in the sample of judicial texts using both the authors' original embed-

ding based dictionary, the NRC trust dictionary, and the NRC trust dictionary using the ambiguity-robust method. Figure 5 reports the top 10 most used words for each dictionary, showing that the embedding-based dictionary places a significant weight into uninformative terms such as "him," "who," and "been." Word counts using the NRC trust dictionary, both traditional and robust, seem to have more informative words. Both identify terms related to evidence and factual evaluations such as *police*, *found*, and *fact*. However, the robust method systematically re-weights these instances; while generic legal terms like *law* and *counsel* drop in the rankings, our method elevates both concrete figures like *mother* and *hospital* and evidence-related words such as *witness* and *confirmed*.

We then identify the top 10 and bottom 10 keywords by average cluster membership weight in Figure 6. These weights are informative in that they allow us to identify which words are most and least consistently associated with the target cluster, providing additional interpretability and transparency. The figure confirms that abstract procedural terminology is systematically down-weighted, whereas the highest weights are assigned to concrete societal anchors: *elder*, *nurse*, *neighbor*, and *hospital*. This uncovers a novel substantive finding: judicial trust in this context is built not through abstract legal rhetoric, but by anchoring narratives to highly respected, authoritative societal figures or evidence. By explicitly weighting words, we identify the process used for trust-building, an insight that remains obscured in traditional word counts.

4.3 Donor Behavior and Moral Affect in the US Congress

American legislative politics is increasingly characterized by a shift in the function of congressional speech away from deliberation and toward symbolic, audience-oriented communication (Hill and Hurley, 2002; Park, 2023). Members of Congress use moral, emotive, and intuition-oriented rhetoric to signal values and identities to external audiences rather than to persuade legislative peers (Osnabrügge et al., 2021; Gennaro and Ash, 2022; Dillion et al., 2024; Aroyehun et al., 2025). This shift raises normative con-

Figure 6: Cluster Memberships (Weights) for Trust Keywords.

cerns about both democratic deliberation and political polarization — when political disagreement is framed primarily in intuitive terms, evidence and data are less able to adjudicate between competing positions, narrowing the scope for persuasion and compromise (Aroyehun et al., 2025). The use of moral language by lawmakers can also promote dehumanization and inflate perceptions of intergroup hostility, especially in digital and media-rich environments where such rhetoric is readily amplified (Crockett, 2017; Brady et al., 2023).

Our previous applications illustrate the ability of our method to parse ambiguity and context and demonstrate the associated increased precision while preserving interpretability. In this application, we focus on using the scalability and low cost of our method to generate substantive insights about congressional speech and donor behavior that would be prohibitively expensive to achieve with more sophisticated models because of the sheer scale of congressional speech.

We use both our ambiguity-robust and conventional dictionary methods to measure negative moral language in congressional speech from 1980-2024 using Congressional Record data provided by Gentzkow et al. (2019) and the Congressional Record API. We restrict attention to speeches over 200 words to eliminate procedural items (see

Appendix for additional details), ultimately analyzing $N = 841,622$ speeches. Moral language is often measured using dictionary measures, the most widely used moral dictionary being based on Moral Foundations Theory (MFT) developed by Haidt et al. (2003). However, as with many other dictionary applications, dictionary words carry substantively different moral valences depending on their contexts — the MFT contains words such as "oppose" and "illegal" which carry different moral valences in a legislative debate context than in common usage. We start with all negatively valenced words from the original Moral Foundations dictionary, the anchor words "immoral" and "evil," and set $k = 2$ clusters; results are robust to alternative choices of k (see Appendix). In total, there are $N = 2,716,739$ occurrences of negative moral language keywords in our corpus; extraction and clustering completes in the order of minutes using the authors' 2022 Macbook Air.

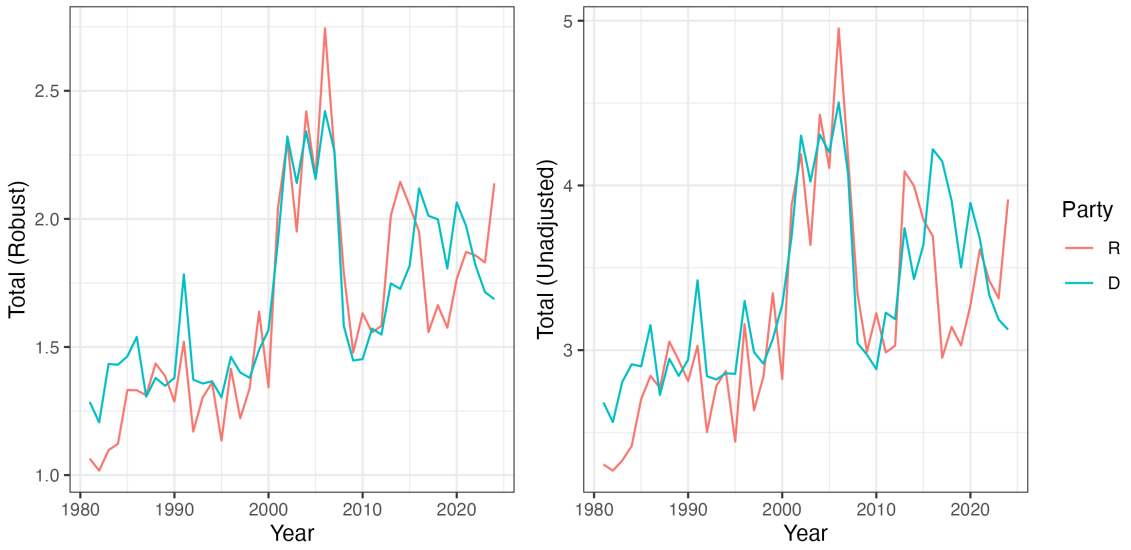


Figure 7: Incidence rates of negative moral language over time, by chamber and party. The Y axis measures the average number of (adjusted) keyword instances per speech. The left panel uses our robust method, weighting moral language by membership probability in the anchor cluster. The right panel uses raw keyword counts to measure total negative moral language.

We compare the average number of negative moral keywords per speech by party in the left panel of Figure 7 and the regularized count of negative moral language per speech. Under both specifications, we observe an increase in negative moral rhetoric

over the last 40 years, although this pattern is slightly more pronounced using the robust measure of negative moral rhetoric. Our results also confirm the pre-existing "opposition effect" of Wang and Inbar (2021) that moral rhetoric is more prevalent among Republicans when Democrats are in power (and vice versa).

To shed light on one potential mechanism that could explain this trend, we analyze the relationship between donor behavior and negative moral rhetoric. Negative moral rhetoric could lead to greater campaign donations because such rhetoric is shown to be more effective in mobilization efforts especially in polarized environments (Simonsen and Widmann, 2025). We collect data on campaign donations from the DIME dataset (Bonica, 2024) and match congressional candidates in the DIME dataset to their speeches in the Congressional record. Using this data, we then construct a legislator-cycle panel of total campaign donations received along with a measure of the mean number of negative moral language usages per speech within each election cycle calculated using both conventional dictionary approaches and the ambiguity-robust method.

Table 3: Individual contributions vs. moral language in floor speeches. Coefficients are change in log(individual contributions) per additional moral keyword per speech. t = election cycle - 2000.

Dependent Variable:	log(Indiv. contribs.)					
	Robust	Unadj.	Robust $\times t$	Unadj. $\times t$	Within (rob.)	Within (unadj.)
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Moral lang. (robust)	0.18***		0.04		0.02	
	(0.05)		(0.07)		(0.05)	
Moral lang. (raw)		0.12***		0.05		0.05*
		(0.03)		(0.04)		(0.03)
Moral lang. (robust) $\times$ t			0.02***			
			(0.005)			
Moral lang. (raw) $\times$ t				0.01***		
				(0.002)		
Fixed-effects						
Legislator					Yes	Yes
Pol. controls	Yes	Yes	Yes	Yes		
Cycle	Yes	Yes	Yes	Yes	Yes	Yes
Observations	8,891	8,891	8,891	8,891	8,891	8,891

Clustered (Legislator) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

We first present a cross-sectional analysis of whether increased moral language is associated with greater individual donations in columns (1–2) of Table 5. We include party, election cycle, and chamber fixed effects. An additional robustness specification including legislator extremism (measured using DIME scores) yields substantively similar results (Table A5). We find that negative moral language, when measured using the ambiguity-robust method, is positively associated with donations: legislators who use one additional negative moral instance per speech receive approximately 20% more contributions on average (0.18 log points). This association is also present using conventional counts but is substantively attenuated (0.12 log points), suggesting that improved measurement precision strengthens the estimated relationship.

Next, we estimate a model that allows the association between donations and negative moral language to vary over time. The results, presented in columns (3–4), show that this relationship is primarily a feature of recent years: before 2000, negative moral language and donation behavior are weakly related. However, the strength of this relationship has increased — a finding that is again more pronounced when using ambiguity-robust measures. By 2020, the cross-sectional relationship is 0.4 log points larger, implying that legislators who use more negative moral language receive roughly 49% more contributions.

Finally, we estimate a model with legislator fixed effects to identify within-legislator changes in moral rhetoric. Columns (5–6) show limited evidence that increases in moral rhetoric are associated with increases in campaign donations. Taken together with the strong and growing cross-sectional relationship, these results suggest that the link between moral rhetoric and donations operates primarily through stable differences across legislators rather than short-term changes in rhetorical style. This pattern is consistent with a sorting process in which donors disproportionately support legislators who consistently employ more negative moral language, potentially contributing to the broader increase in such rhetoric over time.

5 Conclusion

This paper has introduced ambiguity-robust dictionaries, a new method for text analysis that allows dictionary methods to account for context and ambiguity while retaining the advantages of traditional dictionary approaches. By integrating the transparency of researcher-defined lexicons with the contextual power of modern word embeddings, our approach offers a principled way to measure latent concepts while upholding the scientific values of replicability and theoretical grounding. The method preserves the intuitive workflow of dictionary analysis but augments it by systematically identifying and down-weighting word instances that are used in a manner inconsistent with the core concept of interest. This produces a more accurate, de-noised measure that is robust to the inclusion of ambiguous or even unrelated keywords, freeing researchers to build more comprehensive dictionaries without introducing significant measurement error.

The practical advantages of the ambiguity-robust dictionary approach are demonstrated across three distinct empirical domains. In our analysis of environmental discourse in the UN General Debate, the robust measure remains stable and yields consistent regression estimates even when the dictionary is intentionally contaminated with conceptually unrelated noise. In the replication of Choi et al. (2022), the method successfully recovers the subtle emotional signals of in-group favoritism in Kenyan judicial texts that traditional methods fail to reliably detect. Beyond mere replication, our approach uncovers a novel substantive insight: judicial trust-building is a relational process anchored in evidence and concrete societal figures rather than the abstract legal rhetoric that our algorithm systematically down-weights. Finally, in the study of US Congressional speech, the ambiguity-robust approach clarifies trends obscured by noisy word counts. By de-noising context-dependent moral vocabulary, we show that lawmakers who employ more negative moral language receive more campaign contributions and that this relationship is increasing over time. These applications highlight the key methodological contribution of our approach: by explicitly modeling and mit-

igating the impact of polysemy, ambiguity-robust dictionaries produce more valid and reliable measures than conventional word counts. The improved precision reduces measurement error, which not only enhances the descriptive accuracy of the measure but also strengthens the validity of downstream statistical analyses.

Like any method, ours is not without its own set of researcher degrees of freedom. The selection of anchor words and the determination of the number of clusters are critical decisions that require careful justification. While we have provided both theory-driven and data-driven guidance for these choices, they represent potential sources of variation that researchers must consider and address through robustness checks. However, we contend that these challenges are both manageable and a worthwhile trade-off for the significant gains in measurement validity. These decisions, unlike the opaque internal workings of a black-box model, are explicit and auditable, keeping the researcher in full control of the measurement process. Future research could further explore ways to systematize these choices, but as it stands, the ambiguity-robust dictionary offers a significant and practical step forward for political scientists seeking to conduct transparent, replicable, and contextually-aware analysis of text.

References

Aroyehun, S. T., Simchon, A., Carrella, F., Lasser, J., Lewandowsky, S., and Garcia, D. (2025). Computational analysis of us congressional speeches reveals a shift from evidence to intuition. *Nature Human Behaviour*, pages 1–12.

Bailey, C. M., Collins, Jr., P. M., Rhodes, J. H., and Rice, D. (2025). The Effect of Judicial Decisions on Issue Salience and Legal Consciousness in Media Serving the LGBTQ+ Community. *American Political Science Review*, 119(1):108–123.

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

Barrie, C., Palmer, A., and Spirling, A. (2024). Replication for language models problems, principles, and best practice for political science. *URL: https://arthurspirling. org/documents/BarriePalmerSpirling TrustMeBro. pdf*.

Bättig, M. B. and Bernauer, T. (2009). National institutions and global public goods: are democracies more cooperative in climate change policy? *International organization*, 63(2):281–308.

Baturo, A., Dasandi, N., and Mikhaylov, S. J. (2017). Understanding state preferences with text as data: Introducing the un general debate corpus. *Research & Politics*, 4(2):2053168017712821.

Baturo, A., Khokhlov, N., and Tolstrup, J. (2025). Playing the sycophant card: The logic and consequences of professing loyalty to the autocrat. *American Journal of Political Science*, 69(3):1180–1195.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Bonica, A. (2024). Database on ideology, money in politics, and elections: Public version 4.0 [computer file]. *URL: https://data. stanford. edu/dime*.

Brady, W. J., McLoughlin, K. L., Torres, M. P., Luo, K. F., Gendron, M., and Crockett, M. (2023). Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility. *Nature human behaviour*, 7(6):917–927.

Choi, D. D., Harris, J. A., and Shen-Bayh, F. (2022). Ethnic bias in judicial decision making: Evidence from criminal appeals in kenya. *American Political Science Review*, 116(3):1067–1080.

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature human behaviour*, 1(11):769–771.

Dillion, D., Puryear, C., Li, L., Chiquito, A., and Gray, K. (2024). National politics ignites more talk of morality and power than local politics. *PNAS nexus*, 3(9):pgae345.

Feal, A., Gleason, J., Goel, P., Radford, J., Yang, K.-C., Basl, J., Meyer, M., Choffnes, D., Wilson, C., and Lazer, D. (2024). Introduction to national internet observatory. In *Workshop Proceedings of the 18th International AAAI Conference on Web and Social Media*, page 73.

Gelman, A. and Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 348(1-17):3.

Gennaro, G. and Ash, E. (2022). Emotion and reason in political language. *The Economic Journal*, 132(643):1037–1059.

Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019). Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

Goertz, G. (2006). *Social science concepts: A user's guide*. Princeton University Press.

Häffner, S., Hofer, M., Nagl, M., and Walterskirchen, J. (2023). Introducing an interpretable deep learning approach to domain-specific dictionary creation: A use case for conflict prediction. *Political Analysis*, 31(4):481–499.

Haidt, J. et al. (2003). The moral emotions. *Handbook of affective sciences*, 11(2003):852–870.

Hill, K. Q. and Hurley, P. A. (2002). Symbolic speeches in the us senate and their representational implications. *Journal of Politics*, 64(1):219–231.

Jankin, S., Baturo, A., and Dasandi, N. (2017). United Nations General Debate Corpus 1946-2024.

Jankowski, M. and Huber, R. A. (2023). When correlation is not enough: Validating populism scores from supervised machine-learning models. *Political Analysis*, 31(4):591–605.

Khodak, M., Saunshi, N., Liang, Y., Ma, T., Stewart, B., and Arora, S. (2018). A la carte embedding: Cheap but effective induction of semantic feature vectors. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, Melbourne, Australia. Association for Computational Linguistics.

Latura, A. and Weeks, A. C. (2023). Corporate Board Quotas and Gender Equality Policies in the Workplace. *American Journal of Political Science*, 67(3):606–622. [Online; accessed 29. Apr. 2026].

Laver, M., Benoit, K., and Garry, J. (2003). Extracting policy positions from political texts using words as data. *American political science review*, 97(2):311–331.

Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., Apidianaki, M., and Callison-Burch, C. (2023). Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

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

Naftel, D., Green, J., Shoub, K., Edgerton, J., Wagner, M., and Cranmer, S. (2025). Meet the Press: Gendered Conversational Norms in Televised Political Discussion. *Journal of Politics*.

Neumayer, E. (2002). Do democracies exhibit stronger international environmental commitment? a cross-country analysis. *Journal of peace research*, 39(2):139–164.

Ornstein, J. T., Blasingame, E. N., and Truscott, J. S. (2025). How to train your stochastic parrot: Large language models for political texts. *Political Science Research and Methods*, 13(2):264–281.

Osnabrügge, M., Hobolt, S. B., and Rodon, T. (2021). Playing to the gallery: Emotive rhetoric in parliaments. *American Political Science Review*, 115(3):885–899.

Park, J. Y. (2023). Electoral rewards for political grandstanding. *Proceedings of the National Academy of Sciences*, 120(17):e2214697120.

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

Rodriguez, P. L., Spirling, A., and Stewart, B. M. (2023). Embedding regression: Models for context-specific description and inference. *American Political Science Review*, 117(4):1255–1274.

Simonsen, K. B. and Widmann, T. (2025). When Do Political Parties Moralize?: A Cross-National Study of the Use of Moral Language in Political Communication on Immigration. *British Journal of Political Science*, 55:e33.

Spirling, A. (2023). World view. *Nature*, 616:413.

Turpin, M., Michael, J., Perez, E., and Bowman, S. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Wang, S.-Y. N. and Inbar, Y. (2021). Moral-Language Use by U.S. Political Elites. *Psychological Science*.

Watanabe, K. (2021). Latent semantic scaling: A semisupervised text analysis technique for new domains and languages. *Communication Methods and Measures*, 15(2):81–102.

Wirsching, E. M., Rodriguez, P. L., Spirling, A., and Stewart, B. M. (2025). Multilanguage word embeddings for social scientists: Estimation, inference, and validation resources for 157 languages. *Political Analysis*, 33(2):156–163.

Wu, P. Y., Nagler, J., Tucker, J. A., and Messing, S. (2023). Large language models can be used to estimate the latent positions of politicians. *arXiv preprint arXiv:2303.12057*.

Xie, X. L. and Beni, G. (2002). A validity measure for fuzzy clustering. *IEEE Transactions on pattern analysis and machine intelligence*, 13(8):841–847.

Ying, L. (2026). Military Power and Ideological Appeals of Religious Extremists. *Journal of Politics*.

Zollinger, D. (2024). Cleavage identities in voters' own words: Harnessing open-ended survey responses. *American Journal of Political Science*, 68(1):139–159.

Appendix

Contents

A Dictionary Methods in Recent Political Science **A1**

B Environmental Politics **A2**
B.1 Data Collection and Cleaning Details A2
B.2 Full Correlation Matrix A3

C Kenyan Politics **A3**

D American Politics **A4**
D.1 Data Collection and Cleaning Details A4
D.2 Robustness Specifications A5
D.3 Moral Keyword Weights A8

A Dictionary Methods in Recent Political Science

Author	Title	Year	Journal	Role	Dictionary measures
Djourelova et al.	Media Attention and Strategic Timing in Politics	2022	AJPS	Primary	news anticipation vs. surprise
Fouka et al.	Collective Remembrance and Private Choice: German-Greek Conflict and Behavior in Times of Crisis	2022	APSR	Secondary	anti-Germany sentiment in speeches
Graham et al.	An Outbreak of Selective Attribution: Partisanship and Blame in the COVID-19 Pandemic	2022	APSR	Primary	Trump blame in survey responses
Kenwick et al.	You and Whose Army? How Civilian Leaders Leverage the Military's Prestige to Shape Public Opinion	2022	JOP	Secondary	tone of intervention speeches
Müller	The Temporal Focus of Campaign Communication	2022	JOP	Secondary	sentiment of manifesto text
Vallejo Vera et al.	The Politics of Interruptions: Gendered Disruptions of Legislative Speeches	2022	JOP	Primary	speech interruption type
Webster et al.	The Social Consequences of Political Anger	2022	JOP	Secondary	emotion words in survey recall
Ban et al.	How Are Politicians Informed? Witnesses and Information Provision in Congress	2023	APSR	Secondary	analytical content in testimony
Bucchianeri et al.	Legislative Effectiveness in the American States	2023	APSR	Primary	bill significance from newspaper coverage
Esberg et al.	How Exile Shapes Online Opposition: Evidence from Venezuela	2023	APSR	Primary	topic of Venezuelan activist tweets
Latura et al.	Corporate Board Quotas and Gender Equality Policies in the Workplace	2023	AJPS	Primary	gender equality attention in reports
Weiss et al.	How Threats of Exclusion Mobilize Palestinian Political Participation	2023	AJPS	Primary	policy threat salience in posts
Wlezien et al.	Media Reflect! Policy, the Public, and the News	2023	APSR	Primary	direction of budget news
Bailey et al.	The Effect of Judicial Decisions on Issue Salience and Legal Consciousness in Media Serving the LGBTQ+ Community	2024	APSR	Primary	article tone toward court decisions
Baturo et al.	Strategic Communication in Dictatorships: Performance, Patriotism, and Intimidation	2024	JOP	Secondary	rhetoric type in autocrat speeches

Continued on next page

(continued)

Author	Title	Year	Journal	Role	Dictionary measures
Carnegie et al.	Private Participation: How Populists Engage with International Organizations	2024	JOP	Primary	financial sentiment of IMF statements
Chávez	US Military Intervention and Presidential Communication Frames	2024	JOP	Primary	intervention frame in speeches
Fernandes et al.	Political Competition and the Effectiveness of Gender Quotas: Evidence from Portugal	2024	JOP	Primary	gender salience in manifestos
Hunt et al.	How Modern Lawmakers Advertise Their Legislative Effectiveness to Constituents	2024	JOP	Primary	effectiveness language in newsletters
Simmons et al.	Border Anxiety in International Discourse	2024	AJPS	Secondary	border sentiment in UN speeches
Wlezien	News and Public Opinion: Which Comes First?	2024	JOP	Primary	direction of economic news
Ban et al.	Local Orientation in the U.S. House of Representatives	2025	AJPS	Primary	local orientation in floor speeches
Bisbee et al.	'Yellin' at Yellen: Hostile Sexism in the Federal Reserve Congressional Hearings	2025	JOP	Secondary	toxicity in hearing utterances
Castanho Silva et al.	Blending in or Standing Out? Gendered Political Communication in 24 Democracies	2025	AJPS	Secondary	gendered language in parliamentary speeches
Hunter	Credit Claiming in the European Union	2025	JOP	Secondary	EU credit attribution in speeches
Naftel et al.	Meet the Press: Gendered Conversational Norms in Televised Political Discussion	2025	JOP	Secondary	uncivil language in talk shows
Vishwanath	Race, Legislative Speech, and Symbolic Representation in Congress	2025	AJPS	Primary	race-related speech in Congress
Hartmann	Do Voters Value Relief over Preparedness? Evidence from Disaster Policies in Malawi	2026	JOP	Primary	preparedness vs. relief in coverage
Rehmert	Intraparty Competition, Geographic Responsiveness, and Incumbent Deselection in Closed-List Proportional Representation	2026	JOP	Secondary	local place references in questions
Testa et al.	Does the Messenger Shape the Message's Effect? Race, Black Lives Matter, and the Efficacy of Social Movement Messages	2026	JOP	Secondary	sentiment of BLM responses
Ying	Military Power and Ideological Appeals of Religious Extremists	2026	JOP	Primary	religious vs. secular rhetoric
Yoon	Anger Expressions and Coercive Credibility in International Crises	2026	AJPS	Secondary	anger expressions in crisis statements
Weiss et al.	Outgroup Avoidance	2026	JOP	Secondary	intergroup conflict in Facebook posts

B Environmental Politics

B.1 Data Collection and Cleaning Details

We merge UNGD speeches to V-Dem country-year observations on `country_text_id` and `year`, retaining the Boix–Miller–Rosato dichotomous regime indicator (`e_boix_regime`) as the regime measure. Country-years missing either the speech text or the regime score are dropped. Raw speech text is transliterated to Latin-ASCII, lowercased, stripped of non-alphabetic characters, and cleaned of residual HTML/typographic artifacts, underscores, and embedded newline/tab characters. The cleaned corpus is tokenized, removing punctuation, symbols, numbers, and separators. We then drop English stopwords and any token of fewer than three characters. The remaining vocabulary is restricted to features that occur at least 10 times overall and in at least 0.5% of documents; tokens outside this vocabulary are replaced with padding rather than deleted, so that non-adjacent words are not artificially brought into one another's context windows. We

then measure climate attention in each tokenized speech using the dictionaries enumerated in Table A2.

Baseline	**Noisy**	**Ambiguous**
biodiversity, climate, conservation, deforestation, desertification, ecosystem, emission, emissions, environment, environmental, forests, fossil, gases, greenhouse, habitat, mitigation, ocean, oceans, preservation, reforestation, renewable, sustainability, watershed, wildlife, forestation, pollution, recycling, afforestation, agroecology, carbon, co2, contamination, decarbonization, degradation, ecological, ecology, effluent, effluents, environmentalism, greenwashing, landfill, marine, overfishing, renewables, restoration, vegetation, warming, weather	Baseline, plus trade, deficit, development, economic, globalization	Baseline, plus sustainable, green, natural, clean, resources

Table A2: Climate dictionary keywords

B.2 Full Correlation Matrix

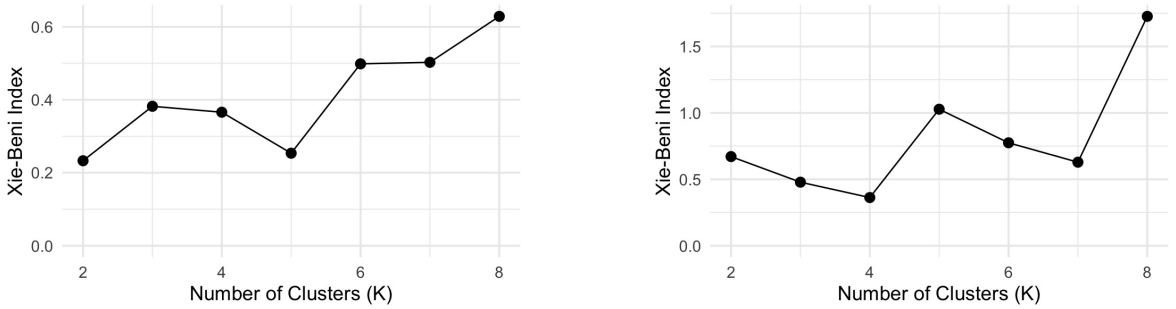
	Base.	Amb.	Noise	Base. (Adj)	Amb. (Adj)	Noise (Adj)
Baseline	-					
Ambiguous	0.91	-				
Noise	0.68	0.73	-			
Baseline (Adj)	0.98	0.91	0.67	-		
Ambiguous (Adj)	0.94	0.98	0.72	0.95	-	
Noise (Adj)	0.82	0.82	0.93	0.84	0.85	-

Table A3: Correlation matrix between different dictionary and method choices (Proportions based on ambiguity-robust (adjusted) counts vs. raw counts)

C Kenyan Politics

We re-estimate the substantive findings from Choi et al. (2022) using two sentiment dictionaries from EmoLex (Mohammad and Turney, 2013), focusing on the categories of trust and disgust. As anchor words, we select *trust*, *truth*, and *truthful* for trust, and *dirty*, *gross*, and *sick* for disgust. Although these terms appear infrequently in the corpus, they are closely tied to the intended concepts in judicial language and thus provide meaningful anchors. Because the EmoLex dictionaries include many entries that are contextually irrelevant or semantically ambiguous in legal texts, we adopt a data-driven procedure to determine the number of clusters. Figure A1 presents the

Figure A1: Selection of the Number of Clusters for Trust and Disgust Dictionaries



(a) Trust

(b) Disgust

results of this procedure, suggesting a small number of clusters is enough to capture the concepts we intend to analyze.

Table A4 replicates the main findings, contrasting a traditional dictionary word count with our context-aware measurement strategy. The results underscore that methodological choices are critical for detecting bias in judicial texts. Columns 2 and 4, which use our ambiguity-robust dictionary, reproduce the substantive conclusions of Choi et al. (2022), while the traditional dictionary in Columns 1 and 3 yields weaker results for trust and contradictory results for disgust. Our approach preserves the transparency of dictionary methods while incorporating the semantic advantages of embedding-based techniques, thereby avoiding the shortcomings of both.

Table A4: Replication of Choi et al. (2022) using dictionaries methods

	Trust Proportion of Words		Disgust Proportion of Words	
Coethnic match	0.0012*	0.0015***	-0.0014**	0.0000
	(0.0007)	(0.0006)	(0.0006)	(0.0002)
Courthouse-year FE	Yes	Yes	Yes	Yes
Individual judge FE	Yes	Yes	Yes	Yes
Case-specific controls	Yes	Yes	Yes	Yes
Number of Clusters	Count	K = 2	Count	K = 4
Observations	9,545	9,545	9,545	9,545
R^2	0.22124	0.24366	0.23975	0.17353

Clustered (courthouse-year) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

D American Politics

D.1 Data Collection and Cleaning Details

We collect floor speeches from the Gentzkow–Shapiro–Taddy Hein-Daily release for the 97th–114th Congresses, extended through the 118th Congress with comparable text from the Congressional Record. For each Congress we merge the speech, description, and speaker-map files, retain speeches longer than 200 words, and restrict the sample

to Democratic and Republican speakers. The speeches are then assembled into a single corpus and tokenized with punctuation and numbers removed, lowercased, with English stopwords and a custom list of Congressional procedural terms (e.g., mr, mrs, speaker, chairman, gentleman, yield, hereby, thereto) dropped to suppress floor-management boilerplate.

The corpus uses two incompatible speaker ID systems — Hein-Daily speakerids for the 97th–114th Congresses (which embed the Congress number, so a member who served across boundaries receives different IDs in different terms) and Bioguide-style identifiers for the 115th–118th. To follow members across Congresses we build a name-based key of the form first—last—state—chamber (lowercased and whitespace-squished), populated from Hein-Daily SpeakerMap files for the early period and from the unitedstates/congress-legislators Bioguide reference for the later period. This key serves as the stable speaker identifier in all within-speaker fixed-effects specifications. Speakers are linked to Bonica's DIME recipient files (1979–2022) via the identical name-state-chamber key, with donations aggregated to the (legislator, election cycle) level.

Negative moral language is identified using the Moral Foundations Dictionary (MFD): we keep the vice (negative-pole) entries from the foundation pairs plus a manually overridden set of unambiguously negative items (e.g., immoral, evil, offend, transgress, wrong). We estimate two-, three-, and four-cluster solutions and use $k = 2$ for the main results chosen using the xie-bini criterion. (k=2, XB = 0.49; k=3, XB = 0.69; k=4, XB=0.51).

D.2 Robustness Specifications

Table A5: Individual contributions vs. moral language in floor speeches, controlling for ideological extremity ($|cfscore|$). Coefficients are change in log(individual contributions) per additional moral keyword per speech. t = election cycle - 2000.

Dependent Variable:	log(Indiv. contribs.)					
	Robust	Unadj.	Robust $\times t$	Unadj. $\times t$	Within (rob.)	Within (unadj.)
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Moral lang. (robust)	0.16***		0.007		0.02	
	(0.05)		(0.07)		(0.04)	
$\|cfscore\|$	0.21	0.18	0.22	0.19	0.21	0.20
	(0.20)	(0.20)	(0.20)	(0.20)	(0.42)	(0.42)
Moral lang. (raw)		0.11***		0.03		0.04*
		(0.03)		(0.03)		(0.03)
Moral lang. (robust) $\times$ t			0.02***			
			(0.005)			
Moral lang. (raw) $\times$ t				0.01***		
				(0.002)		
Fixed-effects						
Pol. controls	Yes	Yes	Yes	Yes		
Cycle	Yes	Yes	Yes	Yes	Yes	Yes
Legislator					Yes	Yes
Fit statistics						
Observations	8,815	8,815	8,815	8,815	8,815	8,815

Clustered (Legislator) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

A5

Table A6: Individual contributions vs. robust moral-language measure across cluster solutions ($K \in \{2, 3, 4\}$). t denotes election cycle centered at 2000.

Dependent Variable:	log(Indiv. contribs.)					
	K=2	K=2 × t	K=3	K=3 × t	K=4	K=4 × t
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Moral lang. (robust)	0.18***	0.04	0.20***	0.02	0.20***	-0.02
	(0.05)	(0.07)	(0.07)	(0.09)	(0.08)	(0.11)
Moral lang. (robust) × t		0.02***		0.03***		0.04***
		(0.005)		(0.007)		(0.009)
Fixed-effects						
Pol	Yes	Yes	Yes	Yes	Yes	Yes
Cycle	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	8,891	8,891	8,891	8,891	8,891	8,891

Clustered (stable_id) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

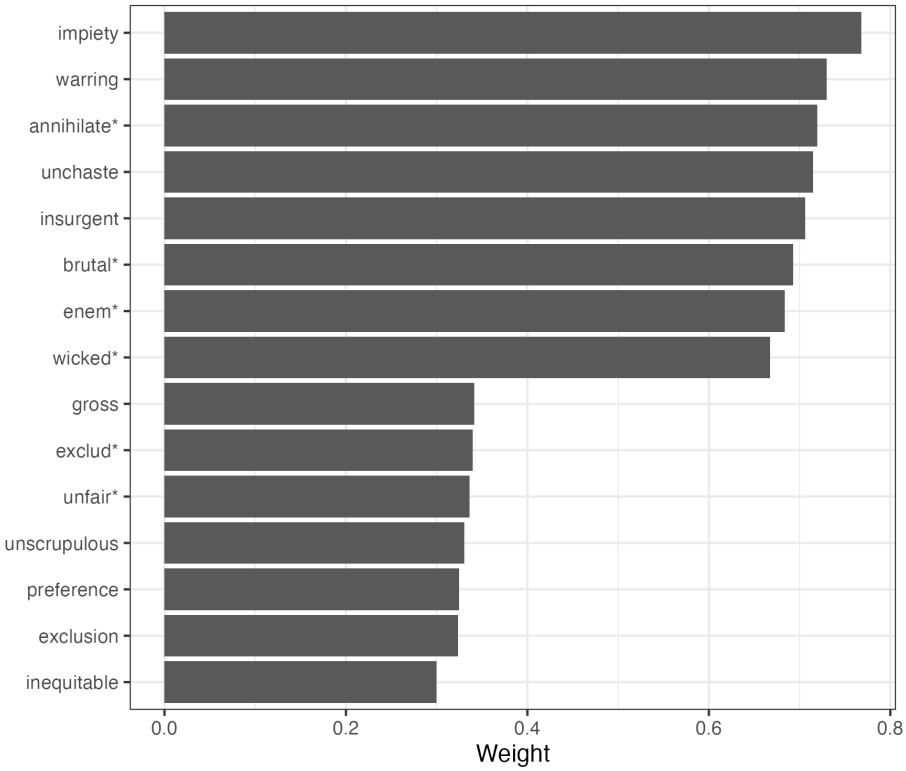


Figure A2: Top and bottom eight moral keywords by average weight. Lower weight keywords are less likely to be counted toward the target concept. Our method successfully assigns less ambiguous words higher weight ("wicked," "brutal," "enemy") and more ambiguous words lower weight ("preference," "exclusion").

Table A5 shows results including a control for absolute legislator ideology measured by cfscore. The results are substantively similar to the main specification, indicating that legislator extremity does not confound the relationship between moral language use and received donations.

Table A6 shows results of the cross-sectional analyses for different cluster solutions. $k = 2$ was chosen as the main specification for parsimony and because it yielded the optimal Xie-Beni index. The table suggests that the results are robust to choice of $k \in [2, 3, 4]$.

D.3 Moral Keyword Weights

Beyond aggregate trends, the word-level weights produced by our algorithm clarify which moral vocabularies are most consistently associated with the anchor concept of "immorality," and how these associations differ across parties. Figure A3 shows that members of both parties show sharp increases in negative moral language around words such as terrorist and attack following September 11, reflecting heightened attention to harm and threat. Over time, Republicans place greater weight on terms such as illegal and enemy, emphasizing rule violation, boundary enforcement, and perceived threats to social order. Democrats, by contrast, emphasize terms such as discriminate, violence, sick, and suffering, reflecting greater attention to harm, vulnerability, and injustice.

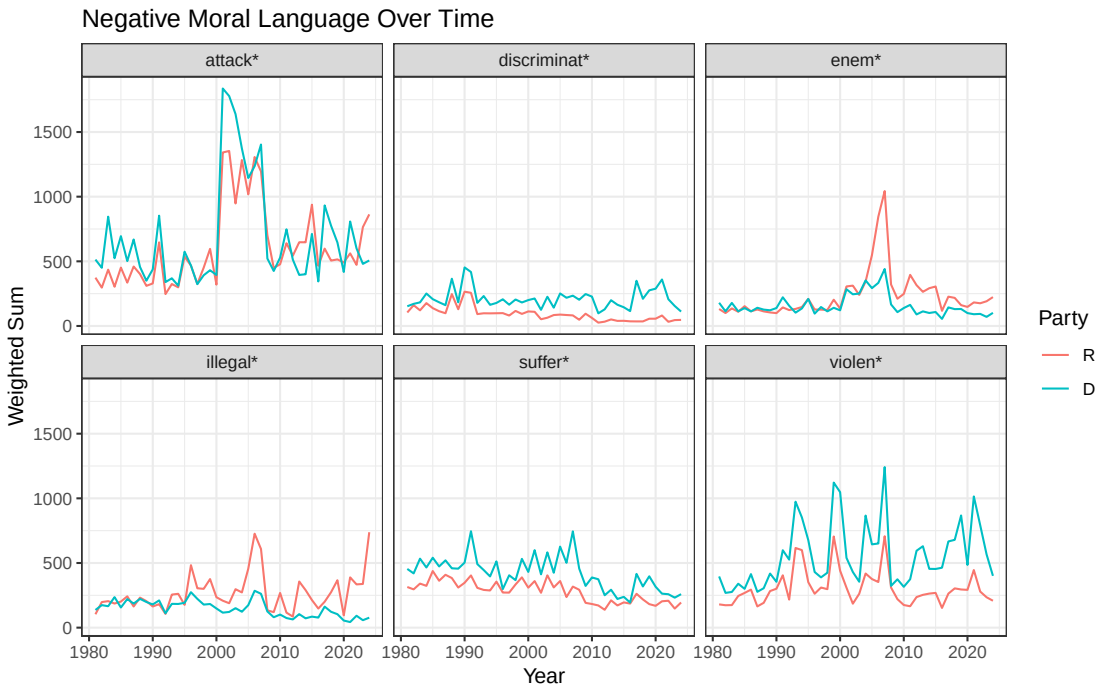


Figure A3: Negative moral language keyword usage over time, regularized using our embeddings-clustering procedure.