Problems with Cosine as a Measure of Embedding Similarity for High Frequency Words

Kaitlyn Zhou[1], **Kawin Ethayarajh**[1], **Dallas Card**[2], and **Dan Jurafsky**[1]

[1]Stanford University, {katezhou, kawin, jurafsky}@stanford.edu
[2]University of Michigan, dalc@umich.edu

Abstract

Cosine similarity of contextual embeddings is used in many NLP tasks (e.g., QA, IR, MT) and metrics (e.g., BERTScore). Here, we uncover systematic ways in which word similarities estimated by cosine over BERT embeddings are understated and trace this effect to training data frequency. We find that relative to human judgements, cosine similarity underestimates the similarity of frequent words with other instances of the same word or other words across contexts, even after controlling for polysemy and other factors. We conjecture that this underestimation of similarity for high frequency words is due to differences in the representational geometry of high and low frequency words and provide a formal argument for the two-dimensional case.

1 Introduction

Measuring semantic similarity plays a critical role in numerous NLP tasks like QA, IR, and MT. Many such metrics are based on the cosine similarity between the contextual embeddings of two words (e.g., BERTScore, MoverScore, BERTR, SemDist; Kim et al., 2021; Zhao et al., 2019; Mathur et al., 2019; Zhang et al., 2020). Here, we demonstrate that cosine similarity when used with BERT embeddings is highly sensitive to training data frequency.

The impact of frequency on accuracy and reliability has mostly been studied on *static* word embeddings like word2vec. Low frequency words have low reliability in neighbor judgements (Hellrich and Hahn, 2016), and yield smaller inner products (Mimno and Thompson, 2017) with higher variance (Ethayarajh et al., 2019a). Frequency also correlates with stability (overlap in nearest neighbors) (Wendlandt et al., 2018), and plays a role in word analogies and bias (Bolukbasi et al., 2016; Caliskan et al., 2017; Zhao et al., 2018; Ethayarajh et al., 2019b). Similar effects have been found in contextual embeddings, particularly for low-frequency senses, which seem to cause difficulties in WSD performance for BERT and RoBERTa (Postma et al., 2016; Blevins and Zettlemoyer, 2020; Gessler and Schneider, 2021). Other works have examined how word frequency impacts the similarity of *sentence* embeddings (Li et al., 2020; Jiang et al., 2022).

While previous work has thus mainly focused on reliability or stability of low frequency words or senses, our work asks: how does frequency impact the semantic similarity of high frequency words?

We find that the cosine of BERT embeddings underestimates the similarity of high frequency words (to other tokens of the same word or to different words) as compared to human judgements. In a series of regression studies, we find that this underestimation persists even after controlling for confounders like polysemy, part-of-speech, and lemma. We conjecture that word frequency induces such distortions via differences in the representational geometry. We introduce new methods for characterizing geometric properties of a word's representation in contextual embedding space, and offer a formal argument for why differences in representational geometry affect cosine similarity measurement in the two-dimensional case.[1]

2 Effect of Frequency on Cosine Similarity

To understand the effect of word frequency on cosine between BERT embeddings (Devlin et al., 2019), we first approximate the training data frequency of each word in the BERT pre-training corpus from a combination of the March 1, 2020 Wikimedia Download and counts from BookCorpus (Zhu et al., 2015; Hartmann and dos Santos, 2018).[2] We then consider two datasets that include

[1]Code for this paper can be found at https://github.com/katezhou/cosine_and_frequency

[2]Additional tools used: https://github.com/IlyaSemenov/wikipedia-word-frequency;

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 401 - 423

pairs of words in context with associated human similarity judgements of words: Word-In-Context (WiC) (expert-judged pairs of sentences with a target lemma used in either the same or different WordNet, Wiktionary, or VerbNet senses) and Stanford Contextualized Word Similarity dataset (SCWS) (non-expert judged pairs of sentences annotated with human ratings of the similarity of two target terms). Using datasets with human similarity scores allows us to account for human perceived similarities when measuring the impact of frequency on cosine (Pilehvar and Camacho-Collados, 2019; Huang et al., 2012).

2.1 Study 1: WiC

Method and Dataset The authors of WiC used coarse sense divisions as proxies for words having the same or different meaning and created 5,428[3] pairs of words in context, labeled as having the same or different meaning:

- same meaning: "I try to avoid the company of gamblers" and "We avoided the ball"
- different meaning: "You must carry your camping gear" and "Sound carries well over water".

To obtain BERT-based similarity measurements, we use `BERT-base-cased`[4] to embed each example, average the representations of the target word over the last four hidden layers, and compute cosine similarity for the pair of representations.[5]

Relation between frequency and similarity in WiC We want to use ordinary least squares regression to measure the effect of word frequency on the cosine similarity of BERT embeddings. First, we split the WiC dataset into examples that were labeled as having the "same" or "different" meanings. This allows us to to control for perceived similarity of the two words in context — any frequency effects found within these subsets cannot be explained by variation in human judgements. Next, we control for a number of other confounding factors by including them as variables in our OLS regression. For each target lemma we considered:

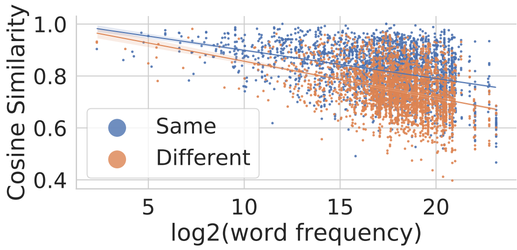

Figure 1: Ordinary Least Squares regression of cosine similarity against frequency, for examples with the same meaning (blue) and different meaning (orange). Both regressions show a significant negative association between cosine similarity and frequency.

frequency: $\log_2$ of the number of occurrences in BERT's training data

polysemy: $\log_2$ of number of senses in WordNet

is_noun: binary indicator for nouns vs. verbs

same_wordform: binary indicator of having the same wordform in both contexts (e.g., *act/act* vs. *carry/carries*) (case insensitive)

An OLS regression predicting cosine similarity from a single independent factor of $\log_2(\text{freq})$ shows a significant negative association between cosine and frequency among "same meaning" examples (R^2 : 0.13, coeff's $p < 0.001$) and "different meaning" examples (R^2 : 0.14, coeff's $p < 0.001$) (see Figure 1). The same negative frequency effect is found across various model specifications (Table 1 in Appendix), which also show significantly greater cosine similarity for those examples with the same wordform, a significant negative association with number of senses, and no difference between nouns and verbs. In summary, we find that using cosine to measure the semantic similarity of words via their BERT embeddings gives systematically smaller similarities the higher the frequency of the word.

Results: Comparing to human similarity To compare cosine similarities to WiC's binary human judgements (same/different meaning), we followed WiC authors by thresholding cosine values, tuning the threshold on the training set (resulting threshold: 0.8). As found in the original WiC paper, cosine similarity is somewhat predictive of the expert judgements (0.66 dev accuracy, comparable to 0.65 test accuracy from the WiC authors).[6]

Examining the errors as a function of frequency reveals that cosine similarity is a less reliable predictor of human similarity judgements for common

https://github.com/attardi/wikiextractor

[3] We used a subset of 5,423 of these examples due to minor spelling differences and availability of frequency data.

[4] https://huggingface.co/bert-base-cased

[5] Out-of-vocabulary words are represented as the average of the subword pieces of the word, following Pilehvar and Camacho-Collados (2019) and Blevins and Zettlemoyer (2020); we found that representing OOV words by their first token produced nearly identical results.

[6] The test set is hidden due to an ongoing leaderboard.

terms. Figure 2 shows the average proportion of examples predicted to be the same meaning as a function of frequency, grouped into ten bins, each with the same number of examples. In the highest frequency bin, humans judged 54% of the examples as having the same meaning compared to only 25% as judged by cosine similarity. This suggests that in the WiC dataset, relative to humans, the model underestimates the sense similarity for high frequency words.

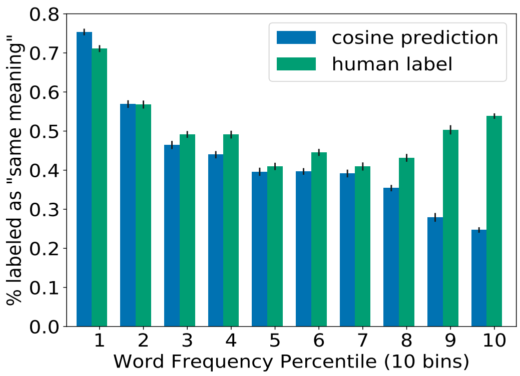

Figure 2: Percentage of examples labeled as having the "same meaning". In high frequency words, cosine similarity-based predictions (blue/left) on average **under**-estimate the similarity of words as compared to human judgements (green/right).

2.2 Study 2: SCWS

Our first study shows that after controlling for sense, cosine will tend to be lower for higher frequency terms. However, the WiC dataset only has binary labels of human judgements, and only indicates similarity between occurrences of the same word. We want to measure if these frequency effects persist across different words and control for more fine-grained human similarity judgements.

Method and Dataset SCWS contains crowd judgements of the similarity of two words in context (scale of 1 to 10). We split the dataset based on whether the target words are the same or different ($break/break$ vs $dance/sing$); this both allows us to confirm our results from WiC and also determine whether frequency-based effects exist in similarity measurements across words.[7] We use the same embedding method as described for WiC, and again use regression to predict cosine similarities from the following features:

frequency: average of $log_2(freq)$ of both words
polysemy: average of $log_2(sense)$ of both words
average rating: average rating of semantic similarity as judged by humans on a scale of 1 to 10 (highest).

Results If we only use frequency, we find that it mildly explains the variance in cosine similarity both within (R^2 : 0.12, coeff's $p < 0.001$) and across words (R^2 : 0.06, coeff's $p < 0.001$). Adding in human average rating as a feature, frequency is still a significant feature with a negative coefficient. High frequency terms thus tend to have lower cosine similarity scores, even after accounting for human judgements. When using all features, the linear regression models explain 34% of the total variance in cosine similarity, with frequency still having a significant negative effect (Table 2 in Appendix). Finally, we verify that for a model with only human ratings, error (true - predicted cosine) is negatively correlated with frequency in held out data (Pearson's $r = -0.18$; $p < 0.01$), indicating an underestimation of cosine in high frequency words (see Figure 5 in Appendix).

This finding suggests that using frequency as a feature might help to better match human judgements of similarity. We test this hypothesis by training regression models to predict human ratings, we find that frequency does have a significant positive effect (Table 3 in Appendix) but the overall improvement over using cosine alone is relatively small ($R^2 = 44.6\%$ vs $R^2 = 44.3\%$ with or without frequency). We conclude that the problem of underestimation in cosine similarity cannot be resolved simply by using a linear correction for frequency.

3 Minimum Bounding Hyperspheres

In order to understand why frequency influences cosine similarity, we analyze the geometry of the contextual embeddings. Unlike static vectors – where each word type is represented by a single point – the variation in contextualized embeddings depends on a word's frequency in training data. We'll call embeddings of a single word type *sibling embeddings* or a *sibling cohort*. To measure variation, we'll use the radius of the smallest hypersphere that contains a set of sibling embeddings (the minimum bounding hypersphere). We tested many ways to measure the space created by high-dimensional vectors. Our results are robust to various other

[7]For consistency across word embeddings, we only use SCWS examples where the keyword appeared lower-cased in context. We reproduced our results with all SCWS examples and found our findings to be qualitatively the same.

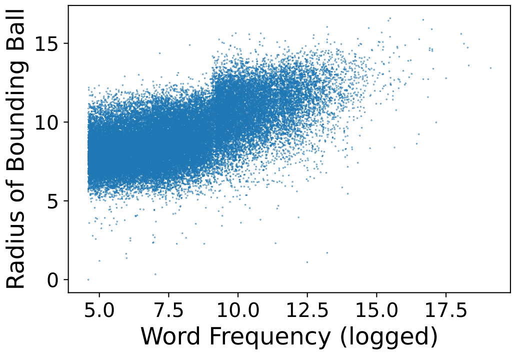

Figure 3: The radius of the minimal bounding ball of sibling embeddings of words is correlated with log(word frequency). (Pearson's $r = 0.62, p < .001$)

measures of variation, including taking the average, max, or variance of pairwise distance between sibling embeddings, the average norm of sibling embeddings, and taking the PCA of these vectors and calculating the convex hull of sibling embeddings in lower dimensions (see Table 29 in the Appendix). Here we relate frequency to spatial variation, providing both empirical evidence and theoretical intuition.

For a sample of 39,621 words, for each word we took 10 instances of its sibling embeddings (example sentences queried from Wikipedia), created contexutalized word embeddings using Hugging Face's `bert-base-cased` model, and calculated the radius of the minimum bounding hypersphere encompassing them.[8] [9] As shown in Figure 3, there is a significant, strong positive correlation between frequency and size of bounding hypersphere (Pearson's $r = 0.62, p < .001$). Notably, since the radius was calculated in 768 dimensions, an increase in radius of 1% results in a hypersphere volume nearly 2084 times larger.[10]

Since frequency and polysemy are highly correlated, we want to measure if frequency is a significant feature for explaining the variance of bounding hyperspheres. Using the unique words of the WiC dataset, we run a series of regressions to predict the radius of bounding hyperspheres. On their own, frequency and polysemy explain for 48% and 45% of the radii's variance. Using both features, frequency and polysemy explains for 58% of the radii's variance and both features are significant – demonstrating that frequency is a significant feature in predicting radii of bounding hyperspheres (Tables 25, 26, 27 in Appendix).

Among the unique words of the WiC dataset, the radii of the target word correlates with training data frequency (Pearson's $r : 0.69, p < 0.001$). Across the WiC dataset, the radii explains for 17% of the variance in cosine similarity (Table 28 in Appendix).[11]

3.1 Theoretical Intuition

Here, we offer some theoretical intuition in 2D for why using cosine similarity to estimate semantic similarity can lead to underestimation (relative to human judgements). Let $\vec{w} \in \mathbb{R}^2$ denote the target word vector, against which we're measuring cosine similarity. Say there were a bounding ball B_x with center $\vec{x_c}$ to which $\vec{w}$ is tangent. If we normalize every point in the bounding ball, it will form an arc on the unit circle. The length of this arc is $2\theta = 2 \arcsin \frac{r}{\|x_c\|_2}$:

- Let θ denote the angle made by x_c and the tangent vector $\vec{w}$.
- $\sin \theta = \frac{r}{\|x_c\|_2}$, so the arc length on the unit circle is $r\theta = \arcsin \frac{r}{\|x_c\|_2}$ (normalized points).
- Multiply by 2 to get the arclength between both (normalized) tangent vectors.

Since the arclength is monotonic increasing in r, if the bounding ball were larger—while still being tangent to $\vec{w}$—the arclength will be too.

The cosine similarity between a point in the bounding ball and $\vec{w}$ is equal to the dot product between the projection of the former onto the unit circle (i.e., somewhere on the arc) and the normalized $\vec{w}$. This means that only a certain span of the arclength maps to sibling embeddings $\vec{x_i}$ such that $\cos(\vec{x_i}, \vec{w}) \geq t$, where t is the threshold required to be judged as similar by humans (see Footnote 3 and Figure 4). If B_x were larger while still being tangent to w, the arclength would increase but the span of the arc containing siblings embeddings

[8] Words were binned by frequency and then sampled in order to sample a range of frequencies. As a result, there is a Zipfian effect causing there to be slightly more words in the lower ranges of each bin. We used `https://pypi.org/project/miniball/`

[9] Given the sensitivity of minimum bounding hypersphere to outliers, we'd imagine that frequency-based distortions would be even more pronounced had we chosen to use more instances of sibling embeddings.

[10] the n-dimensional volume of a Euclidean ball of radius R:

$$V_n(R) = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)} R^n$$

[11] We used 1,253 out of the original 1,265 unique WiC words and 5,412 out of the original 5,428 WiC examples due to availability of frequency data and contextual examples for target words.

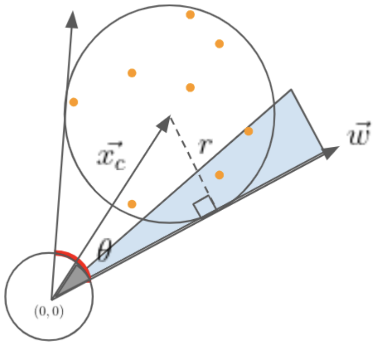

Figure 4: An illustration of how using cosine similarity can underestimate word similarity. The cosine similarity between a contextualized representation (orange) and $\vec{w}$ is the dot product of the former's projection onto the red arc of the unit circle (with length 2θ) and $\hat{w}$. Only points in the blue region are close enough to $\hat{w}$ to be deemed similar by humans. As the bounding ball grows (e.g., with higher frequency words), if it remains tangent to $\vec{w}$, the fraction of points in the blue region will shrink, leading to underestimation.

sufficiently similar to w would not. This means a greater proportion of the sibling embeddings will fail to meet this threshold, assuming that the distribution of sibling embeddings in B_x does not change. Because, in practice, more frequent words have larger bounding balls, depending on how the bounding ball of a word x grows relative to some $\vec{w}$, the similarity of x and w can be underestimated. This helps explain the findings in Figure 2, but it does not explain why more frequent words have lower similarity with themselves across different contexts, since that requires knowledge of the embedding distribution in the bounding ball. The latter is likely due to more frequent words having less anisotropic representations (Ethayarajh, 2019).

4 Discussion and Conclusion

Cosine distance underestimates compared to humans the semantic similarity of frequent words in a variety of settings (expert versus non-expert judged, and within word sense and across words). This finding has large implications for downstream tasks, given that single-point similarity metrics are used in a variety of methods and experiments (Reimers and Gurevych, 2019; Reif et al., 2019; Zhang et al., 2020; Zhao et al., 2019; Mathur et al., 2019; Kim et al., 2021). Word frequency in pre-training data also affects the representational geometry of contextualized embeddings, low frequency words being more concentrated geometrically. One extension of this work might examine how variables such as sentiment and similarity/dissimilarity between sentence contexts could impact both human-judged and embedding-based similarity metrics.

Because training data frequency is something that researchers can control, understanding these distortions is critical to training large language models. Frequency-based interventions might even be able to correct for these systematic underestimations of similarity (e.g., by modifying training data), which could be important where certain words or subjects may be inaccurately represented. For example, Zhou et al. (2022) illustrates how training data frequencies can lead to discrepancies in the representation of countries, and—since frequency is highly correlated with a country's GDP—can perpetuate historic power and wealth inequalities. Future work could also examine how and if frequency effects could be mitigated by post-processing techniques which improve the correlation between human and semantic similarities (Timkey and van Schijndel, 2021).

The semantic similarity distortions caused by the over-and under-representation of topics is another reason why documentation for datasets is critical for increasing transparency and accountability in machine learning models (Gebru et al., 2021; Mitchell et al., 2019; Bender and Friedman, 2018; Ethayarajh and Jurafsky, 2020; Ma et al., 2021). As language models increase in size and training data becomes more challenging to replicate, we recommend that word frequencies and distortions be revealed to users, bringing awareness to the potential inequalities in datasets and the models that are trained on them. In the future, we hope to see research that more critically examines the downstream implications of these findings and various mitigation techniques for such distortions.

Acknowledgements

We sincerely thank Isabel Papadimitriou and our anonymous reviewers for their support, insights, and helpful feedback. This research has been supported in part by a Hoffman-Yee Research Grant from the Stanford Institute for Human-Centered AI, award IIS-2128145 from the NSF, Stanford Data Science, a Stanford Graduate Fellowship, a Facebook Fellowship, and Canada's NSERC.

References

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

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.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019a. Towards understanding linear word analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3253–3262, Florence, Italy.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019b. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Luke Gessler and Nathan Schneider. 2021. BERT has uncommon sense: Similarity ranking for word sense BERTology. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*.

Nathan Hartmann and Leandro Borges dos Santos. 2018. NILC at CWI 2018: Exploring feature engineering and feature learning. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 335–340, New Orleans, Louisiana.

Johannes Hellrich and Udo Hahn. 2016. Bad Company—Neighborhoods in neural embedding spaces considered harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2785–2796, Osaka, Japan.

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea.

Ting Jiang, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Liangjie Zhang, and Qi Zhang. 2022. Prompt-BERT: Improving BERT sentence embeddings with prompts. *arXiv preprint arXiv:2201.04337*.

Suyoun Kim, Duc Le, Weiyi Zheng, Tarun Singh, Abhinav Arora, Xiaoyu Zhai, Christian Fuegen, Ozlem Kalinli, and Michael L. Seltzer. 2021. Evaluating user perception of speech recognition system quality with semantic distance metric.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130.

Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. *Advances in Neural Information Processing Systems*, 34.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2873–2878, Copenhagen, Denmark.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota.

Marten Postma, Ruben Izquierdo Bevia, and Piek Vossen. 2016. More is not always better: balancing sense distributions for all-words word sense disambiguation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3496–3506, Osaka, Japan. The COLING 2016 Organizing Committee.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic.

Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2092–2102, New Orleans, Louisiana.

Tianyi Zhang, Varsha Kishore, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China.

Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. 2022. Richer countries and richer representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

A Appendix

For readability, we've summarized the key results from the regressions in 1 and 2. Table 1 contains results from our WiC experiments where we measure frequency's impact on cosine similarity. We control for human judgements of similarity by splitting the dataset by human labels of "same" and "different" meaning words. The same trends hold for the whole dataset as well.

Table 2 contains results from the SCWS experiments we measure frequency's impact on cosine similarity within and across word similarities. Similar to the WiC results, we see that frequency does impact cosine similarity, with higher words having lower similarities.

Table 3 contains results from the SCWS experiments where we measure frequency's impact on human ratings. We see that frequency does not explain human ratings but when used in a model with cosine similarity, frequency has a positive coefficient, indicating it is correcting for the underestimation of cosine similarity.

B Regression results from WiC experiments

Tables 4, 5, 6, 7, 8, 9, 10, 11.

C Regression results from SCWS experiments

Tables 12, 13, 14, 15, 16, 17, 18, 19

D Regression results from SCWS experiments, explaining for the difference between cosine similarity and human judgements

Tables 20, 21, 22, 23, 24.

Cosine similarity is partially predictive of human similarity judgements. The full model shows a significant positive effect of frequency 24 indicating that for a given level of cosine similarity, more frequent terms will judged by humans to be more similar, again demonstrating that cosine under-estimates semantic similarity for frequent terms.

The effect is relatively small, however; for a word that is twice as frequent, the increase in human rating will be 0.0989 (See table 23). Removing frequency from the model reduces R^2 from 40.8% to 40.4%. Polysemy shows the opposite effect; those words with more senses are likely to be rated as less similar. In a model with only cosine and polysemy factors, however, frequency has no relationship with human judgements, indicating that including frequency is correcting for the semantic distortion of cosine in the full model.

E Regression results from minimum bounding hyperspheres

Using frequency and polysemy to explain for the variability in bounding ball radii. Tables 25, 26, 27. Using radius of the bounding ball to explain for the variability of cosine similarity. Table 28.

F Other ways of measuring the space of sibling embeddings

Using a smaller sample of words (10,000 words out of the initial ∼39,000 words), we calculate the space occupied by these sibling embeddings using a variety of other metrics. In each metric, we find strong correlations between (log) frequency and the metric in question (see table 29).

G Residual of Predicted Cosine

For the SCWS dataset, use 1,000 samples as the train set and use the rest as the development set. We train a linear regression model to predict cosine similarity using only human ratings. Taking the difference between cosine similarity and the predicted similarity, we plot this error relative to frequency. We see a negative correlation between this error and frequency $r = -0.18, p < 0.001$, indicating that there is an underestimation of cosine similarity among the high frequency words. Results are shown in Figure 5.

OLS predicting cosine similarity								
WiC	Different Sense Meaning				Same Sense Meaning			
	Model 1	Model 2	Model 3	Model 4	Model 1	Model 2	Model 3	Model 4
$log_2(freq)$	**-0.014**	**-0.012**	**-0.013**	**-0.013**	**-0.011**	**-0.009**	**-0.009**	**-0.010**
$log_2(sense)$	-	**-0.012**	**-0.008**	**-0.009**	-	**-0.006**	**-0.004**	-0.002
same_wordform	-	-	**0.045**	**0.047**	-	-	**0.059**	**0.056**
is_noun	-	-	-	-0.006	-	-	-	**0.008**
R^2	0.127	0.144	0.203	0.204	0.136	0.142	0.241	0.242
Table Number	4	5	6	7	8	9	10	11

Table 1: Coefficients for each of the variables when used in a OLS regression. Bolded numbers are significant. The WiC dataset is split across examples that were rated to have the same or different meaning by experts. Other confounders (polysemy, part-of-speech, word form) were accounted for as features. In model 1, for a word that is twice as frequent, the decrease in cosine similarity will be 0.011.

SCWS	Within Word Examples				Across Words Examples			
	Model 1	Model 2	Model 3	Model 4	Model 1	Model 2	Model 3	Model 4
$log_2(freq))$	**-0.020**	-	**-0.018**	**-0.016**	**-0.011**	-	**-0.008**	**-0.008**
average rating	-	**0.022**	**0.021**	**0.02**	-	**0.02**	**0.02**	**0.02**
$log_2(sense)$	-	-	-	**-0.019**	-	-	-	-0.001
R^2	0.120	0.225	0.320	0.343	0.059	0.305	0.336	0.337
Table Number	12	13	14	15	16	17	18	19

Table 2: Coefficients for each of the variables when used in a OLS regression. Bolded numbers are significant. The SCWS dataset is split across examples that use the same (within word) or different (across word) target words. Other con-founders (polysemy and average rating) were accounted for as features. In model 1, for a word that is twice as frequent, the decrease in cosine similarity will be 0.02.

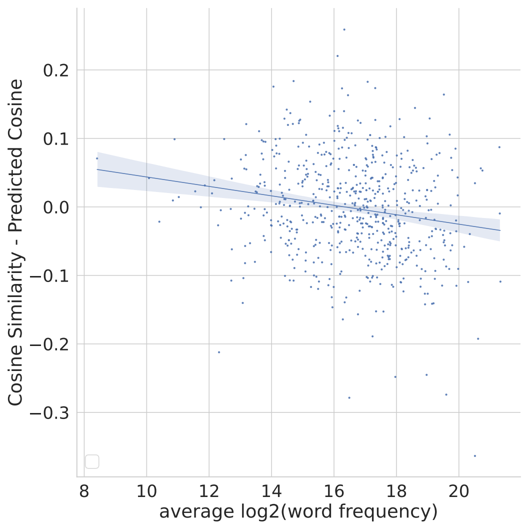

Figure 5: Error in cosine similarity and predicted cosine similarity using human ratings. A negative correlation exists, $r = -0.18, p < 0.001$, indicating an underestimation of cosine similarity among the high frequency words.

OLS Predicting Average Human Rating (Scale of 1 - 10)					
Feature	Model 1	Model 2	Model 3	Model 4	Model 5
avg $log_2(freq)$	-0.057	-	**0.099**	-	**0.076**
avg $log_2(sense)$	-	-	-0.0440	**-0.134**	**-0.189**
cosine	-	**16.345**	**16.665**	**13.513**	**13.809**
same_word	-	-	-	**1.7228**	**1.687**
R^2	0.002	0.404	0.408	0.443	0.446
Table Number	20	21	22	23	24

Table 3: Coefficients for each of the variables when used in a OLS regression. Bolded numbers are significant. Other con-founders (polysemy, same word) were accounted for as features. In model 5, for a word that is twice as frequent, the increase in human rating will be 0.076. Notice that frequency only becomes a significant as a feature when used with cosine, indicating that it is correcting for an underestimation.

Dep. Variable:	Cosine Similarity	**R-squared:**	0.127
Model:	OLS	**Adj. R-squared:**	0.127
Method:	Least Squares	**F-statistic:**	395.1
Date:	Thu, 14 Oct 2021	**Prob (F-statistic):**	3.55e-82
Time:	22:12:38	**Log-Likelihood:**	2947.0
No. Observations:	2713	**AIC:**	-5890.
Df Residuals:	2711	**BIC:**	-5878.
Df Model:	1		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	0.9976	0.013	77.728	0.000	0.972	1.023
log2(freq)	-0.0141	0.001	-19.876	0.000	-0.015	-0.013

Omnibus:	1.261	**Durbin-Watson:**	1.952
Prob(Omnibus):	0.532	**Jarque-Bera (JB):**	1.189
Skew:	0.044	**Prob(JB):**	0.552
Kurtosis:	3.053	**Cond. No.**	149.

Table 4: OLS regression results predicting cosine similarity among "different meaning" senses.

Dep. Variable:	Cosine Similarity	**R-squared:**	0.144
Model:	OLS	**Adj. R-squared:**	0.144
Method:	Least Squares	**F-statistic:**	228.2
Date:	Thu, 14 Oct 2021	**Prob (F-statistic):**	2.48e-92
Time:	22:12:38	**Log-Likelihood:**	2973.7
No. Observations:	2713	**AIC:**	-5941.
Df Residuals:	2710	**BIC:**	-5924.
Df Model:	2		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	0.9997	0.013	78.627	0.000	0.975	1.025
log2(freq)	-0.0115	0.001	-14.624	0.000	-0.013	-0.010
log2(senses)	-0.0118	0.002	-7.330	0.000	-0.015	-0.009

Omnibus:	8.024	**Durbin-Watson:**	1.954
Prob(Omnibus):	0.018	**Jarque-Bera (JB):**	9.222
Skew:	0.060	**Prob(JB):**	0.00994
Kurtosis:	3.259	**Cond. No.**	153.

Table 5: OLS regression results predicting cosine similarity among "different meaning" senses.

Dep. Variable:	Cosine Similarity	**R-squared:**	0.203
Model:	OLS	**Adj. R-squared:**	0.202
Method:	Least Squares	**F-statistic:**	230.2
Date:	Thu, 14 Oct 2021	**Prob (F-statistic):**	5.14e-133
Time:	22:12:38	**Log-Likelihood:**	3070.5
No. Observations:	2713	**AIC:**	-6133.
Df Residuals:	2709	**BIC:**	-6109.
Df Model:	3		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	0.9367	0.013	71.757	0.000	0.911	0.962
log2(freq)	-0.0130	0.001	-16.984	0.000	-0.015	-0.012
log2(senses)	-0.0076	0.002	-4.833	0.000	-0.011	-0.005
same_wordform	0.0447	0.003	14.158	0.000	0.039	0.051

Omnibus:	13.328	**Durbin-Watson:**	1.917
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	14.587
Skew:	-0.123	**Prob(JB):**	0.000680
Kurtosis:	3.261	**Cond. No.**	163.

Table 6: OLS regression results predicting cosine similarity among "different meaning" senses.

Dep. Variable:	Cosine Similarity	**R-squared:**	0.204
Model:	OLS	**Adj. R-squared:**	0.203
Method:	Least Squares	**F-statistic:**	173.4
Date:	Thu, 14 Oct 2021	**Prob (F-statistic):**	2.26e-132
Time:	22:12:38	**Log-Likelihood:**	3071.8
No. Observations:	2713	**AIC:**	-6134.
Df Residuals:	2708	**BIC:**	-6104.
Df Model:	4		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	0.9355	0.013	71.569	0.000	0.910	0.961
log2(freq)	-0.0126	0.001	-15.858	0.000	-0.014	-0.011
log2(senses)	-0.0090	0.002	-5.030	0.000	-0.013	-0.005
same_wordform	0.0467	0.003	13.760	0.000	0.040	0.053
is_noun	-0.0061	0.004	-1.629	0.103	-0.013	0.001

Omnibus:	14.009	**Durbin-Watson:**	1.915
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	15.019
Skew:	-0.135	**Prob(JB):**	0.000548
Kurtosis:	3.244	**Cond. No.**	164.

Table 7: OLS regression results predicting cosine similarity among "different meaning" senses.

Dep. Variable:	Cosine Similarity	**R-squared:**	0.136
Model:	OLS	**Adj. R-squared:**	0.136
Method:	Least Squares	**F-statistic:**	427.3
Date:	Thu, 14 Oct 2021	**Prob (F-statistic):**	2.94e-88
Time:	22:12:38	**Log-Likelihood:**	2926.4
No. Observations:	2710	**AIC:**	-5849.
Df Residuals:	2708	**BIC:**	-5837.
Df Model:	1		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	1.0077	0.009	109.007	0.000	0.990	1.026
log2(freq)	-0.0109	0.001	-20.670	0.000	-0.012	-0.010

Omnibus:	45.476	**Durbin-Watson:**	1.977
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	45.736
Skew:	-0.298	**Prob(JB):**	1.17e-10
Kurtosis:	2.778	**Cond. No.**	103.

Table 8: OLS regression results predicting cosine similarity among "same meaning" senses.

Dep. Variable:	Cosine Similarity	**R-squared:**	0.142
Model:	OLS	**Adj. R-squared:**	0.141
Method:	Least Squares	**F-statistic:**	224.2
Date:	Thu, 14 Oct 2021	**Prob (F-statistic):**	8.17e-91
Time:	22:12:38	**Log-Likelihood:**	2935.6
No. Observations:	2710	**AIC:**	-5865.
Df Residuals:	2707	**BIC:**	-5847.
Df Model:	2		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	0.9974	0.010	104.755	0.000	0.979	1.016
log2(freq)	-0.0090	0.001	-13.270	0.000	-0.010	-0.008
log2(senses)	-0.0063	0.001	-4.283	0.000	-0.009	-0.003

Omnibus:	38.934	**Durbin-Watson:**	1.973
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	39.612
Skew:	-0.283	**Prob(JB):**	2.50e-09
Kurtosis:	2.823	**Cond. No.**	109.

Table 9: OLS regression results predicting cosine similarity among "same meaning" senses.

Dep. Variable:	Cosine Similarity	**R-squared:**	0.241
Model:	OLS	**Adj. R-squared:**	0.240
Method:	Least Squares	**F-statistic:**	285.7
Date:	Thu, 14 Oct 2021	**Prob (F-statistic):**	4.36e-161
Time:	22:12:38	**Log-Likelihood:**	3100.7
No. Observations:	2710	**AIC:**	-6193.
Df Residuals:	2706	**BIC:**	-6170.
Df Model:	3		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	0.8928	0.011	84.562	0.000	0.872	0.914
log2(freq)	-0.0092	0.001	-14.435	0.000	-0.010	-0.008
log2(senses)	-0.0035	0.001	-2.513	0.012	-0.006	-0.001
same_wordform	0.0588	0.003	18.728	0.000	0.053	0.065

Omnibus:	80.675	**Durbin-Watson:**	1.981
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	87.234
Skew:	-0.434	**Prob(JB):**	1.14e-19
Kurtosis:	3.139	**Cond. No.**	130.

Table 10: OLS regression results predicting cosine similarity among "same meaning" senses.

Dep. Variable:	Cosine Similarity	**R-squared:**	0.242
Model:	OLS	**Adj. R-squared:**	0.241
Method:	Least Squares	**F-statistic:**	215.8
Date:	Thu, 14 Oct 2021	**Prob (F-statistic):**	6.75e-161
Time:	22:12:38	**Log-Likelihood:**	3103.2
No. Observations:	2710	**AIC:**	-6196.
Df Residuals:	2705	**BIC:**	-6167.
Df Model:	4		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	0.8952	0.011	84.424	0.000	0.874	0.916
log2(freq)	-0.0096	0.001	-14.547	0.000	-0.011	-0.008
log2(senses)	-0.0022	0.002	-1.457	0.145	-0.005	0.001
same_wordform	0.0560	0.003	16.512	0.000	0.049	0.063
is_noun	0.0078	0.003	2.228	0.026	0.001	0.015

Omnibus:	76.318	**Durbin-Watson:**	1.983
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	82.141
Skew:	-0.421	**Prob(JB):**	1.46e-18
Kurtosis:	3.139	**Cond. No.**	132.

Table 11: OLS regression results predicting cosine similarity among "same meaning" senses.

Dep. Variable:	Cosine Similarity	**R-squared:**	0.120
Model:	OLS	**Adj. R-squared:**	0.115
Method:	Least Squares	**F-statistic:**	28.77
Date:	Sat, 12 Mar 2022	**Prob (F-statistic):**	2.12e-07
Time:	12:16:53	**Log-Likelihood:**	203.87
No. Observations:	214	**AIC:**	-403.7
Df Residuals:	212	**BIC:**	-397.0
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	1.0762	0.063	17.127	0.000	0.952	1.200
avg_freq	-0.0196	0.004	-5.364	0.000	-0.027	-0.012

Omnibus:	7.823	**Durbin-Watson:**	2.040
Prob(Omnibus):	0.020	**Jarque-Bera (JB):**	9.129
Skew:	-0.307	**Prob(JB):**	0.0104
Kurtosis:	3.804	**Cond. No.**	169.

Table 12: OLS regression results predicting cosine similarity among "same" target words

Dep. Variable:	Cosine Similarity	**R-squared:**	0.225
Model:	OLS	**Adj. R-squared:**	0.221
Method:	Least Squares	**F-statistic:**	61.58
Date:	Sat, 12 Mar 2022	**Prob (F-statistic):**	2.07e-13
Time:	12:20:20	**Log-Likelihood:**	217.54
No. Observations:	214	**AIC:**	-431.1
Df Residuals:	212	**BIC:**	-424.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	0.5856	0.021	28.308	0.000	0.545	0.626
average_rating	0.0223	0.003	7.847	0.000	0.017	0.028

Omnibus:	31.336	**Durbin-Watson:**	2.183
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	64.374
Skew:	-0.711	**Prob(JB):**	1.05e-14
Kurtosis:	5.279	**Cond. No.**	25.5

Table 13: OLS regression results predicting cosine similarity among "same" target words

Dep. Variable:	Cosine Similarity	**R-squared:**	0.320
Model:	OLS	**Adj. R-squared:**	0.314
Method:	Least Squares	**F-statistic:**	49.70
Date:	Sat, 12 Mar 2022	**Prob (F-statistic):**	2.06e-18
Time:	12:20:20	**Log-Likelihood:**	231.56
No. Observations:	214	**AIC:**	-457.1
Df Residuals:	211	**BIC:**	-447.0
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	0.8939	0.060	14.907	0.000	0.776	1.012
avg_freq	-0.0176	0.003	-5.434	0.000	-0.024	-0.011
average_rating	0.0211	0.003	7.893	0.000	0.016	0.026

Omnibus:	18.260	**Durbin-Watson:**	2.246
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	27.332
Skew:	-0.524	**Prob(JB):**	1.16e-06
Kurtosis:	4.402	**Cond. No.**	197.

Table 14: OLS regression results predicting cosine similarity among "same" target words

Dep. Variable:	Cosine Similarity	**R-squared:**	0.343
Model:	OLS	**Adj. R-squared:**	0.334
Method:	Least Squares	**F-statistic:**	36.58
Date:	Sat, 12 Mar 2022	**Prob (F-statistic):**	4.63e-19
Time:	12:20:20	**Log-Likelihood:**	235.24
No. Observations:	214	**AIC:**	-462.5
Df Residuals:	210	**BIC:**	-449.0
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	0.9469	0.062	15.214	0.000	0.824	1.070
avg_freq	-0.0161	0.003	-4.983	0.000	-0.022	-0.010
average_rating	0.0198	0.003	7.417	0.000	0.015	0.025
avg_sense	-0.0192	0.007	-2.711	0.007	-0.033	-0.005

Omnibus:	13.882	**Durbin-Watson:**	2.255
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	18.177
Skew:	-0.458	**Prob(JB):**	0.000113
Kurtosis:	4.095	**Cond. No.**	212.

Table 15: OLS regression results predicting cosine similarity among "same" target words

Dep. Variable:	Cosine Similarity	**R-squared:**	0.059
Model:	OLS	**Adj. R-squared:**	0.058
Method:	Least Squares	**F-statistic:**	87.37
Date:	Sat, 12 Mar 2022	**Prob (F-statistic):**	3.41e-20
Time:	12:20:20	**Log-Likelihood:**	1557.3
No. Observations:	1406	**AIC:**	-3111.
Df Residuals:	1404	**BIC:**	-3100.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	0.7858	0.019	42.044	0.000	0.749	0.822
avg_freq	-0.0106	0.001	-9.347	0.000	-0.013	-0.008

Omnibus:	12.804	**Durbin-Watson:**	1.683
Prob(Omnibus):	0.002	**Jarque-Bera (JB):**	16.004
Skew:	-0.130	**Prob(JB):**	0.000335
Kurtosis:	3.453	**Cond. No.**	145.

Table 16: OLS regression results predicting cosine similarity among "different" target words

Dep. Variable:	Cosine Similarity	**R-squared:**	0.305
Model:	OLS	**Adj. R-squared:**	0.304
Method:	Least Squares	**F-statistic:**	614.9
Date:	Sat, 12 Mar 2022	**Prob (F-statistic):**	7.11e-113
Time:	12:20:20	**Log-Likelihood:**	1770.2
No. Observations:	1406	**AIC:**	-3536.
Df Residuals:	1404	**BIC:**	-3526.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	0.5366	0.004	150.800	0.000	0.530	0.544
average_rating	0.0208	0.001	24.796	0.000	0.019	0.022

Omnibus:	32.918	**Durbin-Watson:**	1.861
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	39.508
Skew:	-0.302	**Prob(JB):**	2.64e-09
Kurtosis:	3.556	**Cond. No.**	8.58

Table 17: OLS regression results predicting cosine similarity among "different" target words

Dep. Variable:	Cosine Similarity	**R-squared:**	0.336
Model:	OLS	**Adj. R-squared:**	0.335
Method:	Least Squares	**F-statistic:**	355.7
Date:	Sat, 12 Mar 2022	**Prob (F-statistic):**	1.12e-125
Time:	12:20:20	**Log-Likelihood:**	1803.2
No. Observations:	1406	**AIC:**	-3600.
Df Residuals:	1403	**BIC:**	-3585.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	0.6684	0.016	40.691	0.000	0.636	0.701
avg_freq	-0.0079	0.001	-8.210	0.000	-0.010	-0.006
average_rating	0.0200	0.001	24.238	0.000	0.018	0.022

Omnibus:	35.771	**Durbin-Watson:**	1.832
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	44.869
Skew:	-0.305	**Prob(JB):**	1.81e-10
Kurtosis:	3.628	**Cond. No.**	156.

Table 18: OLS regression results predicting cosine similarity among "different" target words

Dep. Variable:	Cosine Similarity	**R-squared:**	0.337
Model:	OLS	**Adj. R-squared:**	0.335
Method:	Least Squares	**F-statistic:**	237.1
Date:	Sat, 12 Mar 2022	**Prob (F-statistic):**	2.09e-124
Time:	12:20:20	**Log-Likelihood:**	1803.4
No. Observations:	1406	**AIC:**	-3599.
Df Residuals:	1402	**BIC:**	-3578.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	0.6670	0.017	40.027	0.000	0.634	0.700
avg_freq	-0.0076	0.001	-7.044	0.000	-0.010	-0.005
average_rating	0.0199	0.001	23.983	0.000	0.018	0.022
avg_sense	-0.0010	0.002	-0.516	0.606	-0.005	0.003

Omnibus:	36.276	**Durbin-Watson:**	1.832
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	45.556
Skew:	-0.308	**Prob(JB):**	1.28e-10
Kurtosis:	3.632	**Cond. No.**	160.

Table 19: OLS regression results predicting cosine similarity among "different" target words

Dep. Variable:	Human Rating	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	0.001
Method:	Least Squares	**F-statistic:**	3.074
Date:	Sat, 12 Mar 2022	**Prob (F-statistic):**	0.0797
Time:	13:15:45	**Log-Likelihood:**	-3750.9
No. Observations:	1620	**AIC:**	7506.
Df Residuals:	1618	**BIC:**	7517.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	5.0152	0.538	9.330	0.000	3.961	6.070
avg_freq	-0.0568	0.032	-1.753	0.080	-0.120	0.007

Omnibus:	229.333	**Durbin-Watson:**	1.972
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	91.858
Skew:	0.385	**Prob(JB):**	1.13e-20
Kurtosis:	2.124	**Cond. No.**	147.

Table 20: OLS regression results predicting average human ratings.

Dep. Variable:	Human Rating	**R-squared:**	0.404
Model:	OLS	**Adj. R-squared:**	0.403
Method:	Least Squares	**F-statistic:**	1096.
Date:	Sat, 12 Mar 2022	**Prob (F-statistic):**	6.45e-184
Time:	13:15:45	**Log-Likelihood:**	-3333.6
No. Observations:	1620	**AIC:**	6671.
Df Residuals:	1618	**BIC:**	6682.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	-6.2058	0.314	-19.748	0.000	-6.822	-5.589
cosine_similarity	16.3453	0.494	33.101	0.000	15.377	17.314

Omnibus:	25.721	**Durbin-Watson:**	1.974
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	24.246
Skew:	0.260	**Prob(JB):**	5.43e-06
Kurtosis:	2.703	**Cond. No.**	14.7

Table 21: OLS regression results predicting average human ratings.

Dep. Variable:	Human Rating	**R-squared:**	0.408
Model:	OLS	**Adj. R-squared:**	0.407
Method:	Least Squares	**F-statistic:**	371.8
Date:	Sat, 12 Mar 2022	**Prob (F-statistic):**	1.31e-183
Time:	13:15:45	**Log-Likelihood:**	-3327.3
No. Observations:	1620	**AIC:**	6663.
Df Residuals:	1616	**BIC:**	6684.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	-7.9168	0.575	-13.778	0.000	-9.044	-6.790
avg_freq	0.0989	0.028	3.473	0.001	0.043	0.155
avg_sense	-0.0440	0.048	-0.911	0.362	-0.139	0.051
cosine_similarity	16.6654	0.500	33.304	0.000	15.684	17.647

Omnibus:	25.797	**Durbin-Watson:**	1.972
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	22.821
Skew:	0.235	**Prob(JB):**	1.11e-05
Kurtosis:	2.657	**Cond. No.**	252.

Table 22: OLS regression results predicting average human ratings.

Dep. Variable:	Human Rating	**R-squared:**	0.443
Model:	OLS	**Adj. R-squared:**	0.442
Method:	Least Squares	**F-statistic:**	428.7
Date:	Sat, 12 Mar 2022	**Prob (F-statistic):**	7.28e-205
Time:	13:15:45	**Log-Likelihood:**	-3278.2
No. Observations:	1620	**AIC:**	6564.
Df Residuals:	1616	**BIC:**	6586.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	-4.2809	0.379	-11.310	0.000	-5.023	-3.539
avg_sense	-0.1339	0.044	-3.012	0.003	-0.221	-0.047
cosine_similarity	13.5126	0.547	24.707	0.000	12.440	14.585
same_word	1.7228	0.161	10.668	0.000	1.406	2.040

Omnibus:	24.052	**Durbin-Watson:**	2.007
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	20.099
Skew:	0.203	**Prob(JB):**	4.32e-05
Kurtosis:	2.635	**Cond. No.**	46.2

Table 23: OLS regression results predicting average human ratings.

Dep. Variable:	Human Rating	**R-squared:**	0.446
Model:	OLS	**Adj. R-squared:**	0.444
Method:	Least Squares	**F-statistic:**	324.7
Date:	Sat, 12 Mar 2022	**Prob (F-statistic):**	3.91e-205
Time:	13:15:45	**Log-Likelihood:**	-3274.5
No. Observations:	1620	**AIC:**	6559.
Df Residuals:	1615	**BIC:**	6586.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	-5.5590	0.600	-9.258	0.000	-6.737	-4.381
avg_freq	0.0757	0.028	2.738	0.006	0.021	0.130
avg_sense	-0.1892	0.049	-3.881	0.000	-0.285	-0.094
cosine_similarity	13.8092	0.556	24.816	0.000	12.718	14.901
same_word	1.6872	0.162	10.435	0.000	1.370	2.004

Omnibus:	24.612	**Durbin-Watson:**	2.005
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	19.555
Skew:	0.187	**Prob(JB):**	5.67e-05
Kurtosis:	2.612	**Cond. No.**	285.

Table 24: OLS regression results predicting average human ratings.

Dep. Variable:	Radius of Bounding Ball	**R-squared:**	0.477
Model:	OLS	**Adj. R-squared:**	0.477
Method:	Least Squares	**F-statistic:**	1141.
Date:	Sat, 12 Mar 2022	**Prob (F-statistic):**	2.96e-178
Time:	15:46:57	**Log-Likelihood:**	-2045.0
No. Observations:	1253	**AIC:**	4094.
Df Residuals:	1251	**BIC:**	4104.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	5.5878	0.187	29.926	0.000	5.221	5.954
log2(freq)	0.3927	0.012	33.774	0.000	0.370	0.416

Omnibus:	15.637	**Durbin-Watson:**	2.053
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	15.928
Skew:	-0.275	**Prob(JB):**	0.000348
Kurtosis:	3.052	**Cond. No.**	86.0

Table 25: OLS regression results predicting radius of bounding ball using frequency

Dep. Variable:	Radius of Bounding Ball	**R-squared:**	0.448
Model:	OLS	**Adj. R-squared:**	0.448
Method:	Least Squares	**F-statistic:**	1015.
Date:	Sat, 12 Mar 2022	**Prob (F-statistic):**	1.25e-163
Time:	15:46:57	**Log-Likelihood:**	-2078.7
No. Observations:	1253	**AIC:**	4161.
Df Residuals:	1251	**BIC:**	4172.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	9.0630	0.093	97.878	0.000	8.881	9.245
log2(senses)	0.9765	0.031	31.866	0.000	0.916	1.037

Omnibus:	12.796	**Durbin-Watson:**	2.101
Prob(Omnibus):	0.002	**Jarque-Bera (JB):**	13.940
Skew:	-0.193	**Prob(JB):**	0.000940
Kurtosis:	3.344	**Cond. No.**	8.52

Table 26: OLS regression results predicting radius of bounding ball using senses

Dep. Variable:	Radius of Bounding Ball	**R-squared:**	0.583
Model:	OLS	**Adj. R-squared:**	0.582
Method:	Least Squares	**F-statistic:**	872.2
Date:	Sat, 12 Mar 2022	**Prob (F-statistic):**	7.47e-238
Time:	15:46:57	**Log-Likelihood:**	-1903.7
No. Observations:	1253	**AIC:**	3813.
Df Residuals:	1250	**BIC:**	3829.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
constant	6.0781	0.169	35.937	0.000	5.746	6.410
log2(freq)	0.2581	0.013	20.071	0.000	0.233	0.283
log2(senses)	0.5867	0.033	17.784	0.000	0.522	0.651

Omnibus:	21.564	**Durbin-Watson:**	2.097
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	23.741
Skew:	-0.272	**Prob(JB):**	6.99e-06
Kurtosis:	3.398	**Cond. No.**	88.6

Table 27: OLS regression results predicting radius of bounding ball using frequency and senses

Dep. Variable:	Cosine Similarity	**R-squared:**	0.169
Model:	OLS	**Adj. R-squared:**	0.169
Method:	Least Squares	**F-statistic:**	1103.
Date:	Sat, 12 Mar 2022	**Prob (F-statistic):**	2.51e-220
Time:	15:54:04	**Log-Likelihood:**	5534.8
No. Observations:	5412	**AIC:**	-1.107e+04
Df Residuals:	5410	**BIC:**	-1.105e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Constant	1.1096	0.010	111.569	0.000	1.090	1.129
Radius of Bounding Ball	-0.0255	0.001	-33.215	0.000	-0.027	-0.024

Omnibus:	1.512	**Durbin-Watson:**	1.721
Prob(Omnibus):	0.470	**Jarque-Bera (JB):**	1.543
Skew:	-0.027	**Prob(JB):**	0.462
Kurtosis:	2.938	**Cond. No.**	109.

Table 28: OLS regression results predicting cosine similarity using radius of the bounding ball.

	Pearson's R	p
Average Pairwise Euclidean Distance	0.601	< 0.001
Max Pairwise Euclidean Distance	0.584	< 0.001
Variance of Pairwise Euclidean Distance	0.292	< 0.001
Average Norm of Embeddings	0.678	< 0.001
Area of convex hull*	0.603	< 0.001

Table 29: Pearson's correlations for numerous other ways of measuring the space occupied by a sibling cohort of ten instances. *To measure the area of a convex hull, we used PCA to projected the embeddings into 2D space and calculated the area. Measuring the convex hull in 768-dimensional space would have required a lot more data (at least 769 samples).