Regimes in Babel are Confirmed

Report on Findings in Several Indonesian Ethnic Biblical Texts

Hokky Situngkir
hs@compsoc.bandungfe.net
Dept. Computational Sociology
Bandung Fe Institute

Abstract

The paper introduces the presence of three statistical regimes in the Zipfian analysis of texts in quantitative linguistics: the Mandelbrot, original Zipf, and Cancho- Solé-Montemurro regimes. The work is carried out over nine different languages of the same intention semantically: the bible from different languages in Indonesian ethnic and national language. As always, the same analysis is also brought in English version of the Bible for reference. The existence of the three regimes are confirmed while in advance the length of the texts are also becomes an important issue. We outline some further works regarding the quantitative analysis for parameterization used to analyze the three regimes and the task to have broad explanation, especially the microstructure of the language in human decision or linguistic effort – emerging the robustness of them.

Keywords: quantitative linguistics, Zipfian analysis, macro-properties of texts.

iqra'
bismi rabbikal lażī khalaq(a)

1. Prolog

Traditional linguistic studies have not yet really celebrated the quantitative analysis, especially the statistical one. In the other hand, there have been many quantitative tools available particularly in the age of interdisciplinary studies on natural language and on how human languages has evolved in such a way we use to day. Despite such conditions, the quantitative studies to human language have begun that long since the findings of Harvard Professor G. K. Zipf (1949) widely known today as the Zipf's Law. Furthermore, the Zipf's Law is not only recognized well in linguistics, yet it has been stated in a lot broader studies from various disciplines.

This paper reports a further observation continued from Situngkir (2007) tries to see the distinctions as well as similarities of languages widely used among Indonesian people, be it ethnic

and national language. The way to see those is by using statistics in our perceptive that language can be seen as macro phenomena emerged from the socially distributed and in advance, dynamically evolved among human minds and (complex) cognitive systems. The way human mind interpreting, processing, and decide any information elements surrounding her are somehow by using language. Language becomes one of key roles to understand human dynamics since the ability of language acquisition by human civilization are also dynamic and evolve through cultural generations (see for i instance classic work by Hill (1972) and attractive findings in Solé (2006) as well as Bergstorm, *et. al.,* 2001). Thus, it is not at all an exaggeration to say that understanding different languages by means of their structures, use of words, as well as how to create information from them is a challenging steps to deeper understanding to how distinctive civilization and particular social systems might evolve.

Indonesia is an interesting place to observe this for most Indonesian people have capability to talk in at least two languages: their ethnic language (actively or passively) and the national language, Bahasa Indonesia. We analyze the bible texts from various Indonesian ethnic languages – as translated and printed by the Lembaga Alkitab Indonesia (LAI) – Batak Toba (LAI, 1998), Angkola (LAI, 1991), Simalungun (LAI, 2000), Karo (LAI, 2000), Pakpak Dairi (LAI, 1998), Java (LAI, 1994), Sunda (LAI, 1991), and New Translation of Indonesian Bible (LAI, 1974) along with Bible Today's English Version (American Bible Society, 1992) as references. The task of the paper is to find confirmation or any interesting features discovered from those data as we use some advanced statistical findings that have been also seen in some major English texts (see Manning *et. al.,* 1999 and Montemurro, 2004).

The paper is structured as follows. Next section elaborates some understanding about the statistical properties in the nature of Zipfian analysis and several enhancements to find better "laws" of word occurrences in the texts of natural languages. This including the development of interestingly Tsallis' recent statistical mechanics that generalizes thermodynamics (in the sense of non-extensive entropy) of heuristic arguments based on the structure of fractals in symbolic sequences with long-range correlations (see Tsallis, 2004).

This description brought us to the analytical model of word occurrences that is related to different languages, both ethnic and official one used generally in Indonesia to reveal some interesting statistical properties related to its diversity, similarities, and differences. The rest of the paper discusses this issue while trying to see some possibilities of further advanced work in our general endeavor to understand sundry of cultural skins in Indonesia.

2. Regimes in Zipf's Plot

Implementation of Zipfian analysis in linguistics leans on the statistics of the word counts in textual objects, be it a single text (corpus) or some assortments (corpora). The question is then, what kind of words are coming most frequently in texts and what the fact does to the structure of the respective language. It is from this sort of question, what the paper motivates to present: statistically speaking appealing categories arouse and its universality may even be persisting throughout different languages.

As pointed by Zipf (1949), the frequency rank of the words used in a text would follow the power law:

$$f(r) \approx \frac{A}{r^{\alpha}} \qquad (1)$$

and we can write it in logarithmic form where it is easier to (linearly) fitted as

$$\log f(r) \approx \log \mathrm{A} - \alpha \log \mathrm{r} \qquad (2)$$

where $f(r)$ is the normalized number of the occurrence of each word ranked by r, while A is a constant and α the power law exponent acted as the slope of the fitting line in the logarithmic scale.

Table 1
Fifteen Highest Rank Words in the Bible:
the Ethnic Languages and the Ones We Use as Reference

BATAK TOBANESE / SUNDANESE

dominant words	*frequency*	*use*	*dominant words*	*frequency*	*use*
i	33008	determiner	**ka**	19825	preposition
ni	32390	conjunction	**anu**	16133	conjunction
na	28398	determiner	**ku**	16015	preposition
ma	26999	determiner	**jeung**	9536	conjunction
di	20672	preposition	**di**	7986	preposition
tu	19639	preposition	**teh**	7319	determiner
do	18422	complementizer	**ti**	7055	preposition
angka	18203	complementizer	**nu**	7014	conjunction
si	15519	determiner (article)	**eta**	6787	demonstrative
jala	14082	conjunction	**kami**	6624	pronoun
dohot	11664	conjunction	**urang**	6615	pronoun
sian	10839	preposition	**pangeran**	6171	determiner (article)
ibana	9953	pronoun	**geus**	6167	adverb
nasida	9506	pronoun	**allah**	5561	noun
ho	7515	pronoun	**teu**	4202	adverb

ENGLISH / INDONESIAN

dominant words	*frequency*	*use*	*dominant words*	*frequency*	*use*
the	63999	determiner (article)	**dan**	28363	conjunction
and	51777	conjunction	**yang**	24485	conjunction
of	34665	preposition	**itu**	14281	pronoun, demonstrative
to	13583	preposition, verbal infinitive marker	**di**	12843	preposition
that	12930	complementizer, demonstrative	**mereka**	12409	pronoun
in	12664	preposition	**orang**	9807	noun
he	10434	pronoun	**akan**	9003	adverb
shall	9835	modal verb	**aku**	8908	pronoun
unto	9023	preposition	**dari**	8830	preposition
for	8979	preposition	**kepada**	8122	preposition
I	8868	pronoun	**dengan**	7869	conjunction
his	8483	pronoun	**tuhan**	7674	noun
a	8185	determiner	**ia**	7535	pronoun
lord	7842	noun	**tidak**	7427	adverb
they	7392	pronoun	**engkau**	5464	pronoun

In advance, as observed by Mandelbrot (1983), along with a lot of Zipf's plot there exists a kind of hyperbolic curve pattern for the lowest ranked and dominantly used words. As he suggested,

this probably similar with what is also discovered in a lot of power law functions over empirical data in nature. The proposed fitting function to this fact is

$$f(r) = \frac{A}{(1+Br)^{\alpha}} \quad (3)$$

or in the logarithmic form of,

$$\log f(r) \approx \log \mathrm{A} - \alpha \log(1 + \mathrm{Br}) \quad (4)$$

Obviously, the hyperbolic curve will be yielded as $B \neq 1$ and vice versa, the pure straight line is apparent by $B = 1$. The question that might arise among us is, what kind of words they are and how such deviation emerged from the straight line (in log-log scale) as $r \to 0$. The other possible question is whether it is universally observable in any languages.

From the previous work (Situngkir, 2007) we have showed generally that the rounded curve made by the highest used words are robust in any languages we observe. However, some different values of parameterization from the variables used in Mandelbrot's modification possibly becomes a thing that distinct one language to another within exactly the same content of the text. If we take a closer look into the lowest rank words (as also shown in table 1 from texts we scrutinized), apparently those are the functional words sensitive to the grammatical rules governed the construction of the meaning in the text. These words are usually recognized as function words, such as determiners, prepositions, and complementizers (Manning, *et. al.,* 1999:20). Certainly, this is an important thing to find as we related it to the bending curve in the Zipf's plot and the different ways of distinctive languages using them at the same text.

It is clear that the uses of those words are very sensitive to the way a language emerging the meanings in sentences. For example, from fifteen most used words in Batak Tobanese, we have not yet seen any sensitive-to-text words as we discover in Sundanese (e.g.: *pangeran*, *allah*), in Indonesian (e.g.: *tuhan*) or English (e.g.: *lord*) as the corpora is a holy book related to spiritual matters. Of course, there are such things as frequently discussed by linguists about the grammatical differences between languages, but as these somewhat dominant words are bending the straight curve in our Zipfian analysis, this has become interesting for its robustness over languages. Roughly speaking, there is no such word with the highest dominance over other words in other corpus/corpora since the most frequent words – which are recognized as function words – are used within the similar usage frequencies. Some more specific words are left and mostly good enough when we fit them with the power law. So now, we have two regimes in corpus/corpora. One regime is mostly coalesced by the function words sensitive to particular language and the other are words specific to the text – or the "meaningful stuff" brought by the message, sequence of words, sentence.

The observations to these interesting stylized facts of the macro-properties of texts have not yet stopped since Cancho & Solé (2000) and Montemurro (2004) separately discovered another interesting feature. Quite different with the analytical focus of Mandelbrot, these researchers were surprised with the discovery of some other different non-Zipfian power law behavior in the word rank frequency – mostly when the Zipfian analysis was implemented to corpora comprised by large amount of texts/corpuses. In other words, the latest observation found different possible regime in the Zipf plot. However, the findings indicated that is still power law but one with exponent is relatively further from the unity.

The power law regime as it has been showed by Zipf has $\alpha \approx 1$ and this has been confirmed in a lot of places and fields of measurement in the nature (see Newman, 2005). While the findings showed that power law exponents less than unity are rare there are sometimes after some particular values the exponents fall off quickly towards zero – the thing that is also apparent in the observation to large corpora. As discussed in Cancho & Solé (2000), after some value of r, the exponents are deviated faster to zero with $\alpha \approx 1.5 - 2$. Here we raise a question what happens to

the use of words after those high ranks of frequency. Motemurro (2004) used the nonextensive entropy approach to try to statistically explain this. This approach has been used widely in even larger domains of research so far from its origin in statistical mechanics (see Tsallis, 2004).

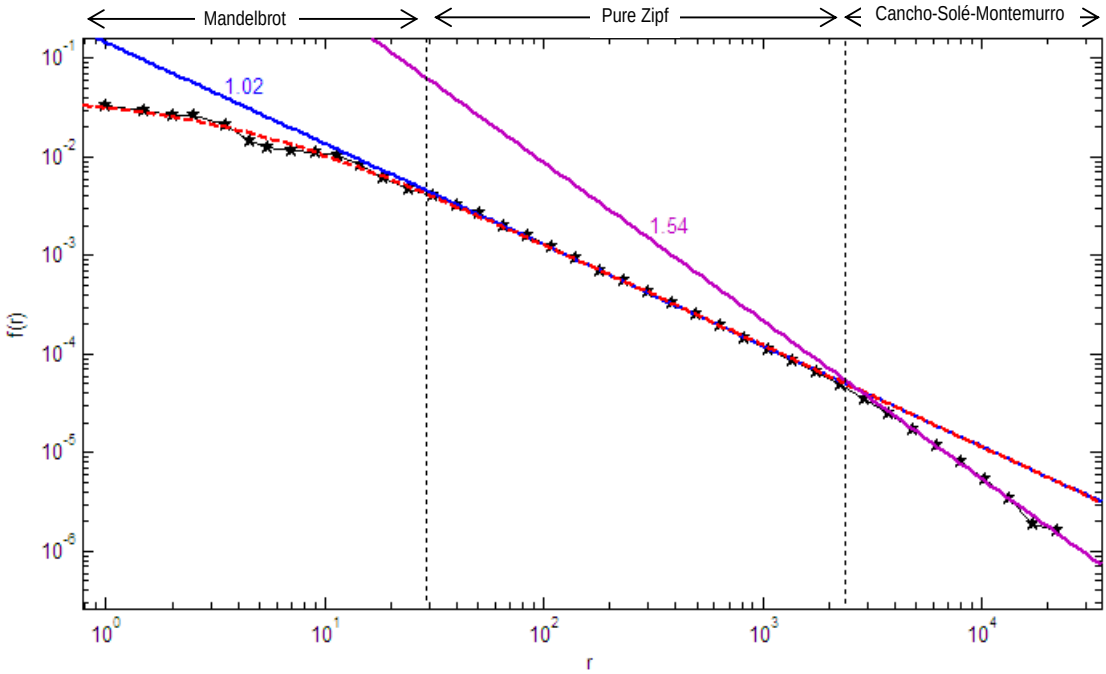

Figure 1. Three regimes observed in the rank frequency plot displaying different use of words in corresponding texts. The data used here is Sundanese Bible (as corpora).

The modification is by adding the possibilities of crossing over between the group of words forming the standard Zipfian power law ($r \approx 1$) to the non-Zipfian one which is exponential. By modified the Mandelbrot's additions to the standard Zipf's fit as differential equation into the nature of Tsallisian nonextensive entropy, the fit parameters have been added and the fitting equation becomes

$$f(r) = \frac{1}{\left[1-\frac{\vartheta}{\varphi}+\left(\frac{\vartheta}{\varphi}\right)\exp\left(\frac{\varphi}{\beta}r\right)\right]^{\beta}} \tag{5}$$

or in logarithmic form,

$$\log f(r) = -\beta \log\left\{1 - \frac{\vartheta}{\varphi} + \frac{\vartheta}{\varphi}\exp\left(\frac{\varphi}{\beta}r\right)\right\} \tag{6}$$

with β as a variable attached to let the crossover. From this equation it is apparent that in such values of $\varphi \ll \vartheta$ and $s \ll$ the equation could become

$$f(r) \sim \frac{1}{\left[1+\frac{\varphi}{\beta}r\right]^{\beta}} \tag{7}$$

which is in the similar form with the Zipf-Mandelbrot equation in eq. 3. This equation is expected to perfectly fit the three regimes introduced. This mission showed in figure 1, showing the well-fitted Sundanese Bible in the regime of Mandelbrot's modification for higher ranks words, the original Zipf law for the aslant curve yielded by the common words, and the fast decaying exponential regime

could be recognized as the Cancho-Solé-Montemurro regime represented by the rarer words. In our observation to multiple language bibles, these latest words are mostly coming in the rank of $r \geq 10{,}000$ or more and most of them are the words known as *hapax legomena*, words with only occur once in the text. The characteristics from these regimes are our focus in discussions of the next section as we empirically see the regimes' persistence over languages.

3. Discussions

Apparently, our examinations to the languages of our case of study conjectured that the three regimes of the texts are robust over Indonesian ethnic languages such as Angkolanese, Karonese,Tobanese, Pakpaknese, Simalungunese, Javanese, and Sundanese, as well as the national language Indonesian and our reference, English. In this sense, we do our examinations in the collections of texts and not just a single book taken from the bible since the persistence of the second power law regime does not always exist in the short texts. It is clear that the presence of the third regime over texts depends on the length of the texts. At first, we do the observation in the first five books of the bible: the Pentateuch (Genesis, Exodus, Leviticus, Numbers, and Deuteronomy). These books are traditionally written by Moses as Genesis is frequently called the first book of Moses, Exodus the second book of Moses, and so forth.

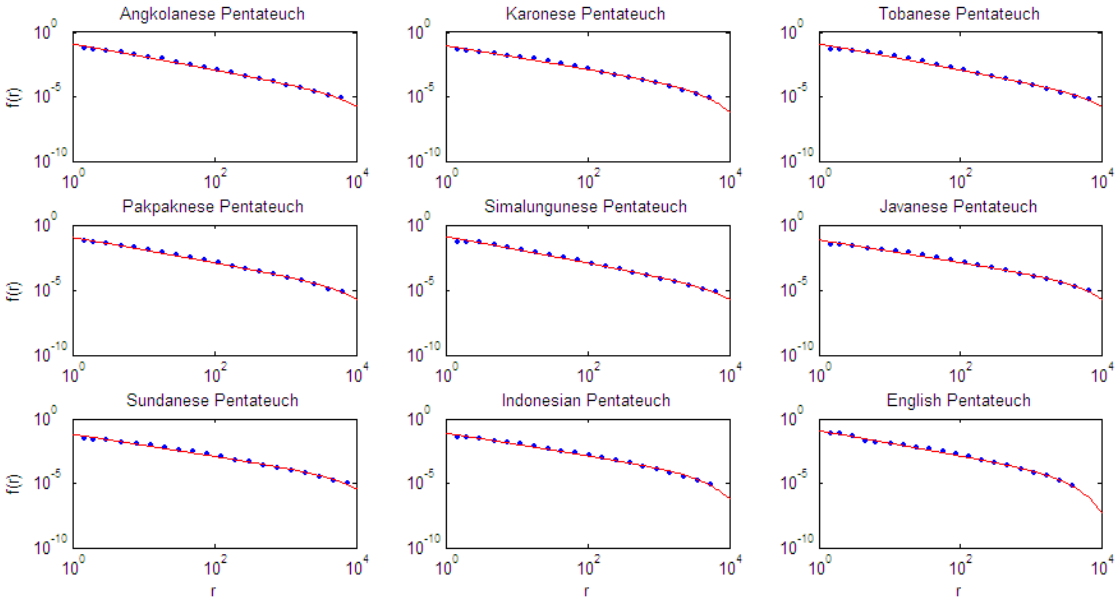

Figure 2. The three regimes as depicted by the eq. 6 in the books of Moses: Pentateuch.

By observing fig. 2, we can see that most of the texts are still not covered the curves lined by the equations providing the three regimes yet – an explanation for this probably lies upon the length of the text still not long or statistically significant enough showing the expected pattern. However, from the figure and as detailed in table 2, the two regimes has been fulfilled well since for the short texts, the eq. 6 has become the eq. 7. To have further explanation about this issue, we plotted the word occurrences of the entire bible for each language. We showed this in fig. 3. It is worth noting that the displayed data points of each figure represent the averages over non-overlapping interval bin on the rank variable (*x*-axis) that is centered at the showing points. Here, the size of each bin is changing on every step in such a way to have the constant value in the logarithmic scale. The aim is of course to smooth the persisting fluctuations in the data in order to ease seeing the emerging pattern.

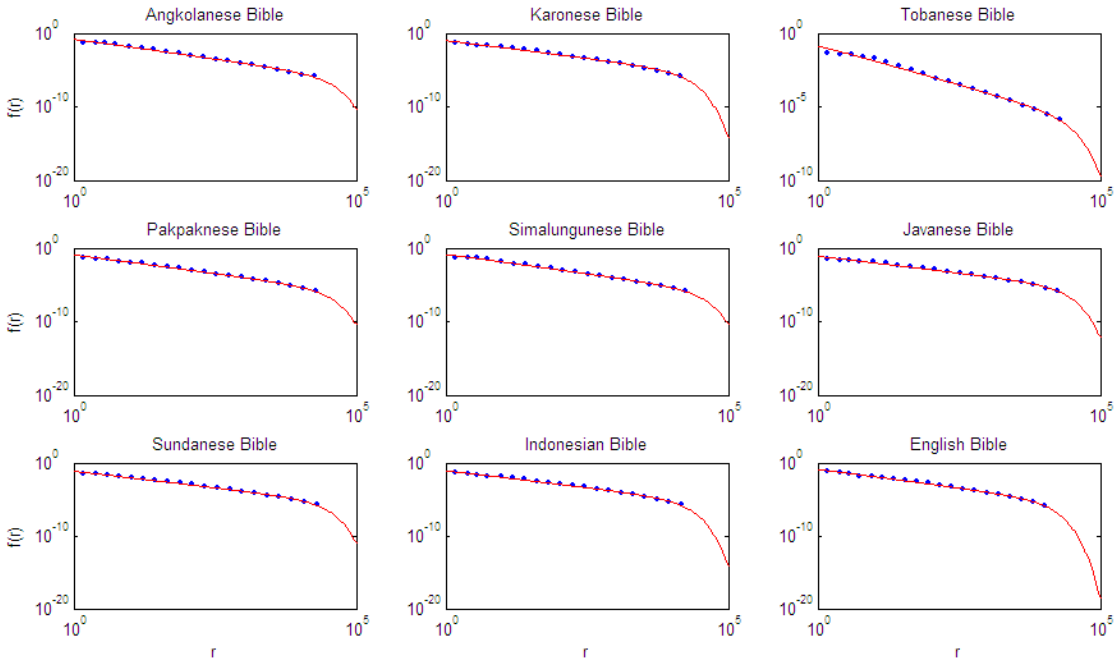

Figure 3. The Zipf plot for the entire bible on each language.

Table 2
Summary of the revealed Statistical Parameters from equation 6

BIBLE		*General Fit Parameters*				*text length*	*used words*
		ϑ	μ	β	R		
ALL	***ANGKOLA***	5.26	0.0001	1.1061	0.998	22375	802508
	KARO	9.5503	0.0002	0.9812	0.997	17989	666634
	TOBA	5.8293	0.0001	1.0855	0.997	23968	761282
	PAKPAK	7.4217	0.0001	1.0338	0.999	23881	699861
	SIMALUNGUN	5.8798	0.0001	1.0806	0.998	22300	705830
	JAWA	13.572	0.0002	0.93	0.997	22200	639806
	SUNDA	15.715	0.0001	0.9127	0.998	25022	609191
	INDONESIA	12.067	0.0002	0.9426	0.998	18123	659943
	ENGLISH	6.7218	0.0003	1.0405	0.998	12688	790664
PENTATEUCH	***ANGKOLA***	6.811	0.0003	1.0352	0.998	7725	135683
	KARO	11.292	0.0005	0.9312	0.997	6509	121549
	TOBA	6.6782	0.0003	1.0419	0.997	8109	149432
	PAKPAK	8.2499	0.0003	0.9957	0.998	7928	126641
	SIMALUNGUN	6.7033	0.0003	1.0396	0.998	8027	144210
	JAWA	17.624	0.0004	0.8734	0.997	8200	112435
	SUNDA	20.264	0.0003	0.8628	0.997	9450	106949
	INDONESIA	15.089	0.0005	0.8849	0.998	6771	128962
	ENGLISH	7.5946	0.0007	0.9945	0.998	4770	157823

Since there is an apparent connectedness between the size of the observed text and the fitting parameters, there is a challenge for further works as pioneered by Dębowski (2002). This further work can be brought from the starting point of our understanding of the existence of the three regimes confirmed in the paper.

Other interesting feature found in our examination to different languages of the bible is varying value of each fitting parameters as the similar approach in our previous work (Situngkir, 2007). It is obvious that the most distinguished parameters of the fitting is ϑ, the constant that governs the ups and downs of the whole fitting curve in the graph. As $\vartheta \gg$ the $f(r) \ll$ and vice versa, and a sure thing that bounded to the size of the texts or numbers of the used words and their frequency. From table 2, we can see that interestingly, while uttering same things (semantically speaking), the syntaxes of each languages vary in big numbers. We denote syntax in a shallow notion as the emerged variables of ratio between the text length and the used words in the whole text. For instance, the ratio between the number of the words used in English bible and its length is *62.316* contrasted to value of *35.9*, *37.1*, *31.7, 29.3, 31.6, 28.8, 24.3,* and *36.4* for Angkolanese, Karonese, Tobanese, Pakpaknese, Simalungunese, Javanese, Sundanese, and Indonesian respectively. This vast difference must be related to the structure of the language since languages used in Indonesian population has also broad difference with English as a European language. We suspect that the big difference in the value of ϑ should be related to this issue. In advance, we left a homework here since the big differences among the value of ϑ still be approached qualitatively and relative differences between ϑ's in different language over the same texts. The same homework give us further challenge to see the discovered value of μ and β among distinct language since these values must come from different language structure too.

Nevertheless, despite the way to see those parameters distinguishable each other over distinctive language, the robustness of the presence of the three regimes over texts over different language is also a challenging further approach that should be carried out in the field of quantitative linguistics. Modestly speaking, this might bring us to an interesting view of how different civilizations may evolve differently while in some other boundaries the evolution is just similar, perhaps persistently following similar rule(s).

4. Concluding Remarks

We show the persistence of three regimes over analysis of same text but different language by using the traditional Zipfian plot. The most frequent words form the hyperbolic curve and becomes a background of the shape of fitting parameters and equations proposed by Mandelbrot. However, interesting feature is also yielded by words with highest ranks. As more specific to text words incorporated a corpus/corpora, there exists crossover from the Zipfian power law (originally discovered with exponents ≈ 1.02 become faster decay following the exponential curve. We use the nonextensive entropy statistics applied to quantitative linguistics proposed by Montemurro (2004) to see the properties of different languages, both ethnic and national language. Up to here, robustness persists.

As the size of the text grows the presence of the three regimes are somehow accentuates. Furthermore, in our analysis we found that there is a qualitatively very big gap of one of the parameters used to fit the macro-properties over different languages. We hypothesize that this might come from the different micro-structure of corresponding language emerging the gap in corpora. As a beginning for further analysis, we see this in the terms of the ratio between the text length and the used words that should be correlated in such a way to the different of the grammatical structure laid upon different language. However, the persistence of the observed statistical pattern also bring us to an understanding that language might built upon the similar thing over human mind using language as well as the economic motive of the least effort proposed by G. K. Zipf (1949) on the use of different words as their frequency follows the power law.

Eventually, some further possible and interesting works are also outlined. The analysis of contrasting (read: correlating) the different quantification of length of text (as well as the number of used words) to the fitting parameters to distinct languages are left for further comprehension. On the other hand, tasks to explain the persistence of the three regimes over human languages are also an attractive issue for further approach. As language becomes the carpet of most aspects of human culture and civilization, our quantitative understanding to this issue is probably an important thing to explain the more general issue: beyond linguistics itself.

Acknowledgement

Author thanks Santi Rahmayuliani for helping the data mining and Surya Research International for the support within the period the paper is written.

Works Cited

American Bible Society. (1992). *Bible Today's English Version 2nd Edition*.

Bergstrom, C. T., Antia, R., Szàmadó, S., Lachmann, M. (2001). "The Peacock, the Sparrow, and the Evolution of Human Language". *Working Paper Series 01-05-027*. Santa Fe Institute.

Cancho, R. F., Solé, R. V. (2000). "Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited". *Working Paper Series* 00-12-068. Santa Fe Institute.

Dębowski, L. (2002). "Zipf's Law Against the Text Size: A Half-Rational Model". *Glottometrics* 4.

Hill, J. H. (1972). "On the Evolutionary Foundations of Language". *American Anthropologist: New Series* 74 (3): 308-17. American Anthropological Association.

Lembaga Alkitab Indonesia. (1974). *Alkitab Terjemahan Baru*.

Lembaga Alkitab Indonesia. (1991). *Alkitab Angkola*.

Lembaga Alkitab Indonesia. (1991). *Alkitab Sunda*.

Lembaga Alkitab Indonesia. (1994). *Alkitab Jawa*.

Lembaga Alkitab Indonesia. (1998). *Alkitab Pakpak Dairi*.

Lembaga Alkitab Indonesia. (1998). *Alkitab Toba Ejaan Baru*.

Lembaga Alkitab Indonesia. (2000). *Alkitab Karo Edisi III*.

Lembaga Alkitab Indonesia. (2000). *Alkitab Simalungun*.

Mandelbrot, B. B. (1983). *The Fractal Geometry of Nature*. Freeman.

Manning, C. D., Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Montemurro, M. A. (2004). "A Generalization of the Zipf-Mandelbrot Law in Linguistics". In Gell-Mann, M. & Tsallis, C. (eds.). *Nonextensive Entropy – Interdisciplinary Applications*. Oxford UP.

Newman, M. E. J. (2005). "Power laws, Pareto distributions and Zipf's law". *Contemporary Physics* 46: 323–351.

Situngkir, H., Hariadi, Y., Suroso, R., Surya, Y. (2004). *Aplikasi Fisika dalam Analisis Keuangan: Mekanika Statistik Interaksi Agen*. Sumber Daya MIPA.

Situngkir, H. (2007). "An Observational Framework to the Zipfian Analysis among Different Languages: Studies to Indonesian Ethnic Biblical Texts". *Working Paper Series WPA2007*. Bandung Fe Institute.

Solé, R. V. (2006). "Scaling Laws in Language Evolution". In C. Cioffi (eds.), *Power Laws in the Social Sciences*. Cambridge UP.

Tsallis, C. (2004)."Nonextensive Statistical Mechanics: Construction and Physical Interpretation". In Gell-Mann, M. & Tsallis, C. (eds.). *Nonextensive Entropy – Interdisciplinary Applications*. Oxford UP.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.