Testing Zipf's meaning-frequency law with wordnets as sense inventories

Francis Bond,♠ Arkadiusz Janz,♢ Marek Maziarz♢ and Ewa Rudnicka♢
♠ Nanyang Technological University, Singapore
♢ Wrocław University of Science and Technology, Poland
bond@ieee.org, {arkadius.janz|marek.maziarz|ewa.rudnicka}@pwr.edu.pl

Abstract

According to George K. Zipf, more frequent words have more senses. We have tested this law using corpora and wordnets of English, Spanish, Portuguese, French, Polish, Japanese, Indonesian and Chinese. We have proved that the law works pretty well for all of these languages if we take - as Zipf did - mean values of meaning count and averaged ranks. On the other hand, the law disastrously fails in predicting the number of senses for a single lemma. We have also provided the evidence that slope coefficients of Zipfian log-log linear model may vary from language to language.

1 Introduction

The dependency between meaning and frequency is undisputable. Since Zipf's discovery of the high correlation between mean sense count and mean rank (Zipf, 1945), the law was confirmed by several research teams. Among many Zipfian laws, the modelling of the law of meaning-frequency dependency is probably the most fascinating one, because it directly concerns semantics. Meaning strongly influences word frequency (Piantadosi, 2014) and it is clear that semantics precedes language form in text generation (Ferrer-i-Cancho, 2018).[1]

Originally, Zipf tested the law on Thorndike's list of 20k most frequent words of standard English[2] and meanings taken from the *Thorndike-Century Senior Dictionary*[3] (Zipf, 1945). The dictionary meaning account was based on the actual usage in English newspapers, so there were no obsolete or rare senses. The corpus itself was 10^7 running words large, the lemmas on the frequency list were divided into bins of 500 and 1,000 words. Zipf proved a very strong correlation between the *average* number of word senses and rank of lemmas (Zipf, 1945, p. 253), formulating the following statistical law (Zipf, 1949, ch. 3):

$$m_i \propto {f_i}^{\delta} \tag{1}$$

where i is a given word's rank, f_i is its frequency, m_i represents the number of lemma meanings, Zipf also claimed that the coefficient $\delta \approx \frac{1}{2}$. Taking the logarithm of both sides leads to the equation in 2:

$$log_{10}(m_i) \propto \delta \cdot log_{10}(f_i) \tag{2}$$

The corresponding equation for the meaning-rank law was formulated as follows:

$$log_{10}(m_i) \propto -\gamma \cdot log_{10}(i) \tag{3}$$

where i is a word rank. Zipf thought that $\gamma = \delta$.

Zipf justified the straight meaning-rank line in log-log scale with the "conflicting Forces of Unification and Diversification". While a lazy speaker would always tend to use only a few highly frequent and strongly polysemous words, a demanding hearer would prefer numerous unequivocal/monosemous words. Since these balancing forces act within each frequency bin, language equips more frequent words with more senses to maintain a constant ('compromise') polysemy ratio (Zipf, 1949).

He argued that the slope coefficient was close to 0.5, which is now called the *strong* Zipf's law

[1] Consider, for instance, a simple model of a random walk on an undirected graph of lexico-semantic relations. The stationary probabilities of lending in each vertex are proportional to the degree of the vertices (cf. Avrachenkov et al. (2015), Lovász and Winkler (1995, p. 5)), that is to the number of sense relations, including the number of interconnected polysemous senses of the same lemma. As a result, one gains more polysemous words being chosen more frequently.

[2] *A Teacher's Word Book of 20,000 Words*, New York: Teachers College, 1932.

[3] New York: Appleton-Century, 1941.

(Ferrer-i-Cancho, 2016), although the exact value was in fact 0.466 (Zipf, 1945).[4]

Although he only proved the dependency between the mean number of senses $\bar{m}$ and the mean rank $\bar{\imath}$ within each frequency bin,[5] Zipf was sure that the law (1) was applicable to every single lemma:

> (...) if we had a rank-frequency distribution of the 20,000 most frequent words of the Thorndike analysis, it would probably be rectilinear (...), at least for the first 10 to 12 thousand most frequent words. (Zipf, 1949, ch.3)

A more recent verification of Zipf's meaning-frequency law has revealed that the relationship is more complex than could have been foreseen in the middle of the 20th century.

The aim of this paper is to provide new broad empirical evidence for the *weakened* version of Zipf's meaning-frequency law[6] based on corpora and wordnets as sense inventories. Five European languages (English, Spanish, Portuguese, French and Polish) and three Asian languages (Mandarin, Indonesian and Japanese) representing four distinct language families (Indo-European, Sino-Tibetan, Japonic and Austronesian) were inspected.

All data sets and source code are available at `https://github.com/MarekMaziarz/Zipf-s-Meaning-Frequency-Law`.

2 Related Work

Despite the fact that most of Zipf's laws, like the law of frequency-rank distribution or the law of abbreviation, were studied thoroughly, the meaning-frequency law itself gained relatively less attention (Casas et al., 2019). Still, some attempts were made by several research teams.

Edmonds (2004) repeated Zipf's experiment on the British National Corpus with the use of Princeton WordNet 2.0. He gathered lemmas in bins of 100 words each and estimated the γ coefficient at 0.40 (cf. Table 1).

experiment	**lang.**	γ
Zipf, 1945, 1949	en	.47
Edmonds, 2004	en	.40
Ilgen & Karaoglan, 2007	tr	.42, .39
Casas et al. 2019	en	.38
Casas et al. 2019	es	.27
Casas et al. 2019	nl	.25

Table 1: The power γ of Zipf's law exponent of Eq. (3) in hitherto experiments, for bins of 500 (with exception of Zipf's paper and Edmonds' article, details in text).

Ilgen and Karaoglan (2007) tested the law on two Turkish corpora (newspapers, novels, short-stories), one of which was manually tagged with word senses, while the second one was compared to an electronic Turkish dictionary. The authors tested different frequency bin sizes ranging from 50 words up to 1000 words, showing gradual predictive power loss while moving from larger to smaller bins. They obtained the slope coefficient slightly lower than that of Zipf's (0.42 and 0.39; Table 1).

Hernández-Fernández et al. (2016) tested the robustness of Zipf's meaning law on two different corpora (child and child-directed speech corpus – CHILDES[7], and the SemCor corpus) and two sense inventories (WordNet and WordNet senses that appear in SemCor).[8] The authors merged the resources in different combinations which, surprisingly, in all cases led to non-zero correlation coefficients. Unlike previous parametric research, the authors did not focus on mean values of sense count and frequency, but estimated direct correlations between row values of the two. They focused on those parts of speech that were present in WordNet (nouns, adjectives, verbs and adverbs). The authors concluded that positive and statistically significant correlation between sense count and frequency seemed to be corpus-independent.

In Casas et al. (2019) the above non-parametric approach was expanded to two other European languages: Dutch and Spanish; English is analysed again. The sources of frequency were the CHILDES corpus and Wikipidia, while the sources of sense inventories were wordnets: Word-

[4] Provided the first 500 words were omitted, which Zipf tended to treat as function words.

[5] Thorndike's frequency list divided words into bins of 500 and 1,000 words without giving any precise information about the exact number of occurrences of each lemma.

[6] *Strong* Zipf's law of meaning-frequency relationship forces the slope coefficient γ to be equal to 0.5, while *weak* version of the law simply states that $\gamma > 0$ (Ferrer-i-Cancho, 2016).

[7] `https://childes.talkbank.org/`

[8] `http://web.eecs.umich` They also took frequency counts from the English part of the CELEX corpus `edu/mihalcea/downloads.html#semcor`

Net, Open Dutch WordNet and the Multilingual Central Repository for Spanish. Each wordnet is "a proxy for the number of meanings of a word" (*ibidem*).

Casas and colleagues (Casas et al., 2019) also did some experiments with parametric modelling with wordnets as sense inventories and CHILDES corpus as a source of frequencies. They calculated slope coefficients and R-squared values for bins of 100 and 500 words (cf. Tab. 1). They also observed that smaller bins gave worse R-squared statistics.

The criticism of the rank-frequency models was addressed by Piantadosi (2014). Piantadosi raised an important question of the explanatory validity of Zipf's law derivation in various theoretical models. Since there are dozens of different ways of deriving the Zipf's equations (such as random-typing, stochastic models, semantic accounts, communicative accounts etc.), the derivation lacks its explanatory power and "[t]he key will be (...) to generate novel predictions and to test their underlying assumptions with more data than the law itself" (*ibidem*).

Altmann and Gerlach (2016) argue that linguistic statistical models should be validated not only by measures of fit like R-squared determination coefficient, but also with additional measures of randomness of model residuals (they propose significance level set at 1%): "A low p-value is a strong indication that the null model is violated and may be used to refute the law (e.g., if p-value < 0.01)." According to them, ordinary Zipfian rank-frequency linear models unfortunately lack this randomness property (p-values $\ll 0.01$). Piantadosi (2014) similarly points that rank-frequency models based on corpus data when analysed in a standard way (i.e., on the same sample), suffer from correlated errors, since the ranking is constructed out of the very same frequency distribution as frequency estimation itself. Luckily, this argument cannot be applied to the same extent to meaning-frequency and meaning-rank distributions, since they are prepared with either frequency, or rank at once. Especially, if the Zipf's meaning-frequency law (or meaning-rank law) is modelled on the basis of different language resources (like a wordnet and a corpus) the problem vanishes.

3 Method

We checked the validity of Zipf's meaning-rank law by collating frequency counts and *corresponding* meaning counts. We did this by comparing general corpora, representing language in usage, and sense numbers taken from wordnets, which represent each language lexical system, cf. Fellbaum (1998). Another way to see these two sides of language reality is to compare frequencies and polysemy count in the very same text (in a widely sense tagged corpus). We did this on the richly annotated Sherlock Holmes subcorpus of Nanyang Technological University Multilingual Corpus, *NTU-MC* (Bond and Tan, 2012).

3.1 Data sets: Wordnets

We treat wordnet as a useful model of human mental lexicon, and wordnet sense numbers as the approximation of real polysemy of a lemma. The choice of wordnets is motivated by their shared properties (e.g. similar relational description models, existence of synsets, glosses) which allow us to directly compare Zipfian curves for different languages. For the purposes of our study, we have chosen eight wordnets. The wordnets include: Princeton WordNet (henceforth, PWN) (Fellbaum, 1998), Polish WordNet (henceforth, plWN) (Maziarz et al., 2016), Wordnet Libre du Français (henceforth, WOLF), Multilingual Central Repository (henceforth, MCR) (Gonzalez-Agirre et al., 2012), Japanese Wordnet (henceforth, WNJA) (Bond et al., 2008), Wordnet Bahasa (WNB) (Bond et al., 2014), and Chinese Open Wordnet (henceforth, COW) (Wang and Bond, 2013). The wordnets are listed in Table 2 together with languages they represent, number of lemmas from wordnets and corpus coverage. They all appear in the Open Multilingual WordNet 1.0[9] (Bond and Foster, 2013) and are thus inter-linked via PWN. The numbers are given with the exclusion of multi-word lexical units and synsets not linked to Princeton WordNet and, hence, not linked to CILI.

3.2 Data sets: Corpora

To test Zipf's meaning-frequency law, we have inspected two types of text data sets: (i) general corpora for English, Spanish, French, Portuguese, Chinese, Japanese and Polish built at Centre for

[9] http://compling.hss.ntu.edu.sg/omw/

wordnet	**lang.**	**#S** [10^3]	**#L** [10^3]	**poly.** #S/#L
COW$^+$	zh	8.1	3.2	2.53
WNJA	jp	158.1	92.0	1.72
MCR	es	57.8	36.7	1.58
OpenWN-PT	pt	74.0	54.0	1.37
plWN$^+$	pl	288.4	191.8	1.50
PWN$^+$	en	218.6	159.4	1.37
WNB	id	95.3	26.9	3.54
WOLF	fr	102.7	55.4	1.85

Table 2: Data sets: OMW wordnets. Symbols: COW - Chinese Open Wordnet, WNJA - Japanese Wordnet, MCR - Multilingual Central Repository, OpenWN-PT - Open Portuguese Wordnet, plWN - Polish WordNet, PWN - Princeton WordNet, WNB - Wordnet Bahasa, WOLF - Wordnet Libre du Français; #S - number of senses, #L - number of lemmas, poly. - average polysemy; $^+$ – wordnet taken in whole. Please note that for most wordnets we have taken only PWN equivalents (connected via (C)ILI). All numbers are given for one-word lexical units only.

Translation Studies, University of Leeds[10], and at Wroclaw University of Science and Technology, Poland,[11] (Broda et al., 2010), as well as (ii) a part of the NTU-MC, containing two Sherlock Holmes stories and their translations into Indonesian, Chinese and Japanese, henceforth: *SH* (Bond and Tan, 2012). All used frequency lists are available under open licences.

Corpus statistics are presented in Table 3. Most general corpora are collections of Web documents (marked as *IC*) gathered by Web crawling within the WaCky project (Baroni et al., 2009), covering 100–300 million running words each. The Web as a Corpus approach was also used to make a corpus of Polish, the largest one, comprising almost 2 billion words, the source of lemma frequency list in the case of Polish (Maziarz et al., 2016). To analyse the impact of the used corpora on our results we made use of frequency lists for Reuters Corpus (a collection of news from Reuters, *RC*) and Gigaword Corpus for Chinese (henceforth: *GC*)[12] Frequency lists for Chinese were word form based,[13] whereas all the rest of the lists contained lemma frequencies.

corpus	**size** [10^9]	**min** f	L [10^3]	**cov.** [%]
en-IC	.18	218	14.5	72
en-RC	.10	1,100	3.8	70
pl-IC	1.80	10,972	8.8	88
es-IC	.14	248	7.2	48
fr-IC	.18	2,080	4.3	86
pt-IC	.19	2,400	4.0	80
zh-IC*	.28	183	11.0	22
zh-GC*	.24	377	7.0	28
jp-IC	.25	567	10.0	67
en-SH	.02	1	2.8	56
id-SH	.01	1	1.8	48
jp-SH	.03	1	3.1	37
zh-SH	.02	1	3.8	67

Table 3: Data sets: corpora. Symbols: en - English, es - Spanish, id - Indonesian (Bahasa), jp - Japanese, fr - French, pl - Polish, pt - Portuguese, zh - Chinese (Mandarin); IC - internet corpus, RC - Reuters Corpus, GC - Gigaword Corpus, SH - NTU-MC subcorpus of Sherlock Holmes stories; * - word frequency list; f – corpus frequency, L – number of lemmas given for frequency lists united with each wordnet list, cov – coverage of the original frequency list as covered by a particular wordnet.

In order to make sure that our lists contain only content words we threw out all words of rank 100 and above (rank $i \leq 100$). On the other hand, all frequency lists were shortened at different cutoff points. For instance, the Reuters Corpus was clipped down to the rank $i = 5{,}000$ (1,100 occurrences in the corpus), while Chinese Gigaword corpus was cut at the rank $i = 25{,}000$ (377 occurrences). Intersections with wordnets' lemma lists gave as a result a similar order of magnitude of the resulting lists for all languages ($5–15 \times 10^3$). These lists had quite good coverage of the original corpora frequency lists (on the average 70-80%, with the exception of Chinese corpora, which had poor coverage of 20-30%).

The final analysis was conducted on a relatively small multilingual SH corpus. The Sherlock Holmes subcorpus of the NTU-MC consists of two of Conan Doyle's short stories (*The Adventure of the Speckled Band*, 1892, and *The Adventure of the Dancing Men*, 1903/1904) annotated with wordnet

[10] http://corpus.leeds.ac.uk/list.html, CC-BY.

[11] In the case of Polish, http://nlp.pwr.wroc.pl/en/tools-and-resources/resources/frequency-list CC BY-NC-SA 3.0

[12] The selection contains only news which makes it comparable to the Reuters Corpus.

[13] Chinese has practically no inflection

senses. The coverage appears low, but this is an undercount, some concepts cover multiple words (especially in Japanese, where the segmenter segments to morphemes).

Apart from nouns, adjectives, verbs and adverbs, the annotation also included pronouns, so to make the lists more comparable to those made out of general corpora, the top of all lemma frequency lists, comprising mostly function words,[14] was cut saving words less frequent than 100 occurrences ($f < 100$).[15] The intersection with wordnets' lemma lists was also comparable to that of general corpora (2–4×10^3, cf. Tab. 3, SH rows).

3.3 Constructing rank bins

The original Zipf's work on meaning-frequency dependency was in fact the research on averaged values of sense number and ranks. Ilgen and Karaoglan (2007) and Casas et al. (2019) proved that the relationship is strong for larger bins, but becomes more and more relaxed for smaller frequency bins.

Since our frequency lists are intersected with wordnet lemma sets, some ranks from the original corpus lists occasionally fall out, so we receive gaps within continuous stream of ranks. If a corpus coverage by each intersection set (see Table 3) is close to 80%, on the average every fifth rank darts out. This is the reason why we cannot take a final intersected corpus-wordnet list and simply divide it into bins of particular size. Instead, we ought to deal with specific rank ranges.

The process of varying word bin sizes was slightly different for general and Sherlock Holmes corpora, thus we give their descriptions separately.

General corpora. We explored only nouns, adjectives, verbs and adverbs, but omitted words of ranks 1–100 in order to avoid introducing non-content words into frequency counts.[16] The bins were collated for specific rank ranges (λ = 1, 50, 100, ..., 500). Since ranks 1–100 were intentionally omitted, we started our rankings in the best case from $i = 101$. Similarly to Casas et al. (2019), we constructed bins of the range λ such that a word with i^{th} rank fitted j^{th} bin if and only if the following inequalities were fulfilled:

$$100 + \lambda \cdot (j - 1) + 1 \leq i \leq 100 + \lambda \cdot j, \quad (4)$$

where $j = 1, 2, 3, ..., \text{round}(\frac{n}{\lambda})$, λ is a rank range, I is a set of ranks i ($i \in \mathbf{N}$, $i > 100$), and n is a maximum rank.

Figure 1 illustrates the process of making the frequency bins smaller and smaller and shows regression lines for some successive rank bins in the Reuters Corpus.

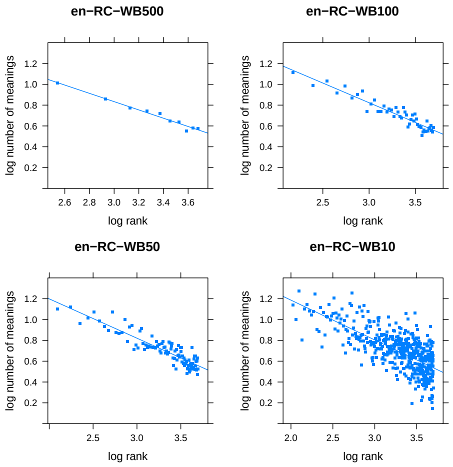

Figure 1: Meaning-rank dependency for the Reuters Corpus and PWN 3.1, with regard to word bins of different rank range. Symbols: WB - word bin λ = 10, 50, 100, 500 ranks. The slope coefficient $-\gamma$ equals -0.42 for λ = 500.

Sherlock Holmes stories. We explored again only nouns, adjective, verbs and adverbs, but threw out words of frequencies greater than a hundred occurrences in a corpus. The bins were collated for the following rank ranges: λ = 1, 3, 5, ... 99. We construct such bins of the range λ that a word with i^{th} rank fitted j^{th} bin if and only if the following inequality was fulfilled:

$$(\lambda \cdot (j - 1) + 1 \leq i \leq \lambda \cdot j) \ \& \ (f_i < 100), \quad (5)$$

where $j = 1, 2, 3, ..., \text{round}(\frac{n}{\lambda})$, λ was a rank range, I was a set of ranks i ($i \in \mathbf{N}$), and n was a maximum rank, f_i was a frequency count for the i^{th} word.

3.4 The log-log model

We investigated the weak version of Zipf's meaning-frequency law in the form of Eq. 3 by changing values of the rank range λ from large bins

[14] In fact they are not strongly polysemous.

[15] For English, e.g., there were words *I*, *be*, *you*, *he*, *we*, *say*, *she*, *this*, *not*; while for Indonesian lemmas – *-nya* (as a pronoun and an article), *itu* (a pronoun/article), *saya* 'I, me, mine', *dia* 'he, she, it', *kami* 'we, our, us'.

[16] In his original paper, Zipf cut off the first 500 words, claiming they were function words (Zipf, 1945). Limiting our analysis to nouns, adjectives, verbs and adverbs would result in a slightly different number of words in each bin.

to small. We aimed at discovering the determination coefficient R^2, as well as the slope coefficient $-\gamma$ for largest bins. R^2 values were used previously as a measure of model fit (Zipf, 1945, 1949; Edmonds, 2004; Ilgen and Karaoglan, 2007; Casas et al., 2019). We checked also the slope coefficient non-zeroness with the t-Student test.

To avoid any possibility of infecting our model with correlated errors, we also inspected residuals with the Shapiro-Wilk statistics, as suggested by Altmann and Gerlach (2016). The Shapiro-Wilk test is the most powerful normality test available now to researchers. Originally designed for small samples, now it is applicable also to samples up to 5,000 observations (Razali and Wah, 2011). Hence, if a model was constructed on a larger sample,[17] we applied sampling 5,000 instances from the original set of observations *without* replacement.

As far as we know, this is the first time when the residuals of the linear Zipfian log-log model for meaning distribution are inspected for non-normality.

4 Results

4.1 Predictive power

General corpora. Seven languages (five Indo-European, Chinese and Japanese) and nine corpora were checked for Zipf's meaning-rank law (Eq. 3) efficiency. Table 4 shows the results for λ = 500, 100, 50 and 1. Clearly the very same pattern that was observed earlier in Ilgen and Karaoglan (2007) and in Casas et al. (2019) is also visible in our data. The bigger rank range λ is, the more efficient is Zipf's law. Making word bins smaller and smaller leads to smaller R^2 values, with a collapse at $\lambda = 1$ (no bins).

Figure 2 presents a more detailed picture of what is happening (λ = 1, 50, 100, 150, ..., 500). The determination coefficient R^2 maintains its values down to quite small bin sizes (λ equals 50-100) and then rapidly lowers to poor percentages of 10-20% of variance explained. The process is accompanied by a non-normal behaviour of residuals (p-value drops below the significance level of 1% at λ in the range of 50–150).

			rank range λ			
data set	γ_{500}	**n**	**500**	**100**	**50**	**1**
en-IC$^+$	.42	40	.98	.94	.90	.22
en-RC$^+$	.40	10	.98	.90	.81	.11
pl-IC$^+$	.22	20	.96	.83	.69	.06
es-IC$^+$	.47	30	.94	.86	.81	.26
fr-IC$^+$	.51	10	.98	.94	.90	.04
pt-IC$^+$	.32	10	.96	.88	.77	.09
zh-IC*	.22	100	.86	.61	.45	.06
zh-GC*	.21	50	.86	.56	.40	.06
jp-IC$^+$	.26	30	.94	.83	.72	.08
tr-B	.42	27	.97	.94	.89	—
tr-G	.39	45	.89	.70	.66	—
en-CH	.38	19	.98	.86	—	—
nl-CH	.25	5	.99	.78	—	—
es-CH	.27	7	.95	.59	—	—

Table 4: Loss of Zipf's meaning-rank law predictive power in terms of determination coefficient R^2 with regard to different rank bin sizes (λ = 500, 100, 50, 1). Symbols: γ_{500} marks the slope coefficient of the regression line for $\lambda = 500$, 'n' is number of rank bins used for calculating γ; 'tr-B' and 'tr-G' denotes BilTD and GozD Turkish corpora, respectively, in Ilgen and Karaoglan (2007), '*-CH' marks the CHILDES corpus in 3 language versions: English (*en*), Dutch (*nl*) and Spanish (*es*), taken from Casas et al. (2019, Tab. 1, 2); we have chosen only values for child language.

In the case of Chinese corpora (the Internet corpus, *IC*, and the news corpus, *GC*) R-squared values are smaller as compared to other languages. It becomes clearer why it is so when one compares the coverage of both corpora by the Chinese Open Wordnet (Tab. 3), which is relatively small (coverage is between 20-30%). For most languages, the coverage is much higher resulting in small difference between real bin size and the face value λ (they differ by one-fifth). For Chinese, the proportion is much worse and the real bin size might be on the average only one-third of the nominal value. Simply when looking at Chinese data we look at much smaller bins.

Sherlock Holmes stories. In Table 5 we provide the actual R^2 values for the NTU-MC subcorpus of Holmesian stories. The gradual loss of Zipf's law predictive power is clear – the smaller a bin is the lower the correlation coefficient becomes. Contrary to the results for general corpora/wordnets coupling, the final variance amount explained by Zipf's model is not very low.

The magnitude of the effect itself might be hid-

[17] It was possible for some general corpora in the case of no-bins, see Table 3, the column L and rows *en-IC*, *pl-IC*, *es-IC*, *zh-IC*, *zh-GC* and *jp-IC*.

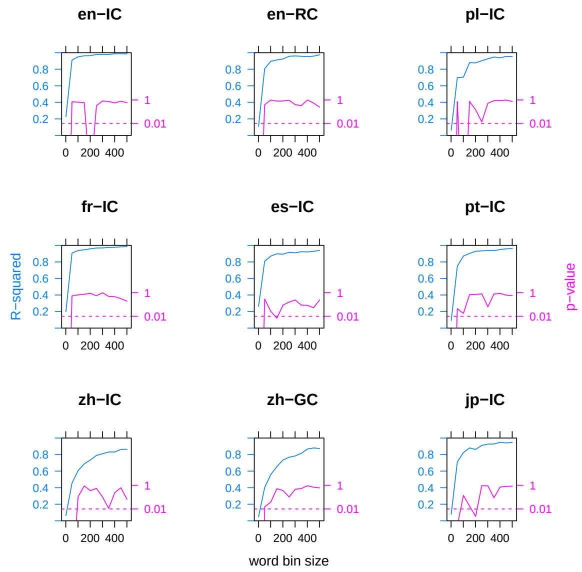

Figure 2: Loss of Zipf's meaning-rank law predictive power in terms of determination coefficient R^2 (blue line) with regard to different frequency bin sizes (λ = 1, 50, 100, ..., 500). With pink line we mark p-values of Shapiro-Wilk normality test for residuals of the model.

data set	γ_{100}	n	rank range λ			
			100	**50**	**10**	**1**
en-SH	.43	28	.96	.94	.86	.44
id-SH	.36	18	.94	.90	.81	.34
jp-SH	.31	31	.96	.94	.83	.37
zh-SH	.33	38	.94	.92	.85	.42

Table 5: Loss of Zipf's meaning-rank law predictive power in terms of the determination coefficient R^2 with regard to different rank bin sizes (λ = 100, 50, 10, 1, $f > 0$ in all cases). The symbol γ_{100} marks the absolute value of the negative slope coefficient of the regression model for λ = 100.

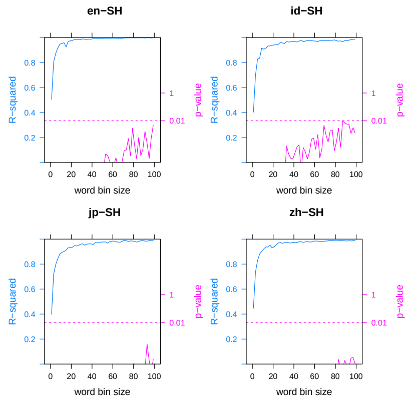

Figure 3: R-squared values of Zipf's meaning-frequency linear model with regard to different word bin sizes (λ = 1, 3, 5, ..., 99) for whole SH corpora (with hapaxes, $f > 0$).

den just by these relatively large correlation values. It remains in agreement with our expectation about how the dependency should act in real texts. The meaning number space is much lower than in the case of comparing general corpora and wordnets, and there is an upper limit imposed on the number of meanings equal to the frequency itself.[18]

The real problem with taking full frequency lists becomes obvious if we inspect the meaning-*frequency* relation for $f > 0$ (Figure 3). Despite the fact that R-squared values are very high, residuals of each model are not normal (p-values $< 1\%$), leading to the presumption that the long tail of words occurring in usage only once per a corpus forces residuals to be correlated.[19] In the case of meaning-*rank* dependencies, this issue is hidden with the common rank ordering practice that we have followed.

All words that occur in a corpus once receive consecutive ranks, rescuing model residuals from total disaster. To be clear, this proves that for *SH* corpora containing *hapax legomena* (like in Table 5) Zipf's law does not function properly, even for mean values.[20]

It is justified to test Zipf's law for words occurring in a corpus more than once. This compromise gives us also an opportunity to compare such shortened lists for Holmes stories with largely abridged lists from general corpora. Figure 4 presents the data. After removing hapaxes, the Zipf's model starts to behave properly: p-values soar above 1%, R-squared values become large when rank ranges are bigger than 20.

4.2 The slope

General corpora. Table 4 provides slope coefficients γ for the largest bins ($\lambda = 500$) in Internet and news corpora. In more detail, we illustrate it with Figure 5 (for different bin sizes). All γ values occurred to be statistically significant in

[18] We cannot get more senses of a lemma than the number of its occurrences in a text.

[19] Since we have a huge amount of points with co-ordinates $f_i = 1$ and $m_i = 1$.

[20] This extraordinary property of Zipf's law does not contradict the results obtainable from general corpora, since they are always shortened with the least frequent lemmas, possibly having hidden this phenomenon out of sight of researchers.

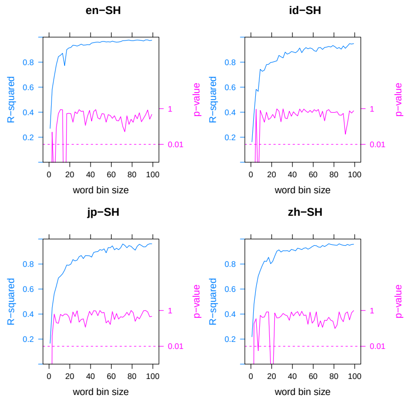

Figure 4: R-squared values of Zipf's meaning-rank linear model with regard to different word bin sizes (λ = 1, 3, 5, ..., 99) for $f > 1$.

t-Student test (p-values are much smaller than the significance level of 1%). It is obvious from the data that the coefficients are mostly less than 0.5 – the value hypothesized by Zipf himself. The values seem also quite stable concerning the vast range of rank bins, however it is not obvious whether slope coefficients are independent of the corpora and frequency lists used.

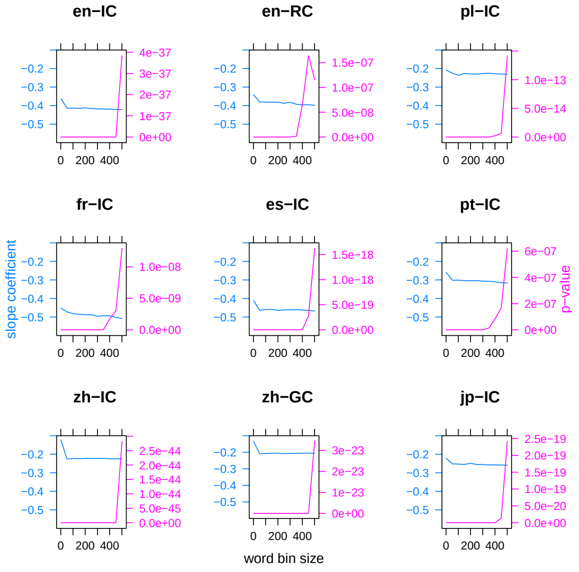

Figure 5: Stability of the slope coefficient $-\gamma_\lambda$ values (blue line) with regard to different frequency bin sizes (λ = 1, 50, 100, ..., 500). Frequency lists are taken in as a whole. With the pink line we mark p-values of t-Student test for non-zeroness of the slope values.

Our frequency lists vary with respect to length, they also come from differently sized corpora (see Table 3). To overcome this problem, we decided to confront languages using relative frequencies. We have chosen the frequency of 12.5 occurrences per million in a corpus as a maximum rank for the need of comparison. The abridged lists cover the most frequent vocabulary of each language, i.e., the top 4000–6000 most frequent lemmas. Table 6 provides Zipfian curve coefficients for $\lambda = 200$ and the shortened frequency lists. The coverage of frequency lists for most languages is very good (80–90%).

Languages differ in terms of regression coefficients. The clear dependency links the slope value and the intercept. The more steep a regression line is, the bigger an intercept becomes. This cross-lingual pattern finds its counterpart in each language data.

corpus	**max. rank**	**cov.** [%]	**poly.** $med(m)$	γ_{200}	I
en-IC$^+$	4856	86	4(5.5)	0.40	2.0
en-RC$^+$	4501	77	4(5.8)	0.38	2.0
pl-IC$^+$	6038	89	3(3.9)	0.22	1.3
es-IC	4575	78	4(4.7)	0.29	1.6
fr-IC	4672	87	5(7.1)	0.48	2.4
pt-IC	4987	79	3(3.4)	0.30	1.5
zh-IC$^+$	6521	48	2(2.6)	0.08	0.7
zh-GC$^+$	6486	45	2(2.7)	0.19	1.1
jp-IC	4681	76	3(4.3)	0.19	1.2

Table 6: Comparison of Zipfian curve coefficients: the slope $-\gamma_{200}$ and the intercept I. The cut-off point is the relative frequency of 12.5 occurrences per million in a corpus. For different languages and corpora the cut-off maximum rank differ. We have chosen $\lambda = 200$ to ensure normality of residuals. Symbols: $^+$ - wordnet taken in whole (not only ILI part), $cov.$ - the coverage of a frequency list by wordnet lemmas, $poly.$ – median of (med) / mean (m) polysemy (senses per lemma).

Figure 6 shows the pattern more thoroughly by presenting regression slope and intercept for different cut-off ranks (maximum ranks) in each language/corpus. Dashed vertical line represents the maximum rank corresponding to relative frequency of 12.5 occurrences per million (chosen as a basis of comparison in Table 6). Coefficients may change their values a lot, as Spanish or French data proves. Yet again, both regression coefficients react inversely to elongating frequency list. While intercepts grow, simultaneously slope coefficients

$-\gamma_{200}$ drop. This reproduces the fact that lengthening frequency lists is the same as adding less and less polysemous lemmas.

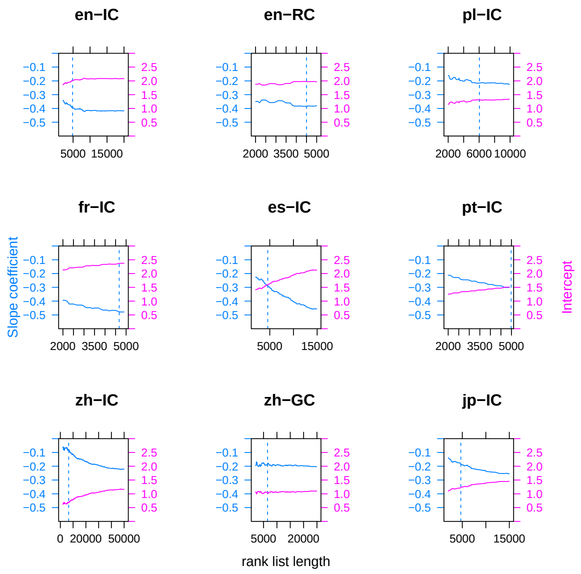

Figure 6: Variation of the slope coefficient $-\gamma_{200}$ values (blue lines) and intercepts (pink lines) with regard to different frequency list lengths (for one particular bin size, λ = 200). With dashed blue lines we mark ranks corresponding to the relative frequency 12.50 occurrences per million.

It is rather unlikely that each language will finally reach its own Zipfian $\gamma = -0.5$ magical zone, even for very long frequency lists. Sherlock Holmes stories give us a unique opportunity to check slope coefficient values under controlled conditions.

Sherlock Holmes stories. As in the case of general corpora, slope coefficients present stable behaviour while changing frequency bin sizes for corpora abridged by *hapaxes* (Fig. 7). Comparing them to values obtained for general corpora and described in the literature shows that although they do change, the change rate is rather moderate (close to $\pm.10$).

Consider the γ values for English. In Zipf's experiment it was .47, Edmonds (2004) estimated it at .40 (Tab. 1), in CHILDES corpus Casas et al. (2019) found it to be close to .38, in Leeds corpora it equals .40/.42 (Internet) and .38/.40 (news), while in Sherlock Holmes stories it is .43. In the case of Japanese, we got .19/.26 in the Leeds corpus and .31 in Sherlock Holmes, not so distant. For Chinese the values change bit more (*zh-IC*: .08/.21, *zh-GC*: .19/.22, *zh-SH*: .33), but this might be caused by the fact that the coverage for Leeds corpora are too low. For Spanish we found the slope coefficient close to .29/.47 in our data, while (Casas et al., 2019) obtained value of .27. The difference might be explained with the poor coverage of the Spanish child speech corpus with Multilingual Central Repository (only 13%, *ibidem*).

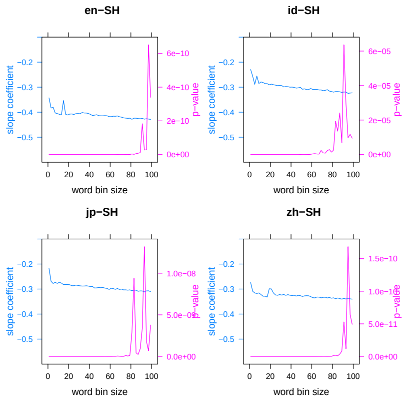

Figure 7: Stability of the slope coefficient $-\gamma_\lambda$ values (blue line) with regard to different frequency bin sizes (λ = 1, 3, 5, ..., 99) for $f > 1$ (no *hapax legomena*). With the pink line we mark p-values of t-Student test for non-zeroness of the slope.

5 Conclusions

We have presented novel, statistically valid, empirical evidence for the weak version of Zipf's law of meaning distribution on eight languages from four distinct language families (Indo-European, Japonic, Sino-Tibetan and Austronesian).

Zipf's law functions pretty well for mean values in terms of high determination coefficient R-squared, and non-zero slope coefficient γ (stable over the vast range of λ values, but changing while altering frequency lists). The law is, however, inefficient for individual lemmas, because of the lack of model residual normality, despite non-zero correlation coefficient R values.

In the case of Sherlock Holmes stories, this Zipfian catastrophe does not manifest itself only while shifting from bins to individual lemmas, but - surprisingly - also within each whole unabridged corpus containing *hapax legomena*, both for smaller and larger bins.

Slope coefficients that Zipf tended to treat being close to -0.5, in fact, vary largely from language to language, and corpus to corpus, ranging from -0.5 to -0.1.

Acknowledgments

This research was financed by the National Science Centre, Poland, grant number 2018/29/B/HS2/02919, and supported by the Polish Ministry of Education and Science, Project CLARIN-PL, and the NTU Digital Humanities Research Cluster.

References

Eduardo G. Altmann and Martin Gerlach. 2016. Lecture Notes in Morphogenesis. *Creativity and Universality in Language*, pages 7–26.

Konstantin Avrachenkov, Arun Kadavankandy, Liudmila Ostroumova Prokhorenkova, and Andrei Raigorodskii. 2015. PageRank in undirected random graphs. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 151–163. Springer.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1352–1362.

Francis Bond, Hitoshi Isahara, Kyoko Kanzaki, and Kiyotaka Uchimoto. 2008. Boot-strapping a WordNet using multiple existing WordNets. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

Francis Bond, Lian Tze Lim, Enya Kong Tang, and Riza Hammam. 2014. The combined Wordnet Bahasa. *NUSA: Linguistic studies of languages in and around Indonesia*, 57:83–100.

Francis Bond and Liling Tan. 2012. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). *Glottometrics*, 22(4):161–174.

Bartosz Broda, Damian Jaworski, and Maciej Piasecki. 2010. Parallel, massive processing in supermatrix - a general tool for distributional semantic analysis of corpus. volume 5, pages 373–379.

Bernardino Casas, Antoni Hernández-Fernández, Neus Català, Ramon Ferrer-i-Cancho, and Jaume Baixeries. 2019. Polysemy and Brevity versus Frequency in Language. *Computer Speech & Language*, 58:207–237.

Philip Edmonds. 2004. Lexical Disambiguation. In *Elsevier Encyclopedia of Language & Linguistics*, pages 43–62. Elsevier.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Ramon Ferrer-i-Cancho. 2016. The meaning-frequency law in Zipfian optimization models of communication. *Glottometrics*, 35:28–37.

Ramon Ferrer-i-Cancho. 2018. Optimization models of natural communication. *Journal of Quantitative Linguistics*, 25(3):207–237.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual Central Repository version 3.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 2525–2529. European Languages Resources Association (ELRA), Istanbul, Turkey.

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. Springer.

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. IEEE.

László Lovász and Peter Winkler. 1995. Mixing of random walks and other diffusions on a graph. In *Surveys in combinatorics, 1995*, pages 119–154. Cambridge University Press.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plWordNet 3.0 – a Comprehensive Lexical-Semantic Resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2259–2268.

Steven Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130.

Nornadiah Mohd Razali and Yap Bee Wah. 2011. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics Vol*, 2(1):21–33.

Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18.

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

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison Wesley Press.