A New Formulation of Zipf's Meaning-Frequency Law through Contextual Diversity

Ryo Nagata
Konan University / 8-9-1 Okamoto, Kobe, Hyogo 658-8501, Japan
RIKEN / 2-1 Hirosawa, Wako, Saitama 351-0198, Japan
nagata-acl2025 @ ml.hyogo-u.ac.jp.

Kumiko Tanaka-Ishii
Waseda University / 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
kumiko@waseda.jp

Abstract

This paper proposes formulating *Zipf's meaning-frequency law*, the power law between word frequency and the number of meanings, as a relationship between word frequency and contextual diversity. The proposed formulation quantifies meaning counts as contextual diversity, which is based on the directions of contextualized word vectors obtained from a Language Model (LM). This formulation gives a new interpretation to the law and also enables us to examine it for a wider variety of words and corpora than previous studies have explored. In addition, this paper shows that the law becomes unobservable when the size of the LM used is small and that autoregressive LMs require much more parameters than masked LMs to be able to observe the law.

1 Introduction

This paper proposes a new way to formulate *Zipf's meaning-frequency law* (Zipf, 1945), known as a relationship between word frequency and the number of word meanings, by contextual diversity measured based on contextualized word vectors[1]. This formulation extends the law both theoretically and practically. Theoretically, it allows us to correlate word frequency to word meaning via contextual diversity, relating Zipf's meaning-frequency law to Harris's (1954) distributional hypothesis. Practically, it enables us to show that the law holds for a wider variety of words and corpora than previous studies have shown.

The meaning-frequency law states that the more frequent a word is, the more meanings it has, which follows a power law. Formally, it is denoted as:

$$m \propto f^{\alpha}, \tag{1}$$

where f and m denote the frequency of a word and the number of meanings it has, respectively. Most previous studies examine the law by regression analysis to Eq. (1) where f and m are respectively obtained from a corpus and from a dictionary such as WordNet (Fellbaum, 1998).

The use of a dictionary causes several problems and limitations in the previous studies as Sect. 2 will describe in detail. It is a difficult task to determine the number of word meanings in the first place; the number of meanings registered for a word can vary greatly from dictionary to dictionary. Conclusions may differ with different meaning counts. Besides, not all the meanings defined in a dictionary necessarily appear in a given corpus, and vice versa. For these reasons (and others as described in Sect. 2), the previous studies use a limited vocabulary list excluding function words, high-frequency words, and inflected/conjugated forms. Therefore, it is not yet known if the law holds for these excluded words.

To overcome these problems and limitations, this paper proposes formulating the meaning-frequency law in a completely different way without using a dictionary. Specifically, the proposed method defines m in Eq. (1) as contextual diversity by using contextualized word vectors obtained from a Language Model (LM). In other words, this formulation measures the quantity related to meaning counts via contextual diversity. In doing so, the proposed method considers all words appearing in a corpus except for infrequent words. It is applicable to a much wider variety of language data than before, including historical and learner corpora. This is the first work to investigate whether the law holds for such corpora.

This paper also explores the relationship between the meaning-frequency law and the lexical capability of LMs with this new formulation. It shows that the value of m, or contextual diversity, measured by the proposed method deviates from

[1]The source codes to reproduce the results in this work are available at https://github.com/nagata-github/meaning_frequency_law_via_contextual_diversity.git

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15323–15335

the meaning-frequency law when the sizes of the LMs used are small. In addition, the comparison between Masked LMs (MLMs) and autoregressive LMs, which predict the next token, reveals that the latter require much more parameters (14 times more in the experiment) to be able to observe the law. From these findings, this paper proposes using the newly-formulated meaning-frequency law as a sanity check for the lexical capability of LMs.

2 Related Work

All previous studies view the meaning-frequency law as a relationship between word frequency and the number of word meanings. This has naturally led researchers in the domain to use meaning counts obtained from a dictionary in their investigations. Examples include Zipf (1945) himself, Edmonds (2005) (English), Ilgen and Karaoglan (2007) (Turkish), Casas et al. (2019) (English, Spanish, and Dutch), Bond et al. (2019) (English, Polish, Spanish, French, Portuguese, Japanese, Chinese, and Indonesian), and Hernández-Fernández et al. (2016) (child language).

Even putting aside the difficulty in determining the number of word meanings, the use of a dictionary limits the previous studies to targeting only the base forms of words. Furthermore, previous studies exclude high-frequency words and function words because meaning counts are not available for these words in the widely-used dictionary, WordNet. As a result, the previous studies only show that the law holds for base forms with limited vocabulary.

Even if meaning counts are available, it is not at all straightforward how to treat words with the same meaning count, but with different distributions. For example, it is questionable whether to treat equally a word with one of its meanings occurring 99% of the time and another word with a uniform distribution of its meanings.

As another approach to examining the law, Ilgen and Karaoglan (2007) and Bond et al. (2019) annotate words in a corpus with their meanings to obtain meaning counts. Even in this approach, word meaning sets are determined based on a dictionary, and thus this approach also suffers from problems stemming from the use of a dictionary. Furthermore, it is not an easy task to annotate words with their meanings and, consequently, it would be difficult to increase the size of the investigation. As the meaning-frequency law is a power low, it is crucial to use data of a size large enough for log-scale.

Our new formulation explained in the next section provides a new view of the law as the relationship between word frequency and contextual diversity. It naturally overcomes all problems and limitations stemming from the use of a dictionary.

3 New Formulation of Meaning Frequency Law

3.1 Measuring Meaning Counts through Directions of Word Vectors

Unlike previous studies where m in Eq. (1) is specified to be the number of meanings listed in a dictionary, the proposed method assumes that every single usage of a word in a different context has a different meaning. It further assumes that the degree of the difference can be continuously measured as the difference of contexts through contextualized word vectors (simply, word vectors, hereafter). These assumptions reflect Harris's (1954) views:

> In other words, difference of meaning correlates with difference of distribution.

and

> ..., the amount of meaning difference corresponding roughly to the amount of difference in their environments.

To be precise, the difference is measured based on the angle between the vectors of the two words in question. It is natural to do so considering the convention that the semantic similarity between two words is measured by the cosine similarity of their corresponding word vectors (i.e., ignoring the norm). Then, the variation of the directions of vectors for a word type can be regarded as the variation of meanings that it has.

In this study, the variation of vector directions is quantified through the von Mises-Fisher distribution (Banerjee et al., 2005), a probability distribution of the random d-dimensional unit vector $\mathbf{x}$; in this study, $\mathbf{x}$ corresponds to a word vector of the word type in question. It is defined as:

$$f(\mathbf{x}; \boldsymbol{\mu}, \kappa) \propto \exp\left(\kappa \boldsymbol{\mu}^{\mathsf{T}} \mathbf{x}\right), \qquad (2)$$

where $\boldsymbol{\mu}$ ($\|\boldsymbol{\mu}\| = 1$) and κ ($\kappa \geq 0$) are parameters called *mean direction* and *concentration*, respectively. It assumes that the unit vector $\mathbf{x}$ distributes on the $(d-1)$-sphere around the mean direction

$\boldsymbol{\mu}$ with the concentration κ. In other words, κ reflects the degree of concentration of vector directions; an intuitive interpretation of κ is described in Appendix A.1. Nagata et al. (2023) show that κ is effective in lexical semantic change detection. Similarly this work needs to estimate the quantity related to meaning counts.

While κ is the degree of concentration of the distribution, the meaning-frequency law is about meaning diversity (and also contextual diversity in this paper). Therefore, we consider the reciprocal:

$$v \equiv 1/\kappa \tag{3}$$

as a measure of contextual diversity in this paper.

To calculate v, we need to estimate κ. Banerjee et al. (2005) show that a simple approximate solution of its maximum likelihood estimate is:

$$\kappa \approx \frac{l(d - l^2)}{1 - l^2}, \tag{4}$$

where d and l denote the dimension of the unit vector (i.e., $\boldsymbol{\mu}$ and $\mathbf{x}$ in Eq. (2)) and the norm of the mean vector of $\mathbf{x}$ for a word type, respectively.

Now, we formulate the meaning-frequency law as:

$$v \propto f^{\alpha}. \tag{5}$$

Eq. (5) states that the more frequent a word is, the more contextual diversity it has and that the relationship follows a power law.

3.2 Overall Procedure for Examining Meaning-Frequency Law

The overall procedure for examination follows Bond et al.(2019). They (and also most previous studies) examine the law by regression analysis where the explanatory and dependent variables are word frequency and meaning counts, respectively. To be able to apply linear regression to this problem, Eq. (5) is turned into:

$$\log(v) = \alpha \log(f) + c, \tag{6}$$

by taking the log of both sides where c is a constant depending on the size of the corpus in question. Most previous studies use $\alpha > 0$ as a criterion of whether the meaning-frequency law holds or not, together with the determination coefficient R^2 as a measure of model fit. Bond et al. (2019) also use the t-Student test for the slope coefficient non-zeroness.

This paper adopts this way to examine the determination coefficient R^2 and the t-Student test for the slope coefficient non-zeroness. It also provides the scatter plots for the obtained f and v with their regression lines.

Previous studies use averaged values of f and m for regression. To calculate the averages, words are divided into a certain number of bins according to their frequency ranks. For example, Bond et al. (2019) construct bins of the range λ such that a word with ith frequency rank is fitted to the jth bin if and only if the following inequalities are fulfilled: $\lambda(j-1)+1 \leq i \leq \lambda j$ where $j = 1, 2, 3, \cdots \text{round}(\frac{n}{\lambda})$, λ is a rank range (also, bin size), and n is a maximum rank[2]. Throughout this paper, $\lambda = 100$ is used. Bond et al. (2019) show more results for this value of λ than the others.

To summarize, the overall procedure consists of the following eight steps:

(1) Count frequency f of every word w in the given corpus

(2) Obtain word vectors for every word w appearing in it

(3) Normalize the obtained vectors so that their norm is one

(4) For each word type, calculate the mean vector of the word vectors and its norm

(5) Calculate $v = 1/\kappa$ using Eq. (4)

(6) Average f and v over each bin

(7) Fit the averaged values of f and v to Eq. (6)

(8) Examine if the meaning-frequency law holds or not with the determination coefficient R^2 and the slope coefficient α.

4 Examination with Canonical Language Data

We now examine the meaning-frequency law with the new formulation of the meaning-frequency law. We use two English corpora: the British National Corpus (BNC)[3] and the 2000s subcorpus in the cleaned version (Alatrash et al., 2020) of Corpus of Historical American English

[2]To be precise, in Bond et al.'s (2019) study, the first 100 most frequent words are excluded and thus they use $100 + \lambda(j-1)+1 \leq i \leq 100 + \lambda j$, instead.

[3]BNC Consortium, The British National Corpus, XML Edition, 2007, Oxford Text Archive, http://hdl.handle.net/20.500.14106/2554. The data were used following the license: http://www.natcorp.ox.ac.uk/docs/licence.html

(CCOHA) (Davies, 2012); and two Japanese corpora: Aozorabunko Corpus[4], a collection of Japanese literature, and Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014). Additionally, we use 27 corpora in 24 languages from five language families from the Bible corpus (Christodouloupoulos and Steedman, 2014). The details of all corpora are shown in Appendix A.2.

We use BERT (Devlin et al., 2019) models to obtain word vectors considering that researchers such as Laicher et al. (2021) and Aida and Bollegala (2023) have shown that BERT models are effective in meaning change detection. Specifically, we use the hidden state of the final layer of BERT (`bert-large-uncased`[5], `cl-tohoku/bert-large-japanese-v2`[6], and `bert-base-multilingual-uncased`[7] for the English, Japanese, and Bible corpora, respectively) as word vectors; we use the uncased models for English, which will be denoted without *-uncased* (e.g., *bert-large*), hereafter. We target words satisfying the following two conditions: (i) its frequency count is more than 100, and (ii) its frequency rank is higher than 20,000. For words split into multiple sub-words, we exclude their middle and final sub-words from the investigation[8].

Fig. 1 and Fig.2 respectively show the results for the English and Japanese corpora; the horizontal and vertical axes respectively correspond to the logs of the frequency f and the contextual diversity v; the line graphs correspond to the regression lines fitted to the scatter plots.

The plots in Fig. 1 and Fig. 2 fit Eq. (6) well, exhibiting a high determination coefficient R^2 in all cases; all slope coefficients are statistically significant ($p < 0.01$, t-Student test). These results show that the meaning-frequency law holds even if meaning counts are measured through contextual diversity. The meaning-frequency law, then, can

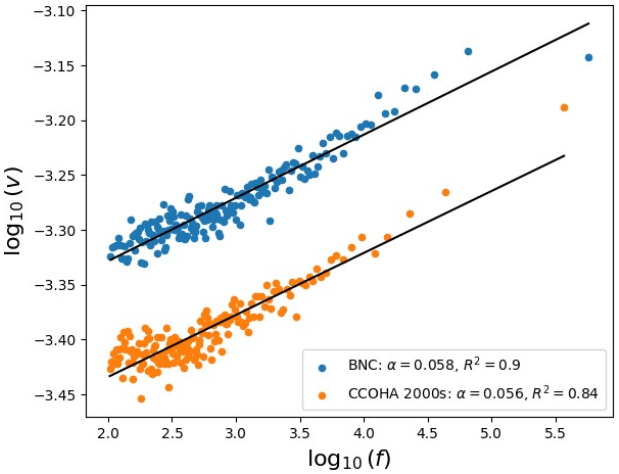

Figure 1: Relationship between word frequency f and word meaning measured as contextual diversity v in English corpora. LM: `bert-large`.

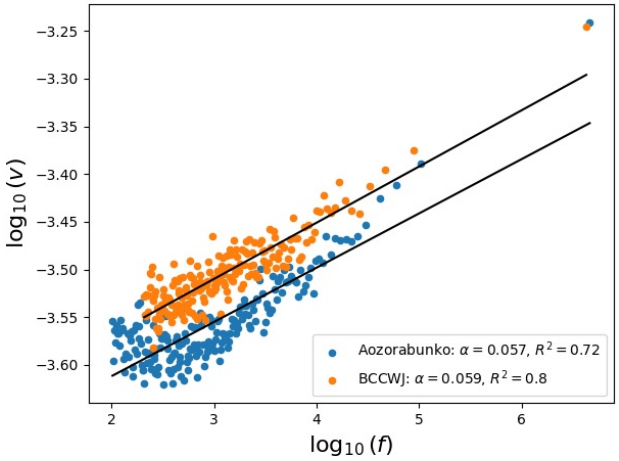

Figure 2: Relationship between word frequency f and word meaning measured as contextual diversity v in Japanese corpora. LM: `cl-tohoku/bert-large-japanese-v2`.

be restated as the context-frequency law: the more frequent a word is, the more contextual diversity it has, following a power law. Although one might predict this from Zipf's meaning-frequency law and Harris's distributional hypothesis, the contribution of this study is that it provides a formulation of this idea, and also empirically shows that word frequency and contextual diversity follow a power law.

Additionally, Table 5 in Appendix A.3 shows the law holds in the 27 Bible sub-corpora; the average R^2 and slope coefficient are 0.85 and 0.043, respectively. These results suggest that the meaning-frequency law might be universal.

The proposed method has a practical advantage

[4] https://github.com/aozorahack/aozorabunko_text, accessed in 3.12.2023. The data were used following the license: https://www.aozora.gr.jp/guide/kijyunn.html

[5] https://huggingface.co/docs/transformers/model_doc/bert, Apache license 2.0.

[6] https://huggingface.co/cl-tohoku/bert-large-japanese-v2, Apache license 2.0.

[7] https://huggingface.co/google-bert/bert-base-multilingual-uncased, Apache license 2.0.

[8] Actually, even with middle and final sub-words, the results are similar to those shown in this section. They are excluded here for vocabulary size comparison, which is shown in Table 1.

over the previous studies. Now that the proposed method is free from the use of a dictionary, it is applicable to words and language data even if their meaning dictionaries are not available or difficult to obtain; all it requires is a corpus of a fair size[9]. Examples of such cases include historical and learner corpora where meanings of words appearing in these corpora are not necessarily defined in standard dictionaries because they contain ancient or erroneous usages. We will discuss the meaning-frequency law with these corpora in Sect. 5. The above investigations include high-frequency words, function words, and inflected/conjugated forms, which are excluded in the previous studies; Table 1 shows target vocabulary sizes in the previous and present studies.

The slopes in Fig. 1 and Fig. 2 tend to be smaller in a low frequency range ($2 < \log_{10}(f) < 3$), especially for Aozorabunko corpus. That is, the contextual diversity v for low-frequency words is estimated to be larger than the regression line predicts. In other words, the MLMs used tend to assign to them word vectors that result in a higher value of contextual diversity than expected. Word vectors for infrequent words might be of lower quality than for frequent words. We will observe and discuss similar tendencies with historical and learner corpora in Sect. 5 and with smaller LMs in Sect. 7.

5 Meaning-frequency Law in Out-of-domain Data

With the new formulation, we explore the meaning-frequency law in out-of-domain data. By out-of-domain, we mean domains on which LMs (specifically, BERT, here) were not trained[10]. We use two historical corpora (the 1800s and 1900s sub-corpora in CCOHA) and a learner corpus (texts written by English learners, which are excerpts from Lang-8[11]); the details of these corpora are shown in Appendix A.2. The proposed method is applicable even to these corpora for which meaning dictionaries are difficult to obtain for ancient usages in historical texts, and unnatural and erroneous usages in the writings of non-native speakers. We use *bert-large* to obtain word vectors as in Sect. 4.

Study	Vocabulary Size
English	
This study	18,450
Bond et al. (2019)	14,500
Casas et al. (2019)	16,200
Japanese	
This study	20,000
Bond et al. (2019)	10,000

Table 1: Target vocabulary sizes in previous and present studies.

Fig. 3 shows the results from the 1800s and 1900s sub-corpora in CCOHA. It also shows the result from the 2000s sub-corpus for comparison.

It turns out that the deviation from the regression line is slightly larger in the older sub-corpora than in the 2000s, reflected in smaller values of R^2. In particular, the slope tends to be smaller in a low-frequency range ($\log_{10}(f) < 3$) as in Aozorabunko corpus in Sect. 4. BERT, which is trained on contemporary English data, may have difficulty in recognizing the meanings of infrequent words in older texts. In such cases, it might assign to them vectors with more diversity, which in turn increases the value of v.

Nevertheless, the plots fit Eq. (6) well in the higher-range ($log_{10}(f) > 3$). Even in the entire range ($\log_{10}(f) > 2$), the slope coefficient for the entire range is statistically significant ($p < 0.01$, t-Student test). These results suggest that the plots follow the meaning-frequency law at least to a certain extent, suggesting that the BERT model captures the meanings of high-frequency words appearing in texts as old as 200 years. Indeed, Aida and Bollegala (2023) report that BERT is effective in detecting lexical semantic change in the SemEval 2020 task 1 (Schlechtweg et al., 2020) where the 1810–1860 and 1960–2010 texts in CCOHA are used, which are also included in our target sub-corpora.

Fig. 4 shows the relationship between f and v obtained from the original and corrected texts in the learner corpus (Lang-8); Lang-8 contains corrections made by volunteers of native speakers. It also shows the relationship between f and v obtained from the CCOHA 2000s sub-corpus for comparison.

[9] As an example, we extracted about 700,000 words from the first line of the 2000s CCOHA sub-corpus. It turned out that the meaning-frequency law held with $\alpha = 0.12$ and $R^2 = 0.93$ although the data points were only seven. In practice, it would be better to use more data for statistically reliability.

[10] Strictly, BERT is trained on Web and book data which may contain historical and learner texts. The vast majority of the data should be contemporary English texts produced by native speakers of English.

[11] https://lang-8.jp/. This data was privately provided to the authors for academic use.

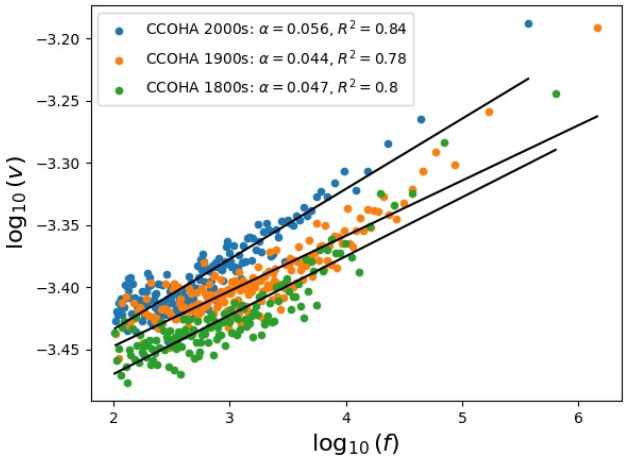

Figure 3: Relationship between f and v in historical corpora. LM: *bert-large*.

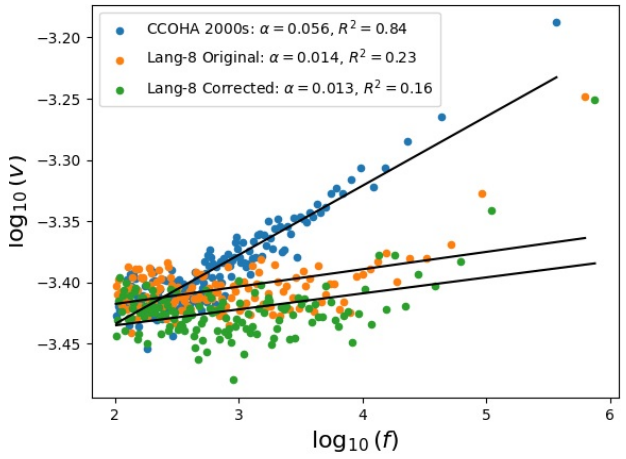

Figure 4: Relationship between f and v in learner corpora. LM: *bert-large*.

The plots for the learner corpora in Fig. 4 exhibit a tendency similar to that found in Fig. 3, only with a stronger tendency of small slopes in a low-frequency range. This is again probably because BERT cannot recognize the different meanings of these infrequent words well. For these deviations from the regression line, it is difficult to tell whether the meaning-frequency law holds for learner corpora or not. Interestingly, the original corpus, which contains errors and unnatural language usages, tends to exhibit a larger value of v for almost all f than its corrected version, even though they have similar sizes. This can be explained from the findings of Nagata et al. (2023) that spelling and grammatical errors increase the value of v; an example is the spelling error *form* (correctly, *from*), which adds usages as the preposition to its noun and verb usages.

6 Meaning-frequency Law in Random Corpora

It would be interesting to explore the meaning-frequency law for random corpora considering that other Zipf's laws such as the frequency-rank law hold even in certain random sequences (Li, 1992; Zörnig, 2015).

To this end, we consider the following three types of random corpus whose details are shown in Appendix A.4:

(1) a shuffled version of the CCOHA 2000s sub-corpus: the words in the sub-corpus are randomly shuffled, and thus the resulting word sequences are mostly unfamiliar to LMs trained on canonical English texts; in contrast, the vocabulary set and the word frequencies are identical to those of the original

(2) a uniformly sampled corpus: words are uniformly sampled from the vocabulary set of the CCOHA 2000s sub-corpus to obtain a corpus of the same size as that of the original sub-corpus

(3) random vectors: in this virtual corpus, the d-dimensional unit vector $\mathbf{x}$ is randomly generated on the $(d-1)$-sphere; the numbers of generated vectors are identical to the numbers (frequencies) of words in the CCOHA 2000s sub-corpus

These randomly generated corpora and vectors are used to calculate f and v.

Fig. 5 shows the relationship between f and v obtained from the random corpora together with those from the original sub-corpus. Note that the plots for the random vectors are shifted by -2 along the vertical line for readability.

It turns out that the meaning-frequency law holds only in the random vectors and the original CCOHA 2000s sub-corpus. The former exhibits the almost perfect fit to the regression line with a slope coefficient of 0.5 while the latter shows a slightly low fit with a much smaller slope coefficient. This suggests that words in human languages use only a small fraction of the $(d-1)$-dimensional sphere to convey meanings compared to the randomly generated vectors.

In contrast, the shuffled corpus does not follow the meaning-frequency law. The plots form a somewhat v-shape line. Most of the word sequences in the shuffled corpus are unfamiliar to the LM

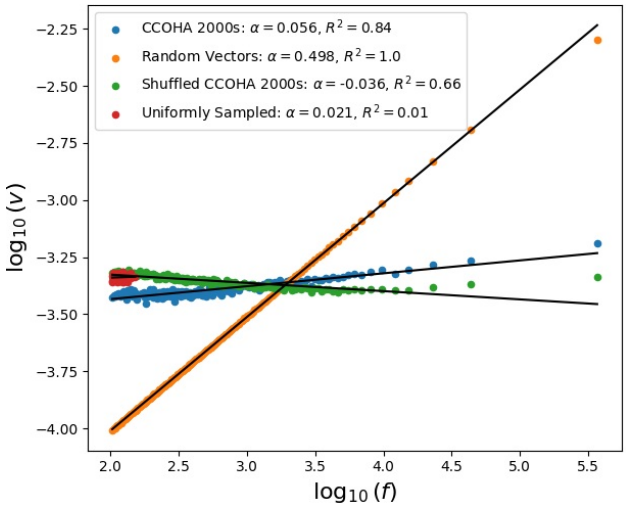

Figure 5: Relationship between f and v in three types of random corpus. LM: *bert-large*

bert-large. As a result, it likely produces considerably different word vectors for each word instance with a certain randomness. Then, the plots would follow a line increasing with respect to f as discussed above. In reality, however, the line for the low-frequency range $\log_{10}(f) < 4$ tends to be decreasing. This implies that there must be another force to decrease the value of v with respect to f. As discussed in Sect. 4, infrequent words will likely introduce randomness into word vectors, which in turn increases the value of v, deviating upward from the regression line. The shuffled corpus simulates this situation. The vocabulary set and the word frequencies in the shuffled corpus are identical to those in the original corpus. Therefore, words tend to co-occur more often with high-frequency words even after shuffling. Then, as the frequency of a word increases, words in its local contexts[12] coincide with each other more often. Note here that the norm of the mean vectors gets larger when there are more similar vectors, which is reflected in a smaller value of the contextual diversity v. As a result, infrequent words tend to receive a larger value of v even in shuffled corpora; conversely more frequent words exhibit the opposite tendency. This is our hypothesis for the v-shape plots found in the shuffle corpus.

The uniformly sampled corpus results in unique plots. In the corpus, words are uniformly sampled and thus, their frequencies are identical. Similarly, v converges to a certain value for all words because their contexts are randomly generated and the degree of their overlaps is similar. Indeed, Fig. 5 reflects this expectation well, showing a rather small grouping of plots rather than a line.

The discussion so far is summarized as follows: What makes the difference between human language data and the random corpora is: (i) the meaning-frequency does not hold in shuffled and uniformly sampled corpora; (ii) it perfectly holds in random vectors; (iii) human language data exhibit a slight deviation from the regression line and its slope coefficient is much smaller than that of random vectors.

7 Meaning-Frequency Law with respect to Model Size and Architecture

When the size of an LM is small, it might not be able to recognize differences in word meanings well. With such small LMs, the meaning-frequency law might not be observable any more.

To investigate this, we compare BERT models of six different sizes[13] as shown in Table 2. Fig. 6 shows the relationships between f and v obtained from the CCOHA 2000s sub-corpus by using the six BERT models. The labels in the legend are shown in descending order of their model sizes.

As expected, Fig. 6 shows that the law becomes unobservable when the model size is small. The slope coefficient gets smaller as the model size gets smaller; *bert-small* and smaller exhibit a negative value.

Looking deeper inside, Fig. 7 shows the results for *bert-base* and *bert-medium*; the upper figure is extracted as it is as in Fig. 6 and the lower figure corresponds to regression for $\log_{10}(f) > 3$. At

[12]Here, *local context* refers to a few words around the word in question. They tend to overlap as the frequency of the word increases with a limited vocabulary set as in the case in this section.

[13]The following implementation was used: https://huggingface.co/prajjwal1/bert-medium, MIT License.

Model	Number of parameters
bert-large	340M
bert-base	110M
bert-medium	41.7M
bert-small	29.1M
bert-mini	11.3M
bert-tiny	4.4M
gpt2-medium	345M
gpt2-xl	1,558M

Table 2: Model sizes. The *uncased* versions are used for all English BERT models (e.g., *bert-large-uncased*).

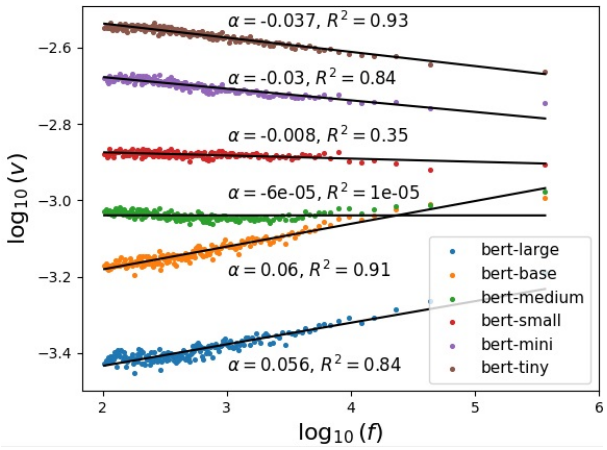

Figure 6: Relationship between f and v calculated by LMs of different sizes. Corpus: CCOHA 2000s.

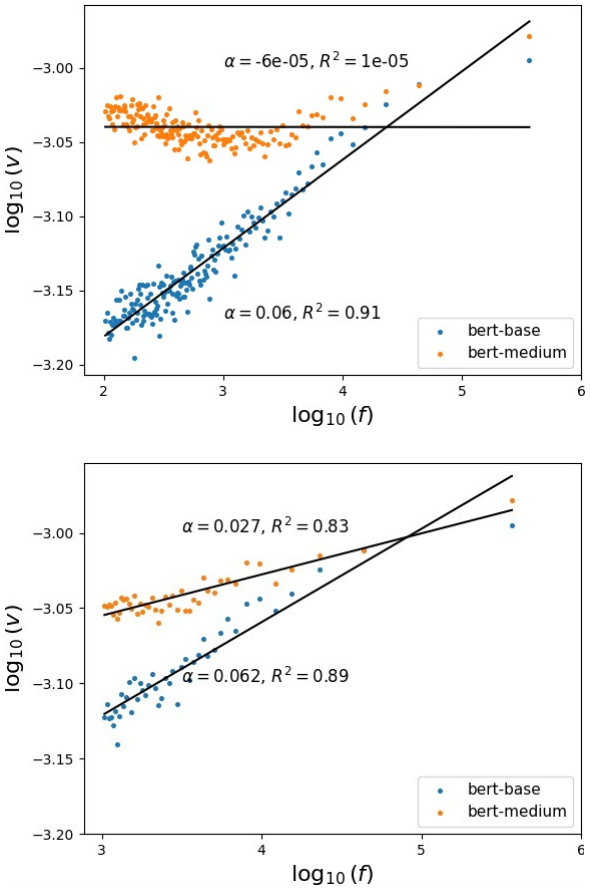

Figure 7: Comparison between *bert-base* and *bert-medium*. Upper: regression fitted to f and v where $\log_{10}(f) > 2$. Lower: regression fitted to f and v where $\log_{10}(f) > 3$. Corpus: CCOHA 2000s.

first sight, the plot for *bert-medium* in the upper figure does not follow the meaning-frequency law, similar to those in the learner corpora, but with almost zero R^2 and slope coefficient. In contrast, those in the lower figure fit Eq. (6) well. A possible reason might be that *bert-medium* and smaller do not have enough parameters to learn all the meanings for all words in the training data. With this limitation, during training, they would focus on words appearing frequently in the training data to reduce the entire loss, instead of trying to learn all words. If so, they should be able to recognize meanings of high-frequency words well, which results in a better model fit in a high-frequency range. In contrast, they are not capable of recognizing those of low-frequency. For this, they would assign very different word vectors to those that are superficially different, but actually have similar meanings. This would increase the variation of the directions of word vectors. This agrees with the plots for *bert-medium* and smaller models.

We can also explore the lexical capability of LMs in terms of their architectures. Here, we compare *bert-base* (MLM) with autoregressive LMs that predict the next token. Specifically, we use two autoregressive LMs (GPT-2 (Radford et al., 2019), *gpt2-medium* and *gpt2-xl*[14]). Fig. 8 shows the relationships between f and v obtained from the CCOHA 2000s sub-corpus by these three LMs.

Fig. 8 reveals that f and v obtained by *gpt2-medium* do not fit Eq. (6) at all. Surprisingly *gpt2-medium* is approximately three times larger than *bert-base* (see Table 2). Nevertheless, its slope coefficient exhibits a negative value; its R^2 is very small. With *gpt2-xl*, which is approximately 14 times larger than *bert-base*, the meaning-frequency law is now observable. In other words, it requires much more parameters to be able to observe the meaning-frequency law with the autoregressive LM.

These results suggest that MLMs have an advantage over autoregressive LMs in distinguishing between differences in meaning through their word vectors. This is mainly because GPT-2, and also other autoregressive LMs, are a decoder that can only use the token in question and its previous context to predict the next token; the information about the following context is not available in their word vectors. In contrast, all the words in the input passage are given to MLMs and thus the information about previous and also following context is available in their word vectors.

The findings in this section and in Sect. 5 suggest

[14] https://huggingface.co/openai-community/gpt2, MIT License.

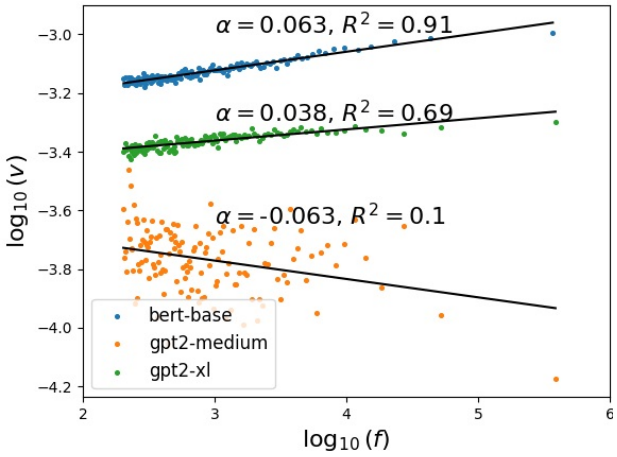

Figure 8: Relationship between f and v calculated by BERT and GPT-2. Corpus: CCOHA 2000s.

that the meaning-frequency law might be used as a sanity check for the lexical ability of LMs. Namely, one can use it to see if LMs perform appropriately on target data; if the meaning-frequency law is not observed, the LM might not perform well on the data in question. Then, it would be better to conduct further checks.

To support this argument, we conducted two additional investigations. In one investigation, we calculated Pearson correlation coefficients between the number of word meanings registered in Wordnet and the values of v obtained from the CCOHA 2000s sub-corpus by using the six models. In the other investigation, we compared the slope coefficients of the six models with their accuracy in Multi-Genre Natural Language Inference (MNLI), which is extracted from the previous study (Bhargava et al., 2021).

Table 3 shows the results. The correlation between the number of word meanings and the contextual diversity v becomes higher as the size of the model increases; their scatter plots are available in Fig. 9 where words are put into bins just as in Fig. 6. Similarly, the correlation between the slope coefficient α and accuracy in MNLI is high.

8 Conclusions

This paper has presented a new formulation of Zipf's meaning-frequency law based on contextual diversity measured using contextualized word vectors. This formulation gives a new interpretation of the law as a relationship between word frequency and contextual diversity. It requires no dictionary, unlike the previous studies, and thus overcomes problems and limitations stemming from the use of a dictionary. With this formulation, this paper has

Model size	α	# meanings	MNLI
large	0.056	0.418	87.5
base	0.060	0.465	83.7
medium	-6×10^{-5}	0.312	79.6
small	-0.008	0.252	76.5
mini	-0.030	0.119	72.3
tiny	-0.037	-0.051	64.5
γ	—	0.92	0.92

Table 3: Relationship between slope coefficient α of six BERT models and task performances. # meanings: Pearson correlation coefficients between the number of word meanings in WordNet and the values of v estimated by each model; MNLI: accuracy of each model in MNLI; γ: Pearson correlation coefficient — correlation between the slope coefficient α and model performances in the two tasks (both statistically significant).

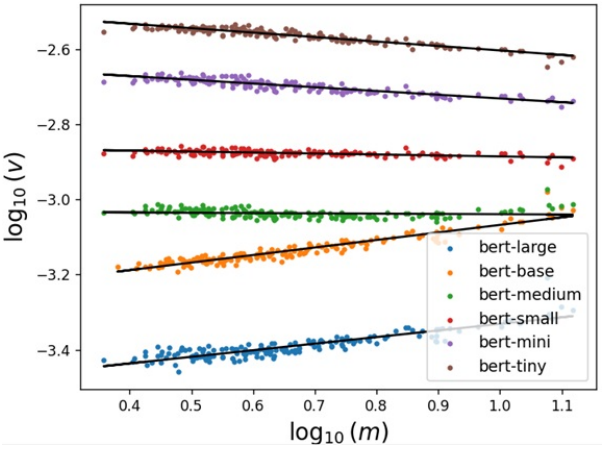

Figure 9: Relationship between meaning count m in WordNet and v estimated from CCOHA 2000s using BERT models of different sizes.

shown that the meaning-frequency law holds for a wider range of words and corpora than the previous studies have shown. This paper has also revealed the differences between human language data and various random sequences. Finally, this paper has explored the meaning-frequency law with LMs of different sizes and of different architectures, which shows how it might be possible to use this method to examine whether an LM will work well on a given dataset.

Limitations

As described in Sect. 3, the proposed method assumes the von-Mises Fisher distribution behind word vectors. This inevitably assumes that the distribution of word vectors is unimodal and isotropic. The true distribution of vectors for certain words may be multimodal and/or anisotropic. A more sophisticated modeling (e.g., a mixture of the von-Mises Fisher distribution (Banerjee et al., 2005)) might achieve more accurate modeling and thus more accurate investigations of the meaning-frequency law, although it is already observable with the von-Mises Fisher distribution.

Correlated with this, the use of the von-Mises Fisher distribution discards norms of individual word vectors. This does not necessarily mean that norms of word vectors are not important for handling word meanings; they might encode some important aspects of word meanings. Therefore, the same argument as above applies to this point.

Theoretically, the proposed method is applicable to all words with certain frequency counts in any language as long as a fair size of a corpus of that language is available. However, the vocabulary size is limited by the LM used to investigate the law. In BERT, for example, its vocabulary size is 30,000 including middle and end sub-words. Words that one desires to include in their investigation may not be in the vocabulary set or be split into multiple sub-words, and thus they may be excluded. One could train an LM with an arbitrary vocabulary set, but it would be costly to train a large LM from scratch.

Acknowledgments

This study was partially supported by JSPS KAKENHI Grant Number JP22K12326 and JST, CREST Grant Number JPMJCR2114, Japan.

References

Taichi Aida and Danushka Bollegala. 2023. Unsupervised semantic variation prediction using the distribution of sibling embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6868–6882, Toronto, Canada. Association for Computational Linguistics.

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean corpus of historical American English. In *Proc. of the 12th Language Resources and Evaluation Conference*, pages 6958–6966.

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. 2005. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6(46):1345–1382.

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in NLI: Ways (not) to go beyond simple heuristics. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 125–135, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Francis Bond, Arkadiusz Janz, Marek Maziarz, and Ewa Rudnicka. 2019. Testing Zipf's meaning-frequency law with wordnets as sense inventories. In *Proceedings of the 10th Global WordNet Conference*, pages 342–352.

Bernardino Casas, Antoni Hernández-Fernández, Neus Català, Ramon Ferrer-i Cancho, and Jaume Baixeries. 2019. Polysemy and brevity versus frequency in language. *Computer Speech and Language*, 58(C):19–50.

Christos Christodouloupoulos and Mark Steedman. 2014. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word corpus of historical American English. *Corpora*, 7(2):121–157.

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

Philip Edmonds. 2005. *Lexical Disambiguation*, pages 43–62. Elsevier, Amsterdam.

Christiane Fellbaum. 1998. Wordnet: an electronic lexical database.

Zellig S. Harris. 1954. Distributional structure. *WORD*, 10(2–3):146–162.

Antoni Hernández-Fernández, Bernardino Casas, Ramon Ferrer i Cancho, and Jaume Baixeries. 2016. Testing the robustness of laws of polysemy and brevity versus frequency. In *International Conference on Statistical Language and Speech Processing*, pages 19–29.

Bahar Ilgen and Bahar Karaoglan. 2007. Investigation of Zipf's 'law-of-meaning' on turkish corpora. In *Proceedings of the 22nd International Symposium on Computer and Information Sciences*, pages 1–6.

Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and improving BERT performance on lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 192–202, Online. Association for Computational Linguistics.

Wentian Li. 1992. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38(6):1842–1845.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written japanese. *Language Resources and Evaluation*, 48:345–371.

Ryo Nagata, Hiroya Takamura, Naoki Otani, and Yoshifumi Kawasaki. 2023. Variance matters: Detecting semantic differences without corpus/word alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15609–15622.

Alec Radford, Jeffrey Wu, Rewon Child 1 David Luan 1 Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

George Kingsley Zipf. 1945. The meaning-frequency relationship of words. *The Journal of General Psychology*, 33(2):251—256.

Peter Zörnig. 2015. Zipf's law for randomly generated frequencies: explicit tests for the goodness-of-fit. *Journal of Statistical Computation and Simulation*, 85(11):2202–2213.

A Appendix

A.1 Interpretation of Concentration Parameter

To begin with, let us first note that the similarity between two words are conventionally measured by the cosine similarity between their word vectors. This is equivalent to measuring the word similarity based only on the directions of word vectors, or to assuming that all word vectors are normalized so that their norms become one.

Under this condition, any word vector appears on the unit hypersphere. As a special case of this, when the dimension of word vectors is two, word vectors appear on the unit circle as in the dashed arrows (vectors) in Fig. 10.

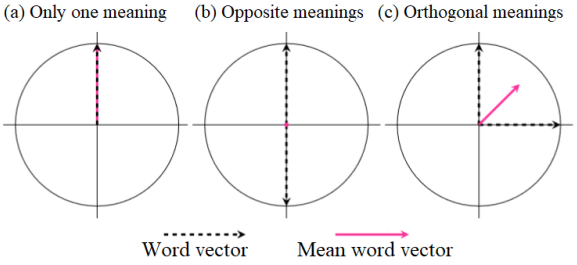

Figure 10: Intuitive Illustration for Mean Norms.

We now examine the norm of the mean word vector for various cases. An extreme case would be that a word is always used in the exact same context, and thus with the same meaning. Its word vectors appear at the same point on the unit hypersphere as in Fig. 10 (a). Then, its mean vector is always identical to the original word vectors, and thus its norm is also always one; recall all word vectors are normalized so that their norms equal one. The other extreme case would be that a word type is represented by two opposite vectors as in Fig. 10 (b), which should cover much wider meanings. In this case, its mean vector becomes the zero-vector with the zero norm. Other cases in between would give a norm between zero and one. For instance, two orthogonal vectors result in the mean word vector whose norm is $\frac{\sqrt{2}}{2}$ as in Fig. 10 (c)[15].

The discussion so far suggests that the concentration of word vectors is related to the norm of its mean vector. This is formalized by the von Mises-Fisher distribution (Banerjee et al., 2005) as described in Sect. 3.

A.2 Details of Used Corpora

Table 4 shows the sizes of the corpora used in the examinations in Sect. 4 to Sect. 7. In Aozorabunko, the size corresponds to the number of characters.

We conducted the following pre-processing for the corpora:

CCOHA: We used the data following the license[16]. Documents containing the string "@@YEAR.txt" (e.g., @1525.txt), which seems to be an erroneously included filename, were removed as noise. The document tags (e.g., <P></P>) were also removed. In CCOHA, 5% of ten consecutive tokens every 200 are replaced by '@' due to

[15] Addition of two orthogonal vectors produces a vector along the diagonal line with a norm of $\sqrt{2}$, and thus the norm of the mean word vector is $\frac{\sqrt{2}}{2}$.

[16] https://licenses.library.ubc.ca/EnglishCorporaCOHA

Corpus	Size
CCOHA 1800s	111,048,657
CCOHA 1900s	262,200,025
CCOHA 2000s	68,678,659
BNC	109,369,848
Aozorabunko	198,755,598
BCCWJ	124,102,859
Lang-8 Original	127,864,912
Lang-8 Corrected	152,681,283

Table 4: Sizes of corpora used in investigations. Sizes are measured by tokens except Aozorabunko where characters are used instead.

copyright regulations. Sentences containing these special tokens were also excluded from the analyses. The remaining sentences were tokenized first by spaCy[17] and then again tokenized by the tokenizer of the corresponding LM.

BNC: We used the data following the license[18]. Only the written part was used. Their sentences were tokenized by the BERT tokenizer.

Lang-8: This data was privately provided from the creator for academic use. It consisted of sentences written by learners. Only English sentences written between 2012 and 2019 were targeted in the investigation. The corrected version of the original sentences was also used as a target corpus; part (but not all) of the sentences were corrected by volunteers of native speakers of English. For those without corrections, the corresponding original sentences were used as correct sentences. The same tokenization process as in CCOHA was applied to these corpora.

Aozorabunko corpus: The data available at the Github site[19] were used. The documents were split into sentences by pySBD[20]. Tokenization was done by `cl-tohoku/bert-large-japanese-v2`.

BCCWJ: The data available at the site[21] were used. The same pre-processing as in Aozorabunko was applied.

Bible corpora: They were used for the additional investigation in Sect. 4. The data available at the Github site[22] were used. The text data were extracted by using the accompanying tool. Tokenization was done by `bert-base-multilingual-cased`.

A.3 Detailed Results with Bible Corpora

To augment the results in Sect. 4, we conducted an additional experiment with the multilingual BERT[23] and the Bible corpora (Christodouloupoulos and Steedman, 2014). The multilingual BERT covered 69 languages in the Bible corpora. Of 69, we targeted 27 corpora that had a vocabulary size of 1,000 or more as a result of the BERT tokenization because the examination of the meaning-frequency law requires a certain vocabulary size to be able to conduct a regression analysis.

Table 5 shows the results. It turns out that the meaning-frequency law holds for all 27 corpora.

A.4 Details of Random Corpora

In Sect. 6, we examined the meaning-frequency law against three types of random corpus. The details of the corpora are as follows:

Shuffle corpus: We created a shuffle corpus from the CCOHA 2000s sub-corpus as follows: (1) all texts in the original sub-corpus were split into sentences and then tokenized into tokens by using spaCy; (2) all sentences were concatenated as a long sequence of words; (3) from the sequence, all tokens were randomly chosen, one at a time, to make another sequence of tokens with the same length; (4) this random process was repeated five times; (5) finally, the sequence was split into sentences by either '.', '?', or '!'. Note that the length and word frequencies of the shuffled sub-corpus are identical to the original although the word orders are considerably different.

Uniformly sampled corpus: We created a uniformly sampled corpus as follows: (1) the vocabulary set was created from the CCOHA 2000s sub-corpus; (2) words were sampled from the vocabulary set of the CCOHA 2000s sub-corpus with the same probability to obtain a word sequence of the same length as that of the original sub-corpus; (3) the sampled word sequence was split into subsequences at the same locations of the sentence boundaries of the original sub-corpus; (4) the re-

[17] https://spacy.io/, the en_core_web_sm model, MIT License.

[18] http://www.natcorp.ox.ac.uk/docs/licence.html

[19] https://github.com/aozorahack/aozorabunkotext, license: https://www.aozora.gr.jp/guide/kijyunn.html

[20] https://github.com/nipunsadvilkar/pySBD, MIT License.

[21] https://clrd.ninjal.ac.jp/bccwj/en/index.html, license: Academic license.

[22] https://github.com/christos-c/bible-corpus, CC0-1.0 license.

[23] https://huggingface.co/google-bert/bert-base-multilingual-uncased, bert-base-multilingual-uncased, Apache license 2.0.

Family	Language	Slope Coefficient	R^2	p-value
Indo-European	Afrikaans	0.054	0.88	0.01
Indo-European	Albanian	0.040	0.97	0.02
Indo-European	Czech	0.030	0.80	0.01
Indo-European	Danish	0.044	0.90	0.01
Indo-European	English	0.047	0.87	0.00
Indo-European	English (WEB)	0.051	0.93	0.00
Indo-European	French	0.046	0.81	0.00
Indo-European	German	0.036	0.88	0.01
Indo-European	Italian	0.049	0.86	0.00
Indo-European	Latin	0.035	0.74	0.00
Indo-European	Lithuanian	0.036	0.86	0.02
Indo-European	Norwegian	0.049	0.83	0.01
Indo-European	Polish	0.038	0.81	0.02
Indo-European	Portuguese	0.048	0.79	0.00
Indo-European	Romanian	0.042	0.93	0.02
Indo-European	Russian	0.037	0.95	0.02
Indo-European	Slovak	0.028	0.87	0.01
Indo-European	Slovene	0.028	0.74	0.01
Indo-European	Spanish	0.051	0.84	0.00
Indo-European	Swedish	0.038	0.75	0.01
Japonic	Japanese	0.050	0.92	0.00
Japonic	Japanese (tokenized)	0.045	0.94	0.01
Sino-Tibetan	Chinese	0.072	0.87	0.00
Sino-Tibetan	Chinese (tokenized)	0.072	0.87	0.00
Turkic	Turkish	0.052	0.92	0.02
Uralic	Finnish	0.042	0.91	0.02
Uralic	Hungarian	0.013	0.44	0.02

Table 5: Regression results on 27 Bible corpora.

sulting sub-sequences were regarded as sentences and were used as inputs to BERT to obtain their word vectors.

Uniformly sampled vectors: In this virtual random corpus, we randomly generated d-dimensional vectors on the $(d-1)$-sphere; the numbers of generated vectors were identical to the numbers (frequencies) of words in the CCOHA 2000s sub-corpus; these randomly generated vectors were used to calculate the values of v.