Information Processing and Management 58 (2021) 102493

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Zipfian regularities in "non-point" word representations

Furkan Şahinuç [a,b], Aykut Koç [a,c,*]

[a] Electrical and Electronics Engineering Department, Bilkent University, Ankara, Turkey
[b] ASELSAN Research Center, Ankara, Turkey
[c] National Magnetic Resonance Research Center (UMRAM), Bilkent University, Ankara, Turkey

ARTICLE INFO

Keywords:
Word variances
Word frequencies
Zipf's law
Meaning-frequency relation
Zipfian regularities
Word entailment
Semantic breadth

ABSTRACT

Being one of the most common empirical regularities, the Zipf's law for word frequencies is a power law relation between word frequencies and frequency ranks of words. We quantitatively study semantic uncertainty of words through non-point distribution-based word embeddings and reveal the Zipfian regularities. Uncertainty of a word can increase due to polysemy, the word having "broad" meaning (such as the relation between broader *emotion* and narrower *exasperation*) or a combination of both. Variances of Gaussian embeddings are utilized to quantify the extent a word can be used in different senses or contexts. By using the variance information embedded in the non-point Gaussian embeddings, we quantitatively show that semantic breadth of words also exhibits Zipfian patterns, when polysemy is controlled. This outcome is complementary to Zipf's law of meaning distribution and the related meaning-frequency law by indicating the existence of Zipfian patterns: more frequent words tend to be generic while less frequent ones tend to be specific. Results for two languages, English and Turkish that belong to different language families, are also provided. Such regularities provide valuable information to extract and understand relationships between semantic properties of words and word frequencies. In various applications, performance improvements can be obtained by employing these regularities. We also propose a method that leverages the Zipfian regularity to improve the performance of baseline textual entailment detection algorithms. To the best of our knowledge, our approach is the first quantitative study that uses Gaussian embeddings to examine the relationships between word frequencies and semantic breadth.

1. Introduction

Zipf's law for word frequencies is a kind of power law relation between frequency of words and their respective frequency ranks (Zipf, 1935, 1949). Since George Kingsley Zipf has introduced it, Zipf's law has been an important empirical regularity within the statistical natural language processing (NLP) and computational linguistics (Zipf, 1935, 1949) fields, and, in a more general sense, within the information, physical and social sciences (Newman, 2005). Despite Zipf's law's simple essence, several Zipfian regularities arise in complex linguistic relationships between word features and frequencies of words. The Zipf's law of meaning distribution and the related meaning-frequency relationship, and the law of abbreviation are common examples (Casas et al., 2019; Grzybek, 2006; Zipf, 1945, 1949). Interest in the Zipfian regularities has been continuing and remains as an important area of study (Altmann & Gerlach, 2016; Baayen, 2002; Casas et al., 2019; Debowski, 2002; Ferrer-i-Cancho & Elvevåg, 2010; Ferrer-i-Cancho & Solé, 2002; Ferrer-i-Cancho & Vitevitch, 2018; Gerlach & Altmann, 2014; Manin, 2008).

* Corresponding author.
E-mail addresses: furkan.sahinuc@bilkent.edu.tr (F. Şahinuç), aykut.koc@bilkent.edu.tr (A. Koç).

https://doi.org/10.1016/j.ipm.2021.102493
Received 29 August 2020; Received in revised form 4 January 2021; Accepted 4 January 2021

The Zipf's law of meaning distribution predicts a relationship between the number of meanings of a word and its frequency rank, (Zipf, 1945, 1949). The meaning-frequency relationship indicates a similar relationship where the predictor variable is the frequency itself, instead of the frequency rank, (Zipf, 1945). Both relations assert that the frequent words tend to have more number of meanings. Later, the relation between the number of meanings and word frequencies is studied by Ferrer-i-Cancho (2005a, 2005b, 2005c) within the scope of communicative optimization. Another recent and important study is Casas et al. (2019), where the authors develop a quantitative analysis to study the meaning-frequency relationship and provide further statistical evidence. The relation between the word length and the word frequencies is also statistically examined to verify the Zipf's law of abbreviation, which states that more frequent words tend to be shorter. All of the aforementioned studies play a significant role in providing quantitative evidence for Zipfian regularities.

In this paper, semantic "breadth" is quantitatively studied within the scope of Zipfian statistics. "Breadth" refers to the extent of meaning of a word when the effects of polysemy are removed. Uncertainty of a word increases due to polysemy, having "broad" meaning (such as the relation between broader *emotion* and narrower *exasperation*) or a combination of both. Manin (2008) refers the notion of meaning extent where words "broad" or "generic" are used on the one hand and "narrow" or "specific" on the other to refer the "extent" or "breadth" of a word's meaning. We quantitatively study the relationship between semantic breadth of a word and its frequency rank and reveal the Zipfian regularities: more frequent words tend to be semantically broader, even when the effects of polysemy are eliminated. Manin (2008) has touched upon this idea and predicted such a relation, where possible semantic volumes are mentioned together with the concepts of "generic" and "specific" words. By an approach using "non-point" word representations, our study presents theoretical results that complement the intuition and predictions in Manin (2008). Different from the previous work where the number of senses of a word is studied with respect to its frequency, we study and reveal the Zipfian regularities between semantic breadth, or uncertainty in the case when polysemy is controlled, of words and their frequency ranks.

In order to computationally study semantic breadth, one needs a mathematical representation (or embedding) that can capture the levels of the uncertainty of words. Continuous dense word embeddings opened a new era in NLP and computational linguistics (Bojanowski et al., 2017; Mikolov, Chen et al., 2013; Mikolov, Sutskever et al., 2013; Pennington et al., 2014). Besides their success in high level NLP tasks and applications, word embeddings are also frequently used in information processing field such as information retrieval from documents, (Bagheri et al., 2018), and social media user profiling, (Lopez-Santillan et al., 2020). Classical word embeddings treat all words as vectors representing points within a high-dimensional semantic vector space. For example, words "emotion" and "exasperation" are both represented as points. However, "emotion" contains a broad extent of meanings and can be used in different contexts and senses. In contrast, "exasperation" stands for only a particular feeling of intense irritation or annoyance, which is quite specific compared to generic "emotion". Although the neighbors of "emotion" come from diverse contexts compared to those of "exasperation" in a semantic vector space, "point-wise" word embeddings are not capable of semantically distinguishing words with broad meanings from words with narrow and specific meanings.

To address this problem, a new line of word embeddings has emerged. The idea is to represent words as Gaussian distributions (Athiwaratkun & Wilson, 2017; Vilnis & McCallum, 2015) within the semantic space. In Gaussian embeddings, words are represented as multivariate Gaussian distributions rather than points. Mean vectors of multivariate Gaussians decide the location of words in the semantic space. Covariance matrices of the multivariate Gaussian embeddings mathematically represent the uncertainty levels of words. The variance of a word is a measure that quantifies the breadth of meaning around the central semantic location of that word, which is represented by the mean vector. In other words, variances indicate to what extent words possess different senses and uncertain/broad meanings. Alternatively, variances can be interpreted as the semantic uncertainty/specificity of words around the central locations of their meanings. Uncertainty and specificity can also be interpreted as identifiers of possible senses of a word and the range of possible contexts within which a word can occur. Generally speaking, words with high uncertainty are more probable to co-occur with various words in many different contexts, whereas words indicating specific concepts can take place in a limited number of contexts. Gaussian embeddings quantitatively reflect this difference between generic and specific words to variance values. At first glance, this result is reasonable. For example, the word "cell" is semantically more generic than "lymphocytes" (a kind of white blood cell). Let alone the contribution coming from its other senses, since the word "cell" can be used with several cell types like "lymphocytes", its extent (variance) within the semantic space is supposed to be higher.

Variance measure of Gaussian embeddings enables us to study the semantic breadth quantitatively. With this tool and by controlling polysemy, one can compare the semantic breadth of words and look for regularities and relations between semantic breadth and frequency. By deploying Gaussian word embeddings, we quantitatively show that there are Zipfian regularities between frequency ranks and semantic breadths of words. We present results for English and Turkish languages, which belong to different language families. We also provide several experiments on sense-controlled corpora and on sub-corpora compiled from special words such as Swadesh List (Swadesh, 1950) and number words. The newly revealed Zipfian regularity is also applied to practical problems related to entailment relations.

Methodologically and conceptually, another closer work to our study is Piantadosi (2014). Piantadosi examined Zipf's law of word frequencies in various aspects of language and showed that the relation between word frequencies and their frequency ranks obeys at least near Zipfian regularities. In Piantadosi (2014), it has also been questioned whether some semantic and syntactic properties of languages exhibit Zipfian characteristics by using different external lexical resources. Similarly, we also perform experiments related to some subsets of our lexicon, such as Swadesh words and number words. Our results indicate that variances of words also display Zipfian patterns in such settings. In our study, we applied corpus related statistical processes in order to obtain more reliable and uncorrelated results from possible corpus-related errors as followed by Piantadosi (2014). To this end, we estimated the semantic breadth level of words (variances) and their frequency ranks from two independent corpora.

In addition to investigating and revealing the Zipfian regularity, we also apply it to improve performance in entailment tests. Since hypernym words have more generic and broader meanings than hyponym words, we expect them to populate higher frequency ranks subject to a Zipfian regularity. Manin (2008) also implemented a study related to frequency and hyponym clusters based on similar insight. Frequencies of target words are estimated from their hyponym frequencies in Russian. Likewise, we incorporate absolute and relative frequency information of entailment test pairs into a baseline method in order to improve entailment test performance.

The paper is organized as follows. In Section 2, related work in the literature will be presented and discussed in detail. The research objectives will be emphasized explicitly in Section 3. The methodology and experimental results will be given in Section 4 and Section 5, respectively. In Section 6, we present the overall discussion of our results and of possible future implications. Finally, the paper will conclude in Section 7.

2. Related work

2.1. Related work on Gaussian embeddings

After the appearance of continuous dense word embeddings such as *word2vec* and *GloVe* (Bojanowski et al., 2017; Mikolov, Chen et al., 2013; Mikolov, Sutskever et al., 2013; Pennington et al., 2014), a new and prominent direction has rapidly emerged within the NLP field. Recently, word representations are merged into language models with transformers such as *ELMo* (Peters et al., 2018), *GPT* (Radford et al., 2019), *BERT* (Devlin et al., 2019) and *ELECTRA* (Clark et al., 2020). Word embeddings have been proven to be very useful and have become common building blocks of neural architectures for addressing high-level NLP tasks. Besides their central role in NLP algorithms for high-level downstream applications, word embeddings have also opened new research directions for better and generalized word representations utilized in computational semantics. In this section, we examine the related work especially in the area of hierarchical word representations. We also cover several works that focus on aspects of word embeddings related to computational linguistics as well as their usage in computational semantics.

As mentioned in the introduction section, there are "non-point" word representations trying to encode word semantics better. This type of embeddings originates from the idea that words are not to be restricted to single points within the embedding space. Erk (2009b) is one of the pioneering studies that mentions this objective. This work has been followed by Erk (2009a), where a model has been proposed with the main objective of attenuating the ambiguities emerging from polysemy to capture hypernym–hyponym relations. Several years later, Vilnis and McCallum (2015) have introduced the multivariate Gaussian distributions into the word embedding research. In this model, the mean values of multivariate Gaussians constitute vectors of words in embedding space, and variance values define a quantitative "uncertainty"/"specificity" level of words around this mean. Correspondingly, one can recognize the variance as a gauge of the semantic breadth of a word. With this model, some hierarchies and entailment relations between words can be captured. Entailment data in Baroni et al. (2012), Baroni and Lenci (2011) contain positive and negative entailment samples. As a positive example, $tree \models plant$ can be given because all trees are plants. Similarly, instances like $pencil \not\models plant$ are negative entailment samples. Additionally, in Vilnis and McCallum (2015), the authors have mentioned relations between word frequencies and variances by sorting variances of 100 nearest neighbors of a set of sample words. They have noted that more specific words have smaller variances and polysemous words have larger variances. However, the authors assert that this pattern is not regular and does not stem solely from word frequencies.

To improve the pioneering Gaussian embedding model, Chen et al. (2015) and Athiwaratkun and Wilson (2017) have independently proposed the Gaussian mixture model (GMM) based embeddings, where words are represented as multimodal Gaussian distributions. The main motivation of "non-point" word embeddings is to better model polysemous words in a semantic vector space by representing them as mixtures of more than one Gaussian distributions, where each mode stands for a sense. Therefore, redundant enlargement of variances of polysemous words and coupling of uncertainty sources coming from polysemy and semantic breadth can be avoided. Mean vectors of multimodal Gaussian embeddings of a particular word are separated such that each vector is positioned in different parts of the semantic space and represents a different sense. Polysemy is a major cause for variance growth in the unimodal Gaussian embeddings. Therefore, word variances represent semantic breadth better when the effects of polysemy are reduced by the multimodal Gaussian embeddings. A drawback of the model in (Athiwaratkun & Wilson, 2017) is that it represents "all" words with multimodal Gaussians with two modes despite most words are not polysemous and several words have more than two senses requiring more than two Gaussian modes to be matched. The model proposed in (Chen et al., 2015) decides which words are to be represented as multimodal Gaussians in the training phase and eliminates redundant multimodal representations of non-polysemous words. GMMs are also utilized in other NLP related tasks such as speech recognition, (Lu et al., 2011).

Besides semantic breadth, Gaussian embeddings have been utilized in computational problems related to ambiguities in the language (Roy et al., 2019). Furthermore, Gaussian embeddings have been used to represent actors, movies, and genres, (Kim, Katerenchuk et al., 2019). In this work, variances are deployed as a measure of versatility of actors to be used in cast prediction applications.

There are previous "non-point" word embeddings other than the Gaussian models. In (Nickel & Kiela, 2017), Poincaré embeddings have been introduced to capture hierarchical relations in an hyperbolic space. Similarly, in (Tifrea et al., 2019), GloVe algorithm (Pennington et al., 2014) has been adapted to learn hyperbolic word representations. Muzellec and Cuturi (2018) have utilized the Wasserstein space of elliptical distributions to represent words. Another study has used the Lorentz model to learn hierarchical relations in hyperbolic geometry, (Nickel & Kiela, 2018). In (Bražinskas et al., 2018), a similar core idea of the Gaussian embeddings has been implemented with the Bayesian skip-gram model.

The common objective of the above previous work is to propose new methods to expand point embeddings to a degree of "regional" embeddings or distributional representations, i.e. "non-point" embeddings, with the principal concern of better representing lexical hierarchies quantitatively.

Other than the aforementioned "non-point" word representations, in a different body of literature, there are studies that focus on diachronic semantic change and semantic breadth of words. In (Mitra et al., 2014), semantic changes across different timelines have been examined. At different points in time corresponding to different decades, sense clusters are created, and the changes of the senses of words are detected and tracked in an unsupervised way with the help of graph models. Different types of sense changes are named as *split*, *join*, *birth* or *death*. After the learning process, detected semantic changes are classified into one of these groups. In (Tang et al., 2016), from entropies of the distribution of possible senses, a proposed semantic feature of a word, called *word status*, is calculated. Possible senses are determined from words that possess the strongest associations with the target word. Association strength is calculated via performing the likelihood ratio test by using co-occurrence statistics. On the other hand, Luo et al. (2019) have proposed the "BREADTH" concept as a kind of definition for semantic breadth. This concept is used to measure how many adjectives with different senses an adverb can pair with. The more the target adverb can be used with many adjectives with different meanings, the semantically broader it is considered. In order to measure "BREADTH" of adverbs, negative cosine similarities between the target adverb and different adjectives are used. As the similarity increases between the adverb and the adjective, the "BREADTH" value of the target adverb would decrease.

2.2. Related work on Zipf's law

Zipf's law is a well-known phenomenon in linguistics (Zipf, 1935, 1949). It is demonstrated that Zipfian regularities can also be observed in several different disciplines. Okuyama et al. (1999) have studied company income distributions and observed Zipfian regularities. Similarly, Soo (2005) has revealed the Zipfian distribution of city populations. Another study has associated internet statistics (e.g., webpage requests) with Zipf's law, (Adamic & Huberman, 2002). On the linguistic side, Mandelbrot (1953, 1961) have generalized Zipf's stimulating study to explain word frequency-rank relation better. As a result of this generalization, word frequency relations are also called Zipf–Mandelbrot law. Later, generalizations and examinations of Zipf's law in various mathematical and information science fields have continued, (Baayen, 2002; Chen & Leimkuhler, 1987; Shan, 2005). Debowski (2002) has proposed a model by combining empirical and rational approaches to show the dependence of the Zipf's law on the text size. The contribution is to show that the additional parameters introduced by Mandelbrot depend on text length. Gerlach and Altmann (2013) has proposed a stochastic model to estimate vocabulary sizes of corpora from different timelines. They create two groups of words which are "core" words and "non-core" words by word frequencies. While core words do not affect the probability of a new word to be joined to vocabulary, usage of non-core words reduces it. The main result of this study is the generalization of Zipf's and Heaps' laws to double power laws. Zipf's law has been used to track diachronic changes in corpora from different timelines and languages in terms of lexical and syntactic properties, (Koplenig, 2018).

Zipf's law has also been expanded and applied to several contextual problems of NLP. For example, many statistical laws including Zipf's have been tested in complex systems that do not carry typical scenarios for data, (Gerlach & Altmann, 2019). These complex systems include earthquake magnitudes, inter-event times of consecutive occurrences of words in texts, frequency rank relationships of words, relation between the degrees of nodes from a network and the ranks of the degrees of the nodes. Deviations of statistical laws against the availability of large datasets have also been examined extensively, (Gerlach & Altmann, 2019). In (Gao et al., 2019), co-occurrence matrices have been subjected to transformations that are compatible with Zipf's distribution to diminish the effects of unreasonable weights of frequent words such as "the" or "is".

Generalizations to Zipf's law and the Zipfian patterns in different knowledge domains are also studied in the literature. Li (1992) asserts that even randomly composed corpora exhibit Zipfian patterns. As a counter argument, Ferrer-i-Cancho and Elvevåg (2010), Ferrer-i-Cancho and Solé (2002) show by performing rigorous statistical tests that random texts do not exhibit the consistent patterns present in regular texts. They conclude that Zipf's law does not emerge from purely random processes, but carries meaningful information concerning language structure and its evolution. Another study reveals that passwords also show Zipfian patterns, (Wang et al., 2017). In (Zhang, 2009), it is asserted that power law regularities can also be discovered in computer programs via empirical analysis of real software systems.

To quantitatively study Zipfian patterns, Ferrer-i-Cancho (2005a, 2005b, 2005c) have associated semantic vagueness of words with the number of links connecting words with meanings. If a word has more links, it would have a high number of interpretations in the context where it appears. Precision and vagueness terms used in Ferrer-i-Cancho (2005b) can be interpreted in analogy with the semantic specificity and uncertainty. Casas et al. (2019) have conducted experiments on different corpora belonging to different languages to quantitatively study the meaning-frequency relationship and the law of abbreviation. Their consistent results from different corpora indicate that word frequencies are highly correlated with the number of senses of words derived from WordNet (Miller, 1995). There also exist important theoretical contributions regarding the origins of Zipfian patterns, (Ferrer-i-Cancho & Vitevitch, 2018; Manin, 2008). In (Ferrer-i-Cancho & Vitevitch, 2018), Zipf's law of meaning has been studied in detail and strong positive correlation between word frequency and the number of meanings has been verified. Zipf's relatively simplistic assumption is replaced by a more general and compact assumption that can fit into the general theory of communication, (Ferrer-i-Cancho & Vitevitch, 2018). According to this assumption, the joint distribution of a word and a meaning can be sufficient to characterize the Zipfian regularities. On the other hand, Manin (2008) has explained the Zipf's law on purely linguistic grounds. Manin (2008) has examined the degree of generality of words and composed a theoretical background by modeling the

semantic space and assuming a measure that can specify the generality. With the help of this modeling and assumptions, Manin (2008) has shown that Zipf's law can be constructed by arrangements of word meanings in the semantic space.

Review of Piantadosi (2014) on Zipf's law presents critical procedures to explore word frequencies. In that study, it has been asserted that word frequencies are complex language features, which cannot be explained by a single power law formula. It is also claimed that although word frequencies exhibit patterns compatible with Zipf's law, they have a complicated structure that can only be revealed by methods beyond Zipf's law. Nevertheless, when Piantadosi (2014) has considered the specific subsets of the lexicon, such as the Swadesh list or number words in the scope of Zipf's law, it has been noticed that those word groups also fit to near-Zipfian patterns, (Piantadosi, 2014).

3. Research objectives

The main objective of this research is to quantitatively study the relationship between semantic breadth of a word and its frequency rank to reveal the Zipfian regularities: more frequent words tend to be semantically broader. Under the mathematical foundation of the multivariate Gaussians embeddings, we calculate variances of words, which are quantitative indicators of the extent of meanings of words. By quantitatively studying the relations between the word variances of Gaussian embeddings and word frequencies, our aim is to show the existence of Zipfian regularities in semantic breadth of words. We also aim to observe whether some distinctive word groups such as Swadesh words and number words, which obey the regular Zipfian patterns, exhibit Zipfian regularities with respect to variances as well.

As an incidental objective of our research, we aim to leverage the revealed Zipfian regularity to improve the performance of word entailment tests. The existence of the Zipfian regularity suggests that frequencies of words are now direct indicators of generality and specificity of meaning. Thus, one can leverage frequencies in entailment problems since they require distinguishing general and specific words from each other. Textual entailment has been subjected to information processing applications. In (Rooney et al., 2014), textual entailment samples have been classified by ensemble learning. Kim, Rabelo et al. (2019) have combined textual entailment with information retrieval on legal texts on the statutory law. Furthermore, in (Clinchant et al., 2006), it has been asserted that lexical entailment contributes to performance in information retrieval tasks related to queries and documents. Therefore, our last objective involves presenting a lexical entailment improvement that can be used in information retrieval tasks. Our proposed method is based on calculating the given entailment pair's frequency score and combining it with scores used in baseline methods. Detailed explanations of our methodology and experimental setup will be given in Section 4 and Section 5, respectively.

4. Methodology

In this section, we first review the Gaussian embedding methods. Then, we present our proposed methodology to quantitatively study the semantic breadth of words through the variances of Gaussian embeddings and to reveal Zipfian regularities. Finally, our proposed method to leverage the observed Zipfian regularities in practical entailment tests will be explained.

4.1. Gaussian word embeddings

Vilnis and McCallum (2015) propose to learn Gaussian embeddings with energy based max-margin objective. As energy function, either expected likelihood kernel (ELK) (Jebara et al., 2004) or well-known Kullback–Leibler (KL) divergence are chosen. In ELK, the inner product for Gaussian distributions is defined as

$$E(P_i, P_j) = \int \mathcal{N}(x; \mu_i, \Sigma_i)\mathcal{N}(x; \mu_j, \Sigma_j)dx = \mathcal{N}(0; \mu_i - \mu_j, \Sigma_i + \Sigma_j). \tag{1}$$

On the other hand, negative energy function with KL divergence is defined as

$$-E(P_i, P_j) = D_{KL}(\mathcal{N}_j, \mathcal{N}_i) = \int \mathcal{N}(x; \mu_i, \Sigma_i) \log \frac{\mathcal{N}(x; \mu_j, \Sigma_j)}{\mathcal{N}(x; \mu_i, \Sigma_i)} dx. \tag{2}$$

The reason for using negative sign in Eq. (2) is that KL divergence measures distance, not similarity between two distributions. After determining the energy function, the max-margin ranking objective is used as

$$L_m(w, c_p, c_n) = max(0, m - E(w, c_p) + E(w, c_n)) \tag{3}$$

In Eq. (3), w is the target word, and c_p and c_n are positive and negative context words co-occurring with the target word, respectively. The max-margin objective tries to make the difference between the energy of positive context and the energy of negative context be at least the margin m (Vilnis & McCallum, 2015). Word2vec in skip-gram mode, (Mikolov, Sutskever et al., 2013), is used to train Gaussian embeddings. There are two options for variance calculation. One is the diagonal variance case where a different variance value is calculated for every dimension. The overall variance is then calculated by the determinant of the covariance matrix. In the spherical variance option, there is only one variance value per embedding since all diagonal values are same. Geometrically, Gaussian components are represented by an ellipsoid in the semantic space. The center of the ellipsoid is designated by the mean vector and the contour surface of that ellipsoid is specified by variances. In this regard, word variances can be interpreted as indicators for the volume of the semantic uncertainty.

4.2. Multimodal Gaussian word embeddings

To enhance Gaussian embeddings by representing them as a multimodal Gaussian mixture, words w_f and w_g are represented as

$$\begin{aligned} f(x) &= \sum_{i=1}^{K} p_i \mathcal{N}(x; \mu_{f,i}, \Sigma_{f,i}) \\ g(x) &= \sum_{i=1}^{K} q_i \mathcal{N}(x; \mu_{g,i}, \Sigma_{g,i}), \end{aligned} \tag{4}$$

respectively, where $\sum_{i=1}^{K} p_i = 1$ and $\sum_{i=1}^{K} q_i = 1$, (Athiwaratkun & Wilson, 2017). In this case, log-energy of two words is calculated as

$$\log E(f, g) = \log \sum_{j=1}^{K} \sum_{i=1}^{K} p_i q_j e^{\xi_{i,j}}, \tag{5}$$

where $\xi_{i,j} = \log \mathcal{N}(0; \mu_{f,i} - \mu_{g,j}, \Sigma_{f,i} + \Sigma_{g,j})$. To measure the similarity between word vectors, ELK, maximum cosine similarity, and minimum Euclidean distance metrics are used, (Athiwaratkun & Wilson, 2017). When Gaussian word embeddings are inspected, it can be seen that mean vectors of polysemous words fall into different locations in the vector space. Since words are divided into two modes, their respective variances are not as large as the variances of the pioneering single mode version of the Gaussian embeddings.

4.3. Uncovering the Zipfian regularities of word variances

While determining the variances of words to measure semantic breadth levels of words, the overall variance of the modes of the multimodal Gaussians should be considered accordingly. For the bimodal case, the mean and variance moments of the mixture can be calculated as:

$$f(x) = p g_1(x) + (1 - p) g_2(x), \tag{6a}$$

$$\mu = p\mu_1 + (1 - p)\mu_2, \tag{6b}$$

$$v_2 = p[\sigma_1^2 + \delta_1^2] + (1 - p)[\sigma_2^2 + \delta_2^2], \tag{6c}$$

where

$$\begin{aligned} \mu &= \int x f(x) dx, \\ \delta_i &= \mu_i - \mu, \\ v_r &= \int (x - \mu)^2 f(x) dx. \end{aligned} \tag{7}$$

In Eq. (6), $g_i(x)$ stands for the components of mixture distribution. p is the mixture probability of the first component in the bimodal case. μ and v_2 are the first moment (mean) and the second moment (variance) of the mixture, respectively. Lastly, σ_i and μ_i are the variance and mean of ith component of the mixture, respectively. Before comparing the variances of words, the following approximations are made. Since mixture components consist of uncorrelated multivariate Gaussians, they are independent multivariate at the same time. From independence, the mean of a mixture component can be calculated by multiplying the mean of each Gaussian multivariate component. This corresponds to the multiplication of every entry of the mean vector of each word. Note that the mean vectors are normalized. Therefore, the result of the multiplication is very close to zero. This leads to approximating the variance of a word as the weighted sum of variances of each mixture component from Eq. (6c).

Before sorting the words according to their frequencies, a final operation needs to be performed. Note that using the same corpus to obtain variances and frequency ranks of word may yield correlated errors. Even in the case that all words are equally probable, some words would appear more than others by chance. When the variance-rank relationship is plotted based on this standard measurement, a decreasing pattern of words will not reflect their actual distributions. Using two independent corpora offers a solution to this problem. One of the two corpora is to estimate variances of words, and the other one will be used to estimate frequency ranks. This can also be achieved by splitting the single corpus into two corpora by applying a binomial split. To sum up, using independent corpora or making corpus-splitting let us perform appropriate statistical analysis to calculate deviations in the Zipfian patterns, (Piantadosi, 2014).

4.4. Use of Zipfian regularities in entailment improvement

The existence of Zipfian regularities can also be exploited in some practical NLP applications. For demonstration purposes, we present a simple example in this paper. To this end, we propose a method to improve entailment tests by injecting the knowledge of this regularity to simple point word embeddings. It should be noted that our proposed method does not need training of complex "non-point" word embeddings but only leverages the knowledge of Zipfian patterns.

While deciding whether there is an entailment relation between a given word pair, we incorporate frequency information of the word in the pair to a baseline method for entailment score calculation. Given the Zipfian regularity between semantic breadth and frequency rank of words, it is reasonable to say that hypernym words tend to have higher frequencies and higher generality in accordance with a quantitative relation. Similarly, hyponym words are supposed to have lower frequencies and higher specificity.

We use the cosine similarity between the pair of words as a baseline method. We then infuse frequency information to this baseline to calculate the overall entailment score for a given pair:

$$CS_{(m,n)} = \frac{\vec{w}_n^T \vec{w}_m}{\|\vec{w}_m\| \|\vec{w}_n\|}, \tag{8}$$

$$FS_{(m,n)} = \frac{f_n}{N} \times 100 + \frac{\log_{10}(f_n/f_m)}{k}, \tag{9}$$

$$OS_{(m,n)} = CS_{(m,n)} + FS_{(m,n)}, \tag{10}$$

where $CS_{(m,n)}$ stands for the cosine similarity between word vectors $\vec{w}_m$ and $\vec{w}_n$. When $CS_{(m,n)}$ is used by itself, it constitutes a baseline method for the entailment test. $FS_{(m,n)}$ as given in Eq. (9) indicates the frequency score of the word pair (m, n). The hypernym candidate is word m, and the hyponym candidate is word n. f_m and f_n are frequencies of m and n, respectively. N stands for the sum of frequencies of all words in the vocabulary. The constant k can be chosen and fine-tuned according to test type. In general, values between 20 and 40 give favorable results. Here, one can calculate the frequency score by using two approaches. The first is related to the respective frequency of the candidate hypernym word according to total word frequencies of all words in the vocabulary. As we will experimentally show in Section 5 that the probability of being simultaneously both a low-frequency word and a hypernym word, which is semantically broader, is small. Therefore, the frequency score of the given pair with low frequency hypernym candidate will also be small. In the second approach, relative frequencies of hypernym and hyponym words are considered. If the frequency gap between the second word and the first word becomes small, the probability of the entailment relation will decrease. If the first word frequency is greater than that of the second, it contributes to the frequency score negatively. In order to decrease the frequency score, in this case, the logarithm of the ratio between frequencies of the pair is taken. Overall entailment score for word pair (m, n) is calculated by adding the frequency score to the cosine similarity of the given word pair as in Eq. (10).

5. Experiments and Results[1]

In this section, we present the results of our experiments. First, experiments that use multimodal Gaussian embeddings are presented in Section 5.1. Then, we will give extensive experimental results and discussions regarding isolating the effect of polysemy in Section 5.2. To isolate the effect of polysemy, we first present results for unimodal Gaussians to quantitatively show that multimodal Gaussians are better models in eliminating the effects of polysemy. Then, we give results obtained from sense-controlled corpora so that the effect of polysemy is completely removed and the variances represent only the extent of semantic breadth of words. In Section 5.3, we will present results of experiments performed for Swadesh and number words. Analysis of the effects of frequency threshold used in constructing the vocabulary is given in Section 5.4. The results for Turkish corpus is presented in Section 5.5. Finally, we will present results regarding the entailment tests in Section 5.6. To diversify our experiments, we have performed experiments both with or without corpus splitting. In Sections 5.1 and 5.3 , we have performed corpus splitting while Sections 5.2, 5.4 and 5.5 utilize the entire corpus to provide more data for training the Gaussian embeddings.

5.1. Examining Zipfian regularities on multimodal Gaussian embeddings

In this subsection, we use the model of multimodal Gaussian embeddings trained on a concatenated corpus of UKWAC and Wackypedia (Baroni et al., 2009) to obtain variance values of the words, (Athiwaratkun & Wilson, 2017). As a regular statistical process, words whose frequencies are less than the threshold 100 are removed from the vocabulary. Final vocabulary size is 314,129. We have experimentally studied the effects of the threshold and showed that its value is inconsequential in Section 5.4. The dimensionality of mean vectors of all the models used throughout this paper is 50. We use a second corpus to determine the frequency ranks of words. The second corpus is the English Wikipedia with a vocabulary of size 229,922. In order to be compatible with the trained model's vocabulary, words with occurrences of less than 100 in the second corpus are also discarded. For pre-processing, downloaded Wikipedia dump is cleaned from all redundant materials such as document numbers, URLs and HTML syntax. All non-alphanumeric characters are also removed. Lastly, all characters are made lowercase. In total, there are 2,127,511,369 tokens in the Wikipedia corpus. 229,922 of them are types and have frequency more than 100. After pre-processing steps, we rank words according to their frequencies by using the Wikipedia corpus. Then, the acquired rank orders and variance values are combined to obtain the relation between variances and frequency ranks.

In the standard Zipf's law, we expect to have frequency-rank relation that is compatible with the power law relation, as expressed in Eq. (11):

$$f_r \propto b/(r^a) \tag{11}$$

Here, r and f_r stand for word rank and the corresponding frequency value of word whose rank is r, respectively. In standard Zipf's law, parameters a and b are approximately 1. Note that there are deviations from the above formula for infrequent words, (Ferrer-i-Cancho & Solé, 2001). One can obtain more accurate results by using a second exponent for higher ranking words to deal with

[1] Source codes and data are available at https://github.com/koc-lab/w2gm-zipfian.

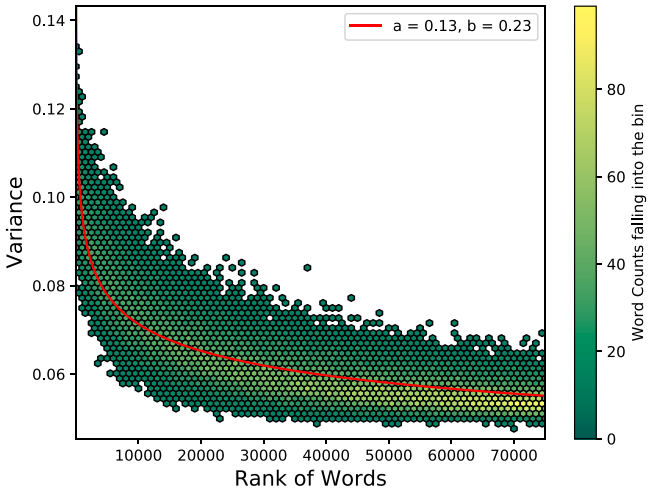

Fig. 1. Two dimensional histogram of word variances and work ranks.

the deviations caused by infrequent word. However, since addressing deviations due to infrequent words is beyond the scope of our study, we use the standard formula of Zipf's law.

Throughout this section, we will present several Zipfian plots of variance versus frequency rank of words in both regular or log–log forms. In the insets of figures, regression parameters (a and b) according to the expression given in Eq. (11) can also be found. Figs. 1 and 2 demonstrate the pattern between variance and word rank. In these figures, mixture weighted average of two variances is calculated as stated in Eq. (6). For practical reasons, around the first seventy five thousand words of the vocabulary, which sufficiently demonstrate the general pattern, are displayed. A two-dimensional histogram is used because there are many ripples between variances of successive words when the gap between frequency ranks is small. As shown in figures, variances of words display a Zipfian pattern where they decrease following the Zipf's law as the order of word rank increases noticeably. Regression parameters to fit the curve are not divergent from classical Zipf's law parameters.

It is important to emphasize that although the regression curves exhibit strict Zipfian regularity, the relationship between variance and frequency rank does not entirely satisfy Zipf's law. There are fast changes in variances of adjacent words. Our study's main aim is to indicate the general pattern of the variance of words at a similar rank order. Although there may be an arbitrary increase/decrease in variance for words of close frequency ranks, it is possible to see a decrease in variance when the word frequency decreases considerably.

The results reported above are intuitively meaningful. The observed Zipfian regularity can be elucidated by using the word "animal" as a qualitative example. "Animal" points out relatively broader concepts compared to the name of a specific animal. In the corpus, it co-occurs with a variety of different animals in specific word windows. It is reasonable to say that the word "animal" has both high frequency and uncertainty (or variance) because the word "animal" cannot specify any species (let alone a family of species) when it is used in a sentence all alone. This leads to an increase in its variance. Although it can gain different side meanings or be used metaphorically in a sentence, note that "animal" is not a polysemous word. Therefore, associating the increased variance only with polysemy would not be a reliable inference. Multimodal Gaussians eliminate the impact of polysemy on variance. On the other hand, a specific animal "antelope" denotes a way more specific entity. Since it is a very specific word, its word frequency is naturally low. Consistently, its uncertainty is low, and it can be seen that its variance value is low through our results. We will present comprehensive experimental results in the following subsections to quantitatively verify our qualitative analysis by also eliminating the effects of polysemy.

In order to make experimental results more reliable and involved, in addition to averaged variances, we examined maximum and minimum variances of the multimodals. Figs. 3, 4 demonstrate the behavior of maximum and minimum variances of words with respect to frequency ranks. Consistently, a similar Zipfian pattern can be observed in both figures. Independent of which mode of the multimodal Gaussians are used, Zipfian pattern remains. These experimental results from multimodal Gaussians further support the Zipfian regularity between uncertainty and frequency ranks.

5.2. Effects of polysemy and experiments on sense-controlled corpora

The main motivation of multimodal Gaussians is to minimize the impact of polysemy on word variances. In this model, Gaussian embeddings with more than one mode give extra degrees of freedom to treat different senses of words as distinct words. Different senses are learned more independently in the semantic space. Therefore, excessive increase on variances can be eliminated such that the variance values represent the uncertainty coming from semantic breadth of words other than contributions coming from different senses due to polysemy. For example, the word "rock" takes place in the same context with the words that are related to

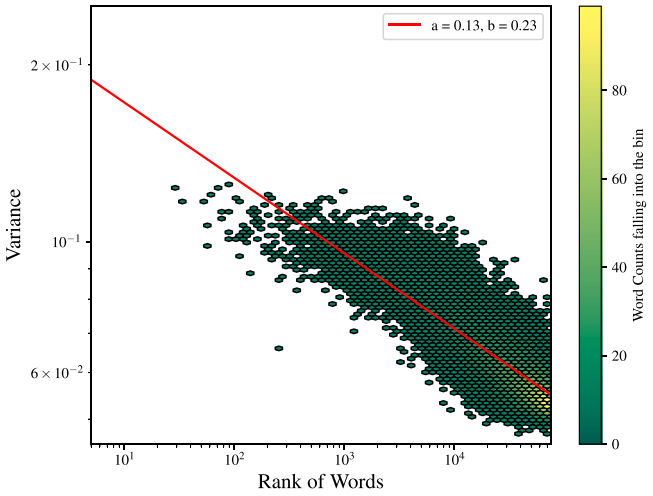

Fig. 2. Two dimensional histogram of word variances and word ranks in logarithmic scale.

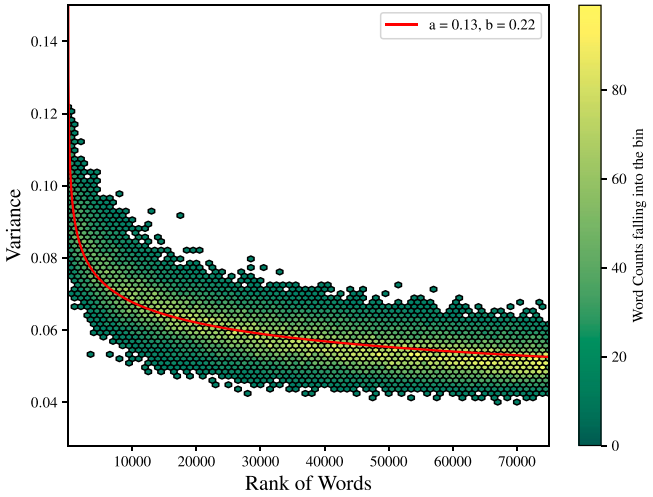

Fig. 3. Minimum variances of multimodal Gaussian embeddings.

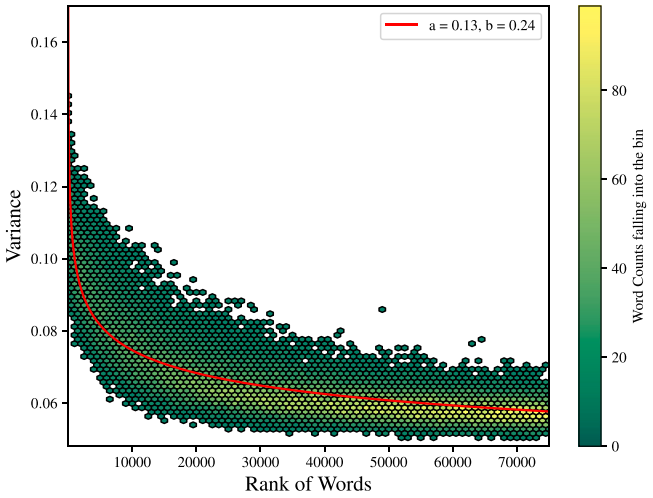

Fig. 4. Maximum variances of multimodal Gaussian embeddings.

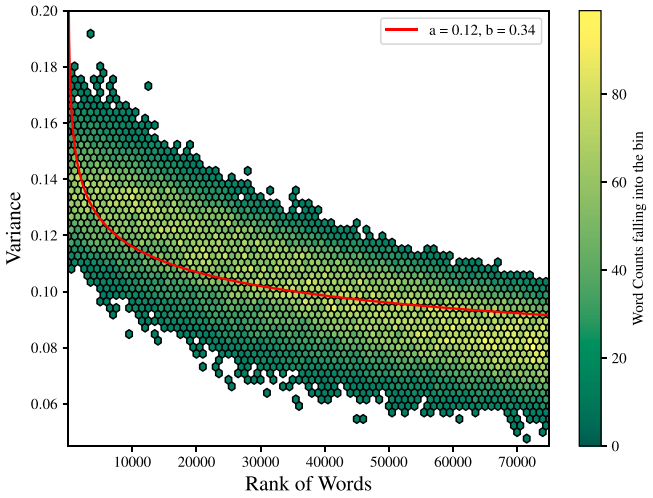

Fig. 5. Variances of unimodal Gaussians embeddings.

Table 1
Logarithmic variance values of the word "bank" for multimodal Gaussian embeddings with different number of modes.

		Variance value
1 Mode	variance("bank",0)	0.147
2 Modes	variance("bank",0)	0.081
	variance("bank",1)	0.085
3 Modes	variance("bank",0)	0.076
	variance("bank",1)	0.077
	variance("bank",2)	0.070

music and also *geography*. If the Gaussian model is unimodal, then variance of "rock" will try to capture different meanings from both *music* and *geography* concepts.

To quantitatively support the qualitative discussions stating that polysemous words tend to have high variances, we first trained a unimodal Gaussian embeddings model on our Wikipedia corpus. The pattern obtained from unimodal Gaussian model is given in Fig. 5. As can be seen in Fig. 5, word variances obtained from unimodal Gaussians are significantly higher than those of bimodals. Additionally, Table 1 also tabulates the variance values of unimodal, bimodal and trimodal Gaussian embeddings of the word "bank" as an additional experimental demonstration. In Table 1, it can be seen that the variance values of the Gaussian embeddings of "bank" decrease as the number of the modes in the model increases. Although variances tend to decrease in unimodal models as frequency decreases, variances do not fit well to Zipfian patterns. The reason is that the Zipfian regularities coming from Zipf's meaning-frequency law and from the semantic breadth couple to each other. This coupling prevents us to conclude that word frequencies are directly related to semantic breadth. Therefore, we can infer that freeing variances from the effects of polysemy plays a significant role to reveal the Zipfian regularities of word variances. As it was mentioned before, Gaussian models with single mode carry excessive amounts of uncertainty, and so variance, due to their attempt to capture too many senses as well as their semantic breadth. That is why utilizing the multimodal Gaussian embeddings more accurately represents the semantic breadth of words.

Now, we move to present our experiments on sense-controlled corpora. In (Casas et al., 2019), the number of senses for each word is imported from WordNet. Following a similar procedure, we labeled the words in our corpus according to their WordNet senses by using the Lesk algorithm provided by NLTK library (Bird et al., 2009; Lesk, 1986). Lesk algorithm receives a sentence as input and returns the WordNet senses of every word in the given sentence. In order to increase the disambiguation success, we also provided the part-of-speech (PoS) information. PoS tags alleviate the complexity of disambiguating senses from WordNet. Nevertheless, we should note that some words are not present in WordNet. We labeled such words with a single dummy label. By treating every sense of a word as distinct words and assigning them to distinct Gaussian embeddings, the effects of polysemy is removed. Then, our sense-controlled corpus is used to train the Gaussian embedding model. In Fig. 6, the Zipfian pattern of the variance values can be observed. We performed another experiment with the manually sense-annotated SemCor corpus (Miller et al., 1994). Since this corpus is manually annotated, it is relatively small with approximately 200,000 tokens. We implemented the same experiment on this corpus. Fig. 7 demonstrates the behavior of variances. Although corpus size is small, manual sense annotation provides us a better fit to Zipfian regularity. It is well-known that although there are several well-performing algorithms, word-sense disambiguation with high performance is still a challenging task, especially in unsupervised way. Therefore, obtaining better fit to Zipfian patterns from the manually sense-annotated SemCor corpus is not surprising.

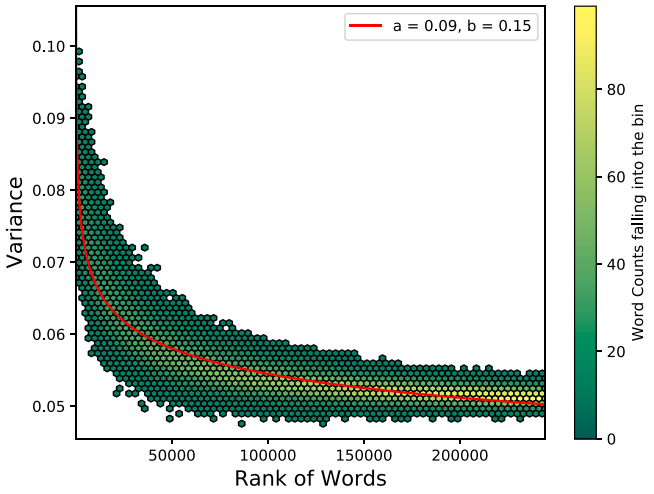

Fig. 6. Variances of Gaussian embeddings trained with the sense-controlled corpus.

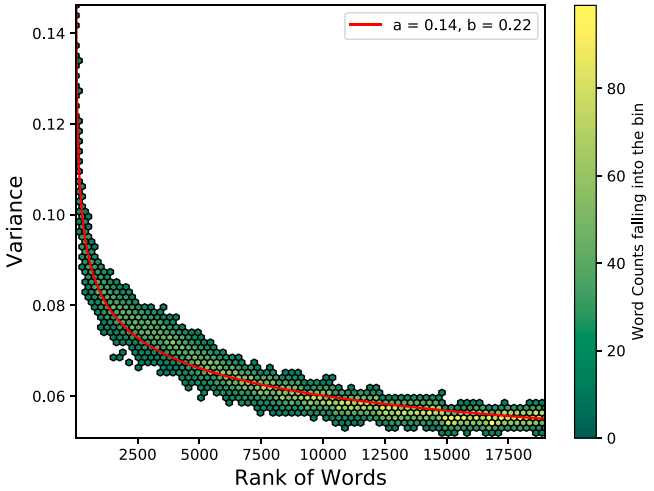

Fig. 7. Variances of Gaussian embeddings trained on the sense-controlled SemCor corpus.

In our experiments on sense-controlled corpora, we showed that the variances of Gaussian embeddings are now only measuring the semantic extents of words without the effects of polysemy. Results of experiments on sense-controlled corpora strongly indicate that there is a Zipfian relation between semantic breadth and frequency.

Lastly, we performed correlation analysis between word variances and word frequencies by following the procedure given in (Casas et al., 2019). We calculated Pearson, Spearman and Kendall correlations between variances and frequency, obtained from both our Wikipedia corpus and its sense-controlled version. Results are tabulated in Table 2, where we also copied the correlation results between the number of senses and frequency from Casas et al. (2019)'s quantitative study for Zipf's meaning-frequency relation. We observe that the variance values are highly correlated with frequencies. Since these independent experiments correlate different variable pairs, our correlation results are not totally comparable with those from Casas et al. (2019). However, in terms of the statistical significance levels that are reached, our results are compatible with those in Casas et al. (2019).

5.3. Zipfian regularities on Swadesh and number words

In (Piantadosi, 2014), many experiments are implemented to investigate whether there are other features of language satisfying Zipf's law, other than just word frequency rank and word frequencies. One of these experiments is related to the Swadesh list (Swadesh, 1950). Swadesh list is a collection of certain words reflecting basic frequent concepts in the language such as "mother", "we", "one", walk". In different languages, Swadesh lists reflect the same basic concepts. In (Piantadosi, 2014), frequency-rank curves of Swadesh lists from different languages are examined. In this experiment, ranks of words pointing to the same concepts across different languages are fixed. Acquired patterns are highly similar to each other, and every language sample shows a Zipfian

Table 2
Correlation scores. Variance vs. frequency for both bimodal and sense-controlled corpora. WordNet number of senses vs. frequency from Casas et al. (2019).

	Pearson		Spearman		Kendall	
	r	p-value	ρ	p-value	τ	p-value
Variance (bimodal)	0.083	$<10^{-323}$	0.828	$<10^{-323}$	0.647	$<10^{-323}$
Variance (sense-controlled)	0.042	$<10^{-94}$	0.861	$<10^{-323}$	0.677	$<10^{-323}$
Number of senses	0.068	$<10^{-323}$	0.422	$<10^{-323}$	0.307	$<10^{-323}$

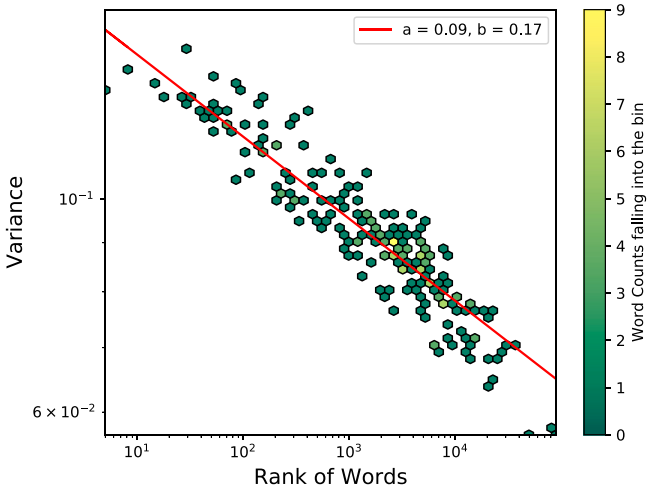

Fig. 8. Two dimensional histogram of average variances for the "Swadesh" list.

pattern. Likewise, we also investigated the behavior of variances of words in the Swadesh list. Results are given in Fig. 8, where the weighted average of variances is used. Indeed, like the original Zipfian patterns of Swadesh list given in (Piantadosi, 2014), variances of Swadesh words display similar linearity in the logarithmic scale.

Another interesting subset that can be studied to strengthen our analysis is the number words, e.g., "one", "two", "ten", "twenty". In this experiment, word ranks of number words are not used. Instead, words are ordered by their cardinality. One critical remark is that decades (e.g., "ten", "twenty") are not included in this experiment since their frequencies are exceedingly higher than the common usage of any arbitrary number. Although word ranks based on word frequencies are not used, a near-Zipfian distribution can still be observed in results reported by Piantadosi (2014). We extend this experiment to variance analysis with our proposed methodology. As can be seen in Fig. 9, the linear pattern of the Zipfian regularity is present in the log–log plot. One drawback of this experiment is that there are not so many numbers consisting of a single word. However, available words are still sufficient to observe the pattern.

5.4. Experiments with different frequency thresholds

Threshold for the minimum word frequency count in determining the vocabulary is an important hyperparameter. Especially in our experiments, frequencies of words play the central role in Zipfian regularities. In the original model of the multimodal Gaussian embeddings that we deployed, the frequency threshold was chosen as 100. We used this value in our experiments. However, to observe the behavior of word variances under different thresholds, we also performed experiments with two additional threshold parameters, 50 and 200. Fig. 10, 11 demonstrate that the Zipfian patterns of word variances. Although the relative frequencies of words change due to threshold selection, Zipfian patterns are preserved for both models. In particular, after thresholding with 50 and 200, the difference between vocabulary sizes of two corpora is approximately 200,000. Nevertheless, Zipfian regularities are clearly observed in both models. These results further validate the consistency of the relation between word frequencies and word variances.

5.5. Results on the Turkish Corpus

To enrich our demonstration of Zipfian regularities with another language not belonging to the Indo-European language family, we study Turkish, which is a member of the Altaic language family. We compiled Turkish corpus from the Turkish Wikipedia. The same experiments were repeated with the same pre-processing procedures. All non-alphanumeric characters and stop words were removed. Since Turkish is an agglutinative language, several words can be derived from the same stem. Therefore, we also applied

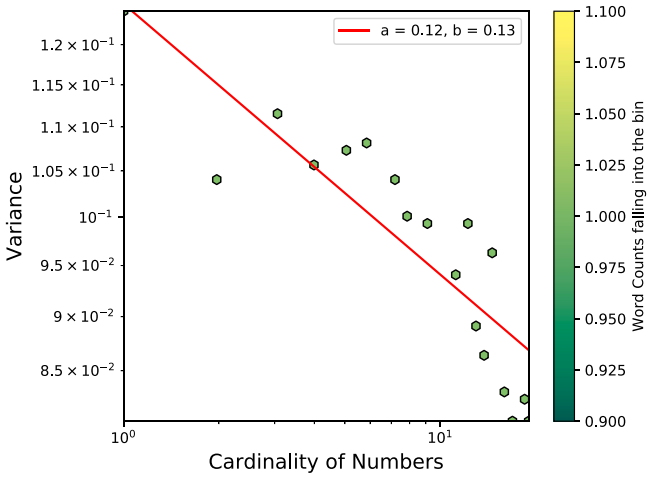

Fig. 9. Two dimensional histogram of average variances of "number words".

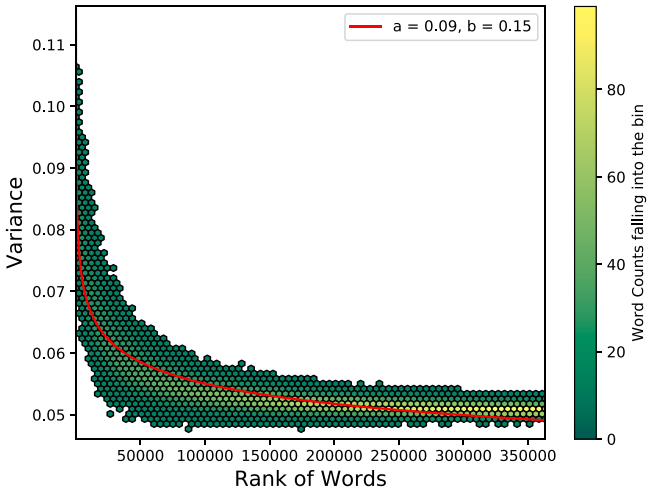

Fig. 10. Average variances of Gaussian embeddings (frequency threshold = 50).

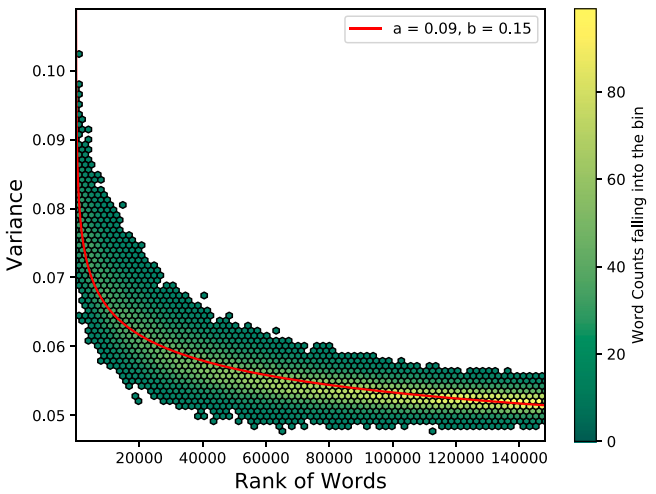

Fig. 11. Average variances of Gaussian embeddings (frequency threshold = 200).

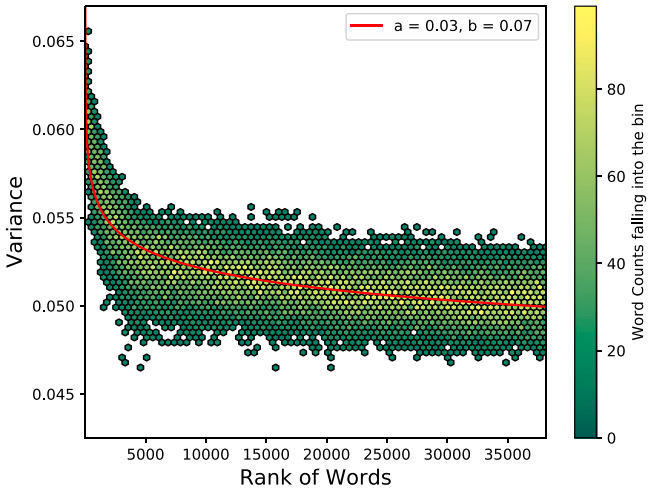

Fig. 12. Average variances of Gaussian embeddings for Turkish corpus.

Table 3
Variance vs. frequency correlations for Turkish Corpus.

	Pearson		Spearman		Kendall	
	r	p-value	ρ	p-value	τ	p-value
Variance vs Frequency	0.518	$<10^{-323}$	0.535	$<10^{-323}$	0.383	$<10^{-323}$

Table 4
Precision, Recall and F1 scores of the entailment test on BBDS dataset.

BBDS	Baseline	w2gm (KL)	BSG (Cos)	Proposed
Precision	0.671	0.742	0.662	0.696
Recall	0.887	0.815	0.891	0.891
F1	0.764	0.776	0.759	0.781

Table 5
Precision, Recall and F1 scores of the entailment test on BLESS dataset.

BLESS	Baseline	w2gm (Cos)	BSG (Cos)	Proposed
Precision	0.131	0.545	0.124	0.135
Recall	0.618	0.134	0.647	0.546
F1	0.216	0.204	0.208	0.217

stemming to Turkish corpus.[2] In total, the Turkish corpus includes 271,317,649 tokens. Since this corpus is relatively smaller than the English Wikipedia corpus, the frequency threshold was determined as 20. The size of the vocabulary with this threshold is around 38,000. The results are presented in Fig. 12 where Zipfian pattern is clearly seen. This experiment provides further evidence that the Zipfian patterns can be observed across different languages and also in relatively smaller corpora (see also Table 3).

5.6. Entailment

For entailment tests, two datasets are utilized: BBDS (Baroni et al., 2012) and BLESS (Baroni & Lenci, 2011) as given in (Bražinskas et al., 2018). Word2vec embeddings are used to calculate the cosine similarity between words, (Mikolov, Sutskever et al., 2013). After calculating overall scores based on the baseline and our proposed method, the threshold that results in the best F1 score is obtained. Parameter k in Eq. (9) is set to 20 and 40 for BBDS and BLESS datasets, respectively. Results from the baseline cosine similarity, the multimodal Gaussian (w2gm) (Athiwaratkun & Wilson, 2017), the Bayesian Skip-gram distribution (BSG) (Bražinskas et al., 2018), and the proposed Zipfian-frequency-injected cosine similarity methods are all tabulated in Tables 4 and 5. For w2gm and BSG, the results of the highest-performing implementation (with KL-divergence or cosine similarity) are presented.

As can be observed from Tables 4 and 5, our proposed method offers improvements over both datasets. The entailment pairs of BLESS are more detailed compared to BBDS. Therefore, extracting the correct entailment relations from BLESS is a harder task.

[2] www.cmpe.boun.edu.tr/tr/content/boun-nlp-morphological-analysis-system-turkish.

Additionally, the number of positive pairs and the number of negative pairs on BBDS are equal. In contrast, the ratio of the numbers of positive and negative pairs is about 0.10 in BLESS. This is why it is more difficult to discriminate positive and negative samples from each other in BLESS compared to BBDS. By this experiment, we showed that capturing entailment relations can be improved by using frequency. Utilizing frequency may not be successful in all scores metrics and sometimes yields mixed results. However, the presented method displays a better trade-off between precision and recall to get higher F1 scores. Due to the difficulty of the BLESS dataset, this improvement can be small compared to other BBDS dataset. Capturing entailment relations is still a challenging and unsolved problem. The main purpose of this experiment is to show the potential of using frequency regularities to reinforce the methods in the literature, not to present state-of-the-art results.

6. Discussion of results and implications

As a result of a trade off between the sender and the receiver in language, which is named "the principle of least effort", words with both broad and narrow meanings emerged, (Zipf, 1949). In (Ferrer-i-Cancho, 2005a, 2005c; Ferrer-i-Cancho & Solé, 2003), the compromise between sender and receiver is studied quantitatively. It is shown that Zipf's law should not be considered as a null hypothesis of a language feature. There is a plausible motivation to study the existence of a relation between semantic breadth of words and word frequency rank. Our contribution comprises the first exhibition of Zipfian patterns between generality and specificity of words and their frequency ranks when polysemy is controlled. We also employ the Gaussian embedding models to study Zipfian statistics, where the variance of Gaussian embeddings is used as a measure for semantic uncertainty of words.

Needless to say, mathematical and quantitative modeling of semantic breadth by variances of "non-point" word embeddings is not as straightforward as simply counting word frequencies within a corpus. Language and word meaning are very flexible concepts, and quantization of meaning may not reflect all aspects of word sense. Thus, ripples in variance values between adjacent words in rank order lists are expected to arise. However, ripples do not constitute a hindrance to apprehend the general Zipfian regularity occurring between word variances and word frequencies. Furthermore, frequency information helps determine entailment relations between words. Assuming that hypernym words have larger frequencies compared to hyponym words, existence of a Zipfian regularity in the semantic breadth of words can be leveraged to enhance baseline methods used to the detect entailment.

Hierarchical relations between concepts can be compelling to some information retrieval tasks. Particularly, information content based studies adopt hierarchical relations. For example, there are studies working on semantic similarities between Wikipedia concepts by using hyponym and hypernym links of categories (Hussain et al., 2020; Jiang et al., 2017, 2015). Therefore, our findings of a fundamental empirical regularity of words and improvements in the detection of entailment, are natural candidates for being tools that may be utilized in information science.

The main results of our paper are the presented Zipfian regularity between the semantic breadth and frequency ranks and the utilization of Gaussian embeddings in a quantitative study of Zipfian statistics and computational semantics. Gaussian embeddings are tools capable of modeling semantics and semantic uncertainty simultaneously. We believe that these findings can also generate new ideas and research directions in computational linguistics as well as several possible utilizations of the presented Zipfian regularity in NLP, information science and computer science applications.

7. Conclusion & future work

We studied the relation between the uncertainty of words and word frequencies by using "non-point" Gaussian embeddings, which are capable of modeling semantics along with semantic uncertainty of words. Gaussian embeddings can handle polysemy as well by assigning different Gaussian modes to different senses. Our quantitative study and experiments showed that variances (or uncertainty) of word embeddings possess a close connection with word frequency ranks, and there are Zipfian regularities in non-point word representations modeling the semantic breadth of words. We performed experiments where the polysemy is eliminated and showed that the variances of Gaussian embeddings only measure the semantic extents of words irrespective of the effects of polysemy. We also showed that our results are valid for different languages and valid for special vocabularies such as Swadesh list and number words.

Our computational study revealed the existence of new Zipfian patterns: more frequent words tend to be generic while less frequent ones tend to be specific. Our contributions are complementary to Zipf's law of meaning distribution and the related meaning-frequency law. Uncovering such a Zipfian regularity present in the semantic breadth of words is essential in linguistic studies, information retrieval, and common NLP tasks. Our results related to the uncertainty of words and the usage of Gaussian embeddings can also be very helpful for further quantitative studies of various semantic properties of words.

To leverage our results, we also presented a method where an immediate utilization of Zipfian regularities in the semantic breadth of words is leveraged. Zipfian regularities are merged with baseline methods to improve the performance of practical word entailment tests, which are important for several NLP and information retrieval tasks.

As a future work, diachronic changes of variances can be examined by using corpora compiled from text samples belonging to different time periods. Hamilton et al. (2016) examine words in a diachronic framework and study how meaning and embeddings in vector spaces change over time. The evolution of language over time in the probabilistic framework is studied as well, (Bamler & Mandt, 2017; Rudolph & Blei, 2018). "Non-point" word embeddings can open up new research directions for diachronic analysis since they can model the dynamic semantic structures of words that may enlarge or shrink through time.

CRediT authorship contribution statement

Furkan Şahinuç: Data curation, Software, Validation, Formal analysis, Investigation, Visualization, Resources, Writing - original draft. **Aykut Koç:** Conceptualization, Methodology, Formal analysis, Supervision, Resources, Writing original draft, Writing - review & editing.

Acknowledgments

We sincerely thank the anonymous reviewers for their detailed and insightful reviews that have significantly improved our manuscript.

References

Adamic, L. A., & Huberman, B. A. (2002). Zipf's law and the internet. *Glottometrics*, *3*(1), 143–150.

Altmann, E. G., & Gerlach, M. (2016). Statistical laws in linguistics. In M. Degli Esposti, E. G. Altmann, & F. Pachet (Eds.), *Creativity and universality in language* (pp. 7–26). Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-24403-7_2.

Athiwaratkun, B., & Wilson, A. (2017). Multimodal word distributions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1645–1656). Vancouver, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P17-1151.

Baayen, R. H. (2002). *Word frequency distributions*. Springer Science & Business Media, http://dx.doi.org/10.1007/978-94-010-0844-0.

Bagheri, E., Ensan, F., & Al-Obeidat, F. (2018). Neural word and entity embeddings for ad hoc retrieval. *Information Processing & Management*, *54*(4), 657–673. http://dx.doi.org/10.1016/j.ipm.2018.04.007.

Bamler, R., & Mandt, S. (2017). Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 380–389). International Convention Centre, Sydney, Australia: PMLR.

Baroni, M., Bernardi, R., Do, N.-Q., & Shan, C.-c. (2012). Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 23–32). Avignon, France: Association for Computational Linguistics.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–226. http://dx.doi.org/10.1007/s10579-009-9081-4.

Baroni, M., & Lenci, A. (2011). How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEOmetrical Models of Natural Language Semantics* (pp. 1–10). Edinburgh, UK: Association for Computational Linguistics.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc..

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. http://dx.doi.org/10.1162/tacl_a_00051.

Bražinskas, A., Havrylov, S., & Titov, I. (2018). Embedding words as distributions with a Bayesian skip-gram model. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1775–1789). Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Casas, B., Hernández-Fernández, A., Català, N., Ferrer-i-Cancho, R., & Baixeries, J. (2019). Polysemy and brevity versus frequency in language. *Computer Speech & Language*, *58*, 19–50. http://dx.doi.org/10.1016/j.csl.2019.03.007.

Chen, Y.-S., & Leimkuhler, F. F. (1987). Analysis of Zipf's law: An index approach. *Information Processing & Management*, *23*(3), 171–182. http://dx.doi.org/10.1016/0306-4573(87)90002-1.

Chen, X., Qiu, X., Jiang, J., & Huang, X. (2015). Gaussian Mixture embeddings for multiple word prototypes. arXiv preprint arXiv:1511.06246.

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Clinchant, S., Goutte, C., & Gaussier, E. (2006). Lexical entailment for information retrieval. In *European Conference on Information Retrieval* (pp. 217–228). Berlin, Heidelberg: Springer Berlin Heidelberg, http://dx.doi.org/10.1007/11735106_20.

Debowski, L. (2002). Zipf's law against the text size: A half-rational model. *Glottometrics*, *4*, 49–60.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1423.

Erk, K. (2009a). Representing words as regions in vector space. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (pp. 57–65). Boulder, Colorado: Association for Computational Linguistics.

Erk, K. (2009b). Supporting inferences in semantic space: Representing words as regions. In *Proceedings of the 8th International Conference on Computational Semantics* (pp. 104–115). Tilburg, The Netherlands: Association for Computational Linguistics.

Ferrer-i-Cancho, R. (2005a). Decoding least effort and scaling in signal frequency distributions. *Physica A. Statistical Mechanics and its Applications*, *345*(1–2), 275–284. http://dx.doi.org/10.1016/j.physa.2004.06.158.

Ferrer-i-Cancho, R. (2005b). Hidden communication aspects in the exponent of Zipf's law. *Glottometrics*, *11*, 98–119.

Ferrer-i-Cancho, R. (2005c). Zipf's law from a communicative phase transition. *The European Physical Journal B*, *47*(3), 449–457. http://dx.doi.org/10.1140/epjb/e2005-00340-y.

Ferrer-i-Cancho, R., & Elvevåg, B. (2010). Random texts do not exhibit the real Zipf's law-like rank distribution. *PLOS ONE*, *5*(3), 1–10. http://dx.doi.org/10.1371/journal.pone.0009411.

Ferrer-i-Cancho, R., & Solé, R. V. (2001). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, *8*(3), 165–173. http://dx.doi.org/10.1076/jqul.8.3.165.4101.

Ferrer-i-Cancho, R., & Solé, R. V. (2002). Zipf's law and random texts. *Advances in Complex Systems*, *5*(01), 1–6. http://dx.doi.org/10.1142/S0219525902000468.

Ferrer-i-Cancho, R., & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, *100*(3), 788–791. http://dx.doi.org/10.1073/pnas.0335980100.

Ferrer-i-Cancho, R., & Vitevitch, M. S. (2018). The origins of Zipf's meaning-frequency law. *Journal of the Association for Information Science and Technology*, *69*(11), 1369–1379. http://dx.doi.org/10.1002/asi.24057.

Gao, L., Zhou, G., Luo, J., & Huang, Y. (2019). Word embedding with Zipf's context. *IEEE Access*, *7*, 168934–168943. http://dx.doi.org/10.1109/ACCESS.2019.2954691.

Gerlach, M., & Altmann, E. G. (2013). Stochastic model for the vocabulary growth in natural languages. *Physical Review X*, *3*, Article 021006. http://dx.doi.org/10.1103/PhysRevX.3.021006.

Gerlach, M., & Altmann, E. G. (2014). Scaling laws and fluctuations in the statistics of word frequencies. *New Journal of Physics*, *16*(11), Article 113010. http://dx.doi.org/10.1088/1367-2630/16/11/113010.

Gerlach, M., & Altmann, E. G. (2019). Testing statistical laws in complex systems. *Physical Review Letters*, *122*, Article 168301. http://dx.doi.org/10.1103/PhysRevLett.122.168301.

Grzybek, P. (2006). *Contributions to the science of text and language: Word length studies and related issues*. Springer Science & Business Media, http://dx.doi.org/10.1007/978-1-4020-4068-9.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1489–1501). Berlin, Germany: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P16-1141.

Hussain, M. J., Wasti, S. H., Huang, G., Wei, L., Jiang, Y., & Tang, Y. (2020). An approach for measuring semantic similarity between Wikipedia concepts using multiple inheritances. *Information Processing & Management*, *57*(3), Article 102188. http://dx.doi.org/10.1016/j.ipm.2019.102188.

Jebara, T., Kondor, R., & Howard, A. (2004). Probability product kernels. *Journal of Machine Learning Research*, *5*(Jul), 819–844.

Jiang, Y., Bai, W., Zhang, X., & Hu, J. (2017). Wikipedia-based information content and semantic similarity computation. *Information Processing & Management*, *53*(1), 248–265. http://dx.doi.org/10.1016/j.ipm.2016.09.001.

Jiang, Y., Zhang, X., Tang, Y., & Nie, R. (2015). Feature-based approaches to semantic similarity assessment of concepts using Wikipedia. *Information Processing & Management*, *51*(3), 215–234. http://dx.doi.org/10.1016/j.ipm.2015.01.001.

Kim, H., Katerenchuk, D., Billet, D., Huan, J., Park, H., & Li, B. (2019). Understanding actors and evaluating personae with Gaussian embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 33*, (pp. 6570–6577). http://dx.doi.org/10.1609/aaai.v33i01.33016570.

Kim, M.-Y., Rabelo, J., & Goebel, R. (2019). Statute law information retrieval and entailment. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law* (pp. 283–289). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3322640.3326742.

Koplenig, A. (2018). Using the parameters of the Zipf–Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes–a large-scale corpus analysis. *Corpus Linguistics and Linguistic Theory*, *14*(1), 1–34. http://dx.doi.org/10.1515/cllt-2014-0049.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation* (pp. 24–26). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/318723.318728.

Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, *38*(6), 1842–1845. http://dx.doi.org/10.1109/18.165464.

Lopez-Santillan, R., Montes-Y-Gomez, M., Gonzalez-Gurrola, L. C., Ramirez-Alonso, G., & Prieto-Ordaz, O. (2020). Richer document embeddings for author profiling tasks based on a heuristic search. *Information Processing & Management*, *57*(4), Article 102227. http://dx.doi.org/10.1016/j.ipm.2020.102227.

Lu, L., Ghoshal, A., & Renals, S. (2011). Regularized subspace Gaussian mixture models for speech recognition. *IEEE Signal Processing Letters*, *18*(7), 419–422. http://dx.doi.org/10.1109/LSP.2011.2157820.

Luo, Y., Jurafsky, D., & Levin, B. (2019). From insanely jealous to insanely delicious: Computational models for the semantic bleaching of English intensifiers. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change* (pp. 1–13). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W19-4701.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. In W. Jackson (Ed.), *Communication Theory*, 486–502.

Mandelbrot, B. (1961). On the theory of word frequencies and on related Markovian models of discourse. *Structure of Language and Its Mathematical Aspects*, *12*, 190–219.

Manin, D. Y. (2008). Zipf's law and avoidance of excessive synonymy. *Cognitive Science*, *32*(7), 1075–1098. http://dx.doi.org/10.1080/03640210802020003.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems: Vol. 26* (pp. 3111–3119). Curran Associates, Inc..

Miller, G. A. (1995). Wordnet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41. http://dx.doi.org/10.1145/219717.219748.

Miller, G. A., Chodorow, M., Landes, S., Leacock, C., & Thomas, R. G. (1994). Using a semantic concordance for sense identification. In *Human Language Technology: Proceedings of a workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Mitra, S., Mitra, R., Riedl, M., Biemann, C., Mukherjee, A., & Goyal, P. (2014). That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1020–1029). Baltimore, Maryland: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/P14-1096.

Muzellec, B., & Cuturi, M. (2018). Generalizing point embeddings using the Wasserstein space of elliptical distributions. In *Advances in Neural Information Processing Systems* (pp. 10237–10248). Curran Associates, Inc..

Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351. http://dx.doi.org/10.1080/00107510500052444.

Nickel, M., & Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems* (pp. 6338–6347). Curran Associates, Inc..

Nickel, M., & Kiela, D. (2018). Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 3779–3788). Stockholmsmässan, Stockholm Sweden: PMLR.

Okuyama, K., Takayasu, M., & Takayasu, H. (1999). Zipf's law in income distribution of companies. *Physica A. Statistical Mechanics and its Applications*, *269*(1), 125–131. http://dx.doi.org/10.1016/S0378-4371(99)00086-2.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/D14-1162.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (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)* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N18-1202.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112–1130. http://dx.doi.org/10.3758/s13423-014-0585-6.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *Technical report, OpenAI, 2019*.

Rooney, N., Wang, H., & Taylor, P. S. (2014). An investigation into the application of ensemble learning for entailment classification. *Information Processing & Management*, *50*(1), 87–103. http://dx.doi.org/10.1016/j.ipm.2013.08.002.

Roy, D., Ganguly, D., Mitra, M., & Jones, G. J. (2019). Estimating Gaussian mixture models in the local neighbourhood of embedded word vectors for query performance prediction. *Information Processing & Management*, *56*(3), 1026–1045. http://dx.doi.org/10.1016/j.ipm.2018.10.009.

Rudolph, M., & Blei, D. (2018). Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference* (pp. 1003–1011). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, http://dx.doi.org/10.1145/3178876.3185999.

Shan, S. (2005). On the generalized Zipf distribution. Part I. *Information Processing & Management*, *41*(6), 1369–1386. http://dx.doi.org/10.1016/j.ipm.2005.03.003.

Soo, K. T. (2005). Zipf's law for cities: A cross-country investigation. *Regional Science and Urban Economics*, *35*(3), 239–263. http://dx.doi.org/10.1016/j.regsciurbeco.2004.04.004.

Swadesh, M. (1950). Salish internal relationships. *International Journal of American Linguistics*, *16*(4), 157–167. http://dx.doi.org/10.1086/464084.

Tang, X., Qu, W., & Chen, X. (2016). Semantic change computation: A successive approach. *World Wide Web*, *19*(3), 375–415. http://dx.doi.org/10.1007/s11280-014-0316-y.

Tifrea, A., Becigneul, G., & Ganea, O.-E. (2019). Poincaré glove: Hyperbolic word embeddings. In *International Conference on Learning Representations*.

Vilnis, L., & McCallum, A. (2015). Word representations via Gaussian embedding. In *International Conference on Learning Representations*.

Wang, D., Cheng, H., Wang, P., Huang, X., & Jian, G. (2017). Zipf's law in passwords. *IEEE Transactions on Information Forensics and Security*, *12*(11), 2776–2791. http://dx.doi.org/10.1109/TIFS.2017.2721359.

Zhang, H. (2009). Discovering power laws in computer programs. *Information Processing & Management*, *45*(4), 477–483. http://dx.doi.org/10.1016/j.ipm.2009.02.001.

Zipf, G. K. (1935). *The MIT paperback series, The psycho-biology of language: An introduction to dynamic philology*. Houghton Mifflin.

Zipf, G. K. (1945). The meaning-frequency relationship of words. *The Journal of General Psychology*, *33*(2), 251–256. http://dx.doi.org/10.1080/00221309.1945.10544509.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press.