Multilayer explorations of Zipf's Law in linguistic structure

Karine Megerdoomian (George Mason University)
May 2014

Abstract

Zipf's Law states that the frequency of word tokens in a large corpus of natural language is inversely proportional to the rank. Zipf's Law has typically been applied in the linguistic domain to single words in a text corpus or to the study of the distribution of letters in written text. These approaches only take into account the surface forms in textual corpora and do not consider the context of use of the words or letters, nor do they consider how the meaning or pronunciation might contribute to the overall word or letter distributions. Thus, Zipf's Law as it has currently been applied typically describes the organization of written text.

In this paper, I explore the existence of Zipf's Law at distinct layers of language phenomena to determine whether a regular distribution of usage holds at the deeper level of linguistic structure. The study performs an analysis of frequency distributions in a number of different textual corpora in Persian language, contrasting the results at the surface token level with data sets that take into account the distinct syntactic categories, compounds, affixal information, and senses of the words. The results show that Zipf's Law is indeed detected at the deeper levels of natural language structure when investigating word frequency distributions, but the Zipf distribution is not observed when considering the frequency of usage of syntactic categories in isolation.

1. Introduction

The frequency distribution of words has been the topic of study in quantitative approaches to linguistics for several decades. This distribution follows a simple mathematical form known as **Zipf's Law**, which states that the size of the rth largest occurrence of the event is inversely proportional to its rank with α close to unity: $f(r) \propto \frac{1}{r^{\propto}}$. This relationship was formulated for linguistics by George Kingsley Zipf who observed a regularity in the distribution of words in the Chinese dialect of Peiping, in Plautine Latin plays and in American English news text, whereby a few words occur with very high frequency while many words occur but rarely (Zipf, 1936, p. 40). In context of word distributions, *r* is known as the frequency rank of a word and *f(r)* is its frequency in a corpus. Zipf's Law is a **Type I power law** representing a **harmonic series**, whereby the frequencies of the words decrease proportionally so that the second most frequent word (r=2) is ½ the size of the largest, the third largest value (r=3) is 1/3 the size of the largest, etc. (Cioffi-Revilla 2008, p. 17, Cioffi-Revilla 2012, Adamic and Huberman 2002).

Analyses of Zipf's Law have targeted the surface wordforms of textual corpora without directly referencing any aspect of the words' meanings. Thus, frequency distribution analyses typically treat the words *bank* and *banks* as different words, each with its own associated frequency value. In addition, no distinction is made between the noun *bank* and the second component in the proper noun *World Bank* or the verbal meaning (e.g., *Legislators bank on $36 billion in savings from cuts to the public sector*). Finally, various senses of the word are not distinguished – frequency distributions do not take into account whether *bank* was used to refer to a financial institution or the sloping land by a body of water. This raises the question of whether the analyses of Zipf's Law capture the properties of the organization of a textual corpus or if they are indeed providing an insight into underlying properties of language and meaning?[1]

Although Zipf's original analysis included several languages, recent research has tested its validity cross-linguistically. Zipf's Law has been shown to be satisfied in English (e.g., Kucera and Francis 1967), while others have challenged the universality of the law (Gelbukh and Sidorov 2001). Ha et al (2003) argue that Chinese does not satisfy Zipf's Law since Chinese characters do not necessarily correspond to units of meaning. In contrast, Xiao (2008) claims that Chinese does exhibit Zipf's Law if the proper segmentation of characters into words (based on meaning) is performed prior to frequency distribution analysis. Calude and Pagel (2011) have observed that frequency distributions are extremely similar across languages and follow a near-Zipfian distribution. Zipf himself suggests, however, that there should be differences in word frequency results depending on whether the language under study is inflected and makes use of high degree of morphology or affixation (e.g., Turkish), or positional where functional elements are represented as distinct words (e.g., 'the' in English) (Zipf 1936, p. 254).

In this paper, I propose to investigate whether Zipf's Law holds at deeper layers of linguistic structure and within lesser studied languages. I analyze the distribution of words in Persian language by taking into account more complex units involving lexical, syntactic and semantic characteristics of words in order to investigate the influence of information content. Persian has a number of linguistic features that are relevant for our study such as productive affixation and compounding that could provide different results if analyzed at the level of surface wordforms. I apply multiple lines of analysis to determine whether the resulting frequency distributions are a plausible representation of a power law, and more specifically if they exhibit Zipf's Law. I perform linearization by applying a base-10 logarithmic transformation to both frequency and rank in order to visually test for a power law, compute a goodness of fit using

[1] Also see Piantadosi (2014) and references therein for an exploration of the influence of meaning and information content in Zipf's Law.

the Kolmogorov-Smirnov statistics (Chakravarti et al 1967), and compare the power law fit to other alternative models using the log-likelihood ratio method. The results show that the word frequency distributions exhibit Zipf's Law but a power law is not observed in the distribution of abstract syntactic categories.

2. Method of Analysis

2.1 Data Sets

In order to investigate the existence of a Zipf distribution in Persian language, I identified three main corpora of Persian, which were used to create the frequency lists for this study (Table 1).

Corpus Name	***Year***	***Domain***	***Total Terms***
Kayhan	2005	Newspaper articles	18 Million
Hamshahri	2009	Newspaper articles	124 Million
Bijankhan	2007	Various written articles	2.6 Million
Bijankhan Reduced	2007	Various written articles	860,000

Table 1 - Persian language corpora used in the study

Both Kayhan and Hamshahri are newspapers published in Iran. The Kayhan corpus was not available for this study, but a list of 10,000 top frequent words from the corpus is available online for download[2]. The Hamshahri collection consists of over 160,000 documents, annotated for 100 distinct topics at the document level (AleAhmad et al 2009, Darrudi et al 2004). The Bijankhan collection is gathered from daily news and common Persian language texts. All documents are classified based on 4,300 different subjects and all words are tagged manually with syntactic category information (Oroumchian et al 2006). There is also a reduced version of the Bijankhan corpus, which was included in the data set. The Hamshahri and Bijankhan collections were developed at the University of Tehran and are available for download[3]. The two versions of the Bijankhan collection come with different tagsets: Bijankhan corpus is annotated with a part-of-speech (POS) tagset consisting of 40 distinct categories; the reduced version of the Bijankhan corpus is instead annotated with a larger tagset consisting of 550 distinct part-of-speech categories. For the purposes of this investigation, all punctuation, document subject tags, and English language words were removed from the Bijankhan collections.

[2] www.ling.ohio-state.edu/~jonsafari/persian_nlp/kayhan_2005_wc_top10000_utf8.tsv.

[3] ece.ut.ac.ir/dbrg/hamshahri/ and ece.ut.ac.ir/DBRG/Bijankhan/, respectively.

The tagsets used to annotate the terms in the two versions of the Bijankhan corpus allow us to study the effects of considering meaning and deeper linguistic structure in computing word frequency distributions. Since this is the main purpose of the current investigation, the different properties of each tagset is described in more detail in the following section.

2.2 Word Frequency Lists

Word frequency lists are derived from each of the data sets described in Section 2.1 and Table 2 provides an overview of each list. The largest frequency list is the one derived from the Hamshahri corpus with close to 600,000 terms, and the smallest list is the one obtained from the Kayhan corpus at 10,000 terms.

No.	***Corpus Name***	***Frequency List Content***	***Term Count***
1	Kayhan	Tokens	10,000
2	Hamshahri	Tokens	599,759
3	Bijankhan	Words/Phrases	75,510
4	Bijankhan Reduced	Words/Phrases	41,183
5	Bijankhan (with tags)	Words/Phrases + POS tags (40)	82,670
6	Bijankhan Reduced (with tags)	Words/Phrases + POS tags (550)	55,689

Table 2 - Persian word frequency lists

The main distinction between the Hamshahri and Kayhan corpora on one side and the Bijankhan collections on the other is in what constitutes the term used in the frequency list formation. In both the Hamshahri and Kayhan collections, the frequency lists consist of the tokens separated by whitespace in the text. In the Bijankhan collections, on the other hand, the meaning of words is taken into account. Hence, if two consecutive tokens represent a single meaning, they are maintained as a single term in the frequency list. To illustrate, the Persian term زمین لرزه ها (*zamin larze hâ*) consists of three distinct tokens separated by whitespace. The literal translation is "earth/ground tremble PLURAL" which is translated as "earthquakes". In the Hamshahri and Kayhan corpora, the word for "earthquakes" will then be split into three distinct tokens ("earth", "tremble", and the plural marker). In the Bijankhan corpora, however, all three tokens will be maintained as a single word in the frequency list.

In addition, the Bijankhan corpus terms have been tagged for POS such as N_SING for singular noun (e.g., *house*), P for Preposition (e.g., *from*) and V_PA for past tense verb (e.g., *laughed*). This tagset includes 40 distinct tags and gives rise to the fifth frequency list in Table 2, which consists of words and phrases along with their corresponding POS tags. The Bijankhan Reduced corpus, however, uses a more elaborate set of tags (totaling 550 distinct POS tags) that closely reflect the Persian morphological or word-level structure, such as N_SING_COM_GEN (common singular noun with possessive marker) or V_PA_PRG_NEG_3 (verb, past progressive tense,

negation, inflected for 3rd person singular). This gives rise to the sixth frequency list in Table 2, consisting of words and phrases along with their associated elaborate set of POS tags.

To summarize, the Kayhan and Hamshahri corpora apply at the surface token level and do not take into account the meaning of words. For the purposes of this investigation, however, the Bijankhan corpora can be used as proxies for studying the effect of word meaning in computing frequencies by identifying phrasal elements and applying POS tags that can help disambiguate based on semantics and grammatical usage.

Tables 3 and 4 illustrate the different terms in each frequency list by comparing the top 20 terms in each file. The main distinction to note is the presence of certain affixes such as the plural markers (PL and PL.GEN) or the present tense marker (PRS) in the Kayhan and Hamshahri collections in Table 3. These affixes can be written either attached or detached in Persian writing, and thus may appear as distinct tokens in text. In contrast, these affixes do not appear in the two versions of the Bijankhan corpus because they are treated as a unit with the noun or verb that they attach to.

Kayhan		***Hamshahri***		***Bijankhan***		***Bijankhan Reduced***	
و	and	و	and	و	and	و	and
در	in	در	in	در	in	در	in
به	to	به	to	به	to	به	to
از	from	از	from	از	from	از	from
که	that	که	that	که	that	که	that
این	this	این	this	این	this	این	this
می	PRS	مي	PRS	است	is	را	DEF
را	DEF	است	is	را	DEF	است	is
با	with	را	DEF	با	with	با	with
است	is	با	with	آن	that/it	آن	that/it
های	PL.GEN	هاي	PL.GEN	براي	for	يك	a/one
برای	for	براي	for	يك	a/one	براي	for
ها	PL	آن	that/it	بر	at	خود	self
آن	that/it	ها	PL	خود	that	بر	at
کرد	did	يك	a/one	کرد	did	بود	was
یک	a/one	شود	becomes	شد	became	کرد	did
شد	became	شده	become	شده	become	ایران	Iran
خود	self	خود	self	ایران	Iran	شد	became
شود	becomes	کرد	did	بود	was	شده	become
شده	become	اي	INDEF/are	سال	year	کشور	country

Table 3 - Top 20 terms in frequency lists (terms only)

Table 4 shows the top terms in the two tagged collections. By virtue of including the distinct POS tags that represent the syntactic category (e.g., noun, verb, adjective) and the affixes (e.g., plural, present tense) of the words, these frequency lists provide a slightly disambiguated word set. For example, the term آن (*ân*) can mean either "that" as in "that day" or it can be a pronoun (meaning "it"). Although this term is ambiguous in the frequency lists in Table 3, it is disambiguated to represent the pronoun form only in Table 4, thus capturing the distinct distributions of the pronoun vs. the determiner usage in Persian text.

Bijankhan with POS tags (40)		***Bijankhan reduced with POS tags (550)***	
و|CON	and	در|P_GENR	in
در|P	in	و|CON_GCOR	and
به|P	to	به|P_GENR	to
از|P	from	از|P_GENR	from
که|CON	that	که|CON_RELC	that
این|DET	this	این|DET	this
است|V_PRE	is	را|P_DEFI	DEF
را|P	DEF	و|CON_GMC	and
با|P	with	است|V_PRE_SIM	is
براي|P	for	با|P_GENR	with
يك|N_SING	a/one	براي|P_GENR_GEN	for
آن|PRO	it	يك|N_SING_CN	one
خود|PRO	self	خود|PRO_DEF_R_XOD	self
بر|P	at	بود|V_PA_SIM_POS_3	was
كرد|V_PA	did	كرد|V_PA_SIM_POS_3	did
شد|V_PA	became	بر|P_GENR	at
ايران|N_SING	Iran	شد|V_PA_SIM_POS_3	became
بود|V_PA	was	ايران|N_SING_LOC_PR	Iran
سال|N_SING	year	است|V_PRE_SIM_3	is
كشور|N_SING	country	گفت|V_PA_SIM_POS_3	said

Table 4 - Top 20 terms for frequency lists (term+tag)

2.3 Tag Frequency Lists

If Zipf's Law is a property of word usage, it may follow that the ranked order of POS frequencies without associated words will also follow a Zipf distribution. Table 5 describes the frequency lists created based on the POS tags and affixes in the two Bijankhan collections.

No.	*Corpus Name*	*Corpus Word Count*	*Frequency List Content*	*Term Count*
1	Bijankhan (with tags)	2,597,937	POS tags (40)	38
2	Bijankhan Reduced (with tags)	858,584	POS tags (550)	454
3	Bijankhan Reduced (with tags)	858,584	Affixes	18

Table 5 - Persian Frequency lists (tags only)

The first list is derived from the larger annotated set (2.6 million words) tagged by the 40 POS tags. The resulting frequency list includes the top POS categories found. Similarly, the frequency of occurrence of the POS categories from the 550 tagset was computed on the reduced Bijankhan corpus, giving rise to the second frequency list. Finally, the third list consists only of the frequency of occurrence of affixes such as plurals, negation, and attached pronouns. Table 6 illustrates the top five categories in these three frequency lists.

Bijankhan with 40 tag set	Bijankhan Reduced with 550 tag set	Affixes (derived from 550 tagset)
N-SING	N_SING_COM_GEN	SG (singular)
P	N_SING_COM	GEN (genitive/possessive)
ADJ_SIM	P_GENR	3SG (3rd person singular)
CON	ADJ_SIM	PL (plural)
N_PL	N_PL_COM_GEN	PST (past tense)

Table 6 - Top 5 terms in the tag frequency lists

Since these POS tags are representing syntactic category and affixation for the two Bijankhan collections, they are able to capture underlying sense distinctions and could therefore be used to investigate whether Zipf's Law holds at deeper layers of meaning.

2.4 Methodology

In order to determine whether the frequency distribution of words and tagsets in Persian conforms to a Zipfian power law, I start by plotting a ranked distribution (frequency of occurrence vs. rank) starting from most frequent term to least frequent. If the data are following a Zipfian distribution, the resulting histogram should follow a straight line on a doubly logarithmic plot.

To further analyze the properties of the frequency distributions of the words and tagsets in Persian corpora, I closely follow the methodology provided in Clauset et al (2009) and use the PowerLaw package provided in Python (Alstott et al 2014) to identify the following:

(i) Estimate the parameters for x_{min} and α of the power-law model. The parameter x_{min} provides a value for the lower bound of the power law behavior in the given data set above which the tail of the distribution is fitted to a power law. The PowerLaw

package uses the maximum likelihood method to obtain a value for the scaling parameter α, as well as for its standard error σ. The value of x_{min} is selected to minimize the Kolmogorov-Smirnov distance between the data points and the power law fit.

(ii) Calculate the goodness-of-fit between the Persian frequency data and the power law using the Kolmogorov-Smirnov statistic. **KS distance** is used to estimate the parameters of the power law that best fit the specified distribution.

(iii) Compare the power law hypothesis with alternative ones via a likelihood ratio test. The likelihood ratio test computes the likelihood of the given data set under two competing distributions and outputs the one with the higher likelihood. This method allows us to determine whether the data set is plausibly drawn from a power law distribution.

3. Results and Findings

3.1 Histograms

Figure 1 shows the frequency distributions of the various collections plotted on base-10 logarithmic axes. The dashed lines in each instance were plotted using the value estimated for α that best fits a power law distribution, using the Kolmogorov-Smirnov method (see Table 5 in Section 3.2).

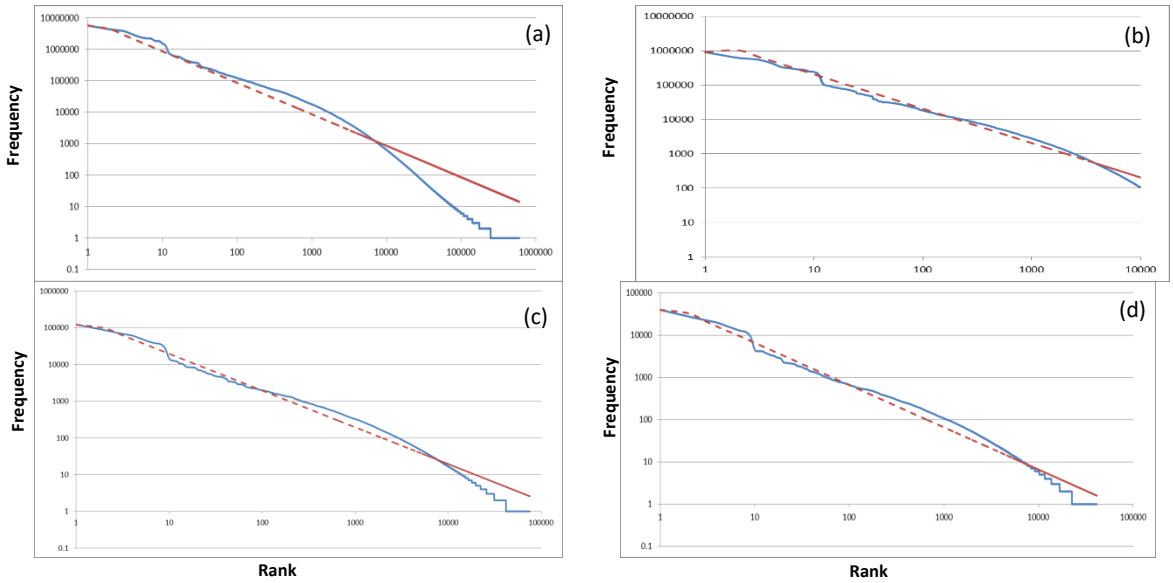

Figure 1 - Fitted power law distributions of word frequency (y-axis) against rank (x-axis) for a) Hamshahri, b) Kayhan, c) Bijankhan, and d) Bijankhan reduced corpora, plotted on a doubly log scale. Dotted lines represent best fit for a Zipfian distribution. The Zipf rank parameter α for the fitted plot is (a) 1.52, (b) 2.26, (c) 1.61, (d) 1.66.

As can be seen from these histograms, the collections seem to be a plausible match for a Zipfian distribution regardless of the linguistic content of the terms, i.e., tokens as in Hamshahri and Kayhan, or words and phrases as in the two versions of the Bijankhan corpus. In the case of the Hamshahri corpus, which is the largest data set, there seems to be a clear drop in the higher ranks (above 10,000). This drop after 10,000 ranks is also visible in the smaller corpora, although to a lesser degree. This is reminiscent of the effect previously observed (Ferrer i Cancho, 2001) and will be further discussed in Section 4. It is interesting to note that the relative shape of the frequency distributions for these four collections are quite similar (even including the squiggle at about rank 10) as shown in Figure 2, even though each distribution appears at a different scale.

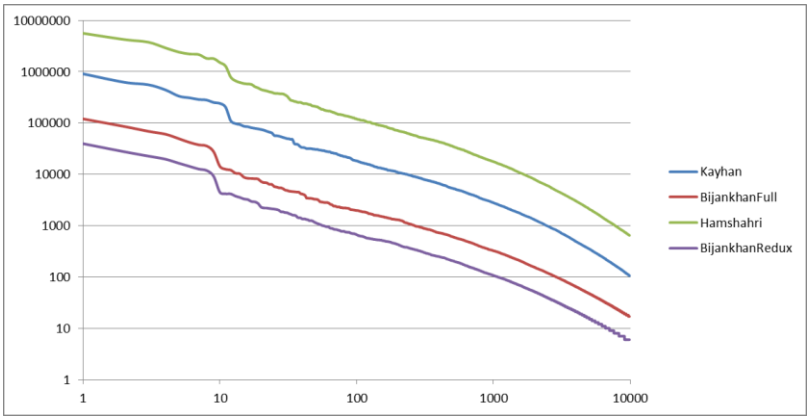

Figure 2 - Log-log plot of the frequency vs. rank distributions of the four Persian collections display similar relative shapes at different scales.

The two other variants of the Bijankhan corpus that included the words with their associated POS tags were also investigated. These frequency sets manage to capture some of the distinctions in word meaning that are ignored in the collections plotted in Figure 1, since the latter do not take into account the syntactic category and morphological information of the terms found in the text corpus. Figure 3 illustrates the histograms for these two collections. As can be seen from the plots, a similar power law fit emerges in these instances.

In fact, a comparison of the Bijankhan corpus set with tags and without tags (for the full and reduced versions) shows very similar results. This is illustrated in Figure 4 where, though data points may vary, the overall histograms seem to match.

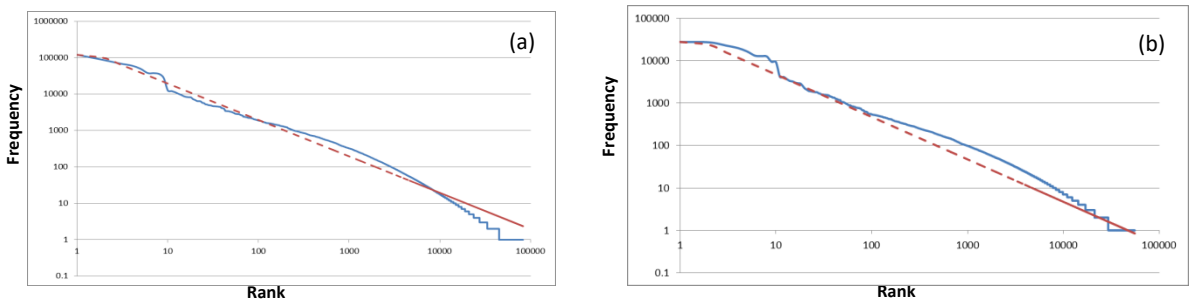

Figure 3 - Fitted power law distributions of word frequency against rank for a) Bijankhan collection with tagset containing 40 distinct POS tags, and b) Reduced Bijankhan collection with tagset containing 550 distinct tags with POS and morphology information, plotted on a doubly log scale. Dotted lines represent best fit for a Zipfian distribution with α = (a) 1.62, (b) 1.71.

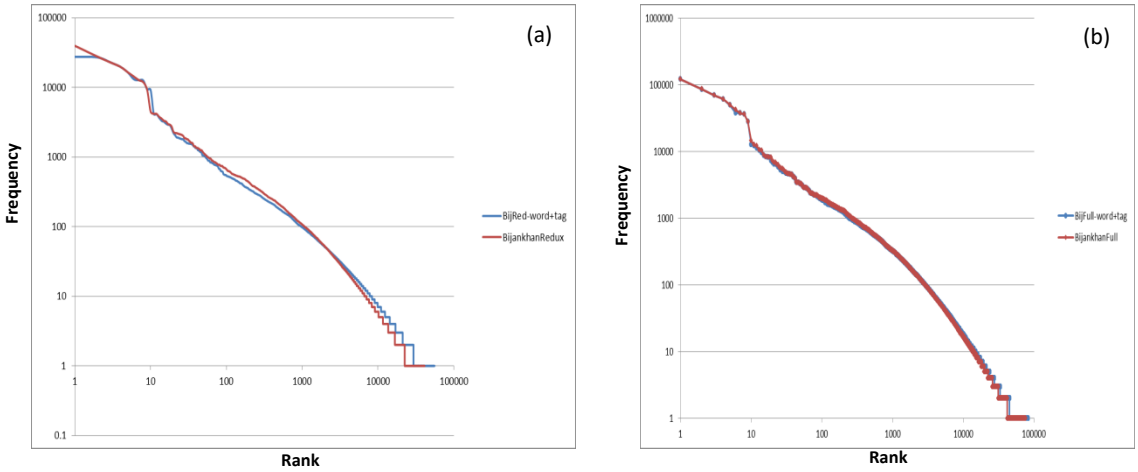

Figure 4 – Log-log plots of word frequency against rank for a) Bijankhan collection with and without tagset, and b) Reduced Bijankhan collection with and without tagset.

Next, I considered the frequency distribution plots of the tagsets described in Section 2.3. The graphs in (a) below represent the frequency of only the syntactic categories and affix information for Persian words encountered in the Bijankhan text collections, while the plot in (b) is the frequency distribution of only the affixes in the reduced Bijankhan corpus. Neither of these data sets seem to be a plausible fit for a power law distribution.

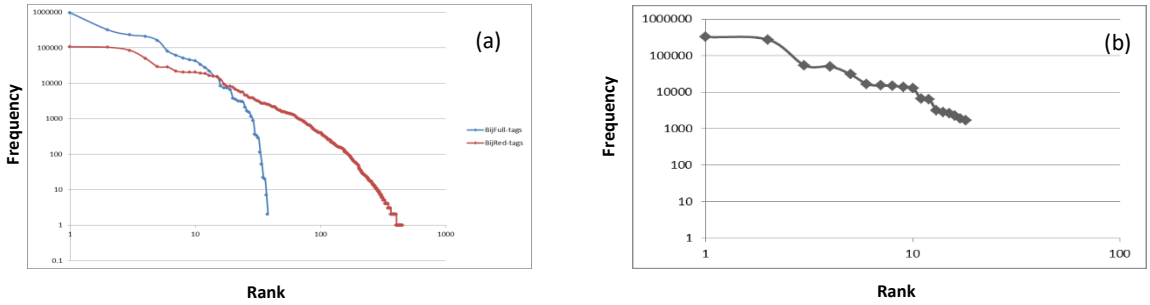

Figure 5 – Log-log plot of word frequency against rank for a) Bijankhan tagsets containing 40 (blue line) and 550 (red line) distinct tags, and b) affixes.

3.2 Statistical Fitting of the Distributions

Applying the Python PowerLaw package to the various data sets, I estimate the lower bound x_{min} value, the Zipf rank parameter α, and its standard error σ for a good fit to a power law. In addition, the package computes the Kolmogorov-Smirnov distance *D* which provides the optimal fit between the data and the power law distribution. The results for all 9 frequency list distributions are provided in Table 5, where *n* represents the total number of terms in the frequency list. The highest KS distance values are provided for the last three data sets which include only the POS tags or affix information, without actual words. These results are in line with the histogram plots discussed in the previous section, suggesting that the POS tags and Affixes lists are not a good fit to a power law, while the other data sets seem to be a plausible fit. The only distinction is perhaps the Kayhan corpus which seems to be a plausible fit only after the high rank of 4,774.

#	*Frequency List*	*n*	x_{min}	α	σ	*D*
1	Hamshahri	124,090,827	7	1.518770557	0.001656349	0.015750086
2	Kayhan	18,219,272	4774	2.264585873	0.052328936	0.029206956
3	Bijankhan	2,339,411	2	1.610369551	0.002984772	0.014926508
4	Bijankhan Reduced	774,202	2	1.656576702	0.004378443	0.024557747
5	Bijankhan with tags	2,339,411	2	1.623240534	0.002937266	0.01411788
6	Bijankhan Reduced with tags	774,202	2	1.709441747	0.004139731	0.021960147
7	POS tags (40)	2,339,411	21900	1.709098641	0.196668578	0.14426131
8	POS tags (550)	774,201	992	1.688001457	0.083432432	0.066441156
9	Affixes	834,228	13485	1.886742319	0.295580731	0.166137583

Table 5 – Estimated results for a power law fit for the Persian text collections (*n* = total number of words, x_{min} = lower bound value, α = Zipf rank parameter, σ = standard error for α, *D* = Kolmogorov-Smirnov distance*).*

In order to determine whether the various distributions are a plausible fit to the power law, I perform a direct comparison of the models. The results are summarized in Table 6. I use the

likelihood ratio test which provides the LR or logarithm $\mathcal{R}$ of the ratio of the power law with respect to another distribution type. This value is positive or negative depending on which distribution is better – if the value is positive, then the data is likely to be a good fit to the power law and if it is negative, then the data might be a better fit to the competing distribution. The higher the value of the LR, the more likely the fit. In addition to the sign of the LR, it is important to take the value of *p* into consideration. The *p*-value associated with the LR tells us whether the observed sign of $\mathcal{R}$ is statistically significant. As described in Clauset et al (2009), if the *p*-value is sufficiently small ($p < 0.1$) then "it is unlikely that the observed sign is a chance result of fluctuations and the sign is a reliable indicator of which model is the better fit to the data". Meanwhile, if the *p*-value is large then it suggests that the observed sign is not reliable, and the test does not favor either model over the other.

#	*Frequency List*	*Log-normal*		*Exponential*		*Stretched Exponential*	
		LR	***p***	***LR***	***p***	***LR***	***p***
1	Hamshahri	-91.48078182	**2.00E-21**	242811.9222	**0**	523543.0134	**1.11E-277**
2	Kayhan	-0.400372747	0.582964	363.1157507	**3.83E-07**	1694.337718	**0.080678**
3	Bijankhan	-300.3960916	**1.95E-62**	56160.26293	**7.33E-128**	121715.8402	**4.35E-63**
4	Bijankhan Reduced	-238.2282324	**1.30E-43**	24702.63686	**3.91E-43**	51653.0662	**5.29E-27**
5	Bijankhan with tags	-296.1885018	**3.05E-62**	59092.0889	**2.88E-137**	124502.7709	**6.59E-66**
6	Bijankhan Reduced with tags	-288.5479342	**2.60E-49**	28861.70712	**4.20E-43**	61150.13987	**4.88E-31**
7	POS tags (40)	-0.58422245	0.489025	2.334801656	0.484645	2.381831994	0.478001
8	POS tags (550)	-1.221798009	0.284652	31.74522374	**0.002645**	31.82001874	**0.002012**
9	Affixes	0.0139624241	0.864194	4.184437218	0.1154654	4.243632600	0.1557768

Table 6 - Comparison of models using the likelihood ratio test. Bolded *p*-values indicate statistical significance.

The results comparing the word frequency lists to an exponential or a stretched exponential suggest that the data are a plausible fit to the power law distribution. The only LR value that is small is the Kayhan frequency list, compared to the other collections, yet it is still pointing to a power law as the most likely fit with statistical significance. On the other hand, comparison to log-normal seems to favor the log-normal distribution to the power law. However, the values of LR are rather small. I will return to this result in Section 4.

In considering the results for the POS tags and affix frequency distributions, we do not seem to have any statistically significant results to determine a best fit. The only small *p*-values obtained are the ones for the larger tag set of about 550 distinct POS and affix tags, which might suggest a better fit to the power law, though the LR values are very low to be reliable.

4. Discussion

The goal of the study was to see whether Zipf's Law holds at deeper layers of linguistic structure that consider differences in meaning and syntactic category, and within a language such as Persian that exhibits productive morphological affixation and compounding. The results clearly show that frequency distribution displays a robust regularity in accordance with Zipf's Law when surface word forms or deeper linguistic units are considered. However, I was not able to identify power law behavior when studying the distribution of abstract syntactic categories such as parts-of-speech and affixes in isolation.

The results are surprising since I expected to find a difference between frequency distributions obtained based on surface word forms or tokens, and distributions derived from data sets that make reference to meaning and syntactic category. The results, however, suggest that the power law behavior observed in word frequency distributions is robust across data sets and might be derived from principles independent of the content or meaning of language.

It is also interesting to note that although power law behavior is obtained for single tokens (Hamshahri and Kayhan), for words and phrases (Bijankhan collections), and words and phrases distinguished based on some level of meaning and syntactic category (Bijankhan collections tagged with POS and affix information), we fail to observe any power law characteristics when we consider the parts-of-speech categories or the affixes in isolation (i.e., without any words). This suggests that the regularity in the organization of language reflects information at the word level and not based on the abstract levels of syntactic category such as singular noun or past tense verb.

It is important to note that the result of the likelihood ratio test against alternative models shown in Table 6 suggested that lognormal might be a plausible alternative for the Persian word frequency distributions. It is a well-known observation that bending can be found in empirical data that approximate a power law, but that it could also reflect a lognormal tail (Cioffi 2008, p. 22). It has also been argued, however, that lognormal distributions with large variance also yield straight lines when plotted on the log-log rank-frequency graph (Downey 2001). Based on this finding, Zhao and Marcus (2012) argue that the straight line considered to be the signature of a power law may suggest a lognormal distribution instead.

Thus, although lognormal might be a potential fit for the frequency distribution observed in our data, Clauset et al (2009) also warn that "in general, we find that it is extremely difficult to tell the difference between log-normal and power-law behavior. Indeed over realistic ranges of α the two distributions are very closely equal, so it appears unlikely that any test would be able to tell them apart unless we have an extremely large data set." (Clauset et al 2009, p. 26). Nevertheless, based on the fact that we are studying a ranked discrete data set that starts at 1,

we don't expect to see lognormal distribution behavior. In addition, the values suggesting a lognormal fit in Table 6 have rather low LR values (though with high significance p-values) compared to the robust power law fit LR values for the same set of corpora.

One observation that is interesting is the deviation observed in Figure 1a of the Hamshahri corpus, where there seems be a distinct bending corresponding to the high rank (low frequency) events. This deviation is similar to an "exponential cutoff" and has sometimes been explained to be the result of the finite nature of the events in consideration. However, comparing this bending effect to the results from the other corpora in Figure 1 shows a correlation where the drop-off occurs approximately after r=10,000 in all plots. Similar deviations have been noted in the literature for large data sets across languages (Ferrer i Cancho and Solé 2001, Tripp and Feitelson 2007). Solé (2008), based on Ferrer i Cancho and Solé (2001), explains this deviation by proposing two distinct power law regimes where the rate at which the frequency of word distribution decreases with rank changes through two scaling regimes – a slower decrease followed by a fast one after the critical rank point[4]. Ferrer i Cancho and Solé (2001) argue that this change in scaling reflects the distinction in the lexicons where the high frequency words form a *kernel lexicon* consisting of the common and versatile words, and a specific lexicon consisting of a large number of domain-specific terms. A closer (but qualitative) look at the actual words in the Hamshahri frequency list shows that although some of the words in this range are in fact rare and literary or domain-specific terms, many of the words are misspellings or distinct words that have been attached to each other[5].

A more striking deviation is observed in the left half of all word frequency plots in Figure 2, where each plot experiences a squiggle at about the same range, representing a strong drop in frequency on the log-log scale. There is clearly an important correlation across all data sets regardless of whether they take syntactic categories into account or if they operate on surface tokens only, indicating statistical irregularities that might be important to understand. One hypothesis might be that this is the range where the very common words that are mainly function words consisting of articles, prepositions and conjunctions are giving way to content words such as nouns and adjectives. This hypothesis is supported by the fact that the drop in log of frequency seems to occur in the Bijankhan collections prior to the one in the token-based corpora (namely, Hamshahri and Kayhan). This would be expected since the Bijankhan collections contain less separated affixes such as the plural or present tense markers, and

[4] Although the critical rank in the English data set seems to occur around 5,000-6,000 words.

[5] This is a common occurrence in Persian text since whitespace is often optional in writing and two distinct words may appear next to each other without an intervening space. These terms are not listed in the Bijankhan collection since the latter has been manually vetted and annotated.

therefore the substantive or content words should have higher relative frequency. This hypothesis needs to be empirically tested and I will leave this for further study.

The study can be further extended to investigate the influence of meaning on word frequency distribution in a corpus. There are minor changes that can be considered, such as taking into account compound verbs that consist of a noun, adjective or prepositional phrase combined with a verb. These constructions are very common in Persian but were not tagged as single units in the Bijankhan collections. These include entries such as *kâr kardand* ('work did.3pl') meaning "(they) worked" or *dar nazar gereftan* ('in observance take') meaning "(to) consider". However, I do not expect the inclusion of these compounds to significantly modify the results. More systematic extensions would include taking into account bigger syntactic constituents such as Noun Phrases (e.g., 'the big house' or 'DET ADJ NOUN').

5. Summary

The empirical law uncovered and popularized by Zipf reveals regular statistical distributions in human language indicating that word frequency distribution in a text corpus follows Zipf's Law whereby only a small number of words have high frequency while a large number of words in the language have low frequency of occurrence. Most applications of Zipf's Law, however, do not consider the information content or meaning of the words, but rely solely on the surface written wordforms. In this paper, I studied the frequency distribution of words and abstract syntactic and meaning categories in Persian text to investigate the influence of meaning on the realization of Zipf's Law. The results show that the frequency distribution of abstract syntactic and morphological categories does not exhibit Zipf's Law. However, Zipf's Law is exhibited across different layers of linguistic and meaning structure, confirming the universal statistical behavior previously noted across languages.

Bibliography

Adamic, L., & Huberman, B. (2002). Zipf's Law and the Internet. *Glottometrics, 3*, 143-150.
AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., & Oroumchian, F. (2009). Hamshahri: A Standard Persian Text Collection. *Journal of Knowledge-Based Systems, 22*(5), 382-387.
Alstott, J., Bullmore, E., & Plenz, D. (2014). Powerlaw: a Python Package for Analysis of Heavy-Tailed Distributions. *PLoS ONE, 9*(1).
Calude, A., & Pagel, M. (2011). How do we use language? shared patterns in the frequency of word use across 17 world languages. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*(1567), 1101-1107.
Chakravarti, I., Laha, R., & Roy, J. (1967). *Handbook of Methods of Applied Statistics, Volume I: Techniques of Computation, Descriptive Methods and Statistical Inference.* New York: John Wiley & Sons.

Cioffi-Revilla, C. (2008). *Power Laws and Non-Equilirium Distributions of Complexity in the Social Sciences.* George Mason University: Unpublished manuscript. Retrieved from http://socialcomplexity.gmu.edu
Cioffi-Revilla, C. (2014). *Introduction to Computational Social Science: Principles and Applications.* London: Springer-Verlag.
Clauset, A., Shalizi, C., & Newman, M. (2009). Power-law Distributions in Empirical Data. *SIAM Review, 51*(4), 661-703.
Darrudi, E., Hejazi, M., & Oroumchian, F. (2004). Assessment of a Modern Farsi Corpus. *Proceedings of the 2nd Workshop on Information Technology and its Disciplines.* Kish Island: ITRC.
Downey, A. B. (2001). The Structural Cause of File Size Distributions. *Proceedings of the Ninth International Symposium in Modeling, Analysis and Simulation of Computer and Telecommunication.* Washington, DC.
Ferrer i Cancho, R. a. (2001). Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law revisited. *Journal of Quantitative Linguistics, 8*(3), 165-173.
Gelbukh, A. a. (2001). Zipf and Heaps Laws' Coefficients Depend on Language. *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics*, (pp. 332-335).
Ha, L. E.-G. (2003). Extension of Zipf's Law to Words and Phrases. *Proceedings of the 19th International Conference on Computational Linguistics*, (pp. 315-320).
Kucera, H. a. (1967). *Computational Analysis of Present-Day American English.* Providence: Brown University Press.
Oroumchian, F., Tasharofi, S., Amiri, H., Hojjat, H., & Raja, F. (2006). *Creating a Feasible Corpus for Persian POS Tagging.* Dubai: University fo Wollongong in Dubai.
Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: a critical review and future directions. *Psychonomic Bulletin & Review*.
Solé, R. V. (2008). Scaling Laws in Language Evolution. In C. Cioff-Revillai, *Power Laws and Non-Equilibrium Distributionsof Complexity in the Social Sciences* (pp. 56-65). George Mason University: Unpublished manuscript. Retrieved from http://socialcomplexity.gmu.edu
Tripp, O., & Feitelson, D. (2007). *Zipf's Law Revisited.* School of Computer Science and Engineering. Jerusalem: The Hebrew University of Jerusalem.
Xiao, H. (2008). On the Applicability of Zipf's Law in Chinese Word Frequency Distribution. *Journal of Chinese Language and Computing, 18*(1), 33-46.
Zhao, Q., & Marcus, M. (2012). Long-tail Distributions and Unsupervised Learning of Morphology. *Proceedings of COLING*, (pp. 3121-3136). Mumbai.
Zipf, G. (1936). *The Psycho-Biology of Language.* London: Routledge.