# DISCLOSING ZIPFIAN REGULARITIES IN SEMANTIC BREADTH OF WORDS VIA MULTIMODAL GAUSSIAN EMBEDDINGS

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

ELECTRICAL AND ELECTRONICS ENGINEERING

By

Furkan Şahinuç

November 2021

Disclosing Zipfian Regularities in Semantic Breadth of Words via Multimodal Gaussian Embeddings
By Furkan Şahinuç
November 2021

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Aykut Koç(Advisor)

Haldun Özaktaş(Co-Advisor)

Murat Saraçlar

Tolga Çukur

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

# ABSTRACT

## DISCLOSING ZIPFIAN REGULARITIES IN SEMANTIC BREADTH OF WORDS VIA MULTIMODAL GAUSSIAN EMBEDDINGS

Furkan Şahinuç
M.S. in Electrical and Electronics Engineering
Advisor: Aykut Koç
Co-Advisor: Haldun Özaktaş
November 2021

Being one of the most common empirical regularities, Zipf's law for word frequencies is a power-law relation between word frequencies and frequency ranks of words. In this thesis, the semantic uncertainty (i.e., semantic coverage) of words is quantitatively studied through non-point distribution-based word embeddings and a new Zipfian regularity is revealed. Uncertainty or semantic coverage of a word can increase due to several reasons such as polysemy, having a **broad** meaning (such as the relation between broader *emotion* and narrower *exasperation*) or a combination of both. Although there are studies that touch upon measuring the generality-specificity levels of words, Zipfian patterns of these features are not shown quantitatively with a theoretical background. Main aim of this thesis is to bridge this gap in the Zipfian literature. To this end, variances of Gaussian embeddings are utilized to quantify to what extent a word can be used in different senses or contexts. Using the variance information embedded in the non-point Gaussian embeddings, Zipfian patterns which exist in the semantic breadth of words are quantitatively shown 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 and uncertain. In contrast, less frequent ones tend to be specific. To verify the generalization of our findings, Zipfian patterns are investigated in the scope of the polysemy neutralization, various language properties and several languages from different language families: English, German, Spanish, Russian, and Turkish. 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 fundamental regularities. A method is also proposed to leverage the Zipfian regularity to improve the performance of baseline lexical entailment detection algorithms. To the best of our knowledge, this thesis is the first quantitative study that uses Gaussian embeddings to examine the relationships between word frequencies and semantic breadth.



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

# ÖZET

# ÇOK MODLU GAUSS KELİME TEMSİLLERİ İLE SÖZCÜKLERİN ANLAMSAL GENİŞLİĞİNDEKİ ZIPF'SEL DÜZENLİLİKLERİN ORTAYA ÇIKARIMI

Furkan Şahinuç
Elektrik Elektronik Mühendisliği, Yüksek Lisans
Tez Danışmanı: Aykut Koç
İkinci Tez Danışmanı: Haldun Özaktaş
Kasım 2021

En yaygın ampirik düzenliliklerden biri olan Zipf kelime frekansları yasası, kelime frekansları ve kelimelerin frekans sıraları arasındaki bir kuvvet yasası ilişkisidir. Bu tezde, sözcüklerin anlamsal kararsızlıkları (anlamsal kapsayıcılıkları), noktasal dağılıma dayalı olmayan kelime temsilleri yoluyla nicel olarak incelenmiştir ve yeni bir Zipf'sel düzenlilik ortaya çıkarılmıştır. Bir kelimenin kararsızlığı ya da kapsayıcılığını; çok anlamlılık, kelimenin **geniş** bir anlama sahip olması (Örneğin, daha geniş *duygu* ve daha dar *kızgınlık* arasındaki ilişki gibi) veya her ikisinin birleşimi gibi nedenlerle artabilir. Sözcüklerin genellik-özgüllük düzeylerini ölçmeye değinen çalışmalar olmasına rağmen, bu tür özelliklerin Zipf'sel düzenlilikleri teorik bir arka plan ile nicel olarak gösterilmemektedir. Bu tezin temel amacı Zipf literatüründeki bu boşluğu doldurmaktır. Bu amaçla bir kelimenin farklı anlamlarda veya bağlamlarda ne kadar kullanılabileceğini ölçmek için Gauss temsillerinin varyansları kullanılmıştır. Noktasal olmayan Gauss temsillerinde gömülü olan varyans bilgisini kullanarak, çok anlamlılığın kontrol altına alındığı durumlarda, kelimelerin anlamsal genişliğinin de Zipf'sel düzenlilikler sergilediği nicel olarak gösterilmiştir. Bu sonuç, Zipf anlam dağılımı yasasını ve onunla ilgili olan anlam-frekans yasasını, Zipf'sel düzenliliklerin varlığını göstererek şu şekilde tamamlar: Daha sık kullanılan kelimeler genel anlamlı olma eğilimindeyken, daha az sıklıkta olanlar özel anlamlı olma eğilimindedir. Bulgularımızın genelleştirilmesini doğrulamak için Zipf'sel düzenlilikler, çok anlamlılığın nötrlenmesinde, farklı dil özelliklerinde ve İngilizce, Almanca, İspanyolca, Rusça ve Türkçe gibi farklı dil ailelerine mensup dillerde araştırılmıştır. Bu tür düzenlilikler, kelimelerin anlamsal özellikleri ile kelime

frekansları arasındaki ilişkileri ortaya çıkarmak ve bu ilişkileri anlamak için değerli bilgiler sağlar. Çeşitli uygulamalarda, bu temel düzenlilikler kullanılarak performans iyileştirmeleri elde edilebilir. Temel sözcüksel gereklilik algılama algoritmalarının performansını geliştirmek için Zipf'sel düzenliliklerden yararlanmak için bir yöntem de önerilmiştir. Bildiğimiz kadarıyla, bu tez kelime frekansları ve anlamsal genişlik arasındaki ilişkileri incelemek için Gauss temsillerini kullanan ilk nicel çalışmadır.

*Anahtar sözcükler*: Kelime varyansları, Kelime frekansları, Zipf yasası, Anlam-frekans ilişkisi, Zipf'sel düzenlilikler, Kelime gerekliliği, Anlamsal genişlik.

# Acknowledgement

First of all, I would like to thank my advisor Asst. Prof. Aykut Koç. His guidance and support have been very beneficial throughout my M.S. journey. What I learned from him is not limited to technical aspects of NLP but involves many skills to be a prominent researcher in academic life. I will always use his advice and feedbacks throughout all my life.

I am also indebted to my co-advisor, Prof. Dr. Haldun Özaktaş for his guidance that led me to study NLP. His unmatched perspective on problems has broadened my horizon and significantly affected my way of thinking.

I would also like to thank the rest of my thesis committee: Prof. Dr. Murat Saraçlar and Assoc. Prof. Tolga Çukur for their time and valuable feedbacks.

In addition, I would like to thank my colleagues at ASELSAN Research Center, especially Dr. Veysel Yücesoy who has never refrained from helping me for even the smallest things, Lütfi Kerem Şenel for his wonderful collaboration, Dr. Çağrı Toraman and Eyüp Halit Yılmaz for their participation in fruitful discussions about every side of NLP, Ümitcan Şahin and Safa Onur Şahin for their valuable help.

My sincerest thanks also go to my friends Fethi Arda Turgut, Ekin Yıldırım, Fethi Taha Çevik and Berk Karadede for their invaluable friendship. They had always cheered me on when things got a bit discouraging and celebrated each accomplishment. A special thanks go to Selim Furkan Tekin. I wish all the best for him. I also thank all my friends from my undergraduate time at Bilkent.

Last but not the least, I would like to thank my family. From my first steps until today, I have always found their support behind me. Without them, none of my accomplishments would exist.

# Contents

# List of Figures

# List of Tables

# List of Publications

This thesis includes content from following publications(s):

1. F. Şahinuç and A. Koç, "Zipfian regularities in "non-point" word representations," *Information Processing & Managament*, vol. 58, no. 3, pp. 102493, 2021.

# List of Abbreviations

| | |
|---|---|
| NLP | Natural Language Processing |
| NNLM | Neural Network Language Model |
| CBOW | Continuous Bag-of-Words Model |
| NMF | Non-negative matrix factorization |
| RNN | Recurrent Neural Network |
| w2g | Gaussian Word Embeddings |
| w2gm | Gaussian Mixture Model |
| ELK | Expected Likelihood Kernel |
| KL | Kullback-Leibler Divergence |
| WSD | Word Sense Disambiguation |
| POS | Part-of-Speech |
| BSG | Bayesian Skip-gram Distribution |

# Chapter 1

# Introduction

People are social creatures by their nature. To meet their necessities, people need to communicate with each other. This constitutes a fundamental reason for the emergence of languages. Just like the development of humankind, languages continuously evolve. Evolutions of languages proceed under the influence of dynamic requirements of humans.

Needless to say, communication is a two-sided phenomenon. There is a message sender at one side and a receiver that understands the received message and takes an action accordingly. Language is the changeable channel between these two participants. As with everything in the universe, using this channel comes at a cost. Both the sender and receiver desire to minimize their costs as much as possible. However, there exists a conflict of interest between the sender and receiver since reducing the cost at one side means an increase in cost at the other side.

The sender always wants to tell the intention with the least effort. For this purpose, s/he needs to use as the smallest number of words as possible. Ideal case for the sender is that a single word that covers every possible meaning tells the whole story to the receiver. In other words, one word should contain all meanings.

On the other side of the coin, the receiver wants every received word to stand for a distinct meaning. Deciding which word matches to which meaning (i.e., decoding the intended meaning of a received word) will increase the effort of the receiver. Therefore, there should be an one-to-one and onto function relation between words and meanings in the best case for the receiver.

One can easily see that minimizing the effort or the cost of one side means maximizing the other side's effort. Languages take shape from this trade-off. None of the natural languages consists of only one word that represents every meaning or innumerable words for each distinct nuanced sense. There is a compromise or equilibrium between these two sides. In his seminal work, George Kingsley Zipf explains this phenomenon in the scope of "the principle of the least effort" [2]. According to Zipf, some words have more than one meanings while others indicate single meanings. The important factor related to the number of meanings of words is the word frequency.

In addition to conventional Zipf's law of word frequencies, Zipf also asserts that there is power law relation between the rank of word frequencies and the number of distinct meanings words possess [3]. This phenomenon is named as the Zipf's law of meaning distribution. According to [3], frequent words (i.e., more used words in a language) carry more meanings than infrequent words (i.e., they have more number of senses). Similar version of this logic behind this property of languages can be seen in the Huffman Coding [4].

Although the Zipf's law of meaning distribution constitutes a basis for the emergence of polysemy, words have several other properties such as side meanings and subordinate-superordinate relations. Such features other than polysemy enrich the meaning of words depending on the context where words occur. In other words, meanings of words highly depend on other words that surround them. Firth summarizes this notion by "You shall know a word by the company it keeps." [5]. From that point, it is apparent that words have a certain degree of uncertainty and this uncertainty indicates how many different contexts the word can be found in.

There are theoretical studies that contribute to Zipfian relations, especially to the Law of Meaning Distribution. In [6], origins of the Zipf's Law of Meaning Distribution is investigated in detail. Zipf's hypothesis on the number of meanings and the frequency is transformed to a suitable version to communication theory [6]. On the other hand, [7] mentions the degree of generality of words and introduces the features of a semantic space where a measure that can specify the generality of words. Main purpose of this thesis is to introduce another theoretical background that merges Zipfian statistics and semantic generality of words (i.e., semantic breadth).

In this thesis, 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 or reduced as much as possible. 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. [7] 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. Previously, [7] 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. However, this notion has not been embodied with quantitative evidence. By an approach using **non-point** word representations, for the first time, our study presents quantitative results that complement the theoretical intuition and predictions from [7]. Furthermore, unlike the previous works where the number of senses of a word is studied with respect to its frequency, to the best of our knowledge, we reveal the Zipfian regularities between semantic breadth and frequency ranks of words for the first time. While inspecting Zipfian regularities of semantic breadth of words, multimodal Gaussian word embeddings are utilized [8]. As will be described in Chapter 2 in detail, Gaussian word embeddings encode words as probability distributions. In this word embedding scheme, every word is represented with a mean vector

and a covariance matrix. The mean vector determines the location of the word in n-dimensional semantic vector space while the covariance matrix includes the variance values indicating the semantic uncertainty of words. In this study, since the variance values are measures of the semantic uncertainty, they are used as indicators of semantic breadth as well while examining Zipfian patterns of the semantic breadth of the words.

There are three main contributions presented in this thesis: (i) To the best of our knowledge, Zipfian regularities in semantic breadth of words are quantitatively shown for the first time. (ii) It has been shown that the reflections of Zipf's law on different features in natural languages continue on word variances as well. (iii) Lexical entailment detection performance can be increased by leveraging the disclosed Zipfian regularity. This thesis is based on the published journal article "Zipfian regularities in non-point word representations" [9] which is an output of the studies conducted during M.S. education.

The organization of the thesis is as follows. In Chapter 2, a review of the word embeddings are given along with the structure of the multimodal Gaussian embeddings which are used as the main tools for demonstrating the Zipfian regularity of semantic breadth. Main framework of Zipf's laws are given in Chapter 3. In the same chapter, related work of Zipf's law including studies that make improvements on conventional Zipf's law or adaptations to various domains, are also given. The methodology followed to uncover Zipfian features of the different language properties and constructed experimented setups are presented along with the obtained results in Chapter 4. Finally, the thesis will conclude by presenting the overall discussion of the results and of possible future implications in Chapter 5.

# Chapter 2

# Word Embeddings

## 2.1 Conventional Word Embeddings

Necessity of mathematical models for words in Natural Language Processing (NLP) and Computational Linguistcs leads to the introduction of various representation methods. [10] is one of the milestone studies for the introduction of the modern word embeddings in language modeling, where word embeddings are referred as distributed representations of words. In that study word embeddings are referred as distributed representation of words. Here, word embeddings are not trained in a standalone training scheme. A neural network language model (NNLM) is proposed and word embeddings are trained jointly along with the language model parameters. On the other hand, [11] becomes one of the leading studies that shows the worth of word embeddings in various NLP tasks.

Later, by the introduction of word2vec, word embeddings become the main tools employed in NLP tasks mostly as input to the underlying neural network architectures [12, 13]. In these studies, two novel model architectures are proposed to compute word representations. These models are named as CBOW (Continuous Bag-of-Words Model) and Skip-gram. Unlike previous NNLM models, hidden layers before the output layer in feed-forward network are not used.

This enables the model to be more efficient in terms of training complexity. At each iteration, these two models manage the words within a window [12]. In CBOW model, words taking place before and after the target word are utilized to predict the target word. The order of the surrounding words in the window are not important, they are projected into same position by averaging their vectors [12]. The Skip-gram model has similar methodology with CBOW. However, the Skip-gram tries to predict words before and after the current word. Objective to be maximized in Skip-gram model is given as follows [13]:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-c\leq j\leq c\\ j\neq 0}}\log p(w_{t+j}|w_t), \tag{2.1}$$

where, $c$ is the size of the context window, $w_t$ is the target word, $T$ is the number of words in the training corpora. Skip-gram model defines the probability $p(w_{t+j}|w_t)$ using the softmax function as follows [13]:

$$p(w_O|w_I) = \frac{\exp\left({v'_{w_O}}^{\top} v_{w_I}\right)}{\sum_{w=1}^{W}\exp\left({v'_w}^{\top} v_{w_I}\right)}, \tag{2.2}$$

where, $v_w$ and $v'_w$ are the input and output vector representations of word $w$ and $W$ is the size of the vocabulary. By learning these embeddings, word similarity relations are encoded easily and some analogy relations in language are captured. Questions involving country-capital, man-woman and some grammatical features can be solved by simple linear operations. For example, the vector $X = v(\text{“}Turkey\text{”}) - v(\text{“}Ankara\text{”}) + v(\text{“}Athens\text{”})$ is calculated. Then, the closest vector to $X$ based on cosine distance gives the vector $v(\text{“}Greece\text{”})$ as expected. Following the same procedure, similar semantic and syntactic relations can be displayed such as `king-man` / `queen-woman` or `large-largest` / `high-highest`.

GloVe (Global Vectors) embeddings are proposed not long after the introduction of word2vec model [14]. Different from word2vec, GloVe model extracts co-occurrence information from the corpus and trains word embeddings on this information. Main motivation of this model is that relevant words take place within

similar contexts and consequently obtain higher co-occurrence counts. For example, the probability that the word `solid` occurs in the context of `ice` is greater than the probability that `solid` occurs in the context of `steam`. Ratios of the co-occurrences of the words from the real corpus confirm this hypothesis in [14]. The cost function used in training of the GloVe model is given as follows:

$$J = \sum_{i,j=1}^{V} f(X_{i,j})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2, \tag{2.3}$$

where, $V$ stands for size of the vocabulary, $i$ and $j$ are the indexes of the words, $w_i$ indicates the vector of word $i$, $b_i$ and $b_j$ are bias terms. Finally, $X_{i,j}$ is the co-occurrence value of the words $i$ and $j$. In addition, the squared cost is weighted. Weighting function is chosen such that rare and frequent co-occurrences are not overweighted. In addition, this function needs to approach 0 fast enough as co-occurrence values approach to 0 such that $\lim_{x\to 0} f(x)\log^2(x)$ is finite [14]. The proposed function is given as follows:

$$f(x) = \begin{cases} (x/x_{max})^a, & x < x_{max} \\ 1, & \text{otherwise} \end{cases} \tag{2.4}$$

where, $x_{max}$ and $a$ are the hyperparameters to be optimized. In original implementation, $x_{max}$ and $a$ are taken as 100 and 0.75 respectively. Like word2vec model, GloVe also achieves successful results in analogy and similarity tests. Despite having different training procedures, word2vec and GloVe models constitute milestones in the conventional word embeddings literature.

Success and practicability of word embeddings lead to the emergence of a vast literature built upon word embeddings. For example, interpretability problem is considered as one of the main research area regarding word embeddings. In general, word embeddings are used as a black-box and their dimensions are not interpretable. [15] investigates the encoded information behind word embedding dimensions. In addition, there are studies generating interpretable word embeddings by using non-negative matrix factorization (NMF) [16], by using sparse encoding [17] and by imparting lexical resource information to word embeddings

[18].

Word embeddings are also used as tools in explaining the relations between different languages [19, 20], learning features of low-resource languages [21] or even non-engineering domains such as legal studies [22, 23, 24].

## 2.2 Transformer-Based Language Models

Although word embeddings usher a new era in NLP by enconding word information into dense vectors, they still have a number of deficiencies. These type of models are very advantageous in word similarity and word analogy tasks. Nevertheless, they include word information via standalone way. Possible different contexts or senses are encoded in a single vector. However, a major defect is that model performance decreases in more complex or high level tasks such as sentence classification or question answering. Therefore, transformer based models that consider context information are started to emerge. Transformers, in the simplest terms, are language models that can encode and decode words along with their contextual features [25, 26, 27, 28, 29].

To increase performance of neural models in high level tasks, attention mechanism was proposed in [30, 31]. Starting point of the attention technique is to obtain better machine translation results. In classical approach, words of source language is fed to a Recurrent Neural Network (RNN) model generating a context vector which is a representation of the sentence from the source language. Policy that is followed up to generate context vector is called encoding. Then, the context vector is decoded and corresponding words in target language are generated. Here, the main bottleneck is the effort to cram all information of source sentence in the context vector. Regardless of the length of the sentence in the source language, information is always encoded in a single context vector. In longer sentences, it is hard to keep all information in a single context vector. In proposed attention mechanism, encoder hidden states coming from RNN model are also fed to the decoder. Therefore, further context information contained

by source sentence is transferred to the decoding phase. Inspired by previous fundamental works, transformer model is introduced in [25]. Transformers boost the speed and performance of the models that use attention. Such models lead to the emergence of the transformed based language models such as BERT [28], GPT and their successors [26, 27, 29]. Transformer based language models are typically subject to pre-training phase followed by fine-tuning. Pre-training is unsupervisedly carried out on a large corpus while fine-tuning is performed by a supervised learning according to specific tasks such as sequence classification, question answering, language generation and named entity recognition. Transformer based language models generally produce state-of-the-art results in the aforementioned high level NLP tasks.

## 2.3 Gaussian Word Embeddings

Although transformer based language models achieve state-of-the art results in many high level NLP tasks, they are not direct representations of words. Therefore, they have not proper solutions for deficiencies of conventional word embeddings. One of the main drawback is that conventional models treat all words as points in semantic vector spaces. For that reason, such models are occasionally called point embeddings. Regardless of semantic extent of words, they are represented as single points in the semantic space. There is no difference between representation of a conjunction which has no meaning alone and a word including many side meanings or covering a broad semantic extent such as `animal`. This leads to a failure in reflecting actual semantic coverage of words. Another example is that 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.

There are studies that aim to encode word semantics better. This type of embeddings originate from the idea that words are not to be restricted to single points within the semantic space. [32] is one of the pioneering studies with this objective. This work has been followed by [33], 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 [34] come up with an idea that differently deals with the disadvantages of conventional word embeddings by representing words with probability density functions. The chosen distribution for word embeddings is Gaussian distribution. They encode word semantics as Gaussian distributions and propose Gaussian word embeddings (w2g), In this model, words are represented as multivariate Gaussians with mean vectors and covariance matrices [34]. Variance values of w2g embeddings are able to represent semantic uncertainty levels of words, unlike traditional word embeddings. In other words, variances indicate to what extent words possess different senses and uncertain/broad meanings. In this context, uncertainty term comprises to both generality-specificity levels of words and polysemy/homonymy relations. Generally speaking, words with high uncertainty or generic words 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. That is why being a generic word increases variance value and uncertainty. Besides being a generic word, polysemy and homonymy also increase uncertainty. For instance, the word `mean` has several senses such as average, intention, or being rude. These diverse senses increase the uncertainty of such words, and this uncertainty is encoded to w2g embeddings through variances.

To improve the pioneering Gaussian embedding model, [35] and [8] have independently proposed the Gaussian mixture model (w2gm), where words are

represented as multimodal Gaussian distributions. The main motivation of these 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 [8] is that it represents **all** words with multimodal Gaussians with fixed number of modes despite most words are not polysemous and several words have differing number of senses requiring different number of Gaussian modes to be matched. The model proposed in [35] decides which words are to be represented as multimodal Gaussians in the training phase and eliminates redundant multimodal representations of non-polysemous words.

While learning word embeddings, similarity measures between word vectors are generally used. In learning conventional word embeddings, dot product is utilized. Since Gaussian word representations are not simply vectors in n-dimensional space but probability distributions, different similarity measures need to be used. [34] proposes to learn Gaussian embeddings with energy based max-margin objective. As energy function, either expected likelihood kernel (ELK) [36] or well-known Kullback-Leibler (KL) divergence [37] are chosen. In ELK, the inner product for Gaussian distributions is defined as:

$$E(P_i, P_j) = \int \mathcal{N}(x; \mu_i, \Sigma_i)\mathcal{N}(x; \mu_j, \Sigma_j)dx = \mathcal{N}(0; \mu_i - \mu_j, \Sigma_i + \Sigma_j), \qquad (2.5)$$

where $P_i$ and $P_j$ indicate the distributions of the words $i$ and $j$, respectively, $\mu$ stands for mean vector and $\Sigma$ is the covariance matrix. Authors state that they do not directly use energy function in training but with its logarithm due

to positivity [34]. The logarithm of the energy function is given as:

$$\begin{aligned}\log \mathcal{N}(0; \mu_i - \mu_j, \Sigma_i + \Sigma_j) = -\frac{1}{2}\log\det(\Sigma_i + \Sigma_j) \\ -\frac{1}{2}(\mu_i - \mu_j)^\top(\Sigma_i + \Sigma_j)^{-1}(\mu_i - \mu_j) - \frac{d}{2}\log 2\pi.\end{aligned} \tag{2.6}$$

Gradients of the energy function with respect to $\mu$ and $\Sigma$ can be calculated in closed form:

$$\begin{aligned}\frac{\partial \log E(P_i, P_j)}{\partial \mu_i} = -\frac{\partial \log E(P_i, P_j)}{\partial \mu_j} = -\Delta_{i,j} \\ \frac{\partial \log E(P_i, P_j)}{\partial \Sigma_i} = \frac{\partial \log E(P_i, P_j)}{\partial \Sigma_j} = \frac{1}{2}(\Delta_{i,j}\Delta_{i,j}^\top - (\Sigma_i + \Sigma_j)^{-1}),\end{aligned} \tag{2.7}$$

where $\Delta_{i,j} = (\Sigma_i + \Sigma_j)^{-1}(\mu_i - \mu_j)$. Furthermore, negative energy function with KL divergence is defined as

$$-E(P_i, P_j) = D_{KL}(\mathcal{N}_j||\mathcal{N}_i) = \int \mathcal{N}(x; \mu_i, \Sigma_i)\log\frac{\mathcal{N}(x; \mu_j, \Sigma_j)}{\mathcal{N}(x; \mu_i, \Sigma_i)}dx. \tag{2.8}$$

The reason for using negative sign in Eq. 2.8 is that KL divergence measures distances, but not similarity between two distributions. Again, gradients can be computed in a closed form:

$$\begin{aligned}\frac{\partial \log E(P_i, P_j)}{\partial \mu_i} = -\frac{\partial \log E(P_i, P_j)}{\partial \mu_j} = -\Delta'_{i,j} \\ \frac{\partial \log E(P_i, P_j)}{\partial \Sigma_i} = \frac{1}{2}(\Sigma_i^{-1}\Sigma_j\Sigma_i^{-1} + \Delta'_{i,j}\Delta'^\top_{i,j} - \Sigma_i^{-1}) \\ \frac{\partial \log E(P_i, P_j)}{\partial \Sigma_j} = \frac{1}{2}(\Sigma_j^{-1} - \Sigma_i^{-1})\end{aligned} \tag{2.9}$$

where $\Delta'_{i,j} = \Sigma_i^{-1}(\mu_i - \mu_j)$. After determining the energy function, the max-margin ranking objective is used as

$$L_m(w, c_p, c_n) = max(0, m - E(w, c_p) + E(w, c_n)) \tag{2.10}$$

In Eq. 2.10, $w$ 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 $m$ [34]. Skip-gram

algorithm in word2vec, [13], is used to train Gaussian embeddings. There are two options to calculate variance. 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.

In addition to training objectives, regularizations are implemented for means and covariances. To prevent means to grow too large, they put hard constraint to the $l_2$ norm:

$$\|\mu_i\|_2 \leq C, \forall i. \tag{2.11}$$

On the other hand, covariance matrices should be positive definite and reasonably sized. To this end, another hard constraint is added such that eigenvalues $\lambda_i$ lies in the hypercube $[m, M]^d$ where $m$ and $M$ are constants:

$$mI \prec \Sigma_i \prec MI, \forall i. \tag{2.12}$$

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

$$\begin{aligned} f(x) &= \textstyle\sum_{i=1}^{K} p_i \mathcal{N}(x; \mu_{f,i}, \Sigma_{f,i}) \\ g(x) &= \textstyle\sum_{i=1}^{K} q_i \mathcal{N}(x; \mu_{g,i}, \Sigma_{g,i}), \end{aligned} \tag{2.13}$$

respectively, where $p_i$ and $q_i$ indicate the probabilities of mode $i$ of words where $\sum_{i=1}^{K} p_i = 1$ and $\sum_{i=1}^{K} q_i = 1$ [8]. $K$ is the predetermined number of modes. In this case, log-energy of two words is calculated as

$$\log E(f, g) = \log \sum_{j=1}^{K} \sum_{i=1}^{K} p_i q_j e^{\xi_{i,j}}, \tag{2.14}$$

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, [8]. 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 multiple modes, their respective variances are not as large as the variances of the pioneering single mode version of the Gaussian embeddings. Although most of the training scheme is similar with initial Gaussian embeddings, there are certain points that separate multimodal version. For instance, mode probabilities $p_i$ should always be kept in the interval [0,1] and their sum should equal to 1. Their optimizations are done over unconstrained $s_i$ values and the softmax function is used to convert $s_i$ values to probabilities:

$$p_i = \frac{e^{s_i}}{\sum_{j=1}^{K} e^{s_i}}. \tag{2.15}$$

In addition, logarithm of variances are optimized due to their positivity constraint. Furthermore, a regularizer term $\epsilon$ is used to prevent instability due to small values of covariance values since inverses of the diagonal covariance matrices in gradients cause instability when diagonal values become too small.

# Chapter 3

# Zipf's Laws

Since George Kingsley Zipf has introduced it, Zipf's law has been an important empirical regularity within the statistical natural language processing, and computational linguistics fields [38, 2], and, in a more general sense, within the information, physical and social sciences [39]. Despite Zipf's law's intuitively simple essence, several Zipfian regularities are observed in relationships between word features and frequencies of words. Zipf's law of meaning distribution and the related meaning-frequency relationship, and the law of abbreviation are common examples [2, 3, 1, 40] Interest in the Zipfian regularities has been continuing and remains as an important area of study [1, 41, 42, 7, 43, 44, 45, 46, 6].

## 3.1 Zipf's Law of Word Frequencies

The law of word frequencies is the most known Zipfian phenomenon. It claims that there is a kind of power-law relation between the frequency of words, and their respective frequency ranks in a document or corpus [38, 2]. Although there exists many variations, the main relation is mathematically expressed as follows:

$$f_r \propto b/(r^a), \tag{3.1}$$

where $r$ and $f_r$ stand for word rank and the corresponding frequency value of word whose rank is $r$, respectively. Generally, power parameter $a$ and the coefficient $b$ are approximately 1. In Fig. (3.1), relative word frequencies are displayed along with their word ranks. Exponential decrease in frequencies is quite noticeable. Besides directly plotting the relation between frequencies and ranks, logarithmic scales are also used. The main reason of choosing the first 100 words is that it is difficult to observe power relation clearly in the presence of the long tail of word frequencies. Power relations give linear appearance in logarithmic scale. Fig. (3.2) displays Zipf's law of word frequencies in log-log scale, where a linear pattern can be observed.

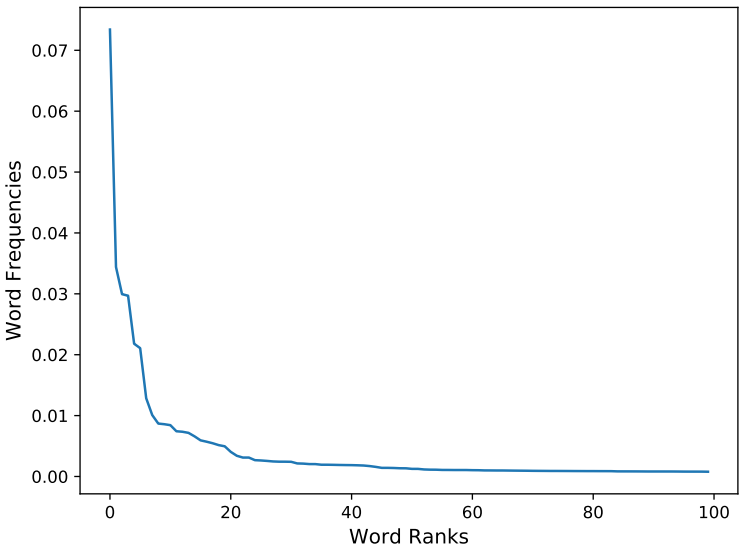


Figure 3.1: Frequency-rank relations of the most frequent 100 words of a corpus.

Besides Zipf's fundamental approach, there are studies that aim to generalize and expand this power-law. For example, [47, 48] generalize Zipf's stimulating study to explain word frequency-rank relation better. As a result of this generalization, word frequency relations are also called the Zipf-Mandelbrot law. Later, generalizations and examinations of Zipf's law in various mathematical and information science fields have continued, [44, 49, 50]. [42] proposes a model by combining empirical and rational approaches to show the dependence of Zipf's law on the text size. The contribution is to show that the additional parameters introduced by Mandelbrot depend on text length. [51] proposes a stochastic

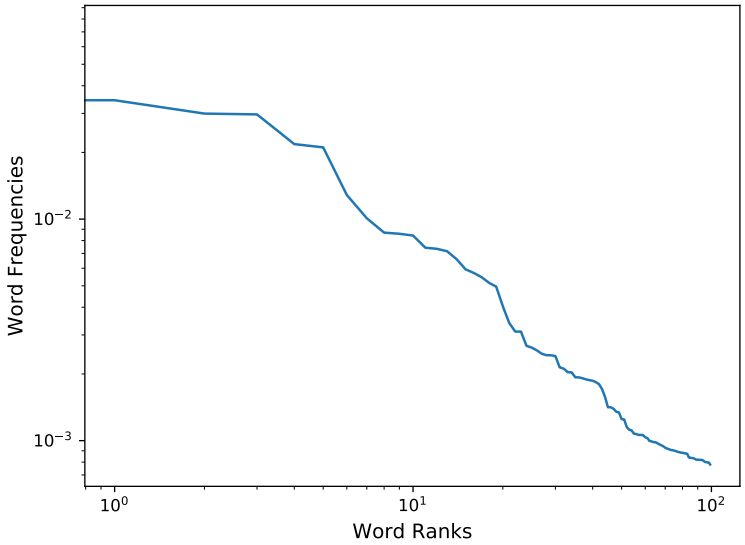


Figure 3.2: Frequency-rank relations of the most frequent 100 words of a corpus in logarithmic scale.

model to estimate vocabulary sizes of corpora from different timelines. They create two groups of words: **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[1] [52] as double power laws. Zipf's law has also been recently used to track diachronic changes in corpora from different timelines and languages in terms of lexical and syntactic properties [53].

## 3.2 Zipf's Law of Meaning Distribution

Another important phenomenon that Zipf proposed is the law of meaning distribution. Main motivation for the emergence of this law mainly leans on "the principle of least effort" given in Chapter 1. Recall that the trade-off between the sender and receiver leads to some of the words being polysemous. Zipf's law of meaning distribution predicts a relationship between the number of meanings of a

[1]Similar to Zipf's law, Heaps' law asserts that there is a power relation between the number of distinct words (i.e., vocabulary size) and the total number of words in a document.

word and its frequency rank, [2, 3]. The law of meaning distribution is expressed mathematically as follows:

$$m \propto r^{-\gamma}, \tag{3.2}$$

where $m$ is the number of meaning of a word and $r$ is its rank as in Eq. 3.1. Here, $\gamma$ at the exponent corresponds to a value which is close to 0.5.

Zipf has also derived the meaning-frequency relation by using the law of word frequencies and the law of meaning distributions. The meaning-frequency relationship indicates a similar relationship where the predictor variable, instead of the frequency rank, is the frequency itself. [3]. Its mathematical expression is as follows:

$$m \propto f^{\delta}, \tag{3.3}$$

where $f$ is frequency value as in Eq. 3.1 and $\delta \approx 0.5$. Both the law of meaning distribution and meaning-frequency relationship assert that the frequent words tend to have more meanings. This law is also valid in other languages similar with the law of word frequencies. For example, the law of meaning distribution is studied on two Turkish corpora to inspect the number of meanings a word can attain in Turkish according to frequency. Later, the relation between the number of meanings and word frequencies is studied by [54, 55, 56] within the scope of communicative optimization. Another recent and critical study is [1], where the authors develop a quantitative analysis to study the meaning-frequency relationship and provide further statistical evidence.

In order to quantitatively study Zipfian patterns, [54, 55, 56] associate semantic vagueness of words with the number of links connecting words with meanings. If a word has a high number of links, it would have many interpretations in the context where it appears. In [55], precision and vagueness terms can be interpreted in analogy with the semantic specificity and uncertainty. [1] conducts 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 [57]. There

also exist significant theoretical contributions regarding the origins of Zipfian patterns, [7, 6]. In [6], Zipf's law of meaning distribution is studied in detail, and a strong positive correlation between word frequency and the number of meanings is verified. Zipf's relatively simplistic assumption is replaced by a more general and compact assumption that can fit into the general theory of communication [6]. According to this assumption, the joint distribution of a word and meaning can be sufficient to characterize the Zipfian regularities. On the other hand, [7] explains Zipf's law on purely linguistic grounds. [7] examines the degree of generality of words and composes 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, [7] shows that Zipf's law can be constructed by arrangements of word meanings in the semantic space.

## 3.3 Zipf's Law of Abbreviation

Another linguistic regularity that is proposed by Zipf is the law of abbreviation, which is also known as the law of brevity. It states that more frequent words tend to be shorter. This law shares the same reasoning with the law of meaning distribution which is "the principle of least effort". Using short words more frequently leads to a decrease in the sender's effort. However, Zipf did not propose any mathematical expression for this phenomenon as in the law of word frequencies or the law of meaning distributions. Nevertheless, there are studies that suggest power-like functions for the relation between word lengths and word frequencies [58].

As an interesting study, [59] shows that even animal communication behaviours agree with the law of brevity. In another work, it is demonstrated that average information content is a better predictor of word length than frequency [60]. Information content is obtained by the n-gram information of words. Experiments on different languages show that information content brings more correlation with word length rather than word frequencies [60].

## 3.4 Other Studies Related to Zipf's Laws

Although Zipf's law is a well-known phenomenon in linguistics [38, 2], it is demonstrated that Zipfian regularities can also be observed in several different disciplines. [61] studies company income distributions and observed Zipfian regularities. Similarly, [62] reveals the Zipfian distribution of city populations. Another study associates internet statistics (e.g., webpage requests) with Zipf's law, [63]. Furthermore, [64] investigates the existence of linguistics law in acoustic signals. Experiments from different languages belonging to different language families show that human voice exhibits features of conventional linguistic laws such as the law of word frequencies, Heaps' law and the law of abbreviation.

Zipf's law is also expanded and applied to several contextual problems of NLP. For example, many statistical laws, including Zipf's, are tested in complex systems that do not carry typical scenarios for data [65]. These complex systems include earthquake magnitudes, inter-event times of consecutive occurrences of words in texts, frequency rank relationships of words, the relation between degrees of nodes from a network, and ranks of the degrees of nodes. Deviations of statistical laws against the availability of large datasets are also examined extensively, [65]. In [66], co-occurrence matrices are 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`.

In the literature, generalizations to Zipf's law and the Zipfian patterns in different knowledge domains are also studied. [67] asserts that even randomly composed corpora exhibit Zipfian patterns. As a counter-argument, [41] and [43] 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 about language structure and evolution. Another study reveals that passwords also show Zipfian patterns [68]. In [69], it is asserted that power-law regularities can also be discovered in computer programs via empirical analysis of original software systems.

Review in [70] on Zipf's laws presents critical arguments to explore word frequencies. In that study, it is asserted that word frequencies are complex language features, which a single power-law formula cannot explain. It is also claimed that although word frequencies exhibit patterns compatible with Zipf's law, they have a complicated structure that only methods beyond Zipf's law can reveal. Nevertheless, when [70] considers the specific subsets of the lexicon, such as the Swadesh list[2] or number words in the scope of Zipf's law, it is noticed that those word groups also fit near-Zipfian patterns, [70].

[2]The Swadesh list is a collection of certain words reflecting basic frequent concepts in languages [71]. More detail will be given in Chapter 4.

# Chapter 4

# Uncovering the Zipfian Regularities of Semantic Breadth

In this chapter, general information about the followed methodology and the designed experimental setup will first be given. Then, results obtained from our experiments will be presented, and results are discussed.

As mentioned in Introduction, the main objective of this thesis is to reveal Zipfian patterns in the generality (i.e., semantic breadth) of words. To this end, a tool that mathematically models the semantic breadth of words is necessary. Multimodal Gaussian embeddings are such tools that provide a quantitative measure of semantic breadth. Our overall picture of experimental design is given in Fig. (4.1). First, a multimodal Gaussian model is trained to obtain mean vectors and covariance matrices. From covariance matrices, overall representative variances are calculated for each word. Then, Zipfian regularities of word variances will be examined by sorting them according to word frequencies. Examination of Zipfian regularities is made in various scopes: behaviours of word variances under the process of neutralizing polysemy, different vocabulary thresholds, Swadesh and number words and different languages. A control experiment is also implemented on our corpus which is randomly created to confirm whether Zipfian regularity is due to the internal dynamics of language but not due to the dynamics of the

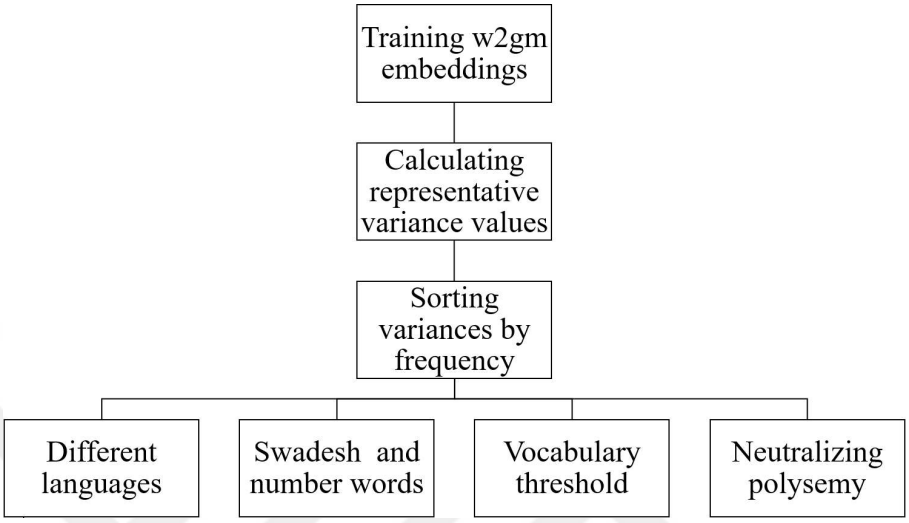


Figure 4.1: Diagram of experimental procedure.

learning mechanism of Gaussian embeddings. In addition to Zipfian experiments, entailment experiments where word frequencies leverage entailment detection algorithms are also explained in the following sections.

## 4.1 Calculating Representative Variance Values

As mentioned in previous chapters, w2gm models offer more than one variance value for each word in the spherical scheme. In order to represent the semantic breadth of words, the overall value, including the information of word variances that belong to different modes of the multimodal Gaussians, should be calculated accordingly. For the bimodal case, the mean and variance moments of the mixture

can be calculated as [72]:

$$f(x) = pg_1(x) + (1-p)g_2(x), \tag{4.1a}$$

$$\mu = p\mu_1 + (1-p)\mu_2, \tag{4.1b}$$

$$\nu_2 = p[\sigma_1^2 + \delta_1^2] + (1-p)[\sigma_2^2 + \delta_2^2], \tag{4.1c}$$

where

$$\begin{aligned} \mu &= \int x f(x) dx, \\ \delta_i &= \mu_i - \mu, \\ \nu_r &= \int (x-\mu)^2 f(x) dx. \end{aligned} \tag{4.2}$$

In Eq. 4.1, $g_i(x)$ stands for the components of mixture distribution, $p$ is the mixture probability of the first component in the bimodal case, $\mu$ and $\nu_2$ are the first moment (mean) and the second moment (variance) of the mixture, respectively. Lastly, $\sigma_i$ and $\mu_i$ are the variance and mean of the $i^{th}$ 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. Each multivariate item can be thought of as separate distribution. The mean of the product of two independent random variables equals to the product of individual means. Note that mean vectors are normalized. Therefore, the result of the multiplication is very close to zero. When each mode is approximated with the product of each multivariate component, the variance of a word can be approximated with the weighted sum of variances in mixture component from Eq. 4.1c.

Before sorting the words according to their frequency ranks, a final operation needs to be performed. Note that using the same corpus to obtain variances and frequency ranks of words may yield correlated errors. Even if 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 estimates variances of words, and the other will be used to estimate frequency

ranks. The same purpose 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, [70].

## 4.2 Leveraging Zipfian Regularities in Entailment Improvement

The existence of Zipfian regularities can also be exploited in some practical NLP applications. For demonstration purposes, an example will be presented. In this thesis, a method is proposed to improve the performance of word embeddings in entailment tests by injecting the knowledge of this regularity into simple point word embeddings. It should be noted that the proposed method does not need training of complex non-point word embeddings but only leverages embeddings with the knowledge of Zipfian patterns.

Entailment relation can be found in two different levels in languages. Those are sentence level and word level entailment. In sentence level entailment relation, the truth of one sentence `A`, requires the truth of the other sentence `B`. In another saying, if there is an entailment relation between the sentences `A` and `B` (if `A` entails `B`), `A` cannot be true without `B` being true. For example, "*I visited Anıtkabir yesterday.*" entails the sentence "*I was in Ankara yesterday.*". One cannot visit Anıtkabir without being in Ankara. The condition for the first sentence being true is that the second sentence is true. Word level entailment relations have the same logic but a simpler version. This relation is also known as subordinate-superordinate or hypernym-hyponym relations. For example, being a `tree` requires being a `plant` simultaneously. Similarly, `football - sport` has the same relationship. Most of the time, word level entailments have `is-a` connection between two words. At early stages of the NLP studies, such patterns or phrases are utilized in entailment detection [73]. Then, there have been studies dealing with the problem in the scope of distributional hypothesis and/or word

embeddings [74, 75, 76, 77]. On the other hand, [8, 78] try to find entailment relations by using probability distribution-based word embeddings.

Besides these studies, [79] approaches the lexical entailment relations from a different perspective. It is stated that hypernym words are semantically more general than hyponym words. Moreover, a hyponym word can only occur in contexts where its hypernym can also occur. For example, `animal` can be used in almost all contexts where `antelope` occurs. However, the opposite case is not possible. This increases the frequencies of hypernym words compared to the frequencies of their hyponyms due to hypernym words' context variety. The perspective of word variances about the word frequency and generality-specificity relations is entirely compatible with the hypothesis in [79]. Starting from findings in [79] and the Zipf's law of meaning distribution [3], we incorporate frequency information of words in the pair to a baseline method for entailment score calculation while deciding whether there is an entailment relation between a given word pair.

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

$$CS_{(m,n)} = \frac{\vec{w}_n^T \vec{w}_m}{\|\vec{w_m}\| \|\vec{w_n}\|}, \tag{4.3}$$

$$FS_{(m,n)} = \frac{f_n}{N} \times 100 + \frac{\log_{10}(f_n/f_m)}{k}, \tag{4.4}$$

$$OS_{(m,n)} = CS_{(m,n)} + FS_{(m,n)}, \tag{4.5}$$

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. 4.4 indicates the frequency score of the word pair $(m, n)$. The hypernym candidate is word $n$, and the hyponym candidate is word $m$. $f_m$ 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 the test type. Based on our experimental results, values between 20 and 40 yield favorable results. We design the frequency score ($FS$) by using two approaches. The first is related to the respective frequency of the candidate hypernym word according to the total word frequencies of all words in the vocabulary. It will be experimentally shown that the probability of simultaneously being a low-frequency word and a semantically broader hypernym word is small. Therefore, the frequency score of the given pair with a 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's frequency is greater than that of the second, it negatively contributes to the frequency score. In this case, the logarithm of the ratio between frequencies is taken to decrease the frequency score. Overall entailment score for word pair $(m, n)$ is then calculated by adding the frequency score to the cosine similarity of the given word pair as in Eq. 4.5.

## 4.3 Experiments

In this subsection, experimental results are presented. First, experiments that use multimodal Gaussian embeddings are given in Section 4.3.1. Then, extensive experimental results and discussions regarding isolating the effects of polysemy will be displayed in Section 4.3.2. For the sake of isolating the effects of polysemy, results for unimodal Gaussians are first presented to quantitatively show that multimodal Gaussians are better models in eliminating polysemy. Then, the results obtained from sense-controlled corpora are given so that the effect of polysemy is completely removed, and the variances represent only the extent of the semantic breadth of words. The details of our control experiment and the corresponding results on the randomly generated corpus are given in Section 4.3.3. In Section 4.3.4, experiments performed for Swadesh and number words are presented. Analysis of the effects of frequency threshold used in constructing the vocabulary is given in Section 4.3.5. The experiments for corpora composed from different languages are presented in Section 4.3.6. Finally, the results regarding

the entailment tests are given in Section 4.3.7. For diversifying the implementations, experiments are performed both with or without corpus splitting. In Sections 4.3.1 and 4.3.4, corpus splitting is performed while, in Sections 4.3.2, 4.3.5 and 4.3.6, the entire corpus is utilized to provide more data for training the Gaussian embeddings.

### 4.3.1 Examining Zipfian Regularities Through Multimodal Gaussian Embeddings

In this subsection, the model of multimodal Gaussian embeddings trained on a concatenated corpus of UKWAC and Wackypedia [80] is used to obtain variance values of words, [8]. This model allocates two modes for each word. In other words, the parameter $K$ is equal to 2. This model and other models employed in this study follows the spherical variance scheme. As a standard statistical process, words whose frequencies are less than 100 are removed from the vocabulary. The final vocabulary size is 314,129. The effects of the threshold are experimentally studied in Section 4.3.5 and it is shown its value is inconsequential. The dimensionality of mean vectors of all models used throughout this study is 50. As an optimizer, Adagrad is employed with initial learning rate 0.05 that decreases linearly [81]. Models are trained along 5 epochs. Mean vectors are initialized with uniform distribution over $[-\sqrt{3/D}, \sqrt{3/D}]$ where $D$ is the number of vector dimensions. Initial variances values are set to 0.05. For the sake of the corpus splitting procedure mentioned in Section 4.1, a second corpus is used to determine the frequency ranks of words. This corpus is the English Wikipedia with a vocabulary of size 229,922. In order to be compatible with the vocabulary of the trained model, words with occurrences of less than 100 in the second corpus are also discarded. The downloaded Wikipedia dump is cleaned from all redundant materials such as document numbers, URLs, and HTML syntax by pre-processing. 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. In total, there are 229,922 types with a frequency of more than 100. After pre-processing steps, words are ranked 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 given in Eq. 3.1. Remind that power parameter $a$ and the coefficient $b$ are approximately 1[1]. In [83], it is shown that there exist some deviations from the above formula for infrequent words. One can obtain more accurate results by using a second exponent for higher ranking words to deal with the deviations caused by infrequent words. However, since addressing deviations due to infrequent words is beyond the scope of this thesis, the standard Zipf's law formula is employed.

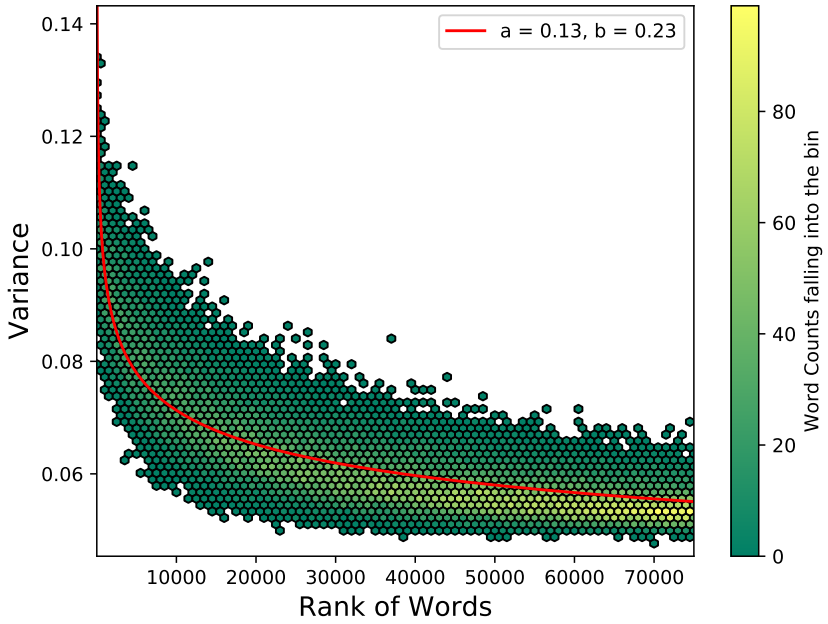


Figure 4.2: Two dimensional histogram of word variances and work ranks.

Throughout this section, several Zipfian patterns of variances versus frequency rank of words are presented in both standard or log-log forms. In the insets of figures, regression parameters ($a$ and $b$) according to the expression given in Eq. 3.1

[1]These parameters are generally determined from regular corpora or texts. In case of texts from fragmented discourse schizophrenia or from children, exponents can deviate. For more information, please refer to [82].

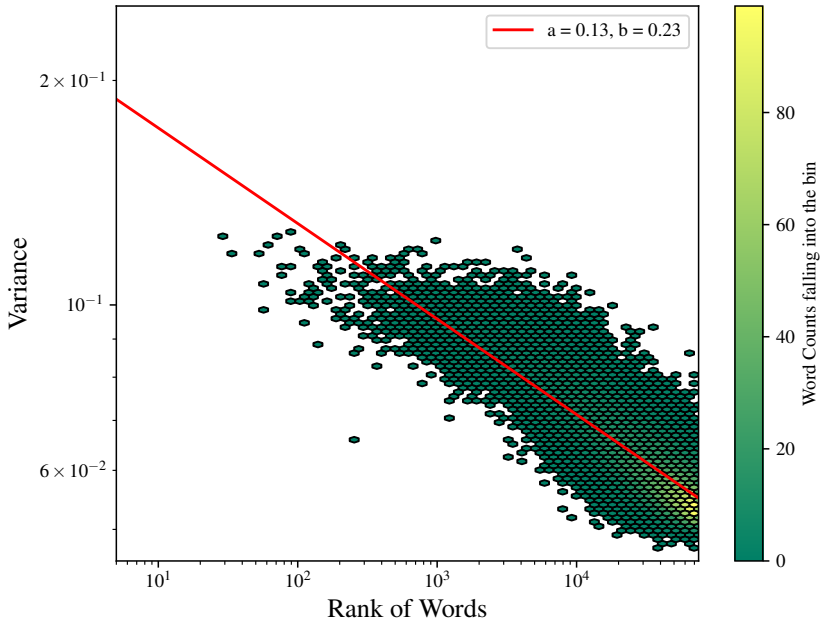


Figure 4.3: Two dimensional histogram of word variances and word ranks in logarithmic scale.

can also be found. Fig. (4.2) and Fig. (4.3) demonstrate the pattern between variance and word rank. Before displaying representative variance value for a word, mixture weighted average of spherical variance values of each mode is calculated as stated in Eq. 4.1. For practical reasons, around the first seventy-five thousand words of the vocabulary, which sufficiently demonstrate the general pattern, are displayed. It should be noted that plotting these values in a conventional way leads to show many ripples between variance values at close ranks, which cannot be appropriately visualized. Two dimensional histograms present results more clearly by gathering similar results into bins where the colors of bins change according to the number of data instances they include. The number of words within every bin equals to the corresponding value at the color-bar multiplied by 100. As shown in Figs. (4.2) and (4.3), variances of words display a Zipfian pattern. Variances decrease as the order of word rank increases noticeably following Zipf's law. Regression parameters to fit the curve are not divergent from conventional Zipf's law parameters.

At this point, 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. The main aim of this study 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 variances for words of close frequency ranks, it is possible to see a decrease in variances when the word frequency decreases considerably.

The results reported above are also 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 various animals or animal-related concepts in specific word windows. It is reasonable to say that the word `animal` has high frequency and uncertainty (or variance) because `animal` does not specify any species (let alone a family of species) when used in a sentence all alone. This leads to an increase in its variance due to the uncertain context. Here, it should be remarked that `animal` is not a polysemous word, although it can carry different side meanings or be used metaphorically in a sentence. Therefore, associating the increased variance only with polysemy would not be a reliable inference. Furthermore, the multimodal Gaussian model eliminates the impact of polysemy on variance to a certain degree. On the other hand, a specific animal `antelope` represents a more specific entity. Since it is a particular word, its word frequency is naturally lower. Consistently, its uncertainty would be low. Indeed, weighted average of variance values of `animal` and `antelope` are 0.090 and 0.064, respectively. Individual quantitative results and general variance pattern of the words confirm the intuitive reasoning given above. Additional and comprehensive experimental results will be presented in the following subsections to quantitatively verify the statistical significance of our analysis by 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 multimodal Gaussians. As stated in Section 4.3.1, the model with two modes is utilized to display Zipfian regularities. Figs. (4.4), (4.5), (4.6) and (4.7) demonstrate the behavior of maximum and minimum variances belonging to 2 modes of words

with respect to frequency ranks. Consistently, a similar Zipfian pattern can be observed in both figures. Regardless of which mode of the multimodal Gaussians are used, the Zipfian pattern remains. These experimental results from multimodal Gaussians further support the proposed Zipfian regularity between uncertainty and frequency ranks.

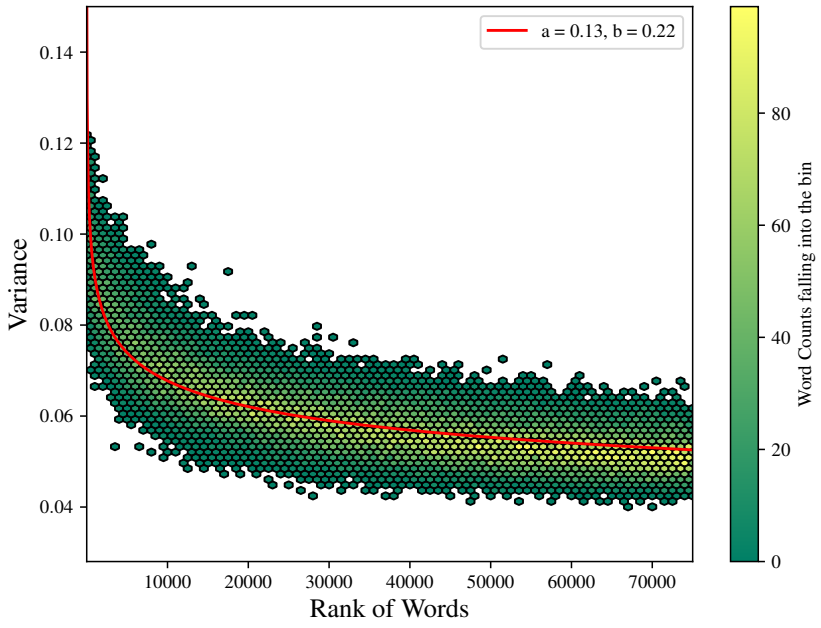


Figure 4.4: Minimum variances of multimodal Gaussian embeddings.

### 4.3.2 Effects of Polysemy and Experiments on Sense-Controlled Corpora

The main motivation behind the introduction of multimodal Gaussians is to enhance the initial version of Gaussian embeddings by minimizing the impact of polysemy on word variances. In this model, Gaussian embeddings with more than one mode give an extra degree of freedom to treat different senses of words as distinct words. Different senses are learned more independently in the semantic space. Therefore, excessive variances can be eliminated. The variance values represent the uncertainty of the semantic breadth of words other than contributions from different senses due to polysemy. For example, the word `rock` takes

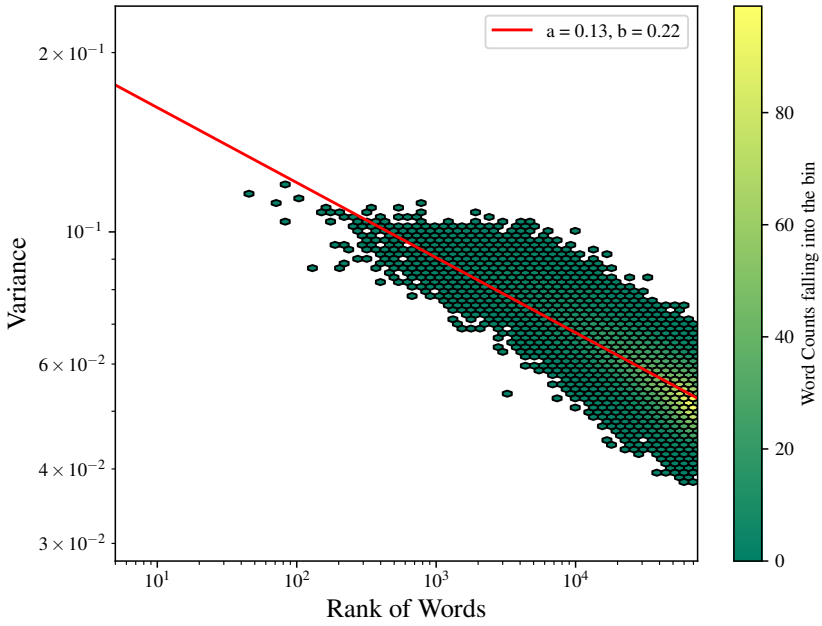


Figure 4.5: Minimum variances of multimodal Gaussian embeddings in logarithmic scale.

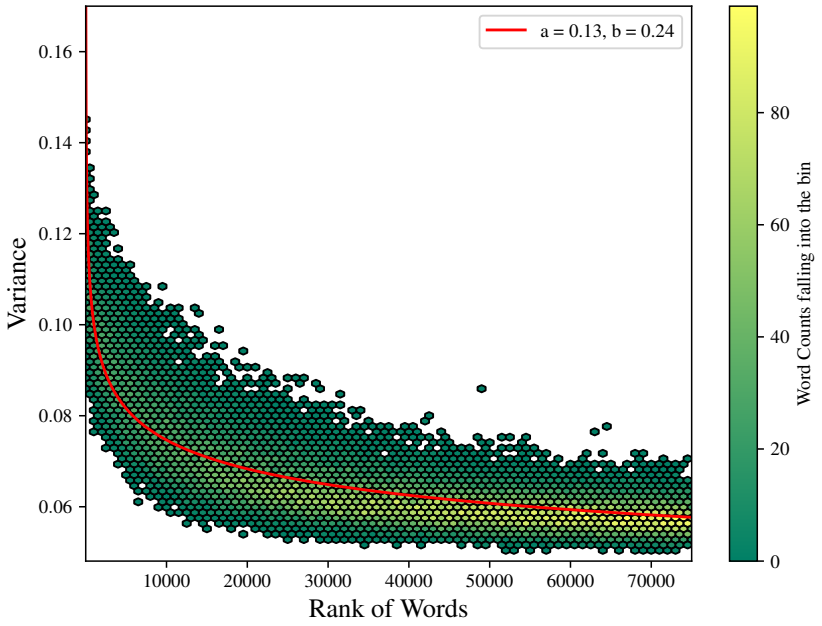


Figure 4.6: Maximum variances of multimodal Gaussian embeddings.

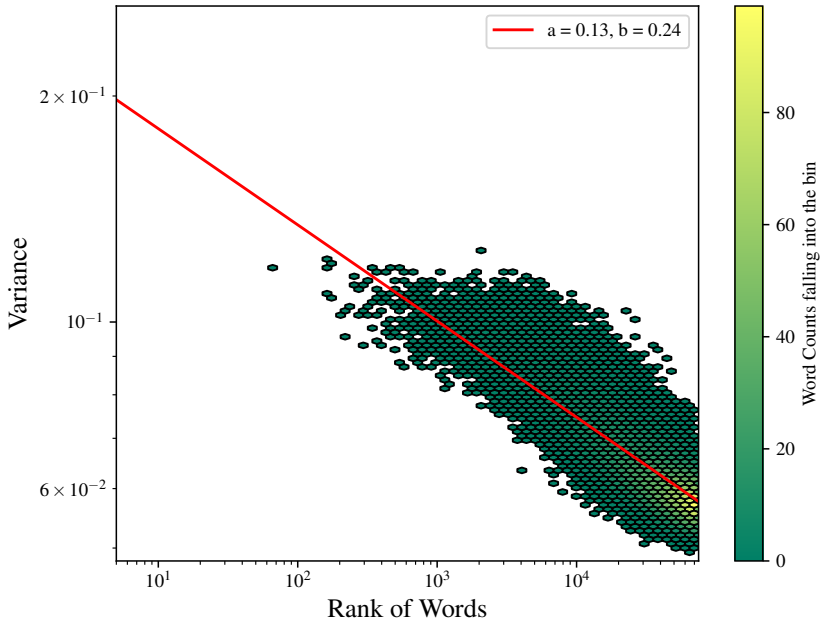


Figure 4.7: Maximum variances of multimodal Gaussian embeddings in logarithmic scale.

place in the same context with the words that are related to *music* and also *geology*. If the Gaussian model is unimodal, then variance of `rock` will try to capture different meanings from both *music* and *geology* concepts. This situation leads to a rise in the uncertainty of the word as well as its variance.

Here, a question arises to inquire about the impact of polysemy on Zipfian patterns. Do Zipfian regularities purely stem from the semantic breadth of words? If not, to what extent does the polysemy affect Zipfian patterns of variances? In order to answer these questions and quantitatively support the qualitative discussions stating that polysemous words tend to have high variances, a unimodal Gaussian embeddings model is trained on the aforementioned Wikipedia corpus at first. The pattern obtained from the unimodal Gaussian model is given in Figs. (4.8) and (4.9). As it can be seen in Figs. (4.8) and (4.9), word variances obtained from unimodal Gaussians are significantly higher than those coming from bimodal one. It is also observed that the unimodal variances do not fit the original Zipf's model as smoothly as the bimodal case. This result illustrates that the effort of the multimodal scheme for diminishing excessive variance has a

Table 4.1: Variance values of the word `bank` for multimodal Gaussian embeddings with different number of modes

| # of Modes | Sense | Variance Value |
|---|---|---|
| 1 Mode | `bank`, 0 | 0.147 |
| 2 Modes | `bank`, 0 | 0.081 |
| | `bank`, 1 | 0.085 |
| 3 Modes | `bank`, 0 | 0.076 |
| | `bank`, 1 | 0.077 |
| | `bank`, 2 | 0.070 |

positive impact on Zipfian regularities on words' variances and the isolating the effects of semantic breadth on them.

In order to clarify variance differences between models with different modes, Table 4.1 tabulates the variances of unimodal, bimodal, and trimodal Gaussian embeddings of the word `bank` as an additional experimental demonstration. In Table 4.1, it can be seen that the variance values of the Gaussian embeddings of `bank` decrease as the number of 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 relation and from the semantic breadth couple to each other. This coupling prevents us from concluding that word frequencies are directly related to semantic breadth. Therefore, one can infer that freeing variances from the effects of polysemy plays a significant role in revealing the Zipfian regularities of word variances. As mentioned before, Gaussian models with single-mode carry excessive amounts of uncertainty and so variance due to their attempt to capture too many senses and their semantic breadth. That is why utilizing the multimodal Gaussian embeddings represents the semantic breadth of words more accurately.

Another way, a more laborious but more effective one, for neutralizing the polysemy is to use sense-labeled (or sense-controlled) corpora in the training phase. Using external lexical sources is a prevalent way to extract the senses of words. In [1], the number of senses for each word is imported from WordNet [57]. Following

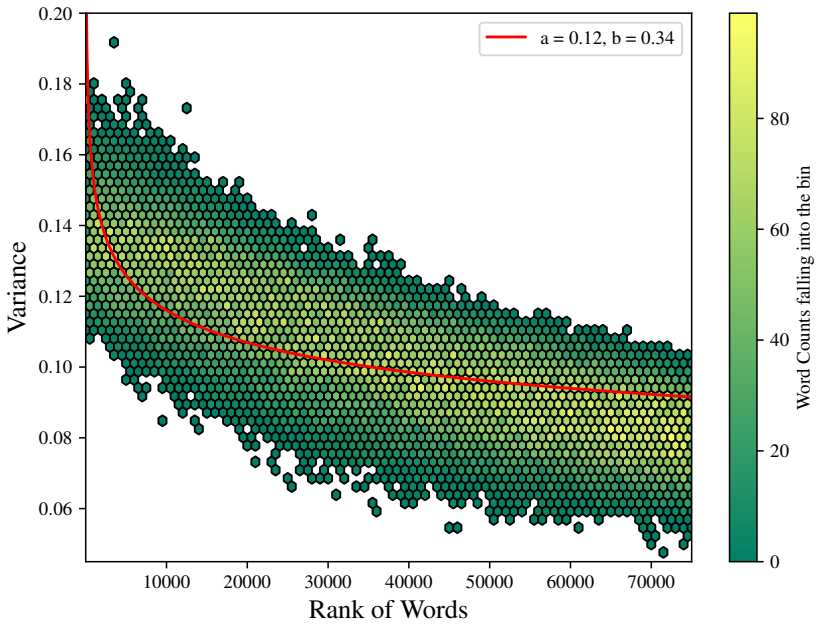


Figure 4.8: Variances of unimodal Gaussians embeddings.

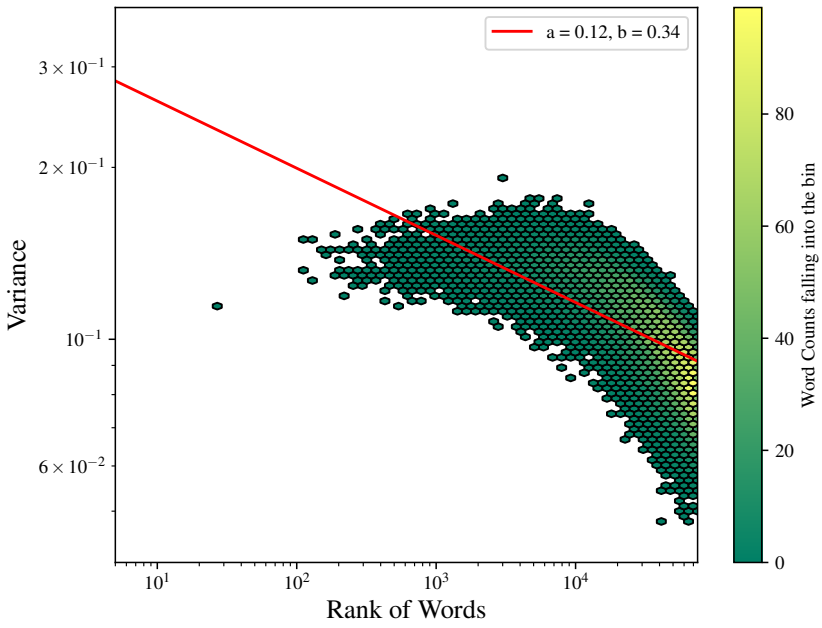


Figure 4.9: Variances of unimodal Gaussians embeddings in logarithmic scale.

a similar procedure, words in our corpus are labeled according to their WordNet senses by using the Lesk algorithm provided by the NLTK library [84, 85]. Lesk is a knowledge-based traditional algorithm for Word Sense Disambiguation (WSD) task [84]. Knowledge-based WSD algorithms utilize sense inventories to extract senses of words rather than supervised learning methods. In particular, the Lesk algorithm receives a sentence as input and returns the WordNet senses of every word in the given sentence. After traversing the entire corpus, every distinct sense becomes as if they are different words. For example, `rock` in *music* context and `rock` in *geology* context are treated as separate words. Therefore, polysemy is totally neutralized in the corpus. Furthermore, to increase the disambiguation performance of the Lesk, the part-of-speech (POS) information of words is also provided to the Lesk algorithm. Having POS information alleviates the complexity of disambiguating senses from WordNet. On the other hand, it should be noted that some words are not present in WordNet. Most of the time, these words are constituted by prepositions, conjunctions, or articles. Such words are labeled with a single dummy label. As a result, by treating every sense of a word as different words and assigning them to distinct Gaussian embeddings, the effect of polysemy is removed. Then, our sense-controlled corpus is used to train the Gaussian embedding model. In Figs. (4.10) and (4.11), the Zipfian pattern of the variance values can still be observed.

In order to provide more experimental evidence, another experiment is performed with the manually sense-annotated SemCor corpus [86]. SemCor corpus is a corpus where humans annotated word senses. Although that kind of annotation is more reliable, it is extremely laborious. Therefore, SemCor is relatively small, with approximately 200,000 tokens. The same experimental strategy is followed while training on this corpus. Figs. (4.12) and (4.13) demonstrate the behavior of variances. Although corpus size is small, manual sense annotation provides a better fit to Zipfian regularity. It is well-known that WSD is an AI-complete problem [87]. Although there are several well-performing algorithms, obtaining high performance in WSD is still a challenging task, especially through unsupervised means. Manual annotation is more effective than unsupervised algorithms. In other words, SemCor provides better polysemy neutralization than

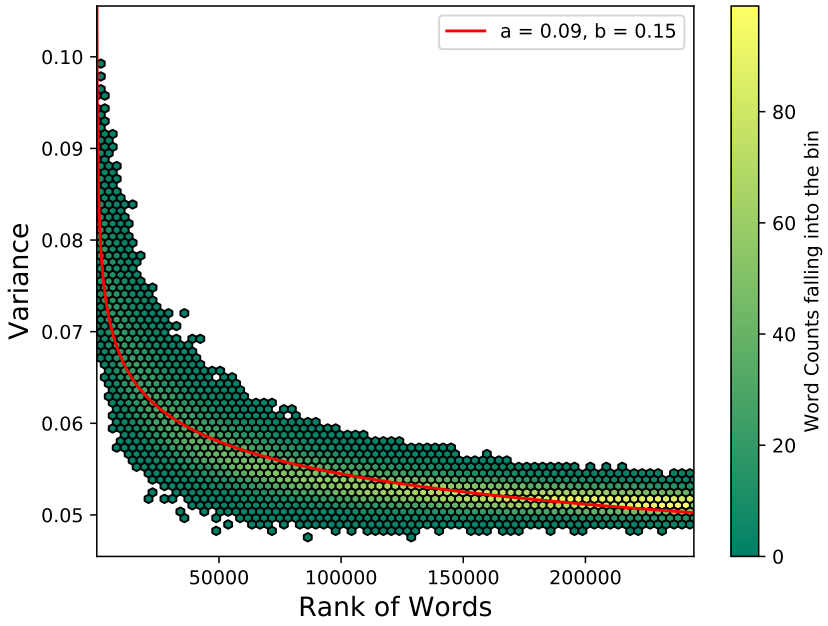


Figure 4.10: Variances of Gaussian embeddings trained on the sense-controlled corpus.

sense-controlled corpus obtained by deploying the Lesk. Therefore, variances obtained by the SemCor corpus fit to the Zipf model better.

Another crucial aspect of these two experiments based on polysemy neutralization is that obtained results show that the variances of Gaussian embeddings are now only measuring the semantic extents of words without the effects of polysemy. In other words, the results of experiments on sense-controlled corpora strongly indicate a Zipfian relation between semantic breadth and frequency.

Lastly, to verify the statistical significance of results, correlation analysis is performed between word variances and word frequencies by following the procedure given in [1]. Pearson, Spearman, and Kendall correlations [88] are calculated between variances and frequency, obtained from both Wikipedia corpus and its sense-controlled version. Results are tabulated in Table 4.2, where the correlation results between the number of senses and frequency from [1]'s quantitative study for Zipf's meaning-frequency relation are added. It is observed that the variance values are highly correlated with frequencies. Here, it should be noted that the

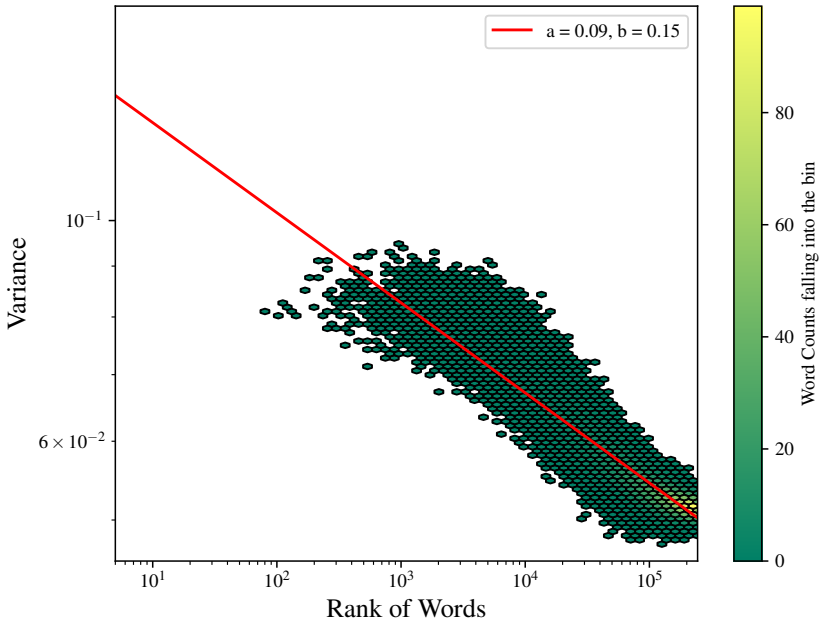


Figure 4.11: Variances of Gaussian embeddings trained on the sense-controlled corpus in logarithmic scale.

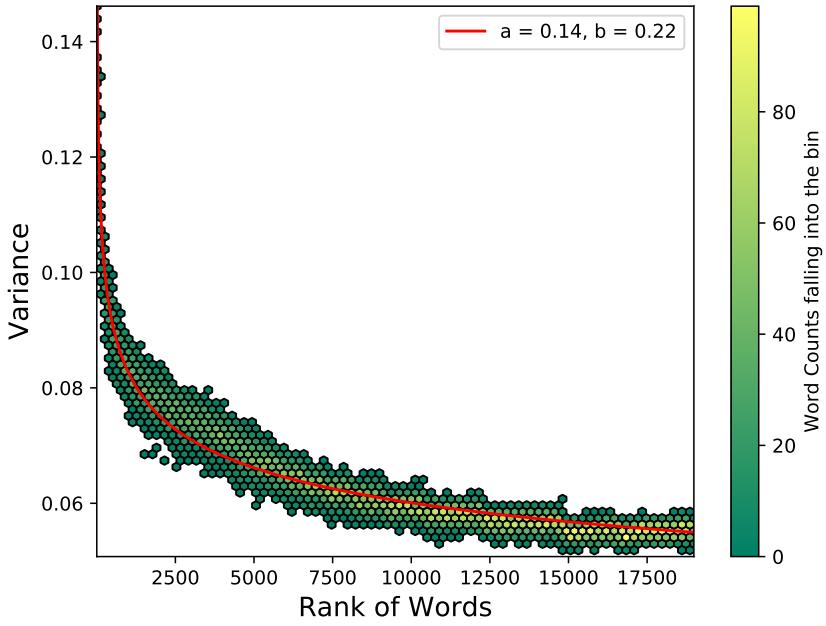


Figure 4.12: Variances of Gaussian embeddings trained on the SemCor corpus.

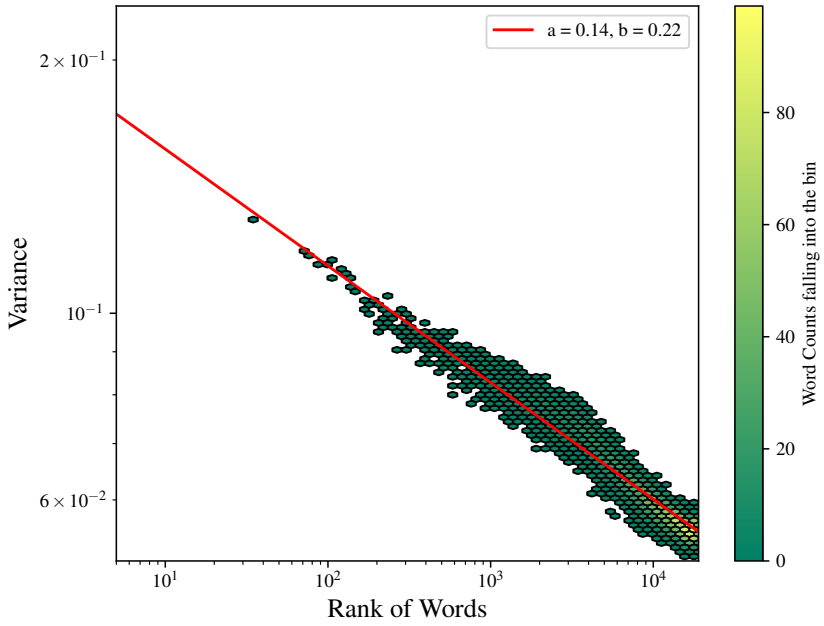


Figure 4.13: Variances of Gaussian embeddings trained on the SemCor corpus in logarithmic scale.

Pearson correlation displays the linear relationship between the variables. It is demonstrated that the actual relationship between variances and word frequencies (or word ranks) is a power relationship rather than linear. Therefore, Pearson correlation results are relatively more minor than other correlation results due to the non-linear relationship between variance values and word frequencies.

On the other hand, these independent experiments correlate different variable pairs; correlation results are not comparable with those from [1]. However, in terms of the statistical significance levels that are reached, results are compatible with those in [1].

### 4.3.3 Control Experiment on Randomly Constructed Corpus

The purpose of this experiment is to inspect whether obtained Zipfian regularities actually reflect the characteristics of words but do not arise due to dynamics of

Table 4.2: Correlations scores. Variance vs. frequency for both bimodal and sense-controlled corpora. WordNet number of senses vs. frequency from [1].

| | **Pearson** $r$ | **Spearman** $\rho$ | **Kendall** $\tau$ |
|---|---|---|---|
| Bimodal Variance vs Frequency | 0.083 | 0.828 | 0.647 |
| Sense-cont. Variance vs Frequency | 0.042 | 0.861 | 0.677 |
| Number of senses [1] vs Frequency | 0.068 | 0.422 | 0.307 |

the Gaussian embedding learning algorithm. To this end, a corpus is randomly constructed. In order to conform with English corpus which is used in previous experiments, the number of tokens and unique words are arranged accordingly. More specifically, 200,000 unique words are composed first. These words are totally random and does not carry any meaning. By randomly choosing from this vocabulary, a corpus of 2,000,000,000 tokens is constructed. Training procedures are kept same as previous experiments. Variance patterns obtained at the end of training are displayed in Figs. (4.14) and (4.15). As observed from Figs. (4.14) and (4.15), variances of words from the random corpus do not exhibit any Zipfian patterns but relatively a uniform pattern. This shows that Zipfian patterns in word variances are beyond a simple random process and can be considered as a representative of language structure in term of word uncertainties.

### 4.3.4 Zipfian Regularities on Swadesh and Number Words

In [70], many experiments are implemented to investigate whether there are other features of languages satisfying Zipf's law, other than just word frequency rank and word frequencies. One of these experiments is related to the Swadesh list [71]. 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 [70], frequency-rank curves of Swadesh lists from different languages are examined. In that experiment, ranks of words

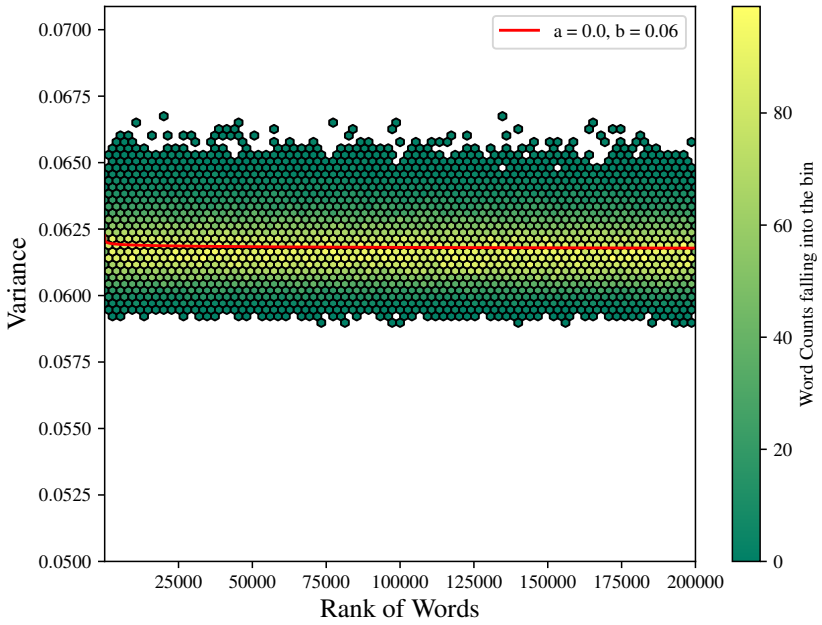


Figure 4.14: Variances of Gaussian embeddings trained on the randomly created corpus.

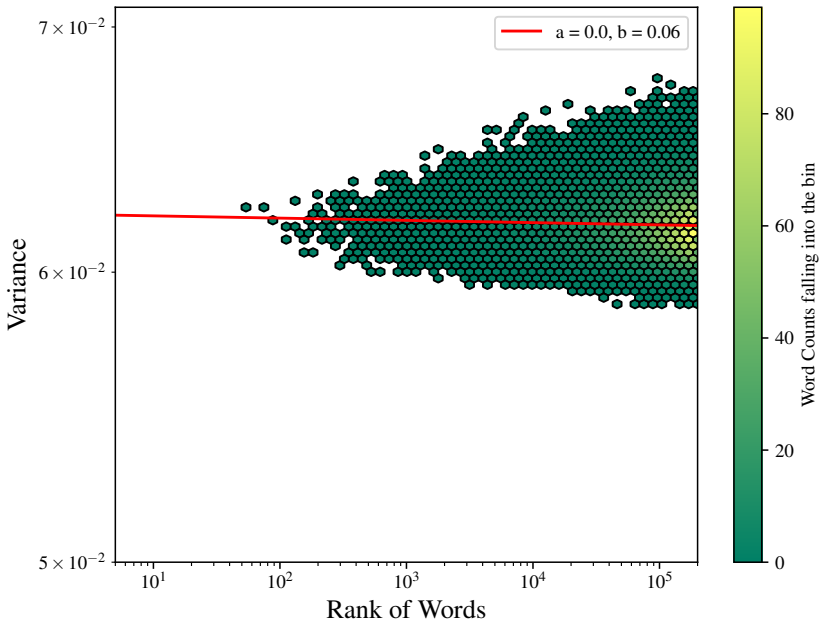


Figure 4.15: Variances of Gaussian embeddings trained on the randomly created corpus in logarithmic scale.

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 pattern. Likewise, the behavior of variances of words in the Swadesh list is investigated to learn whether variances of Swadesh words also display a Zipfian pattern. Results are given in Fig. (4.16), where the weighted average of variances is used again. Indeed, like the original Zipfian patterns of the Swadesh list given in [70], variances of Swadesh words display similar linearity in the logarithmic scale.

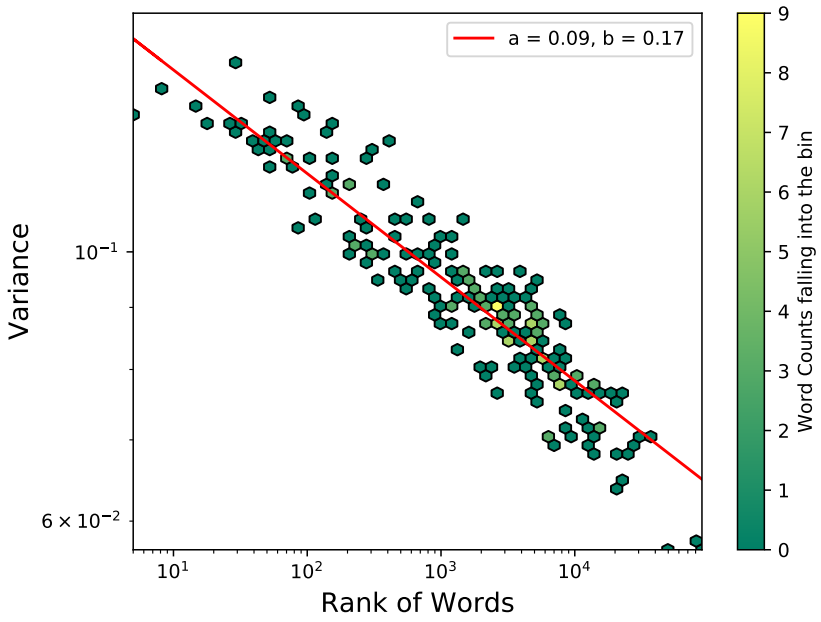


Figure 4.16: Two dimensional histogram of average variances of Swadesh words.

Another interesting vocabulary subset that can be studied to strengthen Zipfian analysis is the number words, e.g., `one`, `two`, `ten`, `twenty`. As mentioned above, [70] investigates different language aspects in order to observe that Zipfian patterns are not limited to only word frequencies. For that purpose, word ranks of number words are not used. Instead, words are ordered by their cardinality [70]. Although word ranks based on word frequencies are not used, a near-Zipfian distribution can still be observed in results reported by [70]. 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

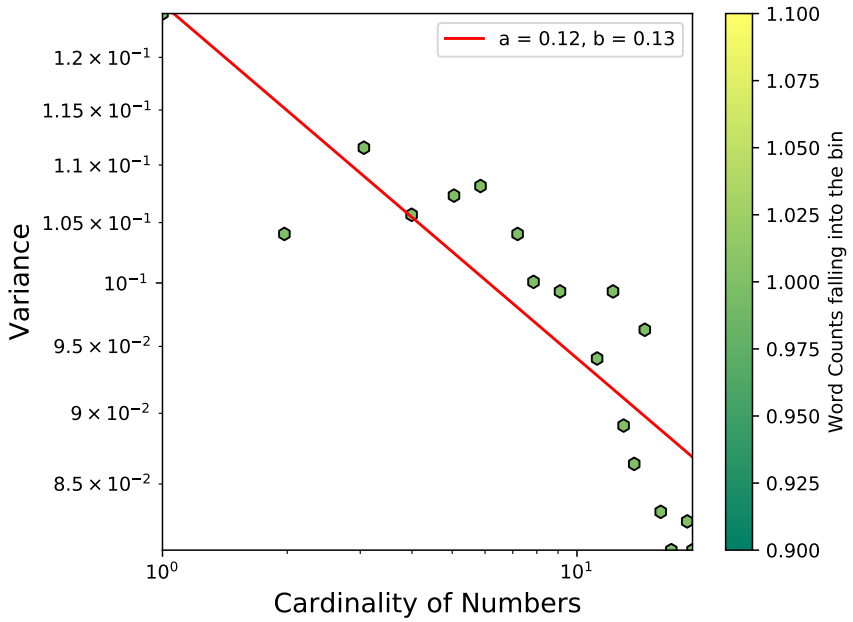


Figure 4.17: Two dimensional histogram of average variances of **number words**.

of any arbitrary number. Therefore, exceeding frequency values may be misleading. With similar motivation, the experiment is repeated with word variances. Number words are ordered according to their cardinality, and corresponding word variances are displayed in Fig. (4.17). As can be seen in Fig. (4.17), 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.

The primary purpose of implementing experiments with Swadesh and number words is to show Zipfian patterns word variances in different domains where original Zipf's law exists. Acquired results demonstrate that word variances, which are indicators of the semantic breadth of words, are compatible with the Zipf's law in a more generalized context. In other words, word variances do not exhibit Zipfian patterns in not only particular cases but also different circumstances where traditional Zipf's law maintains its existence.

### 4.3.5 Experiments with Different Frequency Thresholds

The threshold for the minimum word frequency count in determining the vocabulary is a crucial hyperparameter. Constructing vocabulary without a threshold value causes much noise in vocabulary. A significant part of words in a corpus has a small number of occurrences. This situation can also be understood by looking at the long tail of the rank-frequency graph of a corpus. In order to obtain successful word embeddings in terms of modelling a language, frequency threshold value should always be taken into consideration.

Especially in experiments of this study, word frequencies play a central role in Zipfian regularities. In the original model of the multimodal Gaussian embeddings deployed, the frequency threshold is chosen as 100. This value is continued to be used in the experiments shown in former chapters so far. However, to observe the behavior of word variances under different frequency thresholds, additional experiments are performed with two threshold parameters other than 100. One of them is chosen as 50 to be lower than 100. Similarly, another value is chosen as 200 to observe the effects of both increase and decrease in threshold values. Figs. (4.18, 4.19, 4.20, and 4.21) demonstrate that the Zipfian patterns of word variances for 50 and 200 threshold values, respectively. Changing the frequency threshold does not affect the rank of words, but chosen threshold determines the size of the corpora. If a word's occurrence is below the threshold, that word is not included in the training. In other words, those words are completely ignored in the training phase. Ignoring them changes the context of words in the corpus and affects the information that the model learns. Despite these effects of different thresholds, Zipfian patterns are preserved for both models. In particular, after thresholding with 50 and 200, the difference between the vocabulary sizes of the two corpora is approximately 200,000. This situation can be observed from the long tail of variances of words with low frequencies in Figs. (4.18). Nevertheless, Zipfian regularities are observed in both models. These results further validate the consistency of the relationship between word frequencies and word variances.

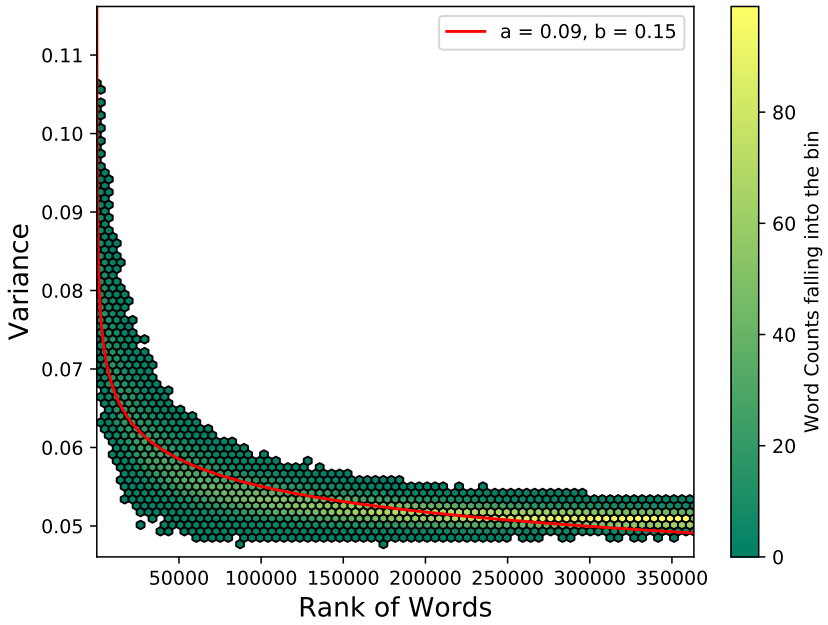


Figure 4.18: Average variances of Gaussian embeddings (frequency threshold = 50).

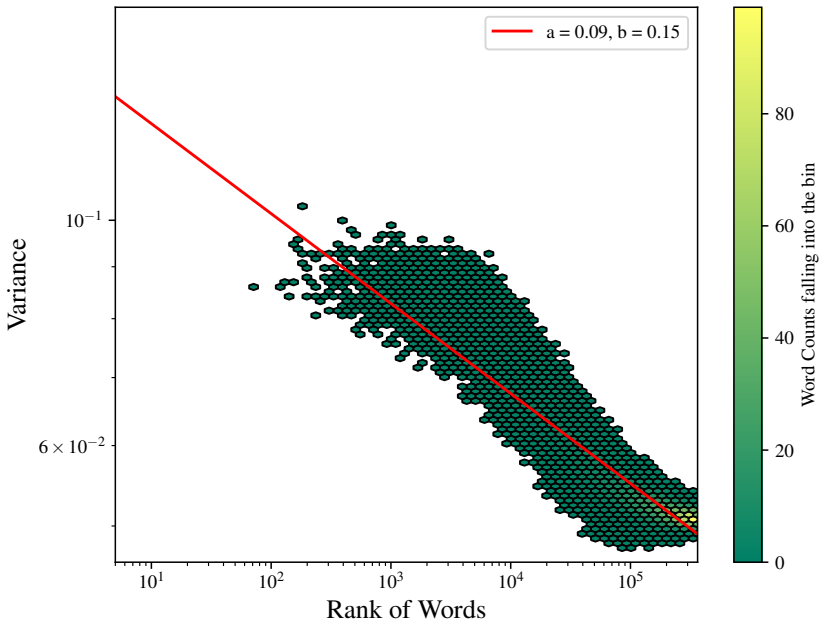


Figure 4.19: Average variances of Gaussian embeddings in logarithmic scale (frequency threshold = 50).

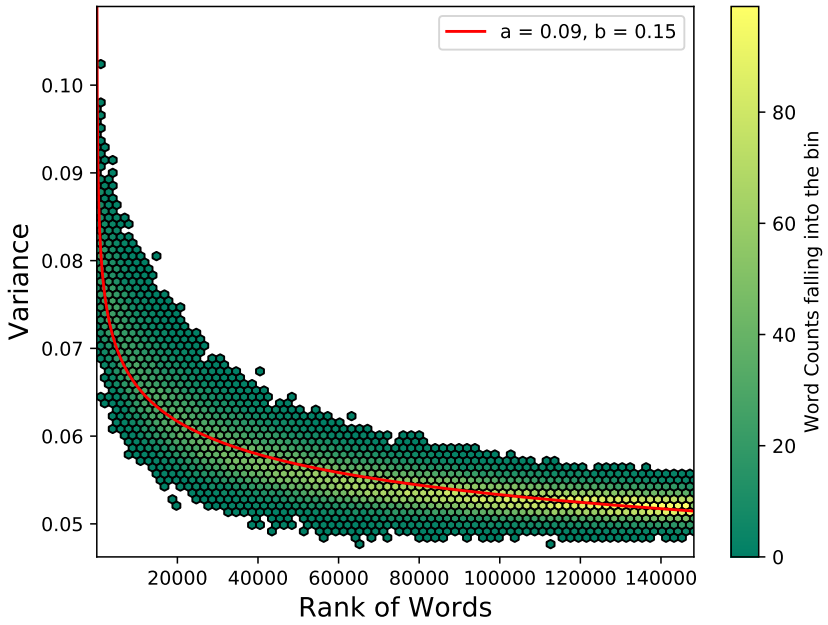


Figure 4.20: Average variances of Gaussian embeddings (frequency threshold = 200).

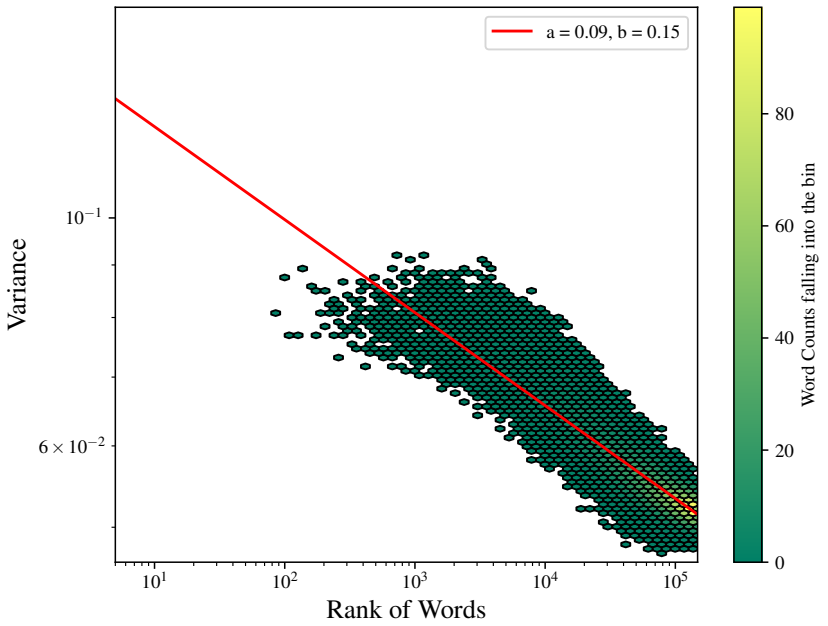


Figure 4.21: Average variances of Gaussian embeddings in logarithmic scale (frequency threshold = 200).

### 4.3.6 Results on the Corpora from Different Languages

One significant property of Zipf's law is that it is independent of language. Here, the word **independent** emphasizes the existence of Zipf's law in all-natural languages. Therefore, examining the behavior of word variances against different languages is a crucial experiment that will show word variances' compatibility of original Zipf's law and their consistency against different languages.

It is important to experiment with distinctive structures and features such as language families, grammatical structures, or the alphabet. In this regard, we study Turkish, German, Russian, and Spanish to enrich the demonstration of Zipfian regularities with various languages from different language families. Turkish is a member of the Altaic language family, while German, Russian and Spanish belong to the Indo-European family. We utilize [89] for organizing the Wikipedia dumps. The same experiments are repeated with the same pre-processing procedures. All non-alphanumeric characters and stop words are removed. Additionally, since Turkish is an agglutinative language, several words can be derived from the same stem. Therefore, we also applied stemming to Turkish corpus [2,3]. Token numbers and vocabulary sizes are given in Table 4.3. The experimental results belonging to different languages are presented in Figs.(4.22, 4.24, 4.25, 4.26, 4.27, 4.28, and 4.29) where Zipfian patterns are clearly seen.

This experiment provides further evidence that the Zipfian patterns of word variances can be observed across various languages belonging to different language families. Another critical aspect of these results is that word variances are compatible with Zipf's law's universality. As mentioned in previous chapters, Zipf's law of word frequencies holds for different languages. In this regard, word variances' attitudes toward different languages support the Zipfian characteristic of word variances. Despite different power parameters, Zipfian patterns maintain their existence in different languages like conventional Zipf's law.

As in Section 4.3.2, correlation scores of the models trained on corpora from

[2]www.cmpe.boun.edu.tr/tr/content/boun-nlp-morphological-analysis-system-turkish
[3]https://github.com/otuncelli/turkish-stemmer-python

Table 4.3: Number of tokens and vocabulary size.

| **Language** | **# of Tokens** | **Vocabulary Size** |
|---|---|---|
| English | 2,127,511,369 | 229,922 |
| German | 780,841,792 | 237,321 |
| Russian | 452,775,788 | 200,665 |
| Spanish | 587,115,302 | 122,452 |
| Turkish | 343,225,727 | 46,801 |

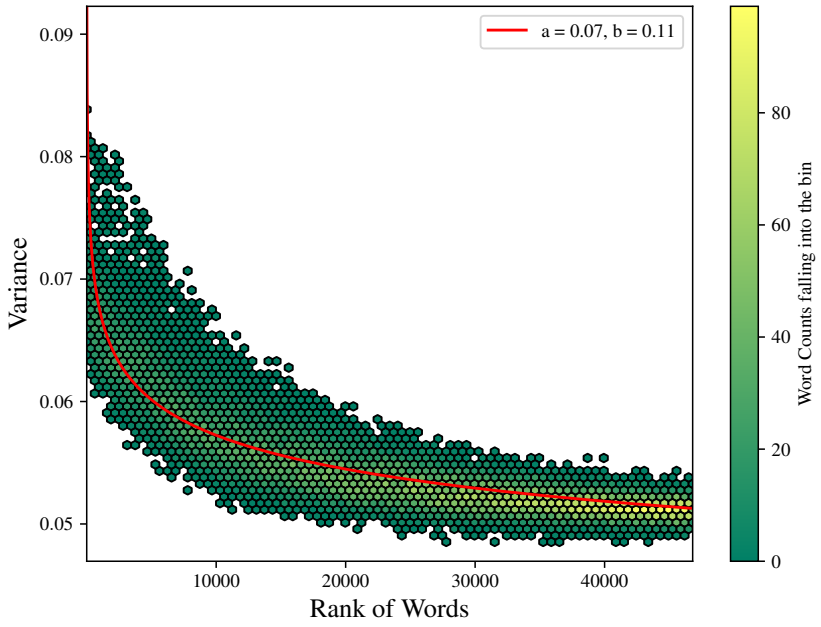


Figure 4.22: Average variances of Gaussian embeddings for the Turkish corpus.

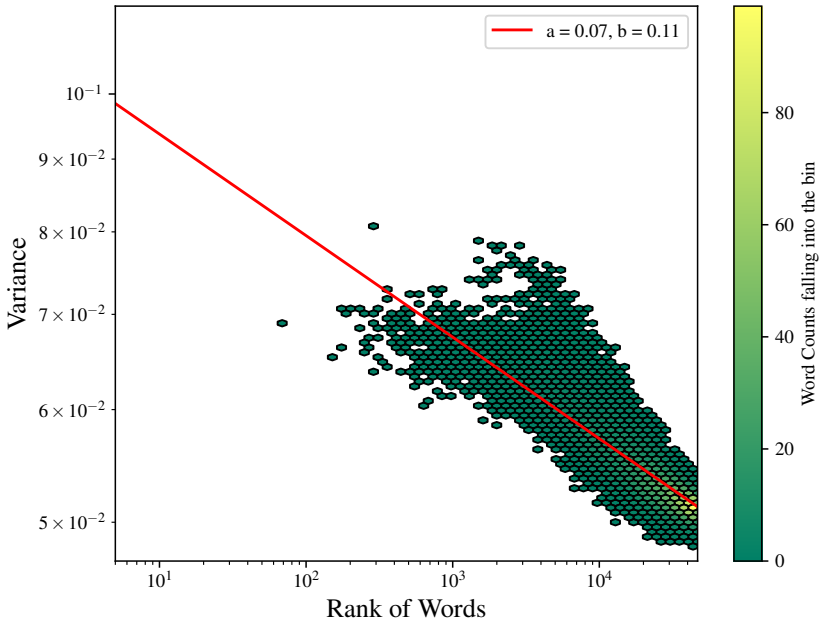


Figure 4.23: Average variances of Gaussian embeddings for the Turkish corpus in logarithmic scale.

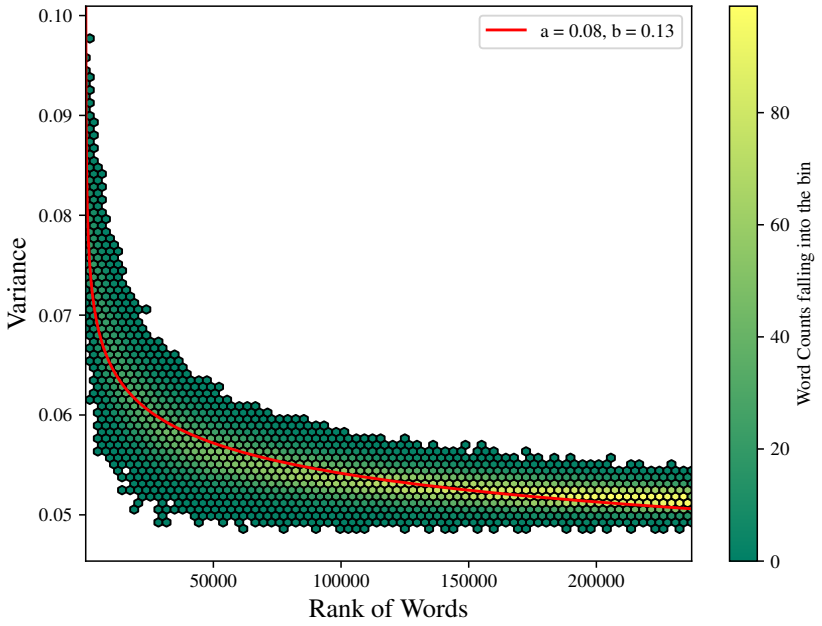


Figure 4.24: Average variances of Gaussian embeddings for the German corpus.

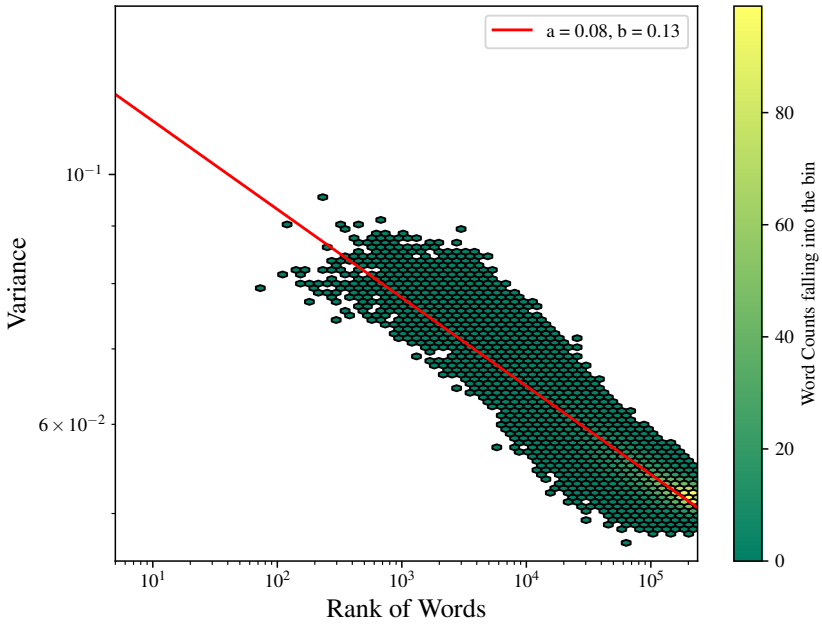


Figure 4.25: Average variances of Gaussian embeddings for the German corpus in logarithmic scale.

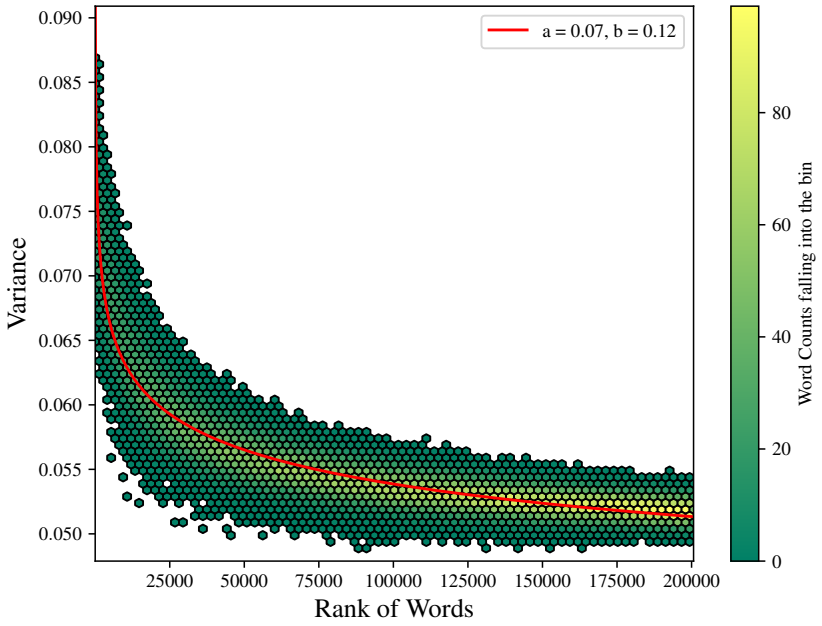


Figure 4.26: Average variances of Gaussian embeddings for the Russian corpus.

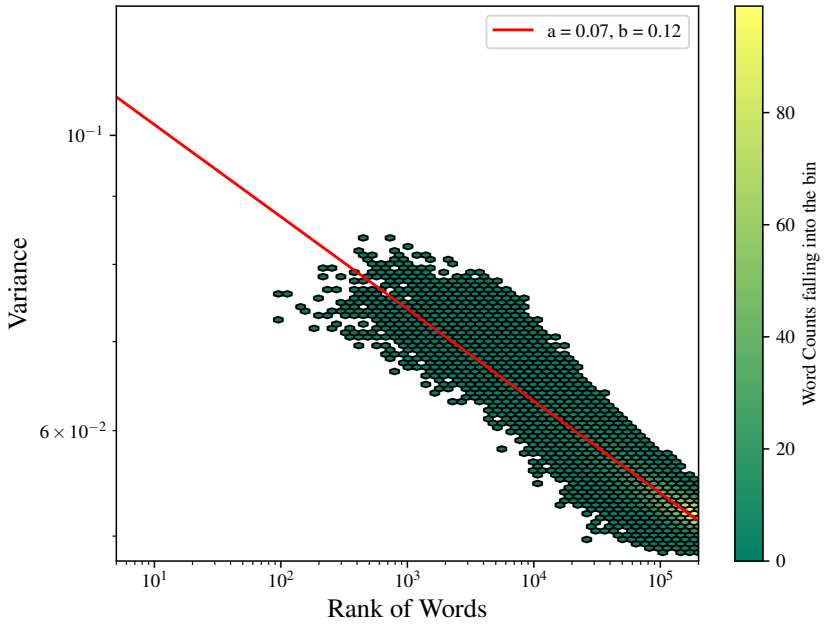


Figure 4.27: Average variances of Gaussian embeddings for the Russian corpus in logarithmic scale.

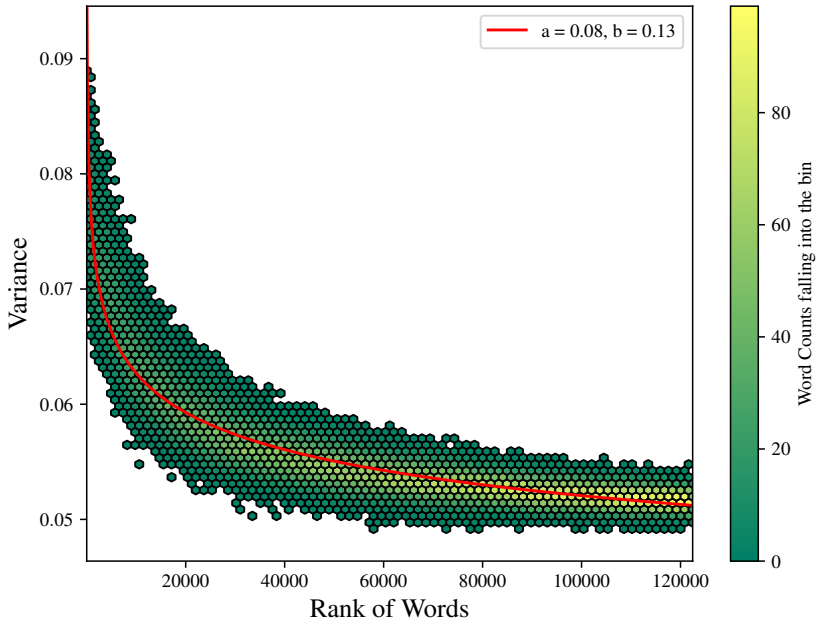


Figure 4.28: Average variances of Gaussian embeddings for the Spanish corpus.

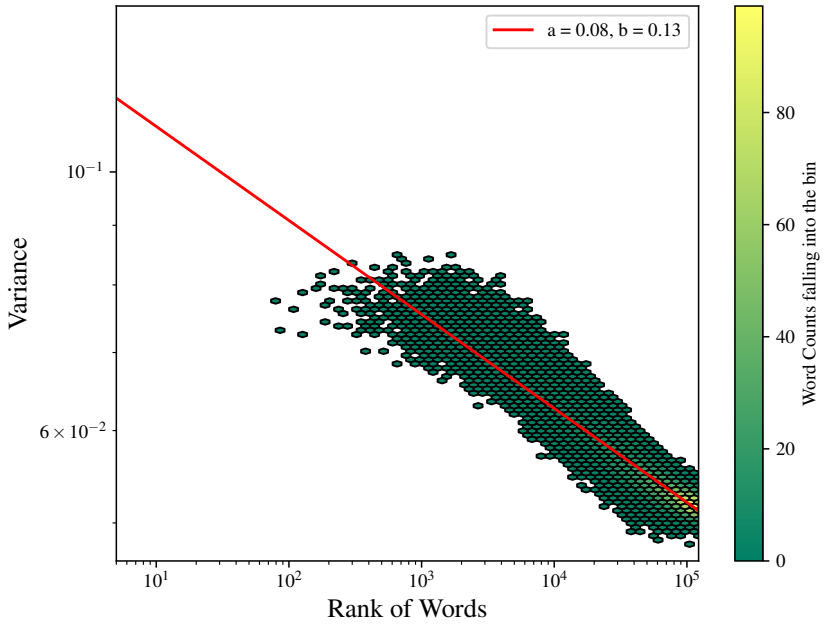


Figure 4.29: Average variances of Gaussian embeddings for the Spanish corpus in logarithmic scale.

different languages are calculated. The same correlation methods are utilized as in Section 4.3.2. Correlation results are given in Table 4.4. It can be seen that all languages provide a high correlation between frequency and variances.

Table 4.4: Frequency-variance correlation comparison of languages.

| **Language** | **Pearson** $r$ | **Spearman** $\rho$ | **Kendall** $\tau$ |
|---|---|---|---|
| German | 0.093 | 0.817 | 0.635 |
| Russian | 0.080 | 0.865 | 0.687 |
| Spanish | 0.075 | 0.878 | 0.704 |
| Turkish | 0.352 | 0.861 | 0.673 |

### 4.3.7 Entailment

In Section 4.2, a method for entailment tests is proposed to improve the performance of word embeddings. In this section, how the proposed method is implemented and evaluated will be described.

For the sake of comparison, studies employing word embeddings based on probability distributions are reimplemented [8, 78]. In experiments, two datasets are utilized: BBDS [76] and BLESS [90] as given in [78]. These datasets include word pairs that can have entailment relations or not. word2vec embeddings are used to calculate the cosine similarity between words, [13].

For each candidate word pair, an overall score is calculated as described in Section 4.2. The main purpose is to make the positive instances have a larger score. Then, a threshold is determined to distinguish positive instances from negative instances. Word pairs whose score is higher than the threshold are considered as positive instances and vice versa for negative instances. After calculating the overall scores of the proposed method and other proposed methods, the threshold is determined to obtain the best F1 score. Parameter $k$ in Eq. 4.4 is set to 20 and 40 for BBDS and BLESS datasets, respectively.

Results from the baseline cosine similarity of word2vec embeddings, the multimodal Gaussian (w2gm) [8], the Bayesian Skip-gram distribution (BSG) [78], and the proposed Zipfian-frequency-injected cosine similarity methods are all tabulated in Tables 4.5 and 4.6. For w2gm and BSG experiments, both KL-divergence and cosine similarity are used to be compatible with original implementations. While reporting their results, measurement with the highest performance is considered.

Table 4.5: 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 |

As can be observed from Tables 4.5 and 4.6, the proposed method offers improvements over both datasets in terms of F1 score. The entailment pairs of BLESS dataset are more detailed and complex compared to the BBDS dataset. Therefore, extracting the correct entailment relations from BLESS dataset is more

Table 4.6: 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 |

challenging. Additionally, the number of positive pairs and the number of negative pairs in the BBDS dataset are equal. In contrast, the ratio of positive and negative pairs is about 0.10 in the BLESS dataset. This is another contribution of BLESS dataset's difficulty in discriminating positive and negative samples from each other. By this experiment, we showed that capturing entailment relations can be improved by using word frequencies. Utilizing frequencies 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 reach higher F1 scores. Due to the difficulty of the BLESS dataset, this improvement can be slight compared to the BBDS dataset. Capturing lexical entailment relations is still a challenging and unsolved problem. The primary purpose of this experiment is to show the potential of using frequency regularities to reinforce the methods in the literature.

# Chapter 5

# Discussion and Conclusions

As a result of a trade-off between the sender and the receiver in natural languages which is named "the principle of least effort", words with both broad and narrow meanings emerged, [2]. In [54, 56, 91], the compromise between the sender and the 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 ranks. Contribution of this study comprises the first exhibition of Zipfian patterns between generality-specificity of words and their frequency ranks when polysemy is controlled. The Gaussian embedding models are employed 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 senses. 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 detect entailment.

Hierarchical relations between concepts can be compelling for 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 [92, 93, 94]. 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 this study are (1) demonstration of Zipfian regularities between the semantic breadth and frequency ranks, (2) the utilization of Gaussian embeddings in a quantitative study of Zipfian statistics and computational semantics. Gaussian embeddings are tools that are 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.

Here, one should note that two main factors promote the semantic uncertainty: direct polysemy and being a generic word. To better observe semantic breadth and frequency relations, the effect of polysemy should be eliminated. Gaussian embeddings with multimodal scheme can handle polysemy up to some point by assigning different Gaussian modes to different senses. In addition to this, experiments are performed on sense disambiguated corpora where the polysemy is eliminated. The results show that the variances of Gaussian embeddings only measure the semantic extents of words irrespective of the effects of polysemy. It is also showed that our results are valid for different languages and valid for special vocabularies such as Swadesh list and number words. Tests for showing the statistical significance of the presented results have also been performed.

To leverage obtained results, we also present 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.

To sum up, this computational study reveals the existence of new Zipfian patterns: more frequent words tend to be generic while less frequent ones tend to be specific. The 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. Acquired results related to the uncertainty of words and the usage of Guassian embeddings can also be very helpful for further quantitative studies of various semantic properties of words.

As a future works, Zipfian regularities in semantic breadth can be investigated in the scope of the basic-level categories [95]. Basic-level categories can be defined as categories that are used generally people choose to use over their superordinate and subordinate categories [96]. For example, most people consider a particular object as `chair`, not as `kitchen chair` or `furniture`. Taking basic-level categories into account can help to understand generality-specificity relations in Zipfian patterns. It is also beneficial to evaluate modeling performance of multimodal Gaussian embeddings. Moreover, diachronic changes of variances can be examined by using corpora compiled from text samples belonging to different time periods. [97] examines the words in a diachronic framework and study how meaning and embeddings in vector spaces change over time. In fact, [98] investigates semantic changes over different timelines via multimodal Gaussian embeddings. The evolution of language over time in the probabilistic framework is studied as well, [99, 100]. 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.

# Bibliography


[1] B. Casas, A. Hernández-Fernández, N. Català, R. Ferrer-i-Cancho, and J. Baixeries, "Polysemy and brevity versus frequency in language," *Computer Speech & Language*, vol. 58, pp. 19–50, 2019.

[2] G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, 1949.

[3] G. K. Zipf, "The meaning-frequency relationship of words," *The Journal of General Psychology*, vol. 33, no. 2, pp. 251–256, 1945.

[4] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proceedings of the IRE*, vol. 40, no. 9, pp. 1098–1101, 1952.

[5] J. R. Firth, "A synopsis of linguistic theory, 1930-1955," *Studies in linguistic analysis*, 1957.

[6] R. Ferrer-i-Cancho and M. S. Vitevitch, "The origins of Zipf's meaning-frequency law," *Journal of the Association for Information Science and Technology*, vol. 69, no. 11, pp. 1369–1379, 2018.

[7] D. Y. Manin, "Zipf's law and avoidance of excessive synonymy," *Cognitive Science*, vol. 32, no. 7, pp. 1075–1098, 2008.

[8] B. Athiwaratkun and A. Wilson, "Multimodal word distributions," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Vancouver, Canada), pp. 1645–1656, Association for Computational Linguistics, July 2017.

[9] F. Şahinuç and A. Koç, "Zipfian regularities in "non-point" word representations," *Information Processing & Management*, vol. 58, no. 3, p. 102493, 2021.

[10] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, Mar. 2003.

[11] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, (New York, NY, USA), pp. 160–167, Association for Computing Machinery, 2008.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, (Scottsdale, Arizona, USA), 2013.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, (Red Hook, NY, USA), pp. 3111–3119, Curran Associates Inc., 2013.

[14] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1532–1543, Association for Computational Linguistics, Oct. 2014.

[15] L. K. Şenel, I. Utlu, V. Yücesoy, A. Koç, and T. Çukur, "Semantic structure and interpretability of word embeddings," *IEEE/ACM Transactions Audio, Speech and Language Processing*, vol. 26, no. 10, pp. 1769–1779, 2018.

[16] H. Luo, Z. Liu, H. Luan, and M. Sun, "Online learning of interpretable word embeddings," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, (Lisbon, Portugal), pp. 1687–1692, Association for Computational Linguistics, Sept. 2015.

[17] M. Faruqui, Y. Tsvetkov, D. Yogatama, C. Dyer, and N. A. Smith, "Sparse overcomplete word vector representations," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Beijing, China), pp. 1491–1500, Association for Computational Linguistics, July 2015.

[18] L. K. Şenel, I. Utlu, F. Şahinuç, H. M. Ozaktas, and A. Koç, "Imparting interpretability to word embeddings while preserving semantic structure," *Natural Language Engineering*, pp. 1–26, 2020.

[19] L. K. Şenel, V. Yücesoy, A. Koç, and T. Çukur, "Measuring cross-lingual semantic similarity across European languages," in *2017 40th International Conference on Telecommunications and Signal Processing (TSP)*, pp. 359–363, 2017.

[20] L. K. Şenel, I. Utlu, V. Yücesoy, A. Koç, and T. Çukur, "Generating semantic similarity atlas for natural languages," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 795–799, 2018.

[21] V. Yücesoy and A. Koç, "Co-occurrence weight selection in generation of word embeddings for low resource languages," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 18, no. 3, 2019.

[22] O. Sulea, M. Zampieri, S. Malmasi, M. Vela, L. P. Dinu, and J. van Genabith, "Exploring the use of text classification in the legal domain," in *Proceedings of the Second Workshop on Automated Semantic Analysis of Information in Legal Texts co-located with the 16th International Conference on Artificial Intelligence and Law (ICAIL 2017)*, vol. 2143 of *CEUR Workshop Proceedings*, (London, UK), CEUR-WS.org, 2017.

[23] D. M. Katz, M. J. Bommarito, II, and J. Blackman, "A general approach for predicting the behavior of the Supreme Court of the United States," *PLOS ONE*, vol. 12, no. 4, pp. 1–18, 2017.

[24] E. Mumcuoğlu, C. E. Öztürk, H. M. Ozaktas, and A. Koç, "Natural language processing in law: Prediction of outcomes in the higher courts of

Turkey," *Information Processing & Management*, vol. 58, no. 5, p. 102684, 2021.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[26] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[29] T. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, Curran Associates, Inc., 2020.

[30] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, (San Diego, CA, USA), 2015.

[31] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Lisbon, Portugal), pp. 1412–1421, Association for Computational Linguistics, Sept. 2015.

[32] K. Erk, "Supporting inferences in semantic space: representing words as regions," in *Proceedings of the Eight International Conference on Computational Semantics*, (Tilburg, The Netherlands), pp. 104–115, Association for Computational Linguistics, Jan. 2009.

[33] K. Erk, "Representing words as regions in vector space," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, (Boulder, Colorado), pp. 57–65, Association for Computational Linguistics, June 2009.

[34] L. Vilnis and A. McCallum, "Word representations via Gaussian embedding," in *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, (San Diego, CA, USA), 2015.

[35] X. Chen, X. Qiu, J. Jiang, and X. Huang, "Gaussian mixture embeddings for multiple word prototypes," *arXiv preprint arXiv:1511.06246*, 2015.

[36] T. Jebara, R. Kondor, and A. Howard, "Probability product kernels," *Journal of Machine Learning Research*, vol. 5, pp. 819–844, Dec. 2004.

[37] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[38] G. K. Zipf, *The Psycho-biology of Language: An Introduction to Dynamic Philology*. The MIT paperback series, Houghton Mifflin, 1935.

[39] M. E. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Physics*, vol. 46, no. 5, pp. 323–351, 2005.

[40] P. Grzybek, *Contributions to the science of text and language: Word length studies and related issues*. Springer Science & Business Media, 2006.

[41] R. Ferrer-i-Cancho and R. V. Solé, "Zipf's law and random texts," *Advances in Complex Systems*, vol. 5, no. 01, pp. 1–6, 2002.

[42] L. Debowski, "Zipf's law against the text size: A half-rational model," *Glottometrics*, vol. 4, pp. 49–60, 2002.

[43] R. Ferrer-i-Cancho and B. Elvevåg, "Random texts do not exhibit the real Zipf's law-like rank distribution," *PLOS ONE*, vol. 5, no. 3, pp. 1–10, 2010.

[44] R. H. Baayen, *Word frequency distributions*. Springer Science & Business Media, 2002.

[45] M. Gerlach and E. G. Altmann, "Scaling laws and fluctuations in the statistics of word frequencies," *New Journal of Physics*, vol. 16, no. 11, p. 113010, 2014.

[46] E. G. Altmann and M. Gerlach, *Statistical Laws in Linguistics*, pp. 7–26. Springer International Publishing, 2016.

[47] B. Mandelbrot, "An informational theory of the statistical structure of language," *Communication Theory*, pp. 486–502, 1953.

[48] B. Mandelbrot, "On the theory of word frequencies and on related Markovian models of discourse," *Structure of Language and Its Mathematical Aspects*, vol. 12, pp. 190–219, 1961.

[49] Y.-S. Chen and F. F. Leimkuhler, "Analysis of Zipf's law: An index approach," *Information Processing & Management*, vol. 23, no. 3, pp. 171–182, 1987.

[50] S. Shan, "On the generalized Zipf distribution. Part I," *Information Processing & Management*, vol. 41, no. 6, pp. 1369–1386, 2005.

[51] M. Gerlach and E. G. Altmann, "Stochastic model for the vocabulary growth in natural languages," *Physical Review X*, vol. 3, p. 021006, May 2013.

[52] H. S. Heaps, *Information Retrieval: Computational and Theoretical Aspects*. USA: Academic Press, Inc., 1978.

[53] A. Koplenig, "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*, vol. 14, no. 1, pp. 1–34, 2018.

[54] R. Ferrer-i-Cancho, "Decoding least effort and scaling in signal frequency distributions," *Physica A: Statistical Mechanics and its Applications*, vol. 345, no. 1-2, pp. 275–284, 2005.

[55] R. Ferrer-i-Cancho, "Hidden communication aspects in the exponent of Zipf's law," *Glottometrics*, vol. 11, pp. 98–119, 2005.

[56] R. Ferrer-i-Cancho, "Zipf's law from a communicative phase transition," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 47, no. 3, pp. 449–457, 2005.

[57] G. A. Miller, "WordNet: A lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[58] U. Strauss, P. Grzybek, and G. Altmann, *Word Length and Word Frequency*, pp. 277–294. Dordrecht: Springer Netherlands, 2007.

[59] R. Ferrer-i Cancho, A. Hernández-Fernández, D. Lusseau, G. Agoramoorthy, M. J. Hsu, and S. Semple, "Compression as a universal principle of animal behavior," *Cognitive Science*, vol. 37, no. 8, pp. 1565–1578, 2013.

[60] S. T. Piantadosi, H. Tily, and E. Gibson, "Word lengths are optimized for efficient communication," *Proceedings of the National Academy of Sciences*, vol. 108, no. 9, pp. 3526–3529, 2011.

[61] K. Okuyama, M. Takayasu, and H. Takayasu, "Zipf's law in income distribution of companies," *Physica A: Statistical Mechanics and its Applications*, vol. 269, no. 1, pp. 125–131, 1999.

[62] K. T. Soo, "Zipf's law for cities: A cross-country investigation," *Regional Science and Urban Economics*, vol. 35, no. 3, pp. 239–263, 2005.

[63] L. A. Adamic and B. A. Huberman, "Zipf's law and the internet," *Glottometrics*, vol. 3, no. 1, pp. 143–150, 2002.

[64] I. G. Torre, B. Luque, L. Lacasa, J. Luque, and A. Hernández-Fernández, "Emergence of linguistic laws in human voice," *Scientific reports*, vol. 7, no. 1, pp. 1–10, 2017.

[65] M. Gerlach and E. G. Altmann, "Testing statistical laws in complex systems," *Physical Review Letters*, vol. 122, p. 168301, Apr 2019.

[66] L. Gao, G. Zhou, J. Luo, and Y. Huang, "Word embedding with Zipf's context," *IEEE Access*, vol. 7, pp. 168934–168943, 2019.

[67] W. Li, "Random texts exhibit Zipf's-law-like word frequency distribution," *IEEE Transactions on Information Theory*, vol. 38, no. 6, pp. 1842–1845, 1992.

[68] D. Wang, H. Cheng, P. Wang, X. Huang, and G. Jian, "Zipf's law in passwords," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 11, pp. 2776–2791, 2017.

[69] H. Zhang, "Discovering power laws in computer programs," *Information Processing & Management*, vol. 45, no. 4, pp. 477–483, 2009.

[70] S. T. Piantadosi, "Zipf's word frequency law in natural language: A critical review and future directions," *Psychonomic Bulletin & Review*, vol. 21, no. 5, pp. 1112–1130, 2014.

[71] M. Swadesh, "Salish internal relationships," *International Journal of American Linguistics*, vol. 16, no. 4, pp. 157–167, 1950.

[72] B. Everitt and D. Hand, *Finite mixture distributions*. Monographs on Applied Probability and Statistics, Chapman and Hall, 1981.

[73] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *The 14th International Conference on Computational Linguistics*, 1992.

[74] M. Geffet and I. Dagan, "The distributional inclusion hypotheses and lexical entailment," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, (Ann Arbor, Michigan), pp. 107–114, Association for Computational Linguistics, June 2005.

[75] L. Kotlerman, I. Dagan, I. Szpektor, and M. Zhitomirsky-Geffet, "Directional distributional similarity for lexical inference," *Natural Language Engineering*, vol. 16, no. 4, p. 359–389, 2010.

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

[77] M. Rei, D. Gerz, and I. Vulić, "Scoring lexical entailment with a supervised directional similarity network," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (Melbourne, Australia), pp. 638–643, Association for Computational Linguistics, July 2018.

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

[79] J. Weeds, D. Weir, and D. McCarthy, "Characterising measures of lexical distributional similarity," in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, (Geneva, Switzerland), pp. 1015–1021, COLING, 2004.

[80] M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta, "The WaCky wide web: A collection of very large linguistically processed web-crawled corpora," *Language Resources and Evaluation*, vol. 43, no. 3, pp. 209–226, 2009.

[81] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. 61, pp. 2121–2159, 2011.

[82] R. Ferrer-i-Cancho, "The variation of zipf's law in human language," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 44, no. 2, pp. 249–257, 2005.

[83] R. Ferrer-i-Cancho and R. V. Solé, "Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited," *Journal of Quantitative Linguistics*, vol. 8, no. 3, pp. 165–173, 2001.

[84] M. Lesk, "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*, (New York, NY, USA), pp. 24–26, Association for Computing Machinery, 1986.

[85] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[86] G. A. Miller, M. Chodorow, S. Landes, C. Leacock, and R. G. Thomas, "Using a semantic concordance for sense identification," in *Proceedings of the Workshop on Human Language Technology*, HLT '94, (Plainsboro, NJ, USA), pp. 240–243, Association for Computational Linguistics, 1994.

[87] J. C. Mallery, "Thinking about foreign policy: Finding an appropriate role for artificially intelligent computers," 1988.

[88] W. Conover, *Practical nonparametric statistics*. New York: Wiley, 3. ed ed., 1999.

[89] G. Attardi, "WikiExtractor." https://github.com/attardi/wikiextractor, 2015.

[90] M. Baroni and A. Lenci, "How we BLESSed distributional semantic evaluation," in *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, (Edinburgh, UK), pp. 1–10, Association for Computational Linguistics, July 2011.

[91] R. Ferrer-i-Cancho and R. V. Solé, "Least effort and the origins of scaling in human language," *Proceedings of the National Academy of Sciences*, vol. 100, no. 3, pp. 788–791, 2003.

[92] Y. Jiang, X. Zhang, Y. Tang, and R. Nie, "Feature-based approaches to semantic similarity assessment of concepts using Wikipedia," *Information Processing & Management*, vol. 51, no. 3, pp. 215–234, 2015.

[93] Y. Jiang, W. Bai, X. Zhang, and J. Hu, "Wikipedia-based information content and semantic similarity computation," *Information Processing & Management*, vol. 53, no. 1, pp. 248–265, 2017.

[94] M. J. Hussain, S. H. Wasti, G. Huang, L. Wei, Y. Jiang, and Y. Tang, "An approach for measuring semantic similarity between Wikipedia concepts using multiple inheritances," *Information Processing & Management*, vol. 57, no. 3, p. 102188, 2020.

[95] L. Hajibayova, "Basic-level categories: A review," *Journal of Information Science*, vol. 39, no. 5, pp. 676–687, 2013.

[96] J. E. Corter and M. A. Gluck, "Explaining basic categories: Feature predictability and information.," *Psychological bulletin*, vol. 111, no. 2, p. 291, 1992.

[97] W. L. Hamilton, J. Leskovec, and D. Jurafsky, "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)*, (Berlin, Germany), pp. 1489–1501, Association for Computational Linguistics, 2016.

[98] A. Yuksel, B. Ugurlu, and A. Koç, "Semantic change detection with Gaussian word embeddings," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. In Press, 2021.

[99] R. Bamler and S. Mandt, "Dynamic word embeddings," in *Proceedings of the 34th International Conference on Machine Learning*, (International Convention Centre, Sydney, Australia), pp. 380–389, PMLR, 06–11 Aug 2017.

[100] M. Rudolph and D. Blei, "Dynamic embeddings for language evolution," in *Proceedings of the 2018 World Wide Web Conference*, (Republic and Canton of Geneva, CHE), pp. 1003–1011, International World Wide Web Conferences Steering Committee, 2018.