A Three-Parameter Rank-Frequency Relation in Natural Languages

Chenchen Ding, Masao Utiyama, Eiichiro Sumita
Advanced Translation Technology Laboratory,
Advanced Speech Translation Research and Development Promotion Center,
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{chenchen.ding, mutiyama, eiichiro.sumita}@nict.go.jp

Abstract

We present that, the rank-frequency relation in textual data follows $f \propto r^{-\alpha}(r+\gamma)^{-\beta}$, where f is the token frequency and r is the rank by frequency, with (α, β, γ) as parameters. The formulation is derived based on the empirical observation that $d^2(x+y)/dx^2$ is a typical impulse function, where $(x, y) = (\log r, \log f)$. The formulation is the power law when $\beta = 0$ and the Zipf–Mandelbrot law when $\alpha = 0$. We illustrate that α is related to the analytic features of syntax and $\beta+\gamma$ to those of morphology in natural languages from an investigation of multilingual corpora.

1 Introduction

Zipf's law (Zipf, 1935, 1949) is an empirical law to formulate the rank-frequency (r-f) relation in physical and social phenomena. Linguistically, Zipf's law can be observed on the distribution of words in corpora of natural languages, where the frequency (f) of words is inversely proportional to its rank (r) by frequency; that is, $f \propto r^{-1}$. Zipf's law is a special form of a general power law, that is, $f \propto r^{-\alpha}$, with $\alpha = 1$.

The Zipf's/power law is usually examined under a log-log plot of rank and frequency, where the data points lie on a straight line. The simple proportionality of the Zipf's/power law can be observed on randomly generated textual data (Li, 1992) and it only roughly depicts the r-f relation in real textual data. A two-parameter generalization of the Zipf's/power law is the Zipf-Mandelbrot law, where $f \propto (r+\beta)^{-\alpha}$ (Mandelbrot, 1965). Li et al. (2010) considered the *reversed rank* of $r_{max}+1-r$, where r_{max} is the maximum of ranking index, and proposed a two-parameter formulation of $f \propto r^{-\alpha}(r_{max}+1-r)^{\beta}$.

As a straightforward observation, the coefficients of proportionality should be distinguished for common and rear words (Powers, 1998; Li et al., 2010). Therefore, an extension of the original Zipf's/power law requires at least two parameters. In this study, a three-parameter formulation of $f \propto r^{-\alpha}(r+\gamma)^{-\beta}$ is derived based on the observation and analysis of multilingual corpora. It is a natural generalization of the power law and the Zipf-Mandelbrot law. The third parameter provides a depiction of the *rigidness* of different coefficients of proportionality. The proposed formulation can also fit non-Zipfian phenomena in natural languages, such as the r-f relation on Chinese characters. Figure 1 shows examples on English words from *Europarl* (Koehn, 2005) [1] and Chinese characters of *Academia Sinica* from the data of Sproat and Emerson (2003).[2]

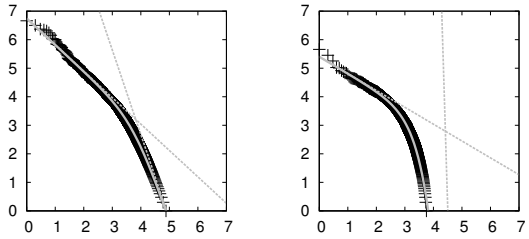

Figure 1: Rank-frequency plots on English words (left) and Chinese characters (right). The x- and y-axes are $\log_{10} r$ and $\log_{10} f$, respectively. The gray curves are the proposed formulation under logarithm: $y = C - \alpha x - \beta \log_{10}(10^x + 10^\gamma)$, where C is a constant. The dashed lines are the asymptotes of $C-(\alpha x+\beta\gamma)$ and $C-(\alpha+\beta)x$. (α, β, γ) is $(0.93, 2.04, 3.82)$ for English words and $(0.59, 32.31, 4.42)$ for Chinese characters.

2 Proposed and Related Formulation

Under a logarithmic form, the Zipf's law states that $x+y=C$, where $(x, y) = (\log r, \log f)$, and C is roughly a constant. We further investigate the

[1] http://www.statmt.org/europarl/v8/europarl.tgz

[2] http://sighan.cs.uchicago.edu/bakeoff2005/data/icwb2-data.zip

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 460–464
July 5 - 10, 2020.

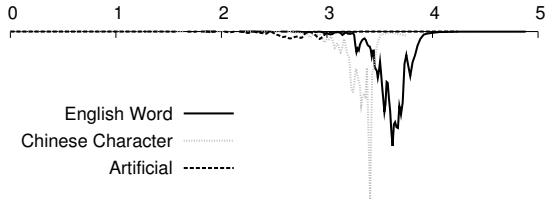

Figure 2: Smoothed second-order differences on the rank-frequency relation. The x-axis is $\log_{10} r$.

property of $C = g(x)$. The first and second-order differences on $g(x)$ are calculated as

$$g'_i = \frac{g_i - g_{i-1}}{x_i - x_{i-1}}, \qquad g''_i = \frac{g'_i - g'_{i-1}}{x_i - x_{i-1}}. \qquad (1)$$

Here (x_i, y_i) is the data point of the i-th frequent token, $g_i = x_i + y_i$ for $i > 1$, and $g'_1 = g''_1 = 0$.[3] Because the differences are intrinsically non-smooth, Bézier curves are applied for smoothing in the investigation.

Figure 2 shows examples of the smoothed g'' on English words and Chinese characters from the same dataset used for Fig. 1. An artificial Zipfian dataset generated in the manner of Li (1992)[4] is also used for comparison. It can be observed that the g'' on English words and Chinese characters has an impulse, but not that on the artificial data. Generally, the impulse becomes more obvious if the data are more non-Zipfian.

If we consider g'' as a general impulse function, then g' is a general *sigmoid* function and g can be modeled by a general *softplus* function in the form of $b\log(\exp(x-c)+1)$. To replace x by a generalized linear form as $ax + d$,

$$y = -d - ax - b\log(\exp(x - c) + 1) \qquad (2)$$

and to substitute (x, y) by $(\log r, \log f)$, we obtain,

$$f = \frac{\exp(bc - d)}{r^a(r + \exp(c))^b} \propto r^{-\alpha}(r + \gamma)^{-\beta}, \qquad (3)$$

where $(\alpha, \beta, \gamma) = (a, b, \exp(c))$. $\exp(bc - d)$ is a constant unrelated to r.

The obtained proportional form is a natural two-component extension of the power law and the Zipf-Mandelbrot law. Because the *softplus* function is a differentiable form of a rigid *ramp* function, Eq. (3) can also be considered as a smoothed piecewise *broken power law*. As shown in Fig. 1, α and $(\alpha + \beta)$ depict the proportional coefficients at the two ends, and the proportional coefficients are switched smoothly around $x = \gamma$.

$f \propto r^{-\alpha}(r_{max} + 1 - r)^{\beta}$ proposed in Li et al. (2010) is also a two-component formulation. One more parameter (i.e., γ) in Eq. (3) is used to identify the location of the impulse observed in g''. Under Li's formulation, we obtain $g = y + \alpha x = \beta\log(r_{max} + 1 - \exp(r))$ and $g'' = -C_1\exp(x)(C_2 - \exp(x))^{-2}$, where C_1 and C_2 are constants. g'' is a monotonically decreasing function with $x = \log(C_2)$ as the asymptote for $x < \log(C_2)$. Therefore, Li's formulation always has a steep tail and lacks the capacity to depict the switching of two stable proportional coefficients. Figure 3 shows examples using Li's formulation to fit data in Fig. 1. It can be observed that the non-Zipfian Chinese characters are fitted well, but not for the tail part in more Zipfian English words. This can be explained from the shape of g'' in Fig. 2. It is reasonable to model the g'' of Chinese characters using a monotonically decreasing function because the γ in Eq. (3) is quite large (around r_{max}). However, it is not proper for English words, where a proper γ is required.

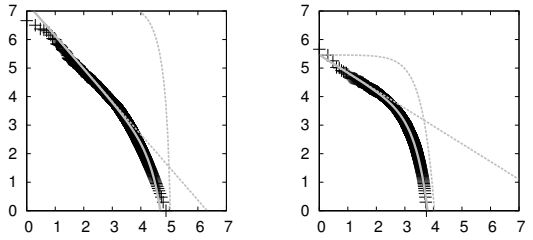

Figure 3: English word (left) and Chinese character (right) data in Figure 1 fitted by the gray curve of $y = C - \alpha x + \beta\log_{10}(r_{max} + 1 - 10^x)$. The dashed lines are of $C - (\alpha x + \beta\log_{10}(r_{max} + 1))$ and $C - \beta\log_{10}(r_{max} + 1 - 10^x)$ for two ends. (α, β) is $(1.15, 9.16)$ for English words and $(0.62, 157.13)$ for Chinese characters.

Based on the analysis, it can be concluded that the formulation $f \propto r^{-\alpha}(r + \gamma)^{-\beta}$ is a generalized form that covers the Zipf's/power law, Zipf-Mandelbrot law, piecewise broken power law, and Li's two-parameter formulation. In the next section, we show the linguistic interpretation of the parameter (α, β, γ).

[3] To avoid too many meaningless zeros in the differences, only the data point with the minimum x is used for data points with the same y, i.e., tokens with the same frequency.

[4] Two letters `a` and `b` are used. The frequency of `a`, `b`, and space is $3:1:1$, and 10^7 characters are randomly generated.

	α	β	γ	$\frac{\gamma}{r_{max}}$
bg	$0.92_{\pm.00}$	$2.05_{\pm.06}$	$4.25_{\pm.02}$	0.85
cs	$0.86_{\pm.00}$	$1.20_{\pm.01}$	$3.89_{\pm.01}$	0.74
da	$0.99_{\pm.00}$	$1.10_{\pm.01}$	$3.85_{\pm.01}$	0.69
de	$0.99_{\pm.00}$	$1.08_{\pm.01}$	$3.94_{\pm.01}$	0.70
el	$0.98_{\pm.00}$	$1.96_{\pm.03}$	$4.43_{\pm.01}$	0.82
en	$0.93_{\pm.00}$	$2.04_{\pm.01}$	$3.82_{\pm.00}$	0.75
es	$0.94_{\pm.00}$	$1.38_{\pm.01}$	$3.82_{\pm.01}$	0.73
et	$0.90_{\pm.00}$	$1.06_{\pm.01}$	$4.13_{\pm.01}$	0.75
fi	$0.87_{\pm.00}$	$0.89_{\pm.01}$	$4.07_{\pm.01}$	0.70
fr	$1.01_{\pm.00}$	$2.05_{\pm.02}$	$4.14_{\pm.01}$	0.80
hu	$0.92_{\pm.00}$	$0.96_{\pm.02}$	$4.16_{\pm.02}$	0.76
it	$0.94_{\pm.00}$	$1.47_{\pm.01}$	$3.84_{\pm.00}$	0.73
lt	$0.84_{\pm.00}$	$1.04_{\pm.01}$	$3.77_{\pm.01}$	0.70
lv	$0.87_{\pm.00}$	$1.69_{\pm.02}$	$4.22_{\pm.01}$	0.81
nl	$0.98_{\pm.00}$	$1.18_{\pm.01}$	$3.73_{\pm.01}$	0.68
pl	$0.87_{\pm.00}$	$1.18_{\pm.01}$	$3.97_{\pm.01}$	0.76
pt	$0.93_{\pm.00}$	$1.33_{\pm.01}$	$3.77_{\pm.01}$	0.72
ro	$0.94_{\pm.00}$	$5.24_{\pm.32}$	$4.78_{\pm.03}$	0.97
sk	$0.89_{\pm.00}$	$1.38_{\pm.01}$	$4.14_{\pm.01}$	0.79
sl	$0.91_{\pm.00}$	$1.77_{\pm.04}$	$4.31_{\pm.01}$	0.84
sv	$0.99_{\pm.00}$	$1.05_{\pm.01}$	$3.86_{\pm.01}$	0.70

Table 1: Fitted parameters on Europarl data.

3 Experiment and Discussion

We used the proposed formulation to fit data of various European languages and typical Asian languages. The *Europarl* corpus (Koehn, 2005) and data from the *Second International Chinese Word Segmentation Bakeoff (ICWB2)* (Sproat and Emerson, 2003) were mentioned in Section 1. We also used English-Japanese patent data from the *7th NTCIR Workshop* (Fujii et al., 2008). The Europarl data and English data from NTCIR were lower-cased and tokenized using the toolkit provided by MOSES[5] (Koehn et al., 2007). Fitting was performed under a logarithmic scale using the `fit` function[6] in `gnuplot`.[7] Specifically, relation-frequency data were used to fit (α, β, γ) and C in $y = C - \alpha x - \beta \log_{10}(10^x + 10^\gamma)$. For the initialization, $(\alpha, \beta, \gamma) = (1, 1, \frac{r_{max}}{2})$ and $C = 3\gamma$ were applied.

Table 1 lists the fitting results for all the languages[8] in the Europarl corpus. The (α, β, γ) with the asymptotic standard error ($\pm$) are listed. Because γ may depend on the vocabulary size, normalized $\gamma_{norm} = \frac{\gamma}{r_{max}}$ is also listed. It can be observed that all the language data were fitted well with an α of around 1.0, which is in accordance with the original Zipf's law. β and γ_{norm} for each language are plotted on the left of Fig. 4.[9] On the β-γ_{norm} plane, we can observe the rough tendency that β and γ_{norm} are linear, in addition to a separation for different language branches. Further principal component analysis on $(\alpha, \beta, \gamma_{norm})$ suggests that α and $\beta + \gamma_{norm}$ can be generally considered as two dominant components.[10] The plot on the right of Fig. 4 shows that the language branches can be separated roughly by lines parallel to the axes of α and $\beta + \gamma_{norm}$. This indicates the linguistic explainability of the two axes.

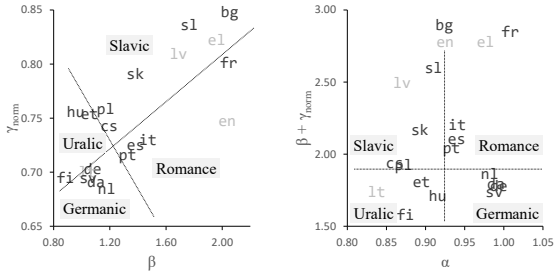

Figure 4: Distribution of languages in Europarl.

From the nature of these languages, we consider that α can be explained as an axis of analysis-synthesis on syntax and $\beta + \gamma_{norm}$ as that on morphology. A large α suggests a couple of extremely frequent words in the corpus. As typical examples, languages with a relatively large α, that is, Romance and Germanic, generally contain abundant prepositions, particles, and determiners to mark syntactic roles, whereas those with a smaller α, that is, Slavic and Uralic, tend to use complex declension and conjugation within words to afford syntactic information. Interesting evidence is that `bg`, as a very analytic Slavic language, has a larger α than other Slavic languages. In another dimension, a large $\beta + \gamma_{norm}$ suggests a dramatic decrease in the frequency of rare words. Hence, lan-

[5] http://www.statmt.org/moses/

[6] An implementation of the nonlinear least-squares Marquardt-Levenberg algorithm was used.

[7] http://www.gnuplot.info/

[8] Bulgarian (`bg`), Czech (`cs`), Danish (`da`), German (`de`), Greek (`el`), English (`en`), Spanish (`es`), Estonian (`et`), Finnish (`fi`), French (`fr`), Hungarian (`hu`), Italian (`it`), Lithuanian (`lt`), Latvian (`lv`), Dutch (`nl`), Polish (`pl`), Portuguese (`pt`), Romanian (`ro`), Slovak (`sk`), Slovene (`sl`), and Swedish (`sv`).

[9] The non-typical Germanic `en`, Baltic `lt` and `lv`, and Hellenic `el` are in gray. `ro` with a large β is excluded.

[10] First principal component: $-0.1\alpha - 0.7\beta - 0.7\gamma_{norm}$, and second principal component: $1.0\alpha + 0.1\beta - 0.3\gamma_{norm}$.

	α	β	γ	$\frac{\gamma}{r_{max}}$
a.w	$0.92_{\pm.00}$	$0.73_{\pm.01}$	$3.73_{\pm.02}$	0.72
c.w	$0.84_{\pm.00}$	$1.09_{\pm.04}$	$3.84_{\pm.03}$	0.79
m.w	$0.80_{\pm.00}$	$1.22_{\pm.04}$	$3.77_{\pm.03}$	0.81
p.w	$0.81_{\pm.00}$	$1.32_{\pm.06}$	$3.76_{\pm.04}$	0.79
a.c	$0.59_{\pm.00}$	$32.31_{\pm2.04}$	$4.42_{\pm.03}$	1.17
c.c	$0.49_{\pm.00}$	$31.30_{\pm2.73}$	$4.32_{\pm.04}$	1.17
m.c	$0.50_{\pm.00}$	$15.51_{\pm0.52}$	$3.95_{\pm.02}$	1.08
p.c	$0.50_{\pm.00}$	$21.02_{\pm1.18}$	$4.10_{\pm.03}$	1.12

Table 2: Fitted parameters on ICWB2 data.

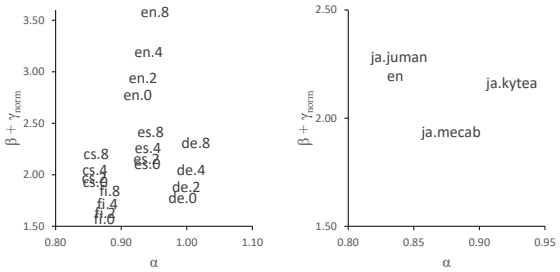

Figure 5: Effects on α and $\beta + \gamma_{norm}$.

guages with a small $\beta + \gamma_{norm}$, that is, Germanic and Uralic, have a more gradual decrease in rare words, which are instances of various phenomena of derivation and compounding from complex morphology. By contrast, languages with a large $\beta + \gamma_{norm}$, such as en and fr, tend to use phrases composed of multiple common words to express complex concepts, so that the drop in frequency of rare words is relatively dramatic. As $\beta + \gamma_{norm}$ is sensitive to the portion of rare words, this dimension may be easily affected by the property of specific data. An example is ro, for which a much larger β than other languages was fitted.

Table 2 lists the fitting results on ICWB2 Chinese data. a.*, c.*, m.*, and p.* denote *Academia Sinica*, *City University of Hong Kong*, *Microsoft Research*, and *Peking University* data, respectively. *.w and *.c denote manually segmented words and characters, respectively. For the results on words, a trade-off on α and $\beta + \gamma_{norm}$ can be observed. Based on the previous analysis, we can consider that a.w has more segmentations on function words. An evidence is the segmentation of the expression *shibushi* (whether or not), which is composed of three characters *shi* (to be) *bu* (not), and *shi* (to be). The expression is segmented into *shi / bu / shi* in most cases in a.w, but always kept together in m.w. Regarding characters, we have small α and huge $\beta + \gamma_{norm}$. Note that both common functional words and rare specific concepts in Chinese are commonly composed of multiple characters. Therefore, the contrast between common and rare characters is not so obvious, which leads to small α (no overwhelmingly functional words in syntax) and huge $\beta + \gamma_{norm}$ (extremely analytic in morphology).

Figure 5 provides further evidence. The data size of typical languages in Europarl is gradually halved and the change of the fitted parameters is shown in the plot on the left of Fig. 5. *.0 denotes the original data and *.n denotes the data of one n-th size. α does not change substantially for smaller data because of the stable syntax features and functional words. However, $\beta + \gamma_{norm}$ becomes larger, which suggests that there are fewer morphological varieties because of the smaller data size. The plot on the right of Fig. 5 shows how different word segmentations in Japanese affect the parameters. There are three common Japanese morphological analysis tools: kytea, mecab, and juman. kytea provides the most fragmentary segmentation and juman tends to attach suffixes to stems. For example, the three tools segment *wakarimashita* (understood, in polite form) as follows: *waka / ri / ma / shi / ta* (5 tokens) by kytea, *wakari / mashi / ta* (3 tokens) by mecab, and *wakari / mashita* (2 tokens) by juman. As the most fragmentary segmentation by kytea contains more functional suffixes as *words*, it has the largest α, and by contrast, the segmentation by juman has the smallest α. Furthermore, mecab has a smaller $\beta + \gamma_{norm}$ because it may keep proper nouns unsegmented, which can be considered as introducing more *compounded words*. For example, *tōkyōdaigaku* (The University of Tokyo) is kept as one word by mecab, but segmented as *tōkyō / daigaku* (Tokyo / university) by the other two tools.

4 Conclusion and Future Work

We have shown that $f \propto r^{-\alpha}(r + \gamma)^{-\beta}$ for the rank-frequency relation in natural languages. This is an explainable extension of several related formulations, with α related to the analytic features of syntax and $\beta + \gamma$ to that of morphology. A more general form, $f \propto \prod_k (r + \gamma_k)^{-\beta_k}$, can be considered for further investigation. The k terms can depict k different proportional coefficients.

References

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. 2008. Overview of the patent translation task at the NTCIR-7 workshop. In *Proc. of NTCIR*, pages 389–400.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. of MT summit*, volume 5, pages 79–86.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. ACL (Demo and Poster)*, pages 177–180.

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

Wentian Li, Pedro Miramontes, and Germinal Cocho. 2010. Fitting ranked linguistic data with two-parameter functions. *Entropy*, 12(7):1743–1764.

Benoît Mandelbrot. 1965. Information theory and psycholinguistics.

David M. W. Powers. 1998. Applications and explanations of Zipf's law. In *Proc. of NeMLaP3/CoNLL98*, pages 151–160.

Richard Sproat and Thomas Emerson. 2003. The first international Chinese word segmentation bakeoff. In *Proc. of the SIGHAN workshop on Chinese language processing*, pages 133–143.

George K. Zipf. 1935. The psycho-biology of language.

George K. Zipf. 1949. Human behaviour and the principle of least-effort.