Adjusting Interpretable Dimensions in Embedding Space with Human Judgments

Katrin Erk
University of Texas at Austin
katrin.erk@utexas.edu

Marianna Apidianaki
University of Pennsylvania
marapi@seas.upenn.edu

Abstract

Embedding spaces contain interpretable dimensions indicating gender, formality in style, or even object properties. This has been observed multiple times. Such interpretable dimensions are becoming valuable tools in different areas of study, from social science to neuroscience. The standard way to compute these dimensions uses contrasting seed words and computes difference vectors over them. This is simple but does not always work well. We combine seed-based vectors with guidance from human ratings of where words fall along a specific dimension, and evaluate on predicting both object properties like size and danger, and the stylistic properties of formality and complexity. We obtain interpretable dimensions with markedly better performance especially in cases where seed-based dimensions do not work well.

1 Introduction

Properties are commonly used in linguistics (Katz and Fodor, 1964; Jackendoff, 1990; Wierzbicka, 1996) as well as in psychology (Murphy, 2002) for representing word meanings and concepts. Those same properties are discoverable as *interpretable dimensions* in word embedding space, and can be used to explore the patterns and regularities encoded by Large Language Models (LLMs) (Mikolov et al., 2013b; Bolukbasi et al., 2016). Because LLMs are trained on texts from many different authors, we can view them as a compact repository of human utterances. This makes them an interesting resource for studying linguistic phenomena, analyzing social contexts of words, or as a stand-in for conceptual knowledge for interpreting brain voxels. Interpretable dimensions provide an attractive and simple way to access this resource (Grand et al., 2022; Kozlowski et al., 2019; Garí Soler and Apidianaki, 2020; Lucy et al., 2022). Compared to probing (Tenney et al., 2019; Conneau et al., 2018), interpretable dimensions allow for a direct exploration of LLM embedding space, without external classifiers.

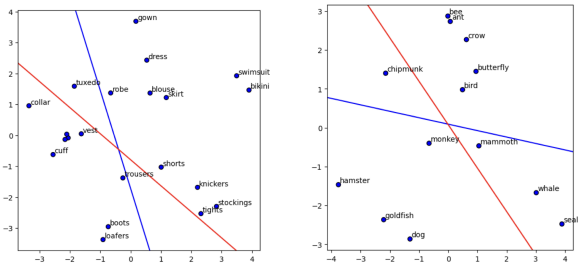

Figure 1: Interpretable dimensions for two object categories and features from Grand et al. (2022): clothes by wealth (left), animals by size (right). PCA projection of embeddings, with seed-based (blue) and FIT+S (red) dimensions.

The most common way to obtain interpretable dimensions is to specify some seed pairs of antonyms, and take the average over their vector differences. But it is unclear what makes good seed pairs, or even how to test whether a particular property corresponds to a discernible dimension in embedding space. Antoniak and Mimno (2021) and Lucy et al. (2022) express concerns about the quality of commonly used hand-curated seed lexicons and propose metrics for evaluating seeds.

In this paper, we take a different approach to addressing the problem of "bad seeds" (Antoniak and Mimno, 2021): We propose a method to augment seed-based interpretable dimensions with additional guidance from human ratings, and we show that this augmentation is particularly impactful when the original seed-based dimensions are problematic.[1] Figure 1 shows words for clothes with dimensions for wealth, and animal names with dimensions for size, blue for seed-based dimensions and red for our new fitted dimensions. The fitted dimensions correct overly high wealth estimates for

[1]Our code and data are available at https://github.com/mariannaapi/interpretable-dimensions.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2675–2686

tights and *stockings*, and exaggerated size estimates for *bee* and *ant*.

While interpretable dimensions are useful both to social science and to cognitive science, there is an important difference between the fields: In social science, crowd-sourced datasets cannot be trusted in absolute terms, because annotators may have social biases of their own; this is the scenario that Antoniak and Mimno (2021) address. In cognitive science however, experimental data from human participants is central (though the method used to solicit data can have an influence on the outcome). We work with data from cognitive science here, so we can use human ratings to improve seed dimensions.

Our new method draws inspiration from a completely separate strand of research on interpretable dimensions, in the context of knowledge graph embeddings (Derrac and Schockaert, 2015; Jameel and Schockaert, 2016; Bouraoui et al., 2022). There, interpretable dimensions are learned using labeled training data. In the current paper, we use a similar learning strategy, and apply it to a combination of seed-based dimensions and labeled training data. We apply this technique to predict human ratings on object properties and stylistic aspects of words, and find that it improves performance particularly in cases where seed-based dimensions underperform, and that in contrast to seed-based dimensions it is able to make predictions at the same scale as the original ratings.

A larger issue is: When can we trust that an interpretable dimension shows us what the LLM truly "knows" about the property in question, that we are not misled by noise in our tool? This same worry also arises for probing classifiers (Hewitt and Liang, 2019; Belinkov, 2022).[2] We take one step towards addressing this issue: By combining two sources of information, seeds and human annotation, we hope to reduce noise present in either source.

2 Related Work

Interpretable dimensions in word embedding space have first been observed in NLP (Bolukbasi et al., 2016), and the idea was then taken up in neuroscience and social science. Grand et al. (2022) discover dimensions for objects' scalar properties (e.g., DANGER, SIZE, SPEED, WEALTH). Kozlowski et al. (2019) identify dimensions including AFFLUENCE, GENDER, RACE and MORALITY, and show that concepts such as sports (e.g., *golf, boxing*) and music genres (e.g., *opera, rap, jazz*) are ordered along these axes in ways that match cultural stereotypes. Garg et al. (2018) explore ethnic stereotypes, and Stoltz and Taylor (2021) go as far as to propose a cultural cartography with word embeddings. An et al. (2018) use a large number of dimensions to characterize sentiment, which Kwak et al. (2021) apply to whole documents. Interpretable dimensions have also been used to represent linguistic notions, such as complexity and scalar adjective intensity (Garí Soler and Apidianaki, 2020, 2021b; Lyu et al., 2023). In our work, we explore dimensions addressing object properties in the Grand et al. (2022) datasets, and the abstract notions of formality and complexity.

In all these studies, dimensions are discovered using the seed-based methodology, where a few seed pairs of antonyms are specified and the dimension is computed as the average over vector differences for these pairs. This method is simpler than alternative representation approaches (e.g., the multi-task learning framework of Allaway and McKeown (2021)).

Seed pair selection has until now been ad hoc; but some choices, such as the selected word pairs, their number and order, and the way they are combined, have a strong impact on the quality of the derived dimension. Antoniak and Mimno (2021) address the "bad seeds" problem by measuring the *coherence* of each seed set pairing after mapping to the bias subspace: When all words in the vocabulary are projected onto the subspace, the two seed sets must be drawn as far apart as possible. Lucy et al. (2022) propose to measure the semantic axis' *self-consistency* using a leave-one-out approach, where each seed is compared to an axis constructed from the remaining seeds. A good seed, when left out, should be closer to the pole it belongs to.

In our approach, we do not test for seed quality. Instead, we use human ratings to improve on seed-based dimensions. Our approach is inspired by work on knowledge graph embeddings (Derrac and Schockaert, 2015; Jameel and Schockaert, 2016; Bouraoui et al., 2020). Drawing on the conceptual spaces of Gärdenfors (2014) for intuition,

[2]There is a related question of whether information encoded in the model is also relevant for downstream performance (Ravichander et al., 2021; Lyu et al., 2024). This question is not so central for linguists or cognitive scientists interested in knowledge reflected in an LLM. What is central for them is that the dimension really encodes the property of interest.

Jameel and Schockaert (2016) learn embeddings of knowledge graph nodes that include interpretable dimensions for properties. Like us, they learn interpretable dimensions using labeled training data. Our objective function is an adaptation of their objective function, but still different as they also learn the space while we have a fixed space.

For constructing interpretable dimensions, most previous work used static embeddings (GloVe (Pennington et al., 2014) and word2vec (Mikolov et al., 2013a)). Recent work extends the methodology to contextualized representations (Garí Soler and Apidianaki, 2020; Lucy et al., 2022). We experiment with both kinds of embeddings.

3 Methods

3.1 Models

Seed-based dimensions (SEED model). The seed-based method is the most commonly used for computing interpretable dimensions (Bolukbasi et al., 2016; Kozlowski et al., 2019; Dev and Phillips, 2019; Garí Soler and Apidianaki, 2021b; Grand et al., 2022). A group of *seed words* are chosen which represent opposite ends of the dimension. For the DANGER dimension in Grand et al. (2022), for example, the seeds are {*safe, harmless, calm*} for the positive side and {*dangerous, deadly, threatening*} for the negative side. For each pair of a positive and negative seed word p, n with vectors $\vec{p}, \vec{n}$, the difference vector $\vec{p} - \vec{n}$ is computed; this is a first estimate of the interpretable dimension, but the vectors can differ substantially across seed pairs. To obtain a more stable estimate, the vector for the interpretable dimension is then computed as the average of the difference vectors from individual seed pairs. The rating of any word a on the property d with interpretable dimension $\vec{d}$ – in our example from above, DANGER – is then predicted as the scalar projection onto the dimension:

$$||\text{proj}_{\vec{a}}(\vec{d})|| = \frac{\vec{a} \cdot \vec{d}}{||\vec{d}||}$$

Fitted dimensions (FIT model). Whenever we have gold ratings on some dimension, like human judgments on degrees of danger of different animals (Grand et al., 2022) or gold ratings for complexity (Lyu et al., 2023), we can estimate a direction in embedding space that best matches the gold ratings. We adapted an idea from Jameel and Schockaert (2016), who learn an embedding space for knowledge graph nodes in such a way that properties of the nodes correspond to dimensions in space. But rather than learning a new space, we need to use an existing space spanned by static or contextualized embeddings, because it is these spaces, and the patterns in human language use that they encode, that we want to analyze.

We proceed as follows. Let $W = \langle w_1, \ldots w_n \rangle$ be an annotated dataset of n words with real-valued gold ratings $\hat{Y} = \langle \hat{y}_1, \ldots, \hat{y}_n \rangle$ for some feature f. Let $\vec{w}_i$ be the embedding of word w_i. For the dimension $\vec{f}$ to be computed for feature f in that same embedding space, we stipulate that the scalar projection of $\vec{w}_i$ onto $\vec{f}$ be proportional to the gold rating $\hat{y}_i$. For example, say the gold rating (average human rating) of *dolphin* on the DANGER scale (on a scale of 1-5) is 2.1, and the gold rating of *tiger* is 4.9. Then we want the length of the projection $\text{proj}_{\vec{\text{dolphin}}}(\vec{\text{DANGER}})$ to be proportional to $c_{\text{DANGER}} \cdot 2.1 + b_{\text{DANGER}}$, and the length of the projection $\text{proj}_{\vec{\text{tiger}}}(\vec{\text{DANGER}})$ to be proportional to $c_{\text{DANGER}} \cdot 4.9 + b_{\text{DANGER}}$, for some weight and bias constants $c_{\text{DANGER}}, b_{\text{DANGER}} \in \mathbb{R}$. So in general, we would like to have

$$\frac{\vec{w}_i \cdot \vec{f}}{||\vec{f}||} = c_f \, \hat{y}_i + b_f$$

We turn this into a loss function for computing a *fitted dimension* f, dropping the denominator $||\vec{f}||$:

$$J_f = \sum_{w_i} \left(\vec{w}_i \cdot \vec{f} - c_f \, \hat{y}_i - b_f \right)^2$$

Fitted dimensions with seed words (FIT+SW model). We also test whether fitted dimensions can be guided by the seed words used to make seed-based dimensions. The first method follows the intuition of Antoniak and Mimno (2021) that the scalar projections of seed words should sit "far out" on an interpretable dimension, further than other words. The FIT+SW model simply extends the collection W of rated words by the seed words. We make synthetic gold ratings for the seedwords, giving them extreme ratings: $\max(\hat{Y}) + o$ for positive seed words, and $\min(\hat{Y}) - o$ for negative seedwords, for an offset o that is a hyperparameter. We optionally add a small amount of random jitter (sampled from the interval $[0.001, 0.005]$) so that the seed words don't all have the same rating.

Fitted dimensions with seed dimensions (FIT+SD model). Our second way of extending fitted dimensions with seed word information is built on

the idea that seed-based dimensions and human ratings both provide useful information for fitting an interpretable dimension, and that they should be combined. So we use an overall loss function of

$$J = \alpha J_f + (1 - \alpha) J_d(D)$$

where J_f is the loss function from above, and α is a hyperparameter. $J_d(D)$ is a loss that measures distance of the fitted dimension $\vec{f}$ from a set D of seed-based dimensions, defined as

$$J_d(D) = \sum_{d \in D} 1 - cosine(\vec{d}, \vec{f})$$

Fitted dimensions with seeds as words and dimensions (FIT+S model). Our final model uses seeds both as seed words, as in FIT+SW, and as seed dimensions, as in FIT+SD.

Baselines. We compare our methods to a baseline which ranks words by frequency (FREQ). Frequency has been a strong baseline for complexity and formality in previous work, given that rare words tend to be more complex than frequently used words (Brooke et al., 2010). We use log-transformed frequency counts in the Google N-gram corpus (Brants and Franz, 2006). We also use a random baseline, which assigns to each word a randomly selected score in the range [-3, 3].[3]

3.2 Evaluation metrics

In contrast to interpretable dimensions computed from seed words, the FIT models use training data: words with human ratings for the property in question. When we evaluate these models, we use up some of the data for training, leaving less for testing. To mitigate the issue, we do cross-validation, and we focus on evaluation metrics that work well with smaller test sets. We do not use the correlation metrics used in Garí Soler and Apidianaki (2020) and Grand et al. (2022), as significance tests become unreliable with small datasets. Instead, we use a variant of pairwise rank accuracy, a metric used in Garí Soler and Apidianaki (2020), Cocos et al. (2018), and Grand et al. (2022).

Pairwise rank accuracy measures the percentage of pairs of words whose ordering in the gold ratings is the same as in the model predictions. We define a new variant which we call **extended pairwise rank accuracy, r$^+$-acc**, which measures pairwise rank accuracy among words in the test set, and additionally pairwise rank accuracy between each test word and each training word. For example, if *tiger* and *butterfly* are in the training set for DANGER, and *cat* is in the test, we check whether the score assigned to *cat* ranks it after *tiger* and before *butterfly*. This metric gives us more evidence on the quality of predictions than pairwise rank accuracy on its own because it includes more comparisons, thus making the metric less sparse. Let $<_g, <_m$ be two complete orderings of the words in W, the gold and model orderings, respectively. For words w, w' in W, we define an auxiliary function *rm* for "rank match":

$$rm_{<_g,<_m}(w, w') = \begin{cases} 1 \text{ iff } (w <_g w' \wedge w <_m w') \\ \quad \vee (w >_g w' \wedge w >_m w') \\ 0 \text{ else} \end{cases}$$

Then standard pairwise rank accuracy on $W = \langle w_1, \ldots, w_n \rangle$ is defined as

$$\text{r-acc}_W(<_g, <_m) = \frac{1}{n(n-1)} \sum_{1 \le i < j \le n} rm_{<_g,<_m}(w_i, w_j)$$

Now assume that $T = \{k_1, \ldots, k_\ell\}$, with $k_j \in \{1, \ldots, n\}$ for all j, is the set of test word indices among the indices of W. Assume both orderings, $<_g$ and $<_m$, are defined on all of W. Then our new extended pairwise rank accuracy is

$$\text{r}^+\text{-acc}_W(<_g, <_m) = \frac{1}{\ell(\ell-1)+\ell(n-\ell)} \sum_{1 \le i < j \le \ell} rm_{<_g,<_m}(w_{k_i}, w_{k_j}) + \sum_{i \in \{1,\ldots,\ell\}, j \notin T} rm_{<_g,<_m}(w_{k_i}, w_j)$$

The first half of this formula measures pairwise rank accuracy among members of the test set; the second half measures rank accuracy of test words with respect to training words.

Pairwise rank accuracy and extended pairwise rank accuracy are similar to correlation metrics in that they measure to what extent gold and model-based rankings agree. And in fact all three metrics are highly correlated: We tested correlation between the three metrics on seed-based dimensions for the Grand et al. (2022) data and obtained highly significant correlations ($p \ll 0.0001, r = 0.972$ for pairwise rank accuracy, $r = 0.971$ for extended pairwise rank accuracy).[4]

[3] This range was chosen because all gold ratings in our study are z-scores.

[4] To measure extended pairwise rank accuracy, the data was split into training and test folds in 5-fold cross-validation, and r$^+$-acc scores were averaged over folds. It is not surprising that the values are almost the same, as in this case extended pairwise rank accuracy is almost the same as standard pairwise rank accuracy except that the latter omits some pairwise comparisons, namely between training data points.

As a second evaluation metric, we test how far off from the gold ratings each individual predicted rating is. We use the **mean square error (MSE)** of predicted ratings compared to gold ratings. We can do this because all FIT models learn to predict ratings on the same scale as the gold ratings. In order to apply the same evaluation to the SEED model and the baselines, we simply use linear regression to map model predictions to ratings on the same scale as the gold ratings. Linear regression models are fit on the training portion of each data set, so that test words remain truly unseen.

3.3 Data and Vectors

We use the ratings collected by Grand et al. (2022) which describe properties of objects in nine categories:[5] animals, clothing, professions, weather phenomena, sports, mythological creatures, world cities, states of the United States, and first names.[6] Each category is matched with a subset of these semantic features: age, arousal, cost, danger, gender, intelligence, location (indoors vs. outdoors), partisanship (liberal vs. conservative), religiosity, size, speed, temperature, valence, volume, wealth, weight, and wetness. For style, we use datasets released by Pavlick and Nenkova (2015) which contain words and phrases with human ratings of formality and complexity. For each dimension, we sample words[7] with high annotation confidence (i.e. where annotators agreed about the word being complex or formal): We calculate the mean standard deviation for words in our sample, and keep words where deviation between human scores is lower than that mean. The filtered datasets contain 1,160 words for complexity, and 1,274 words for formality.

We extract seed words from two other datasets released by Pavlick and Nenkova (2015) which contain pairwise paraphrase judgments of formality and complexity.[8] Annotations reflect which phrase in a pair (e.g., *letter-communication*, *largely-extensively*) is more complex or formal than the other. We collect five pairs of words for each style dimension for which inter-rater agreement is high. For complexity, we obtain the seed pairs *work - employment, further - subsequently, strong - powerful, train - railway, shown - indicated*, where the first member of each pair is the negative seed (the simpler word). For formality, we used *winner - recipient, terrible - disastrous, membership - affiliation, highest - paramount, test - verify*, where again the first member of each pair is the negative seed (the less formal word).

Following Grand et al. (2022), we averaged over human subject ratings for each datapoint, then normalized ratings to z-scores separately for each pair of a category and property.[9] For formality and complexity, ratings were also converted to z-scores.

As embeddings, we use the same representations as Grand et al. (2022), pre-trained GLoVE embeddings trained on the Common Crawl (42b tokens), 300 dimensions, uncased (Pennington et al., 2014). We also use contextualized representations from the BERT (`bert-large-uncased`) and RoBERTa (`roberta-large`) models (Devlin et al., 2019; Liu et al., 2019) with sentences from UkWac (Baroni et al., 2009).[10] For each word instance, we average its contextualized representations from the top 4 layers of the model.[11] If the word is split into multiple wordpieces during tokenization, we average the representations of its pieces in order to obtain a single type-level representation for each word, as is common practice in semantic probing studies (Bommasani et al., 2020; Vulić et al., 2020; Garí Soler and Apidianaki, 2021a). The final representation for a word is the average of its representations from the available sentences. Aggregating representations across multiple contexts is the most common approach for creating word type-level embeddings from contextualized representations which serves to null out, to some extent, the impact of specific contexts (Apidianaki, 2022). It is possible to use more sophisticated ways for sentence selection, such as language modeling criteria (Garí Soler and Apidianaki, 2020) and exclusion of contexts where antonyms could occur (Lucy et al., 2022). However applying such sophisticated context selection methods is not always

[5]The data is available on the Open Science Framework at `https://osf.io/5r2sz/`. No license information is given in the repository.

[6]Most categories consist of 50 items.

[7]We sample words with more than three characters to exclude pronouns, articles, numerals, and multiword phrases.

[8]The data is available at `https://cs.brown.edu/people/epavlick/data.html`, under "Style Lexicons: Human and automatic scores of formality and complexity for words, phrases, and sentences". No license for the data is given.

[9]We would expect the data to show subjective differences between annotators, and in the future we would like to model the subjective ratings directly, following Plank et al. (2014).

[10]Details about sentence selection are given in the Appendix.

[11]Averaging across layer subsets is generally better than averaging across all layers or selecting a single layer (Vulić et al., 2020).

better than random selection, which might be due to the skewed distribution of word senses and the stronger presence of the most frequent sense of a word in randomly selected sentences (Kilgarriff, 2004; McCarthy et al., 2004).

4 Results and discussion

In this section we evaluate the different interpretable dimension models from Section 3.1 on the tasks of predicting human ratings of object properties (Grand et al., 2022), and human ratings of the complexity and formality of words (Pavlick and Nenkova, 2015).

Experimental setup. To make the most of the limited available data, all models were tested in 5-fold cross-validation. In addition, all models that involve randomness (all except SEED) were re-run three times with different random seeds. For Grand et al. object features, we first compute mean r^+-*acc* and median MSE for each category/property pair (averaging over cross-validation runs and random seeds), then we report averages over those. For formality and complexity we report overall mean r^+-*acc* and median MSE.[12] Note that because we split the data into training and test using cross-validation, the numbers reported in this paper are not comparable with those reported in earlier papers on the same dataset. We do however compute SEED dimensions with the same cross-validation setup as the FIT-based dimensions, so that the numbers that we report are comparable to each other.

To set hyperparameters, we sample 6 category/property pairs from the Grand et al. data as development set. Hyperparameters were optimized once per embedding space; there was no separate hyperparameter optimization for the formality and complexity data.[13] Overall, we find that low values of α work well, and that it is beneficial to input a single averaged seed dimension to FIT+SD and FIT+S, rather than individual seed dimensions. The choice of offset and jitter does not matter. More details on hyperparameters and the development set can be found in the Appendix. Results reported below for Grand et al. data are for all category/property pairs not in the development set.

Overall performance. Overall results are shown for object properties in Table 1 and for stylistic features in Table 2. Looking at object properties first, and focusing on extended rank accuracy, the FIT model by itself is not very good, and adding seeds as words (FIT+SW) does not help. FIT+SD is better, and outperforms SEED slightly, but the FIT+S model, which computes fitted dimensions using seed information both as words and dimensions, shows the best performance, outperforming SEED strongly. In terms of MSE, even medians are very high for the SEED model, so many seed-based dimensions were not able to predict ratings on the same scale as the gold ratings. MSE is much lower for both fitted models that make use of seed dimensions, especially FIT+S.[14] Looking at the baselines, the FIT and FIT+SW models with BERT and RoBERTa underperform the frequency baseline, and are on par with random guessing. The frequency baseline is somewhat higher than random, though it is not entirely clear what kind of signal for object properties can be derived from word frequency.

On the stylistic data, the relative performance of the fitted models is similar, but here they mostly do not outperform the SEED dimensions. Overall performance of the SEED dimensions is higher here, which raises the question if fitted models help in particular when seed-based dimensions do not perform well; we explore this further below. Looking at MSE, we confirm that fitted models that use seed dimensions have much lower error than the other models. Comparing embedding spaces, we see consistently the best performance with GLoVE embeddings. The BERT and RoBERTa FIT and FIT+SW models in particular are again at the level of the random baseline. The frequency baseline is reasonably strong, matching previous findings.

Fitted dimensions, by themselves, are underdetermined by human ratings. The FIT model, which computes fitted dimensions from human ratings only, does not perform well, and we suspect that the size of the embedding space allows for too many ways to fit a dimension to ratings, causing the model to overfit. To test this, we first computed FIT dimensions, for Grand et al. object features, from *all* human ratings, obtaining perfectly fit dimensions in every single case. We next train dimensions on all human ratings but scramble the word/rating pairs, making them nonsensical. Again we obtain perfectly fit dimensions in every single

[12] We use median MSE because outliers make the mean uninformatively high.

[13] Human ratings on all datasets were on the same scales as we normalized them all to z-scores.

[14] Though note that the ratings are z-scores, so the MSE is still larger than half a standard deviation.

		SEED	FIT	FIT+SW	FIT+SD	FIT+S	
GLoVE	r^+-acc	0.64 (0.1)	0.54 (0.03)	0.53 (0.03)	0.65 (0.1)	**0.80** (0.06)	FREQ r^+-acc: 0.58
	MSE	$>$ 1000 ($>$ 1000)	113.2 (111.7)	177.1 (125.4)	89.6 (199.6)	**0.7** (0.36)	
BERT	r^+-acc	0.64 (0.1)	0.51 (0.03)	0.52 (0.03)	0.66 (0.10)	**0.71** (0.04)	
	MSE	$>$ 1000 ($>$ 1000)	417.4 (271.7)	597.4 (525.6)	115.0 (437.4)	**2.0** (0.6)	
RoBERTa	r^+-acc	0.57 (0.08)	0.51 (0.03)	0.51 (0.03)	0.60 (0.1)	**0.69** (0.04)	RAND r^+-acc: 0.49
	MSE	$>$ 1000 ($>$ 1000)	392.5 (291.3)	458.0 (284.3)	125.2 (270.5)	**1.9** (0.6)	

Table 1: Results on **object properties**: Extended rank accuracy (abbreviated r^+-acc) and Mean Squared Error (MSE), averaged over category/property pairs. In brackets: Standard error. Shown for 3 embedding spaces. Bolded: best performance for each embedding.

		Complexity						Formality					
		SEED	FIT	FIT+SW	FIT+SD	FIT+S		SEED	FIT	FIT+SW	FIT+SD	FIT+S	
GLoVE	r^+-acc	0.74	0.59	0.57	**0.76**	0.72	FREQ r^+-acc: 0.65	**0.73**	0.53	0.37	0.68	0.69	FREQ r^+-acc: 0.63
	MSE	31.5	24.3	75.1	1.5	**1.2**		60.5	396.5	285.7	1.8	**1.6**	
BERT	r^+-acc	0.69	0.52	0.52	0.71	**0.72**		0.64	0.52	0.51	0.64	**0.69**	
	MSE	123.5	437.2	724.0	3.6	**2.4**		215.6	216.3	617.1	7.8	**3.2**	
RoBERTa	r^+-acc	**0.74**	0.51	0.51	0.71	0.73	RAND r^+-acc: 0.50	0.67	0.53	0.53	0.66	**0.71**	RAND r^+-acc: 0.51
	MSE	82.7	591.9	$>$ 1000	3.9	**2.3**		223.0	325.0	778.1	7.3	**2.4**	

Table 2: Results on **formality** and **complexity**: Extended rank accuracy (r^+-acc) and Mean Squared Error (MSE). Shown for 3 embedding spaces. Bolded: best performance for each embedding.

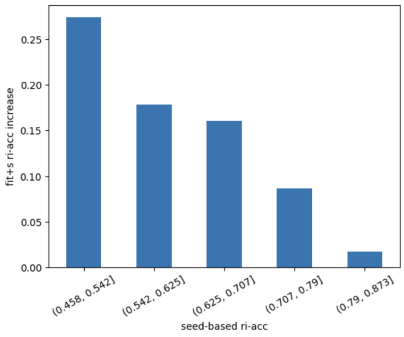

Figure 2: Increase in *r^+-acc* for FIT+S over SEED, object properties grouped by performance of SEED.

case, which confirms our suspicion. The picture that emerges is that FIT by itself does not have enough information to fit a good dimension and overfits to the training data. The seed information provided to FIT+SW, FIT+SD and FIT+S gives the models the additional guidance needed to make good use of the human ratings, and the combination of seeds and human ratings on words leads to overall better dimensions – at least in some cases. We next ask which cases those are.

Human ratings help most when seed-based dimensions underperform. Comparing *r^+-acc* values for seed-based dimensions and FIT+S dimensions on the object property data, we find that FIT+S improves over SEED in every single one of the 50 category/property pairs.[15] The performance increase is highest when performance of the seed-based dimensions is lowest, as shown in Figure 2: For the 20% of category/property pairs with lowest SEED performance, average improvement is 27.3 points, while for the 20% of category/property pairs with the highest SEED performance, average improvement is 1.7 points. This could explain the lack of improvement achieved on stylistic features, as SEED already performs well on this data.

Table 3 further zooms in on the object feature data, showing performance on some category/property pairs with low, medium, and high performance of the SEED dimensions. We see that FIT and FIT+SW underperform throughout. FIT+S shows the overall best performance, but the improvement over SEED is particularly high for the first group of conditions, where SEED dimensions get no traction on the data. FIT+SD shows good extended rank accuracy on the conditions with medium to high SEED performance, but not on the conditions that are particularly poorly modeled by SEED.

FIT+S models are the only ones that predict ratings on the gold scale. We saw above that median MSE values are extremely high for many models, especially for SEED. We now take a closer look, in particular we want to know how often we obtain MSE values that are extremely far off from the gold ratings. We again focus on the object fea-

[15]The analysis includes all conditions not in the development set. For each condition, we again ran 5-fold cross-validation, each repeated over 3 random seeds, then averaged by condition.

Category, Feature	r$^+$-acc SEED	FIT	FIT+SW	FIT+SD	FIT+S
sports/speed	0.46	0.55	0.56	0.52	0.78
states/cost	0.46	0.5	0.5	0.42	0.83
cities/arousal	0.47	0.52	0.5	0.51	0.82
animals/intelligence	0.48	0.55	0.54	0.5	0.79
clothing/cost	0.48	0.52	0.51	0.55	0.76
clothing/wealth	0.62	0.52	0.53	0.6	0.82
states/wealth	0.62	0.55	0.56	0.64	0.82
weather/temperature	0.69	0.56	0.52	0.66	0.76
animals/danger	0.7	0.6	0.57	0.76	0.84
clothing/age	0.71	0.52	0.55	0.71	0.8
weather/danger	0.79	0.55	0.54	0.7	0.82
clothing/gender	0.81	0.56	0.53	0.8	0.82
sports/gender	0.81	0.61	0.56	0.81	0.84
professions/gender	0.85	0.56	0.56	0.87	0.86
names/gender	0.87	0.56	0.51	0.87	0.87

Table 3: Detail results for Grand et al. by SEED performance: lowest performance (top box), middling performance (middle), best performance (bottom).

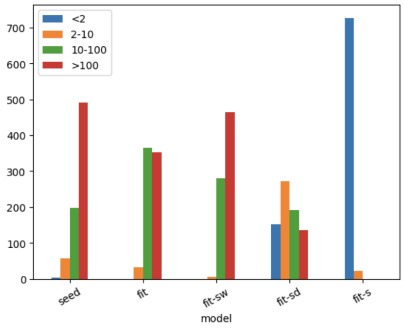

Figure 3: MSE distributions for runs of different models. y-axis: count of runs.

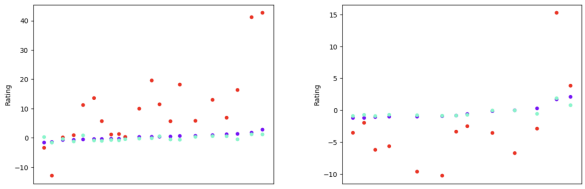

Figure 4: Clothes rated for wealth (left), animals rated for size (right). Gold ratings: dark purple. SEED predictions: red. FIT+S prediction: light blue. Datapoints ordered by gold rank.

Gold	SEED	FIT+S
bee	chipmunk	butterfly
ant	hamster	bee
butterfly	*monkey*	*chipmunk*
goldfish	butterfly	bird
hamster	goldfish	hamster
chipmunk	dog	*ant*
bird	*bee*	crow
crow	bird	*goldfish*
dog	seal	seal
monkey	crow	dog
seal	*ant*	monkey
mammoth	whale	whale
whale	mammoth	mammoth

Table 4: Comparing word rankings by humans, SEED dimensions, and FIT+S dimensions: Animals by size. Italicized: 3 words with highest error in ranking.

ture data as we there have many conditions that we can compare. Figure 3 shows, for each model, how many runs had MSE values in the ranges of $< 2, 2-10, 10-100$, and > 100. Recall that gold ratings are z-scores, so they tend to be in a range of -2 to 2. We again only use the category/property pairs that are not in the development set, but now count separately each cross-validation run and each random seed. We see that many runs of SEED, FIT and FIT+SW have very high MSE values. In FIT+SD we first see a considerable percentage of runs with MSE values below 2 (the blue bar comprises 20% of runs for this model), but strikingly, 97% runs of FIT+S have MSE values below 2, and all have values below 10. So this model is much more consistent than the other models, and in fact is highly consistent in fitting dimensions that deliver predictions in the range of the gold data.

Zooming in: Examples of predictions. We take a closer look at two kinds of object properties: clothes by wealth, and animals by size. In order to obtain sufficiently many test datapoints to look at, we divide the data into 2/3 training and 1/3 test (as opposed to the 1/5 we use with 5-fold cross-validation). Figure 1 shows the test data words, along with seed-based and FIT+S dimensions, downprojected into 2 dimensions using PCA. For the same datapoints, Figure 4 plots gold ratings, SEED predictions, and FIT+S predictions. This plot illustrates how the SEED predictions are on a much larger scale than gold ratings, while FIT+S is the only model whose predictions stay on the same scale. (The next to last datapoint in animals/size is *mammoth*, which *seed* largely overestimates – maybe because *mammoth* is also an adjective indicating gargantuan size.) Tables 4 and 5 show how the test data words for animals by size, and for clothes by wealth, are ranked by humans, by the SEED dimension, and by the FIT+S dimension. The italicized words are the three words whose model rank is furthest off from their gold rank. For the animals data, both models mis-rank *ant*, and overall seem to struggle more with smaller animals. Among the clothes, both models overestimate the wealth projected by wearing hats.

Gold	SEED	FIT+S
sweatshirt	shorts	shorts
shorts	sweatshirt	boots
belt	belt	bikini
boots	blouse	tights
hat	boots	skirt
tights	swimsuit	swimsuit
bikini	skirt	stockings
swimsuit	trousers	trousers
skirt	bikini	*loafers*
blouse	robe	blouse
knickers	cuff	belt
dress	knickers	knickers
collar	*hat*	dress
trousers	collar	*sweatshirt*
stockings	vest	robe
robe	*tights*	collar
vest	loafers	cuff
cuff	stockings	vest
loafers	*dress*	*hat*
gown	gown	tuxedo
tuxedo	tuxedo	gown

Table 5: Comparing word rankings by humans, SEED dimensions, and FIT+S dimensions: Clothes by wealth. Italicized: 3 words with highest error in ranking.

5 Conclusion

In this paper we have proposed a method for constructing high quality interpretable dimensions in embedding spaces. We show that by combining seed-based vectors with guidance from human ratings about properties, it is possible to induce better dimensions than with the seed-based methodology alone. We expect the proposed dimensions to be useful in various areas of study, including linguistics, psychology, and social science.

For the moment, the proposed dimensions address one property at a time. In future work, we are planning to explore multifaceted properties which would be better represented through multiple dimensions. Aside from a more elaborate description of these properties, a space of multiple interpretable dimensions will offer a rich context of comparison for words that might be similar in some respect and not in others (e.g., *tiger* and *spider* with respect to DANGER and SIZE).

6 Limitations

In our experiments we use English models and data. The seed-based methodology has been shown to work well in other languages, so an extension of the proposed methodology to other languages is possible. A limitation regarding this extension is the lack of human ratings which are needed for calculating the fitted dimensions. A possible mitigation would be to translate the annotated English data into other languages.

The ratings we used in our study were averages over individual human ratings, possibly obscuring legitimate differences between raters (Plank et al., 2014). Another limitation of the human ratings used in this study is that they were out of context, possibly obscuring effects of topic and polysemy.

There are many different ways to use contextualized embeddings. We have averaged over all token representations generated by BERT and RoBERTa for a word in a sentence pool, and used the top 4 layers of the models. It is possible that BERT and RoBERTa would do better, or at least differently, if other model layers (or layer combinations) were used.

Our approach is not at all compute intensive. All computations were done on a laptop.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback. This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Emily Allaway and Kathleen McKeown. 2021. A unified feature representation for lexical connotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2145–2163, Online. Association for Computational Linguistics.

Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2450–2461, Melbourne, Australia. Association for Computational Linguistics.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the*

11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1889–1904, Online. Association for Computational Linguistics.

Marianna Apidianaki. 2022. From Word Types to Tokens and Back: A Survey of Approaches to Word Meaning Representation and Interpretation. *Computational Linguistics*, 49(2):465–523.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation*, 43(3):209–226.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems 29*, pages 4349–4357, Barcelona, Spain.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Zied Bouraoui, José Camacho-Collados, and Steven Schockaert. 2020. Inducing Relational Knowledge from BERT. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7456–7463, New York, NY, USA. AAAI Press.

Zied Bouraoui, Víctor Gutiérrez-Basulto, and Steven Schockaert. 2022. Integrating ontologies and vector space embeddings using conceptual spaces. In *International Research School in Artificial Intelligence in Bergen (AIB 2022)*, volume 99 of *Open Access Series in Informatics (OASIcs)*, pages 3:1–3:30, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram Version 1. In *LDC2006T13*, Philadelphia, Pennsylvania. Linguistic Data Consortium.

Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Coling 2010: Posters*, pages 90–98, Beijing, China. Coling 2010 Organizing Committee.

Anne Cocos, Skyler Wharton, Ellie Pavlick, Marianna Apidianaki, and Chris Callison-Burch. 2018. Learning scalar adjective intensity from paraphrases. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1752–1762, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Joaquín Derrac and Steven Schockaert. 2015. Inducing semantic relations from conceptual spaces: A data-driven approach to plausible reasoning. *Artificial Intelligence*, 228:66–94.

Sunipa Dev and Jeff M Phillips. 2019. Attenuating Bias in Word Vectors. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, Naha, Okinawa, Japan.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Aina Garí Soler and Marianna Apidianaki. 2020. BERT knows Punta Cana is not just beautiful, it's gorgeous: Ranking scalar adjectives with contextualised representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7371–7385, Online. Association for Computational Linguistics.

Aina Garí Soler and Marianna Apidianaki. 2021a. Let's Play Mono-Poly: BERT Can Reveal Words' Polysemy Level and Partitionability into Senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Aina Garí Soler and Marianna Apidianaki. 2021b. Scalar adjective identification and multilingual ranking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4653–4660, Online. Association for Computational Linguistics.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2022. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behavior*, 6:975–987.

Peter Gärdenfors. 2014. *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. The MIT Press, Cambridge, MA.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Ray S. Jackendoff. 1990. *Semantic Structures*. The MIT Press, Cambridge, MA.

Shoaib Jameel and Steven Schockaert. 2016. Entity embeddings with conceptual subspaces as a basis for plausible reasoning. In *Proceedings of the Twenty-Second European Conference on Artificial Intelligence*, ECAI'16, page 1353–1361, NLD. IOS Press.

Jerrold J. Katz and Jerry A. Fodor. 1964. The structure of a semantic theory. In Jerry A. Fodor and Jerrold J. Katz, editors, *The Structure of Language*. Prentice-Hall, Englewood Cliffs, NJ.

Adam Kilgarriff. 2004. How Dominant Is the Commonest Sense of a Word? Lecture Notes in Computer Science (vol. 3206), Text, Speech and Dialogue, Sojka Petr, Kopeček Ivan, Pala Karel (eds.), pages 103–112. Springer, Berlin, Heidelberg.

Austin C. Kozlowski, Matt Taddy, and James A. Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

Haewoon Kwak, Jisun An, Elise Jing, and Yong-Yeol Ahn. 2021. Frameaxis: characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science*, 7.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint:1907.11692.

Li Lucy, Divya Tadimeti, and David Bamman. 2022. Discovering differences in the representation of people using contextualized semantic axes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3477–3494, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2023. Representation of lexical stylistic features in language models' embedding space. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM)*, Toronto, Canada.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards Faithful Model Explanation in NLP: A Survey. arXiv preprint:2209.11326.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 279–286, Barcelona, Spain.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Gregory L. Murphy. 2002. *The Big Book of Concepts*. MIT Press, Boston, Mass.

Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 218–224, Denver, Colorado. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Dustin S. Stoltz and Marshall A. Taylor. 2021. Cultural cartography with word embeddings. *Poetics*, 88:101567. Measure Mohr Culture.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Anna Wierzbicka. 1996. *Semantics: Primes and Universals*. Oxford University Press, New York.

A Appendix

Details on computing. All experiments were conducted on a MacBook Pro laptop using Python 3.8, with huggingface version 4.35.2, torch version 2.0.1, sklearn version 1.2.1, numpy version 1.22.4 and scipy 1.10.0.

Hyperparameter estimation. Development set: 6 conditions sampled at random from the object features dataset: cities-danger, states-political, animals-wetness, cities-intelligence, animals-weight, names-age.

As said above, only hyperparameters that made a difference: averaging, always good, and mixing parameter alpha. Chosen values:

Best parameters:

- α for FIT+SD: GLoVE 0.02
- α for FIT+S: GLoVE 0.05

Embedding spaces, and sentence selection. The GLoVe embeddings used were trained on Common Crawl (42B tokens, 1.9M vocab, uncased, 300d vectors), downloaded from `https://nlp.stanford.edu/projects/glove/`.

In order to generate embeddings for contextualized instances of words in our datasets using BERT (`bert-large-uncased`) and RoBERTa (`roberta-large`) models (Devlin et al., 2019; Liu et al., 2019), we used sentences from the UkWac corpus (Baroni et al., 2009). We collected ten sentences for each word, when available. We filtered out sentences with more than 100 tokens in order to avoid including noisy contexts such as webpage menus crawled from the web. If a word had less than 10 occurrences in UkWac, we used as many sentences as were available. This was the case for 10 words in the Grand et al. dataset (*nairobi* (6), *seoul* (4), *taipei* (5), *lahore* (2), *baghdad* (9), *peyton* (9), *tehran* (4), *johannesburg* (4), *jaime* (5), *karachi* (7); and for one word (*jazeera* (6)) in the formality dataset. For hyphenated words in the Grand et al. dataset (e.g., *new-york, south-carolina, south-dakota*), we collected sentences where they occur without the hyphen.