To Know by the Company Words Keep and What Else Lies in the Vicinity

Anonymous ACL submission

Abstract

The development of state-of-the-art (SOTA) Natural Language Processing (NLP) systems has steadily been establishing new techniques to absorb the statistics of linguistic data. These techniques often trace well-known constructs from traditional theories, and we study these connections to close gaps around key NLP methods as a means to orient future work. For this, we introduce an analytic model of the statistics learned by seminal algorithms (including GloVe and Word2Vec), and derive insights for systems that use these algorithms and the statistics of co-occurrence, in general. In this work, we derive—to the best of our knowledge—the first known solution to Word2Vec's softmax-optimized, skip-gram algorithm. This result presents exciting potential for future development as a direct solution to a deep learning (DL) language model's (LM's) matrix factorization. However, we use the solution to demonstrate a seemingly-universal existence of a property that word vectors exhibit and which allows for the prophylactic discernment of biases in data—prior to their absorption by DL models. To qualify our work, we conduct an analysis of independence, i.e., on the density of statistical dependencies in co-occurrence models, which in turn renders insights on the distributional hypothesis' partial fulfillment by co-occurrence statistics.

1 Motivation

Suppose one wished to randomly optimize a Rube Goldberg machine (RGM) over many Dominoes with the intent of accomplishing a small downstream task. Should the RGM be initialized to a random state, with dominoes scattered haphazardly, i.e., with no prior? Or would it help more to constrain the RGM to initializations with all dominoes standing on end? Perhaps less effort could be used to modify the dominoes-on-end state for the goal—but that depends on the goal and how dominoes can be used to transfer energy over long ranges.

Pre-trained models are often used as *initializations*, eventually applied to downstream NLP tasks like part-of-speech tagging or machine translation. This means model pre-training is a lot like initializing an RGM to a highly-potentiated state, while retaining a flexibility/generality to optimize sharply for the diversity of phenomena which can depend on statistical, linguistic information. A challenge partly met by big data pre-training is with the need for models to remain useful on a large diversity of data and tasks. Under the RGM theory, pre-training over big data simply potentiates more dominoes, in more-usefully correlated ways, where 'useful' is hands-off defined by a model's parametric ability to explain language, i.e., which words were where. However, if we knew how many dominoes should be on end at the start and how many dominoes should be in configurations that make stairs, etc., it seems plausible to initialize the RGM with distributionally-useful tools, given what we know about how humans use dominoes to transfer energy, i.e., the statistics of how humans use vocabularies to communicate. We investigate these questions, replacing 'domino' with 'parameter', and lay the groundwork for provisioning statistical priors to efficiently meet model pre-training needs for future research, while uncovering a cost-effective method for probing the biases that DL models will learn if they train on specific data.

2 Introduction

GPT-3 is an off-the-shelf AI that is perhaps the pinnacle of LMs, and compared to GPT-2, was basically just trained on more data and with more parameters. The data that trained GPT-3's SOTA unsupervised machine translation performance were simply an "unfiltered distribution of languages reflected in internet text datasets" (Brown et al., 2020). To uncover how a blend of training data like this aligns the semantics of, e.g., French and English vocabularies requires explaining *what* (statis-

tically) NLP tasks are teaching to SOTA algorithms, e.g., via the language modeling or masked language modeling tasks. However, while off-the-shelf AIs like GPT-2 are becoming ubiquitous in applications, they're also being shown to contain dangerous biases that emerge from training data (Wallace et al., 2019; Heidenreich and Williams, 2021).

2.1 Related Work

Within the last decade, there have been major shifts in representation learning from context-independent word vectors (Mikolov et al., 2013a,b; Pennington et al., 2014), to context-dependent word representations (Howard and Ruder, 2018; Peters et al., 2018), to pre-trained language models (Devlin et al., 2019; Radford et al., 2018, 2019). These trends have been accompanied by large architectural developments from the dominance of RNNs (Hochreiter and Schmidhuber, 1997), to the appearance of attention (Bahdanau et al., 2015) and the proliferation of the Transformer architecture (Vaswani et al., 2017). Our work seeks to open a path towards the *efficient* engineering of SOTA NLP technologies. We aim to compute the natural statistics to which model parameters converge, and towards this our work analyzes the older, static-representations that preceded large LMs.

Despite gains on empirical benchmarks, recent works suggest surprising findings: word order may not matter as much in pre-training as previously thought (Sinha et al., 2021), random sentence encodings are surprisingly powerful (Wieting and Kiela, 2018), one can replace self-attention operations in BERT (Devlin et al., 2019) with unparameterized Fourier transformations and still retain 92% of the original accuracy on GLUE (Lee-Thorp et al., 2021), and many modifications to the Transformer architecture do not significantly impact model performance (Narang et al., 2021). There's no denying increases in empirical performance, but these confounding results raise questions about these models and the processing needed to perform NLP tasks.

3 Harmonically-Distributed Data

Historically, research has naïvely approached the characterization of language statistics by counting the occurrence of symbols. While *occurrence frequency* can be measured at different levels, e.g., characters, tokens, or phrases, a statistical ubiquity was discovered early on for tokens—specifically the harmonic relationship which exists in the usage of a document's vocabulary (Zipf, 1935, 1949). To understand the harmonic relationship, suppose a vocabulary $\mathcal{V}$ of $|\mathcal{V}| = N$ distinct *types* is used to convey a collection of documents, $\mathcal{D}$, containing M tokens. A harmonic analysis of $\mathcal{D}$ first *ranks* each $t \in \mathcal{V}$ with a positive integer r_t that sorts the vocabulary from high-to-low by frequency. Intuitively a rank, r_t indicates the number of *other* types which occur at least as often as t (without loss of generality). Via this ranking, the empirical occurrence frequency for any type, f_t, can be mathematically approximated by harmonically-proportioned values: $f_t \approx N \cdot r_t^{-1}$, where N scales models to have least-frequent types occur once. Crudely, harmonic distributions describe the bulk of statistical structure in token-frequency distributions.

3.1 Co-Occurrence and Context

Token co-occurrence matrices, i.e., co-frequency distributions, measure the number of times tokens appear, but specifically, 'near' one another. In general, for types t and s, we denote the occurrence of s in a fixed window of size $\pm m$ tokens around t across a collection, $\mathcal{D}$, by $F^m_{t,s}$. Most of the seminal representation-learning algorithms (including LSA, Word2Vec, and GloVe) rely on such empirical, m-sampled 'data' of co-occurrence. Here, it's important to note that totality in co-occurrence distributions is dependent on the size of the context window, i.e., co-occurrence marginalization, which we denote by M^m_F, exhibits how the distribution 'inflates' with larger values of m: $M^m_F = \sum_{t,s\in\mathcal{V}} F^m_{t,s} = \mathcal{O}\left(2mM\right)$. This m-window inflation thus slightly re-defines unigram statistics along marginals, denoted: $f^m_t = \sum_{l\in V} F^m_{t,l}$.

Generally, word co-occurrences define a specific family of word-context joint-distributional models, or, *context distributions*, which can be tuned, e.g., to count only forward, backward, or any uncentered 'windows' of context. These can likewise be generalized to n-gram context models (Piantadosi et al., 2011). While the over-counting effects of co-occurrence and n-gram contexts can be alleviated to form integrated higher-order models via weighted context distributions (Williams et al., 2015), no representations have to-date used these models. Here, our work is again retrospective, focusing on building solid foundations from the standard, symmetrically-centered word co-occurrence model of context, which has been used across the seminal word vector-learning algorithms.

4 Representation and Co-Occurrence

Harmonic distributional structures have long been observed, but applications of them to NLP systems have largely not emerged. We can juxtapose this lack of application to the transformative impact on NLP by representation learning's *embeddings*, or, *word vectors*. These allow modern DL systems to approximate the meanings of tokens. Since Latent Semantic Analysis (LSA) was introduced (Dumais et al., 1988), vector representations of tokens have been used to predict and retrieve synonyms and analogies (Mikolov et al., 2013a,b; Pennington et al., 2014). The fact that word vectors exhibit linear semantic relationships between tokens, i.e., predict analogies, is heralded as a success in their capture of meaning, but exists without solid understanding of how these meanings are captured. LSA has influenced theories about human cognition (Landauer and Dumais, 1997) and been used to measure association of concepts during free recall (Howard and Kahana, 2002; Zaromb et al., 2006). Word vectors are limited in representing polysemous words. However, as demonstrated in (Arora et al., 2018), polysemous words lie in a superposition of their senses within a linear semantic space, and one can approximately recover underlying sense vectors (Arora et al., 2018).

4.1 Modeling Co-Occurrence

The statistical dynamics of co-occurrence strongly depend on the hyper-parameter m, whose effect can be seen from a low-complexity model. Specifically, one can crudely sample from an empirical, harmonic-frequency distribution to retain some realistic structure. To compute a model $\hat{F}^m$ for a document collection, $\mathcal{D}$, a token t samples f_t windows of $\pm m$ other tokens s that are also distributed by f. This makes the sampling proportional to frequency ratios with t: $\hat{F}^m_{t,s} = C_t f_s / f_t$. To physicalize the model, one need only assert: $\sum_{s \in V} \hat{F}^m_{t,s} = 2m f_t$ and solve for the constant of proportionality, $2m f_t^2 / M$, allowing for a closed-form specification:

$$\hat{F}^m_{t,s} = \frac{2m f_t f_s}{M} \tag{1}$$

We refer to **Eq. 1** as the *independent frequencies model* (IFM), which forms a dense co-occurrence matrix that is computable from *any* set of unigram frequencies. To view this model, we present **Fig. 1**, which exhibits the IFM against co-occurrences of the word 'they' in the Georgetown University Multilayer (GUM) Corpus (Zeldes, 2017).

4.2 Co-Occurrence Factorizations

There is a deep connection between word representation algorithms and the factorization of token co-occurrence matrices. This connection is perhaps most transparent for the GloVe algorithm (Pennington et al., 2014), whose loss function is *defined* to factor the positive values of the log-co-occurrence matrix, and is minimized under frequency-dependent weights, W, to produce word vectors $\vec{u}_t, \vec{v}_s$ and bias parameters a_t, b_s that predict the values of F^m:

$$\sum_{t,s \in \mathcal{D}} W_{t,s} \left(\vec{u}_t \vec{v}_s^T + a_t + b_s - \log F^m_{t,s} \right)^2 \tag{2}$$

Under GloVe's loss function (Eq. 2), a perfect model's point of convergence would have zero-valued squared terms (Kenyon-Dean et al., 2020):

$$\log F^m_{t,s} = \vec{u}_t \vec{v}_s^T + a_t + b_s \tag{3}$$

Observing this point of convergence, (Kenyon-Dean et al., 2020) remark upon the variation exhibited by GloVe's vector products and bias terms, but provide little insight into *how* word vectors interact via inner products to produce PMI-like values. We investigate these details and discover critical, mechanical insights that will be used to produce a bias-probing methodology.

4.2.1 Clamped GloVe

Separating the effects of bias terms and vector products is essential for understanding GloVe's connection to other models, and can be achieved by introducing a 'clamping' hyper-parameter, $\kappa \in \{0, 1\}$, to turn on/off the bias terms. Multiplying this Boolean factor into the bias terms, GloVe's general factorization is:

$$\log F^m_{t,s} = \vec{u}_t \vec{v}_s^T + \kappa(a_t + b_s) \tag{4}$$

So, suppose GloVe is clamped ($\kappa = 0$) and that its data follow the IFM (Eq. 1). In this case, *vector differences*, e.g., between $\vec{u}_t$ and $\vec{u}_s$, act on *every* other token $w \in \mathcal{V}$'s v-vector as a constant: $(\vec{u}_t - \vec{u}_s) \cdot \vec{v}_w^T = \log(f_s / f_t)$. This then indicates that pairs of vectors with the same frequency ratio: $\frac{f_{s_x}}{f_{t_x}} = \frac{f_{s_y}}{f_{t_y}}$ have representations which operate semantically equivalently, under the GloVe model. We now emphasize the importance of this *frequency-ratios property* in describing model mechanics *across all* classical word-vector models.

4.2.2 The Frequency-Ratios Property

As it will be regularly discussed throughout the remainder of this document we formally define the *frequency-ratios property* for *any* classical word-vector representation, below.

Definition: Given two words from a vocabulary $t, s \in \mathcal{V}$ and any set of classical, IFM-trained word vectors: $U, V \in R^{|\mathcal{V}| \times k}$ $(k \leq |\mathcal{V}|)$, the *frequency-ratios property* exists when the action of *vector differences*, e.g., between U-vectors on *any* other token $w \in \mathcal{V}$'s V-vector is equal to the log-frequency ratio of t and s, regardless of w's choice:

$$(\vec{u}_t - \vec{u}_s) \cdot \vec{v}_w^T = \log \frac{f_s}{f_t} \tag{5}$$

We'll use the frequency-ratios property to efficiently measure the semantic bias in data that representations *would* learn. To get there, we will ultimately ask: does the linear-semantic analogy property (completing analogies by addition/subtraction) relate to a relationship of comparable unigram frequency ratios (and products)? For example, this asks if "man is to king as woman is to queen" is described in data by: $f_{\text{king}}/f_{\text{man}} \approx f_{\text{queen}}/f_{\text{woman}}$.

4.2.3 Un-Clamped GloVe

While a clamped model is technically less complex (having fewer predictive parameters), GloVe is often defined without clamping. In this case $(\kappa = 1)$, the connections between vector differences and model parameters become less clear. Current conjecture inclines bias parameters will converge to log-unigram-frequency values, leaving the vector products to model the point-wise mutual information (PMI) between tokens (Kenyon-Dean et al., 2020). Provided GloVe's inner products model the PMI, training GloVe on the IFM (Eq. 1) should force all vector products to zero: $\vec{u}_t \vec{v}_s^T = 0$. From this view, *more-independent word co-occurrences should produce less informative vector products, i.e., poorer GloVe models.* While the evidence for the bias-parameters' dependence on frequency is compelling, we note that in (Kenyon-Dean et al., 2020)'s comparison of bias terms with unigram frequency exhibited a *super-linear* trend, which from the logarithmic scale of presentation allows rough approximation by a power-law. Denoting a model exponent by $\gamma > 1$, one can estimate the un-clamped bias parameters' behavior as $e^{a_t}, e^{b_s} \propto (f_t^m)^\gamma, (f_s^m)^\gamma$. With an IFM defined by F^m-marginal frequencies, GloVe's

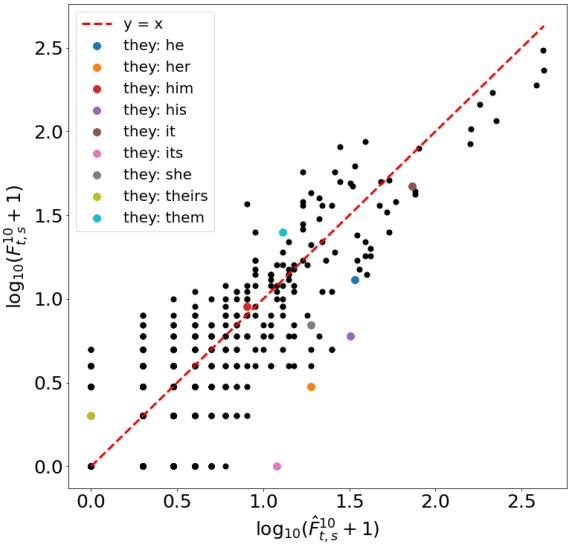

Figure 1: Comparison of the IFM and empirical co-occurrences for the word "they" within the GUM corpus. Statistical dependencies between words are the distances between points and the red dashed line. Unity is added to all points to clarify non-occurrent pairs.

γ-scaled PMIs exhibit a frequency-ratios property: $(\vec{u}_t - \vec{u}_s) \cdot \vec{v}_w^T = (\gamma - 1) \log(f_s^m / f_t^m)$. So GloVe's optimization away from *true* PMI avoids inner-product singularities under the IFM, ensuring the frequency-ratios property's presence.

4.2.4 Word2Vec Softmax

Prior to GloVe's development, the Word2Vec algorithm first emerged as a seminal advancement for word representation. While Word2Vec is perhaps most commonly applied under the skip-gram with negative sampling (SGNS) objective (Mikolov et al., 2013c), negative sampling objectives were originally developed to approximate more computationally complex softmax objectives (Mikolov et al., 2013a). Here, we investigate the effects of the IFM's co-occurrences on both objectives.

While it has been known for some time that the SGNS Word2Vec objective factorizes a shifted PMI matrix (Levy and Goldberg, 2014), the implicit matrix factorization behind Word2Vec's softmax objective to-date has not been derived. While this could be due to the softmax objective's mathematical complexity or the perceived lack of a factorization's utility (given unfactorized softmax's computational complexity), we show that neither is truly an obstacle and now derive the softmax factorization. While providing insight into *Word2Vec as an LM*, this presents an optimization strategy that makes the softmax objective much more computa-

tionally feasible, opening new potential for large scale applications, which we leave to future work.

Theorem: Under the log-softmax objective:

$$\mathcal{L}_{soft} = -\sum_{t \in V} \sum_{s \in V} F^m_{t,s} \log \varphi(\vec{u}_t \vec{v}_s), \qquad (6)$$

the Word2Vec algorithm implicitly converges towards a matrix factorization for all non-zero co-occurrences of the form:

$$\vec{u}_t \vec{v}_s^T = \log \frac{F^m_{t,s}}{f^m_t}, \qquad (7)$$

which is equal to the log-conditional probability matrix of the co-occurrence model.

The proof of this theorem is provided in **Appendix A**, which is the first known—to the best of our knowledge—proof of the softmax objective's factorization. This facorization produces a true LM, while Word2Vec's SGNS objective and GloVe do not. Historically, the softmax objective hasn't been utilized for pre-training applications due to computational complexity, and because SGNS has been seen as a partial approximation of softmax. However, our softmax solution provides a low-complexity strategy for pre-training—more powerful—*LM*-representations efficiently, via aggregated co-occurrences, just like GloVe's regression-based loss, e.g., replacing the rightmost term of **Eq. 2** with **Eq. 7**. This could have far reaching consequences, but devising new pre-training techniques was not the explicit intention of this proof. For us, the factorization elucidates the existence of the frequency-ratios property for what is arguably the most fundamental/influential classical word vector algorithm as a corollary (proof, **Appendix B**).

Corollary: When trained on the IFM, Word2Vec's softmax objective, $\mathcal{L}_{soft}$, exhibits the frequency-ratios property asymmetrically for differences of V-vectors acting on U (only).

Similar to *un-clamped* GloVe, the softmax skip-gram objective for Word2Vec only supports the frequency-ratios property on one side. The other side of its parameters could be responsible for maintaining the softmax's normalization and/or contrast. This could possibly explain why one of the U vs. V matrices' parameters have traditionally been preferentially retained, i.e., since only the V-vector differences are guarenteed to exhibit the frequency-ratios property when acting on U. However, we note that both U and V are intrinsically intertwined as two complimentary parts of the factorization. This also indicates that both U and V matrices should probably be retained for later use, and perhaps only ever combined by concatenation, since distribution in this form would allow other researchers apply full models. For example, this would allow for *vectors* trained by **Eq. 7** to be used *as a low-compute LM*.

4.2.5 Word2Vec-SGNS

We ask if the SGNS-Word2Vec objective also exhibits any frequency-ratios property. Here, we find asymmetric support again, and which is strikingly similar to that of un-clamped GloVe:

Theorem: The Word2Vec SGNS objective trains vectors which exhibit a frequency-ratios property scaled by one minus its sampling parameter: $\vec{u}_w(\vec{v}_t - \vec{v}_s)^T = (1-\alpha)\log(f^m_s / f^m_t)$.

This theorem (proof, **Appendix C**) shouldn't be too surprising, since SGNS *also* factors a PMI-like matrix (Kenyon-Dean et al., 2020). What is perhaps most surprising about this result is that SGNS' frequency-ratios property emerges directly from hyper-paramaterization via $\alpha > 0$, which tempers the negative-sampling rate as a power-law scaling of frequency. While α is generally presented with limited theoretical justification, its intent is accelerated learning, and its effect is biased (high-entropy) sampling during learning. Reflecting on this, it seems possible that un-clamping Glove induces γ in lieu of receiving a biased sample of contrastive information via α, as is done with SGNS. We likewise note that the piece-wise construction of W in GloVe's formulation complicates analysis, which could explain γ's limited presence over only the largest frequencies (Kenyon-Dean et al., 2020).

Considering how the frequency-ratios property appears ubiquitously across the diversity of classical word vector models *under the IFM*, we will examine the degree to which independence pervades co-occurrence models, below. However, with the frequency-ratios property in the focus, we now exhibit its immediate capacity to profile the semantic biases present in data.

5 Probing Data for Semantic Biases

The experiments described here draw from several publicly available data sets and intend to exhibit how analogies and token frequencies interact. Token-frequency distributions are taken from two corpora denoted by $\mathcal{G}$ and $\mathcal{W}$, correspond-

ing to Google Books' most recent N-grams release (Google, 2006) and a controlled collection of Wikipedia articles, described in detail below. Our interest with analogies is not in their prediction, and rather in developing a bias-probing methodology for evaluating *data*. So while a number of analogical test sets exist—including from the well-studied MSR collection (Mikolov et al., 2013a,b)—we utilize the Bigger Analogy Test Set (BATS) for its size, organization, and diversity, providing a total of roughly 10^5 analogical comparisons across categories (Gladkova et al., 2016).

Critically, BATS contains analogical comparisons for multiple encyclopedic groups. While analogical prediction experiments often perform well at the country-capital relationship, the more acute geographic category comparing UK counties and cities appears more challenging (Gladkova et al., 2016). We ask if this lower performance is due to poor representation in source data, i.e., if the relative abundance of language which discusses UK cities and counties *is low* in the data used to train word vectors that have been studied in the past. To examine this question, we will study the extent to which an intentionally-biased sample exhibits support for the UK city-county analogies.

5.1 Bias Measurement via Analogies

Stepping back, there should be no surprise if analogies can be used to directly measure bias in data. The WEAT test for measuring bias in word vectors (Caliskan et al., 2017) is based on four same-sized sets of words, which are referred to as target, e.g., gender-related words; and attribute, e.g., role-related words. Sampling one word from each of these sets essentially forms an analogy (even if nonsensical or offensive), and the WEAT formula measures bias via similarity statistics averaged across all comparisons. Furthermore, more recent methods for controlling bias in modern, self-attending systems retain this formulation (Karve et al., 2019), generalizing WEAT to four potentially-different-sized word sets, but again, with two for target words and two for attributes that can be used to constitute the dyads of hypothetical analogies.

To measure bias directly in data using analogies, denote each of the dyads, e.g., $(\text{man}, \text{woman})$, within a given analogy, $x : y$, as: $x = (t, s)$ and $y = (\tilde{t}, \tilde{s})$. On any given pair of dyads, we introduce the absolute difference of log-frequency ratios as a measurement of the *dissonance*, Δ, expressed towards the dyads, given a corpus, $\mathcal{D}$:

$$\Delta\left(x, y \mid \mathcal{D}\right) = \left|\log \frac{f_t f_{\tilde{s}}}{f_s f_{\tilde{t}}}\right| \Big/ \max_{l \in \mathcal{V}} \{\log f_l\} \qquad (8)$$

This quantity is entropicly 'normalized' by the largest value that the absolute difference could possibly take, occurring when a dyad of ratio 1 is compared to one with least-most frequent words. This places Δ in $[0, 1]$ and makes it possible to compare dissonance *between* corpora, i.e., the expressions of bias that data exhibit.

Musically, Δ, measures the degree to which the dyads are consonantly/dissonantly equivalent, i.e., whether the dyads play the same 'chord' (regardless of pitch). This is because Δ can be considered in terms of physical waves, i.e., modeling a document as a superposition of unit-amplitude square waves, whose peaks approximate the positions of each type's occurrences. Since physical waves of constant amplitude and velocity will have powers proportional to the squares of their frequencies, each token-frequency ratio becomes equivalent to the square root of two waves' power ratio. Thus when un-normalized, the dissonance can be understood by its units of *decibels*, which more broadly informs us that Δ measures an absolute difference in decibels expressed by each dyad, or, *the difference in loudness between the dyads' overtones*.

5.2 Analogical-Bias Probing Experiment

While it is customary to train word vectors on Wikipedia articles, we hypothesize that historical samples have had relatively few descriptive passages relating the UK cities to their counties, as compared to national capitols with their countries. The latter are likely more-broadly discussed on Wikipedia, and we hypothesize the former would be if their relative associations were more adequately represented in data. To ensure this, we composed our sample of Wikipedia articles, $\mathcal{W}$, from the collection of *all pages leading to and from any UK city's Wikipedia page,* where UK cities are strictly defined according to those listed on Wikipedia in its presentation of the UK's 69 officially-designated cities (as of 2021). This resulted in a Wikipedia corpus of roughly $200,000$ articles that is linked to the subjects of UK cities and counties.[1] By comparison of frequencies, $\mathcal{W}$ is about one-thousand times smaller than $\mathcal{G}$.

[1] Accessed 10/31/21: en.wikipedia.org/wiki/List_of_cities_in_the_United_Kingdom

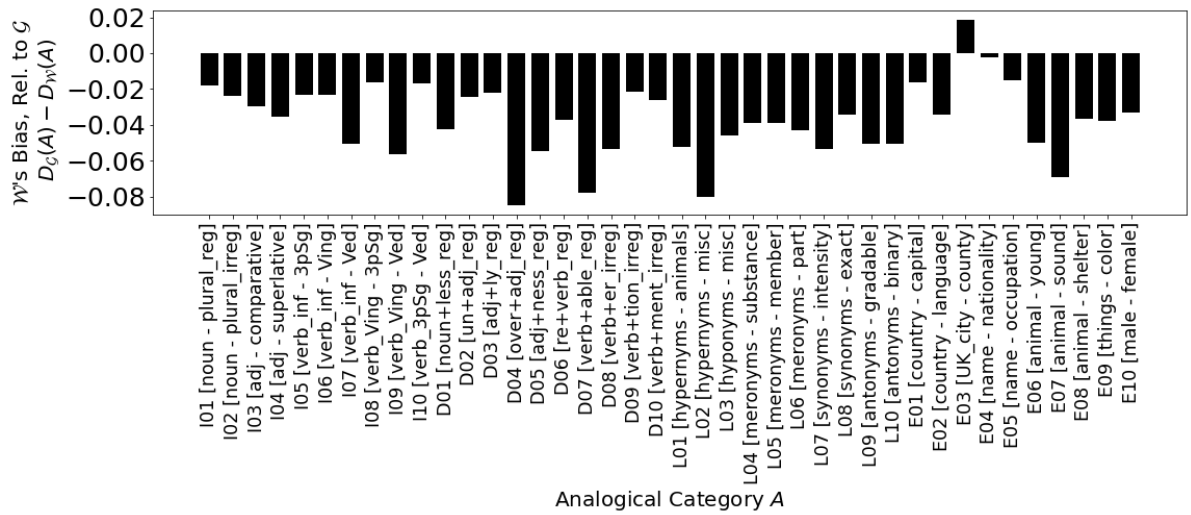

Figure 2: Comparison of the dissonance towards the different BATS analogy categories for the Google Book corpus, $\mathcal{G}$, and a much smaller corpus of Wikipedia articles that connect to pages discussing the UK cities. Positive bars indicate categories towards which $\mathcal{W}$ is more biased, i.e., which contain analogies that $\mathcal{W}$ supports more.

In application, low-dissonance values indicate which analogies are supported by corpus frequency ratios. To determine the *overall* support a corpus has for a *set* of analogies, A, an average, $D_{\mathcal{D}}(A)$, of Δ-dissonance values can be computed. Methodologically, we weight averages by the corpus frequencies of the tokens within each analogy's dyads. In **Fig. 2**, we compute the difference of averages between $\mathcal{G}$ and $\mathcal{W}$: $D_{\mathcal{G}}(A) - D_{\mathcal{W}}(A)$ for the different analogical categories of BATs, i.e., so that positive bars indicate where $\mathcal{W}$ supports a category's analogies more than $\mathcal{G}$. Average values on their own (no differences) for this experiment can likewise be observed within **Tab. 1** in the Appendices. Either view exhibits how the *only* category for which $\mathcal{W}$ exhibits less dissonance (more bias) than $\mathcal{G}$ is the UK city-county category, and furthermore, that this bias is clustered amongst related categories, e.g., name-nationality, which are elevated. We view these results as quite sensible for a bias metric, and indicating a promising pathway towards developing low-cost and -compute bias probes for data. This will feature centrally in our final discussion, after investigating the IFM's relevance to real-world corpora, both for its central role in elucidating the analogical bias probe methodology, and the paths it lays toward future discoveries.

6 Co-occurrence and Independence

Our study of the IFM and measurement of bias in data with analogical test sets raises an important question: *how relevant is the IFM to real-world data?* While it's not possible to objectively state if co-occurrences are independent or not, empirical systems do express independence on a spectrum. Determining the prevalence of independence in co-occurrence statistics requires control over the context model, i.e., m affects independence. To compare co-occurrence frequencies between the data and those sampled independently from marginal distributions, one can compare to independence, as measured by the IFM (**Fig. 1**). In the figure, the extent to which the empirical frequencies are equal to the IFM's can be quantified by how close points fall to the line $y = x$. Intuitively, this expresses how independent the empirical co-occurrences are when $m = 10$, and grounds the subject (independence) that we wish to study at larger scales of data.

6.1 Quantifying Independence

To measure statistical dependencies one can take the PMI's expectation over its joint probabilities and compute the co-occurrence mutual information (MI). MI measures how dependent the statistics of a joint distribution are. When MI is normalized by the joint information its values fall in $[0, 1]$ and define the information quality ratio (IQR): $\mathcal{I}_{k,m} = -\sum_{t,s \in V} P^m_{t,s} \log \frac{P^m_{t,s}}{p^m_t p^m_s} / \sum_{t,s \in V} P^m_{t,v} \log P^m_{i,j}$. Each of P^m and p^m are probabilistic forms of F^m and f^m (divided by M^m_F), and we will used k to record the number of documents in a given sample. Intuitively, $\mathcal{I}_{k,m}$ describes how close to independent co-occurrences are, and $\mathcal{I}_{k,m} \to 0$ indicates co-occurrences becoming more independent.

Linguistic dependencies are reported to exhibit power-law relationships between dependence length-d (the number of other words up through the dependence) and frequency (Chen and Gerdes, 2019). Through preliminary analysis of the GUM corpus', we find $\mathcal{I}_{k,m}$ values—statistical dependencies—*also* appear to decay as a power law function of the window size, m (**Appendix E.2**). Critically, we observe that the sentence-length distribution sets a bound on the background of contrastive information, saturating with increasing values of m (**Fig. 3**). However, for even the smallest-k corpora and *any* values of m, $\mathcal{I}_{k,m}$ appears less than 0.5, suggesting in some sense that *co-occurrences are more independent than dependent*. However, this view is reductive and leaves some critical questions, which we begin to address below.

Does the IQR measure linguistic dependencies? Beyond observing $\mathcal{I}_{k,m} < 0.5$, we evaluate $\mathcal{I}_{k,m}$ more broadly in **Appendix E**, where a full profile of $\mathcal{I}_{k,m}$-values is provided for GUM. We likewise model the IQR as a power-law function of the context-window size, m, to elucidate if and how linguistic dependencies contribute to the statistical dynamics of co-occurrence dependence, as measured by $\mathcal{I}_{k,m}$. Seemingly, $\mathcal{I}_{k,m}$ can be modeled strongly, when properly modulated by the sentence length distribution. This means sentence tokenization plays a central role in defining co-occurrence statistics. Likewise, lower-quality sentence tokenizations seem to result in more-complex distributions, optimization challenges, and noisy $\mathcal{I}_{k,m}$-profiles. Perhaps most surprisingly, the power-law which models $\mathcal{I}_{k,m}$'s *statistical dependencies* appears to exhibit a scaling exponent, ν, which parametrically fits the density of *linguistic dependencies* annotated in GUM (**Fig. 3**, inset). Further experimentation on different, parsed corpora is clearly required to determine if this model and relationship are robust.

Does the IQR have a lower bound? Determining this requires measuring and fitting $\mathcal{I}_{k,m}$ for larger data sets. Combinatorality imposes significant computational challenges for large values of m, so lower-m values (a smaller window) were used to measure the IQR for larger values of k (with more data). As there is no way to measure $\mathcal{I}_{k,m}$ for arbitrary corpus sizes, limiting arguments are ultimately required (**Appendix. E.4**). We find that the k-limiting dynamics of $\mathcal{I}_{k,m}$-values appear non-zero and convergent with bounds that can be solved (**Eq. 12**) and computed (**Fig. 4**).

7 Discussion and Conclusions

The gravity of the IQR's lower bound should not be understated: even a countably-infinite collection of documents will retain a definite portion of dependent statistical information in its co-occurrences. In some sense, this assures the statistical need for large corpora to 'chip away' at the underlying statistical dependencies recorded in linguistic data. However, while convergence is rapid at first, it slows considerably for larger corpora, indicating ever-diminishing returns from bigger data. From our bias-probing experiments, we exhibit how more data isn't necessarily more representative (**Fig. 2**). Thus, we ask if the IQR's limiting behavior is a process of document structure washing out in favor of more-local relationships. If so, we might then interpret **Fig. 2** as exhibiting a corpus whose co-occurrences have been unusually constrained for its document distribution, providing another interpretation of semantic bias.

Seemingly, statistical dependencies are sparse in sentences, and m-word sliding-window context models can't separate these from independent variation while absorbing co-occurrences. This interpretation can be intuitively stated by modifying Firth's famous quote (hence the paper's title): *You shall know a word by the company it keeps, and what else lies in the vicinity.* However, we know less if it is us who know words by this truism, as much as it is AIs who *know* how to use words by it.

Arriving at this point has entailed the development of novel techniques for probing unstructured, linguistic data for semantic biases using data sets of analogies. On their own, these results appear positive, and exhibit their own methodological value. However, they likewise emerged from another discovery, of the universal, frequency-ratios property for word vectors. The substantiation of that property required deriving a limiting factorization for the *original* Word2Vec objective, whose apparent natural formulation as *the* contextualizing LM produced by the co-occurrence conditional-probably-matrix underpins its importance. This production of a closed-form solution to Word2Vec could perhaps produce the biggest impacts of this work by providing rich new pathways towards efficiently deriving representation statistics. To this end, we highlight the IFM's derivation as another core outcome of our work, both for its central roles in analysis, and it's potential to warm-start embedding layers efficiently via unigram statistics.

References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

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

Xinying Chen and Kim Gerdes. 2019. The relation between dependency distance and frequency. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 75–82, Paris, France. Association for Computational Linguistics.

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.

Susan T Dumais, George W Furnas, Thomas K Landauer, Scott Deerwester, and Richard Harshman. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–285.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.

Google. 2006. All our n-gram are belong to you. https://ai.googleblog.com/2006/08/all-our-n-gram-are-belong-to-you.html. Accessed: May 27th, 2021.

Hunter Scott Heidenreich and Jake Ryland Williams. 2021. The earth is flat and the sun is not a star: The susceptibility of gpt-2 to universal adversarial triggers. In *Proceedings of the Fourth International AAAI/ACM Conference on Artificial Intelligence, Ethics and Society*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Lstm can solve hard long time lag problems. *Advances in neural information processing systems*, pages 473–479.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Marc W Howard and Michael J Kahana. 2002. When does semantic similarity help episodic retrieval? *Journal of Memory and Language*, 46(1):85–98.

Saket Karve, Lyle Ungar, and João Sedoc. 2019. Conceptor debiasing of word representations evaluated on WEAT. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 40–48, Florence, Italy. Association for Computational Linguistics.

Kian Kenyon-Dean, Edward Newell, and Jackie Chi Kit Cheung. 2020. Deconstructing word embedding algorithms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8479–8484, Online. Association for Computational Linguistics.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2021. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27:2177–2185.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *CoRR*, abs/1609.07843.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pages 3111–3119.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Sharan Narang, Hyung Won Chung, Yi Tay, William Fedus, Thibault Fevry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, et al. 2021. Do transformer modifications transfer across implementations and applications? *arXiv preprint arXiv:2102.11972*.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv preprint arXiv:2104.06644*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162, Hong Kong, China. Association for Computational Linguistics.

John Wieting and Douwe Kiela. 2018. No training required: Exploring random encoders for sentence classification. In *International Conference on Learning Representations*.

Jake Ryland Williams, Eric M. Clark, James P. Bagrow, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Identifying missing dictionary entries with frequency-conserving context models. *Phys. Rev. E*, 92:042808.

Franklin M Zaromb, Marc W Howard, Emily D Dolan, Yevgeniy B Sirotin, Michele Tully, Arthur Wingfield, and Michael J Kahana. 2006. Temporal associations and prior-list intrusions in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(4):792.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

G. K. Zipf. 1935. *The Psycho-Biology of Language*. Houghton-Mifflin.

G. K. Zipf. 1949. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley.

A Word2Vec's Softmax Factorization

Theorem: Under the log-softmax objective:

$$\mathcal{L}_{soft} = -\sum_{t\in V}\sum_{s\in V} F^m_{t,s} \log \varphi(\vec{u}_t\vec{v}_s), \qquad (9)$$

the Word2Vec algorithm implicitly converges towards a matrix factorization for all non-zero co-occurrences of the form:

$$\vec{u}_t\vec{v}_s^T = \log \frac{F^m_{t,s}}{f^m_t}, \qquad (10)$$

which is equal to the log-conditional probability matrix of the co-occurrence model.

Proof: The softmax function is computed by row: $\varphi(\vec{u}_t\vec{v}_s^T) = e^{\vec{u}_t\vec{v}_s^T} / \sum_{l\in V} e^{\vec{u}_t\vec{v}_l^T}$. To solve for $\vec{u}_t\vec{v}_s^T$, we must determine all components of $\mathcal{L}_{soft}$'s gradient which depend on $\vec{u}_t\vec{v}_s^T$, and which arise from different portions of $\mathcal{L}_{soft}$'s Jacobian. This includes the positive, differential portion from the softmax's numerator: $-F^m_{t,s}\left(1-\varphi\left(\vec{u}_t\vec{v}_s^T\right)\right)$ as well as the negative, differential portion emerging from the softmax denominators: $\sum_{l\in V\setminus\{s\}} F^m_{t,l}\varphi\left(\vec{u}_t\vec{v}_s^T\right)$, which sums over all $l \neq s$, since softmax's derivative is vector valued.

By combining the negative and positive portions, the partial derivative of $\mathcal{L}_{soft}$ with respect to $\vec{u}_t\vec{v}_s^T$ is a sum which ranges over the entire vocabulary:

$$\frac{\partial\mathcal{L}_{soft}}{\partial\left(\vec{u}_t\vec{v}_s^T\right)}\left(\vec{u}_t\vec{v}_s^T\right) = -F^m_{t,s} + \sum_{l\in V} F^m_{t,l}\varphi\left(\vec{u}_t\vec{v}_s^T\right)$$

When set equal to zero, the sum is easily solved:

$$\varphi\left(\vec{u}_t\vec{v}_s^T\right) = \frac{F^m_{t,s}}{\sum_{l\in V} F^m_{t,l}} = \frac{F^m_{t,s}}{f^m_t} \tag{11}$$

where the co-occurrence m-window 'inflation' defines the unigram statistics as: $f^m_t = \sum_{l\in V} F^m_{t,l}$ by the t^{th} marginal sum, i.e., pushing the factorization towards the log-conditional probability matrix

Eq. 11 almost provides the main result, but only factorizes the softmax's application. Due to normalization, there will necessarily be error from the log-conditional probability matrix, which we handle by defining some β_t close to 1 in $(0, 1)$ for each $t \in V$. Selecting these values can used to produce an ansatz solution, which can be used to understand the limiting matrix being factorized, and hence algebraically solve for arbitrarily-well optimized softmax representations. First, define the ansatz's positive-occurring elements by:

$$\vec{u}_t\vec{v}_s^T = \log \beta_t \frac{F^m_{t,s}}{f^m_t}$$

Then define t's n_t negative-occurring elements by:

$$\vec{u}_t\vec{v}_l^T = \log \frac{1-\beta_t}{n_t}$$

Under this initialization, the error for the positive-occurrence pairs, $\varepsilon_{t,s}$, is determined as:

$$\varepsilon_{t,s} = \frac{F^m_{t,s}}{f^m_t} - \varphi\left(\vec{u}_t\vec{v}_s^T\right) = \frac{F^m_{t,s}}{f^m_t}(1-\beta_t)$$

Likewise, we can also now easily observe the error for the non-occurrent pairs:

$$\varepsilon_{t,l} = -\varphi\left(\vec{u}_t\vec{v}_l^T\right) = \frac{1-\beta_t}{k_t}$$

Critically, these errors diminish as $\beta_t \to 1$. Furthermore, driving $\beta_t \to 1$ reduces the negative log likelihood as it pushes the true co-occurrent factorized values towards the claimed limiting solution. This also indicates that the softmax model likely has no exact algebraic solution for its factorization. Specifically, while positive-occurring entries converge toward:

$$\lim_{\beta_t\to 1} \vec{u}_t\vec{v}_s^T = \log \frac{F^m_{t,s}}{f^m_t}$$

the non-occurring pairs in the factorized matrix have values which become ever more negative:

$$\lim_{\beta_t\to 1} \vec{u}_t\vec{v}_l^T = -\infty$$

This is generally the case for GloVe and Word2Vec's SGNS objective, too, as neither is defined on negative-occurring values and would require a similar, negative-diverging ansatz for an algebraic solution to their factorizations. This concludes the main proof, and now allows for investigation of how unigram-frequency ratios interact with vector differences. ■

B Softmax Frequency Ratios

Corollary: When trained on the IFM, Word2Vec's softmax objective, $\mathcal{L}_{soft}$, exhibits the frequency-ratios property asymmetrically for differences of V-vectors acting on U (only).

Proof: Substituting the IFM into the solved softmax-Word2Vec factorization (**Eq. 10**), we find:

$$\vec{u}_t(\vec{v}_s - \vec{v}_{\tilde{s}})^T = \log \frac{f_s}{f_{\tilde{s}}}$$

which is precisely the frequency-ratios property. However, when we apply this analysis symmetrically we find something different:

$$(\vec{u}_t - \vec{u}_{\tilde{t}})\vec{v}_s^T = \log \frac{\beta_t f_t f^m_{\tilde{t}}}{\beta_{\tilde{t}} f_{\tilde{t}} f^m_t},$$

which depends on co-occurrence 'inflation', as well as the ansatz's choice of β_t values. This is yet another, different form of the frequency-ratios property, where if the β_t values are chosen proportional to their respectively-*inflated* unigram frequencies: $\beta_t = f^m_t / M^m_F$, the exact frequency-ratios property is recovered to a full symmetry. ■

C SGNS Frequency Ratios

Theorem: The Word2Vec SGNS objective trains vectors which exhibit a frequency-ratios property scaled by one minus its sampling parameter: $\vec{u}_w(\vec{v}_t - \vec{v}_s)^T = (1-\alpha)\log(f^m_s / f^m_t)$.

Proof: In (Kenyon-Dean et al., 2020)'s work, the noise distribution was assumed different from convention, which utilizes a hyper-parameter, $\alpha \in R$, commonly set to $\alpha = 3/4$. Its general effect will modify the PMI-convergence points into:

$$\vec{u}_t\vec{v}_s^T = -\log\left[\frac{F^m_{t,s}}{f^m_t}\frac{M^{m,\alpha}_f}{(f^m_s)^\alpha}\right] + \log k.$$

where $M^{m,\alpha}_f$ is the corresponding normalization constant for an α-power, m-inflated unigram frequency distribution: $M^{m,\alpha}_f = \sum_{s\in V}(f^m_s)^\alpha$.

Observing the V-vector-difference action on U and When one defines an IFM by the inflated unigram statistics and substitutes the corresponding $\hat{F}^m$in for F^m another frequency-ratios property emerges, but this time with effect scaled by $1 - \alpha$:

$$\vec{u}_w(\vec{v}_t - \vec{v}_s)^T = (1 - \alpha) \log \frac{f_s^m}{f_t^m}$$

So when one sets $\alpha \neq 1$, the frequency-ratios property appears again, for Word2Vec's V-vectors on the U matrix. Note that the only value of α for which this property doesn't exist ($\alpha = 1$) is generally not utilized in applications, with most discussion usually asserting that $\alpha = 1$ produces less adept models.

For SGNS, a frequency-ratios property is only clearly entailed for U-vector differences on the V matrix. However, it appears that the frequency-ratios property for U-vector differences on V should be absent, as SGNS's negative information/normalization is noisily rigid, based entirely on independent sampling at a fixed rate of k-to-1:

$$(\vec{u}_t - \vec{u}_s)\vec{v}_w^T = -\log \frac{\hat{F}_{t,w}^m f_s^m}{\hat{F}_{s,w}^m f_t^m} = 0$$

A fuller analysis of the frequency-ratios property for the SGNS objective (as well as for softmax) would ultimately benefit from limiting analysis of the gradient descent process. While this is partly considered for softmax in the context of an ansatz solution, further discussions of limiting effects and optimization is left for future work. ■

D Bias Probing Experiment

BATS Category	Dissonance (D)	
Inflection	$\mathcal{G}$	$\mathcal{W}$
I01: noun-plural_reg	**0.035**	0.053
I02: noun-plural_irreg	**0.044**	0.068
I03: adj-comparative	**0.038**	0.068
I04: adj-superlative	**0.032**	0.068
I05: verb_inf-3pSg	**0.033**	0.057
I06: verb_inf-Ving	**0.03**	0.054
I07: verb_inf-Ved	**0.025**	0.076
I08: verb_Ving-3pSg	**0.048**	0.064
I09: verb_Ving-Ved	**0.036**	0.092
I10: verb_3pSg-Ved	**0.04**	0.057
Derivation	$\mathcal{G}$	$\mathcal{W}$
D01: noun+less_reg	**0.076**	0.118
D02: un+adj_reg	**0.048**	0.072
D03: adj+ly_reg	**0.045**	0.067
D04: over+adj_reg	**0.066**	0.151
D05: adj+ness_reg	**0.074**	0.129
D06: re+verb_reg	**0.073**	0.11
D07: verb+able_reg	**0.1**	0.178
D08: verb+er_irreg	**0.061**	0.114
D09: verb+tion_irreg	**0.061**	0.082
D10: verb+ment_irreg	**0.039**	0.065
Lexicography	$\mathcal{G}$	$\mathcal{W}$
L01: hypernyms-animals	**0.159**	0.212
L02: hypernyms-misc	**0.1**	0.181
L03: hyponyms-misc	**0.119**	0.165
L04: meronyms-substance	**0.084**	0.123
L05: meronyms-member	**0.072**	0.111
L06: meronyms-part	**0.123**	0.166
L07: synonyms-intensity	**0.098**	0.152
L08: synonyms-exact	**0.091**	0.125
L09: antonyms-gradable	**0.113**	0.163
L10: antonyms-binary	**0.122**	0.173
Encyclopedia	$\mathcal{G}$	$\mathcal{W}$
E01: country-capital	**0.051**	0.067
E02: country-language	**0.101**	0.135
E03: UK_city-county	0.081	**0.063**
E04: name-nationality	**0.06**	0.062
E05: name-occupation	**0.065**	0.08
E06: animal-young	**0.097**	0.147
E07: animal-sound	**0.094**	0.163
E08: animal-shelter	**0.095**	0.132
E09: things-color	**0.073**	0.11
E10: male-female	**0.072**	0.105

Table 1: Comparison of D for $\mathcal{G}$ and $\mathcal{W}$ (lower values mean less dissonance/more bias) over BATS analogies.

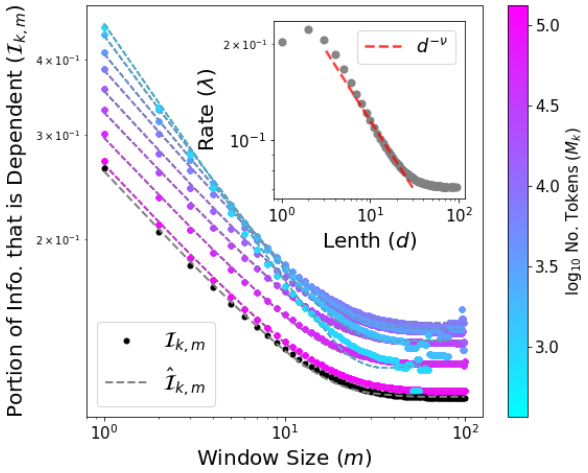

Figure 3: IQR profile for GUM (points), measured on 75 randomizations, scanning *all* values of m and k at powers of two, alongside dependence model fits (dashed lines). Color indicates the log-average number of tokens for each sample size, k. Inset shows the whole corpus' (black points/dashed lines) scaling exponent, ν, as a natural fit (red dashed line) for the cumulative-rate of dependencies, λ, that one can observe against the co-occurrence background (gray points), as the dependence length *and* co-occurrence window size ($d = m$) increase.

E Quantifying Independence

Here, we first empirically review the IQR's overall shape in the context of the Georgetown University Multilayer (GUM) Corpus (Zeldes, 2017), which affords opportunity to model the IQR as a function of the context-window size, m, allowing for determination of if and how linguistic dependencies contribute to the statistical dynamics of co-occurrence dependence, as measured by $\mathcal{I}_{k,m}$.

E.1 Measuring Dependencies Empirically

We perform samples amongst GUM's $k_{\max=150}$ documents. In Fig. 3, we observe that even for the highest-IQR (1-document) samples, $\mathcal{I}_{1,m} < 0.5$, i.e., the IQR's values are less than one half for all window and sample sizes. For now, we'll forego the effects of k and focus on how $\mathcal{I}_{k,m}$ is a decreasing function of m. This should be expected, i.e., that lower-m values exhibit less independence, and as we now show, this can be understood according to the dashed-line models that Fig. 3 exhibits.

E.2 Modeling the Density of Dependencies

In one sense, dependency parsing grammatically determines a rule for 'who' each given word's company is. By annotators, each dependency must be determined from the full range of co-occurrences available in the given sentence. As it turns out, dependencies are believed to have their own power-law statistical relationships between dependence length-d (the number of other words up through the dependence) and frequency (Chen and Gerdes, 2019). Previously, raw counts of dependencies were observed to form a power-law-like distribution that scaled as $d^{2.5}$. However, we wish to model dependencies against their background of co-occurrences in sentences. This means the sentence-length distribution modulates a critical bound on co-occurrence IQR. Next, we use the nature of how this background of contrastive information in sentences saturates with increasing m to model the lower limit of IQR values observed in Fig. 3. To produce this model, we first formally state our conjecture, and then derive the model.

Conjecture. Linguistic dependencies are 'the company words keep' from the distributional hypothesis, and underpin the statistical dependencies one can measure against the co-occurrence background via $\mathcal{I}_{k,m}$. We find support for this conjecture by developing a parametric model for $\mathcal{I}_{k,m}$, fitting it over the GUM corpus, and exhibiting how its fit corresponds to the density of linguistic dependencies against their co-occurrence background as the same power-law of m.

E.2.1 Forming Dependence Models

Define g_m to be the number of dependencies of a given length: $g_m \propto m^{-\nu}$, where ν is a positive, power-law scaling exponent. Alongside ν, we define a maximum dependence length, $m_{\max}$, as a model parameter to form the IQR's estimator as a function of the context window, m. First, we approximate the ν-power-law's cumulative distribution function over dependencies covered by the window m:

$$G_m = \int_0^m g_\ell d\ell \;=\; \frac{m^{1-\nu}}{m_{\max}{}^{1-\nu}}.$$

Here, totality requires setting $G_m = 1$ for all $m > m_{\max}$. This definition for G_m allows us to accurately show how independent statistics saturate co-occurrence models as m becomes large. However, predicting the IQR for empirical co-occurrences depends heavily on the sentence-length distribution, which determines how many co-occurrences exist per each center word. Any sentence of length L will induce $L(L-1)$ co-occurrences if $m \geq L$.

If $m < L$, the longer-range co-occurrences are ignored, making the general, total number of co-occurrences per sentence of length L equal to:

$$T_{L,m} = (\min(m, L) - 1) \cdot (2L - \min(m, L))$$

Supposing there are S_L sentences for each L and that the longest sentence length is $L_{\max}$, the total number of co-occurrences in an m-radius sliding-window model will be $T_m = \sum_{L=1}^{m} S_L T_{L,m}$, which defines the limiting-m co-occurrence model—with the longest range dependencies—by $T_{L_{\max}}$. Setting $T_0 = 0$ and denoting the total unigram frequency in a sample of size k as M_k allows us to define the co-occurrence sampling rate for all $m = 1, \cdots, L_{\max}$ as:

$$q_m = \frac{T_m - T_{m-1}}{2M_k}$$

Given any m, G_m, represents the cumulative portion of all dependences sampled from a $2m$-window, according to our base model. This base model is then modulated by the sampling rate, q_m, and scaled by a constant ρ:

$$\hat{\mathcal{I}}_{k,m} = \rho \cdot \left[q_m \frac{G_m}{2m} + (1 - q_m) \frac{M_k}{T_m} \right]$$

which parametrically defines the average number of dependencies word. Overall, this formulation can intuitively by understood to transfer—via the sampling rate—power-law varying dependence density smoothly, into a limiting ρ number of dependences per word in the sample: $\frac{M_k}{T_{L_m}}$. Via ρ, the model assumes on average that each word depends on ρ of the other words in the same sentence (co-occurrences), meaning $\rho \in [0, \frac{T_{L_{\max}}}{M}]$.

E.3 Fitting Dependence Models

As can be seen in Fig. 3, $\hat{\mathcal{I}}_{k,m}$ can be parameterized to fit the IQR quite strongly. Compared to the scale of co-occurrences absorbed by large representation models, the GUM data set is quite small. However, it affords a critical opportunity to observe how the power-law exponent, ν, corresponds to the density of *linguistic* dependencies against the co-occurrence background. Critically, we find in Fig. 3's inset a correspondence between ν and the rate of decay of linguistic dependencies against the co-occurrence background for full order of magnitude in the dependence-length distribution. While the scale of these results are small inside of GUM, they strongly support our conjecture. However, a number of questions and challenges emerge from these experiments. Practically, sampling from GUM has exhibited how the longest-range dependencies are simply not available to model in shorter sentences—the effects of small samples and varying sentence lengths can be seen in the empirical roughness for larger-m windows in smaller-k samples. Furthermore, larger samples pose combinatoral scaling for larger-m co-occurrences that makes direct measurement of the IQR prohibitive. To compound these issues, we seek to know if any limiting bounds on the IQR exist, i.e., we ask: "would the IQR from an infinite number of documents be zero?"

E.4 Bounding Dependence from Below

To see the effects of scale on the IQR, more data are required for experimentation than are available within the GUM corpus. Even if we can't expect linguistic dependencies to be annotated, we still wish to control for tokenization. Hence, to bound the IQR we work next with the well-known WikiText language modeling benchmark (Merity et al., 2016), which was expanded in v103 to over $30,000$ documents (roughly 200-times the size of GUM). Our objective in this section is to bound the IQR, and large-m measurement becomes intractable for large k-document samples. Hence, we will restrict $m \leq 10$ for all bounding experiments.

Suppose we fix any window size, m, let k be the number of documents in a sample, and $\mathcal{I}_{k,m}$ denote the IQR's average value. For any sample size, k, then let M_k denote the expected number of tokens in the sample, and observe that doubling k will double the expected number of tokens: $M_{2k} = 2M_k$. We're most interested in *the rate of IQR reduction incurred from doubling the number of sampled documents*, which we denote by $\delta_{2k,m}$. Averaging across samples, we have generally observed the IQR to fall into a well-ordered—perhaps power-law—pattern of decay (Fig. 4, left). This means that for large values, $k_1 \leq k_2$, one can expect: $\delta_{2k_2,m} \leq \delta_{2k_1,m}$. Next, we prove that these observed conditions result in the existence of a positive lower bound for the IQR, which exists below the IQR one could measure for *any* document-sample size, k.

Theorem. If the k-limiting behavior of the IQR-reduction rate is power-law decay: $\delta_{2k,m} \propto M_k^{\gamma}$, the IQR's limiting, $\mathcal{I}_{\infty,m}$, values are positive.

Proof. Supposing we fix any window size, m, let k be the number of documents in a sample,

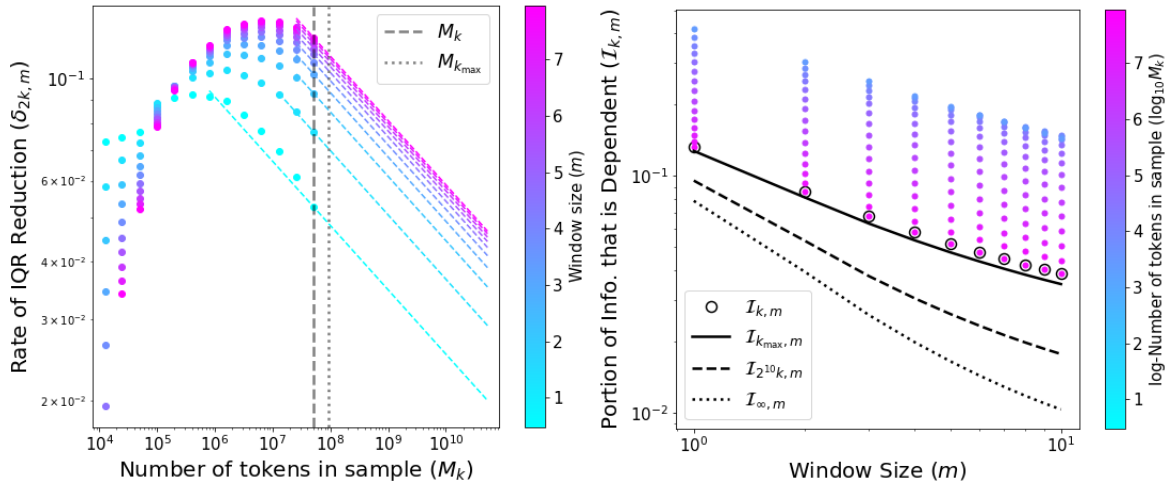

Figure 4: (Left) Rate of IQR reduction presented against the number of tokens in 75 samples of of size k taken at powers of two from the training articles of the WikiText-103 corpus for context-window sizes m up to 10 (indicated by color). Past the size of the whole corpus (black dotted line), the reduction rates are extrapolated with a power-law to model limiting behavior (spectrum of dashed lines). (Right) IQR profile for the WikiText-103 training corpus up to the largest two-power sample size (most pink) for context-window sizes up to $m = 10$. Past the largest-sampled size (circled points), the rate-reduction power law is used to extrapolate (e.g., the black dashed line) beyond the corpus (black dashed line) to compute non-zero limiting IQR values across m-window sizes.

and $\mathcal{I}_{k,m}$ denote the IQR's average value. For any sample size, k, let M_k denote the expected number of tokens in the sample, and observe that doubling k will double the expected number of tokens: $M_{2k} = 2M_k$. We next express the rate of IQR reduction incurred from doubling the number of sampled documents by $\delta_{2k,m}$. Averaging across samples, we have generally observed the IQR to fall into a well-ordered—perhaps power-law—pattern of decay. This means that for large values, $k_1 \leq k_2$, one can expect: $\delta_{2k_2,m} \leq \delta_{2k_1,m}$.

We wish to know about any limiting dynamics of $\mathcal{I}_{k,m}$ for large k, which under the observed pattern of decay ammounts to asking if the IQR converges to zero or a positive limit. In either case, we'll refer to any limiting quantity as $I_{\infty,m}$, which describes the portion of information that is dependent in a co-occurrence model of context for a *population* of data, i.e., an arbitrarily-large sample. Critically, this limit expresses the dependence in *how* a population of language was used, separately from the dependence on *what* its samples convey.

Assuming k is large enough to well order the reduction rate at a window size of m, we use $\delta\mathcal{I}_{2k,m}$ to write an IQR-update rule for doubled samples:

$$\mathcal{I}_{2k,m} = \mathcal{I}_{k,m}(1 - \delta_{2k,m}),$$

Applying recursion over this equation allows us to express IQR values for arbitrarily large samples:

$$\mathcal{I}_{2^n k,m} = \mathcal{I}_{k,m} \prod_{l=1}^{n} (1 - \delta_{2^l k,m})$$

However, to study a limiting value for the IQR we apply the reduction rate in series: $\mathcal{I}_{\infty,m} = \mathcal{I}_{k,m} - \sum_{n=0}^{\infty} \mathcal{I}_{2^n k,m} \delta_{2^{n+1} k,m}$. With this, we can substitute the product form for $\mathcal{I}_{2^n k,m}$ into our expression for $\mathcal{I}_{\infty,m}$ to produce:

$$\mathcal{I}_{\infty,m} = \mathcal{I}_{k,m} \left[1 - \sum_{n=0}^{\infty} \delta_{2^{n+1} k,m} \prod_{l=1}^{n} (1 - \delta_{2^l k,m}) \right]$$

Decreasing monotonicity in the reduction rate implies that the fastest-decaying extreme occurs when the reduction rate is a constant. Supposing this to be the case, we assume a critical sample size, k_m, past which a constant, δ_m, describes the reduction rate. When one substitutes this into our expression for the IQR's limit, a geomtric series emerges which unsurprisingly brings the IQR's limit to its low (0-valued) extreme:

$$\mathcal{I}_{\infty,m} = \mathcal{I}_{k_m,m} \left[1 - \delta_m \sum_{n=0}^{\infty} (1 - \delta_m)^n \right] = 0$$

One can in fact approximate the reduction rate empirically by computing a quotient of expected IQR values from samples of documents:

$$\delta_{2k,m} \approx \frac{E\left[\mathcal{I}_{k,m} - \mathcal{I}_{2k,m}\right]}{E\left[\mathcal{I}_{k,m}\right]}$$

When measured, we find that for k-samples larger than some critical sample size, k_m (dependent on m), the reduction rate appears to scale like a power law in the number of tokens sampled:

$$\delta_{2k,m} \approx \frac{M_k^{-\gamma}}{10^{b_m}}$$

This models $\delta_{2k,m}$ with a scaling exponent, γ, and constant of proportionality, 10^{b_m}, the latter of which is dependent upon the window size, m. Utilizing this empirically-motivated power-law, we obtain a different form for the limiting IQR:

$$\mathcal{I}_{\infty,m} = \mathcal{I}_{k,m}\left[1 - \sum_{n=0}^{\infty} \frac{M_{2^{n+1}k}^{-\gamma}}{10^{b_m}} \prod_{l=1}^{n}\left(1 - \frac{M_{2^l k}^{-\gamma}}{10^{b_m}}\right)\right]$$

To bound $\mathcal{I}_{\infty,m}$ from below one can replace each of the products of n with unity. The identity: $M_{2k} = 2M_k$ allows further generalization of the doubling numbers as: $M_{2^{n+1}k} = 2M_{2^n k}$, whose substitution into the bound produces a convenient form and infinite geometric series of ratio $2^{-\gamma}$:

$$\mathcal{I}_{k,m}\left[1 - \frac{M_{2k}^{-\gamma}}{10^{b_m}(1 - 2^{-\gamma})}\right] < \mathcal{I}_{\infty,m} \qquad (12)$$

To assure this lower bound is positive we now need only require:

$$\frac{M_{2k}^{-\gamma}}{10^{b_m}(1 - 2^{-\gamma})} < 1$$

which amounts to asserting that the average number of tokens, M_k, from the original sample size of k is sufficiently large to bound the the a positive-valued function parameterized by m via γ and b_m that notably has no dependence on k:

$$10^{-b_m/\gamma}(2^{\gamma} - 1)^{-1/\gamma} < M_k$$

Hence, positive lower bounds on the IQR could be confirmed experimentally for any given m by increasing the initial sample size of an the analysis. This likewise provides a means for the observing the elimination of transient independence in a co-occurrence model, where specifically, as k is increased, the positive, lower bound tightens to the limiting IQR value from below. ■

E.4.1 Computing a Bound for Dependence

Returning to the WikiText Corpus, we repeatedly sample the available powers of $k = 2^n$, which for WikiText-103 allows $n \leq 14$, since the total number of documents in the collection is roughly: $k_{\max} = 2^{15}$. These doubling samples are used to empirically compute our approximations of the reduction rate's average behavior in Fig. 4 (left), and the value of γ is optimized over only those values for this the power-law decay is apparent. Moreover, we constrain the values of b to satisfy a continuous model projecting from the data, i.e., optimization is only performed over γ. Once γ and the different values of b_m are established, 2^{10} iterative updates to , for 2^{15} to produce the large-k IQR values needed for stable computation the IQR's limit, $\infty,$. Each of these modeling components used in computing the IQR's bound is exhibited in Fig. 4 (right), and confirm the nature of our bounding result.