An Observational Framework to the Zipf's Analysis among Different Languages

Studies to Indonesian Ethnic Biblical Texts

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

Abstract

The paper introduces the used of Zipfian statistics to observe the human languages by using the same (meaning) corpus/corpora but different in grammatical and structural utterances. We used biblical texts since they contain corpuses that have been most widely and carefully translated into many languages. The idea is to reduce the possibility of noise came from the meaning of the texts in distinctive language. The result is that the robustness of the Zipfian law is observable and some statistical differences are discovered between English and widely used national and several ethnic languages in Indonesia. The paper ends by modestly propose further possible framework in interdisciplinary approaches to human language evolution.

Keywords: statistical processing of natural language, Zipf's law, Zipf-Mandelbrot fit, corpus, evolution of language.

in principio erat Verbum
et Verbum erat apud Deum
et Deus erat Verbum

1. Introduction

There have been a lot of places beyond many scientific domains exhibited the signatures of power law as an interesting statistical properties emerged from complex systems (SItungkir & Surya, 2003). Concerning the classic work of Harvard linguistic professor, G. K. Zipf (1947), the power law seems to be not that outlandish among linguist, as it has been showed that the power law in the Zipf plot is linked to the writer and reader natural behavior to minimizing their effort while communicating. In this case, the writer's

efforst is conserved by using a small vocabulary of common words and the reader's effort is reduced by having a large vocabulary of individually rarer words – a way to make the messages less vague (Manning, *et. al.,* 1999). Here, Zipf argued that the Zipf's law is supported by the maximally economical compromise between the competing needs between writer (or speaker) and reader (or listener).

Interestingly, recent studies on astonishingly different fields of research have shown us that the Zipf's Law is not only emerged in statistical analysis of textual objects. For instance, the Zipf's law is also discovered in the DNA sequences (Mantegna, *et. al*., 1994), population of cities (Mulianta, *et. al.*, 2004), rank of general elections (Situngkir & Surya, 2004), the daily precipitation series in geophysics (Primo, *et. al*., 2007) and even inspired an alternative to investment strategies (Situngkir & Surya, 2005). Whether or not the persistence of the law in those fields related to its origin in linguistics are basically left to philosophical issues may arise but yet, we can now see that the (statistical) application of the Zipf's law are not limited to a single discipline of study.

However, despite its broad implementation, some of dominant voices were risen in linguistics regarding to the nature of such quantitative probabilistic and statistical approach on natural language. MIT linguist, Noam Chomsky (1957), stated:

> One's ability to produce and recognize grammatical utterances is not based on notions of statistical approximation and the like.
> (Chomsky, 1957:16)

It is not, of course, the aim of the paper to stand in front of such notices in linguistics even though it is clear that any advanced works on things presented in the paper could possibly change the way we see of any application of quantitative analysis in linguistics. The Chomskyian notion and its proponents might be right for conventional probabilistic and statistical approaches to human language while this paper steps forward and invigorated by the understanding of complex phenomena in many fields using recent understanding of complex system through, for example statistical mechanics and complexity studies. Nonetheless, the paper cannot be seen as a presentation of such limitless aim, since the paper is a single step for general framework in a possible long road to our understanding about natural language.

The paper presents what we can see in the Zipf's plot as we draw the rank frequency of words used in ethnic biblical text as the observed corpus. We are motivated to see the robustness of Zipf's law in different language corpus while the message refers to similar content. We do this by using bible translation to some of Indonesian ethnic language. The next section of the paper reviews some aspects of the model: the classical Zipfian and the Mandelbrot-Zipf analysis on texts. It is followed by discussions of what we could learn from the statistical analysis and other possible works in advancement of what we have done in the paper.

2. Indonesian Ethnic Linguistic Data

Probably biblical texts are those the world most translated into languages for their theological urgencies in Christian missionaries. Indonesia is a well-known country for its diversity in ethnics yet keeps its unity. Here, there are more than 400 ethnic languages and it is a matter of fact that most of them have read bibles in their own language. This becomes interesting for a great deal of linguistic analyses trying to make distinctions among

languages. For instance, the plenty of ethnic languages were arisen from the evolution of language that in some ways related to the first populations in Indonesia as they migrated from other parts in Asia. Thus, any analysis regarding to the variations of language, in some cases could lead us to our further understanding about the evolution of the various Indonesian ethnics since language plays a very important role in the evolution of civilization.

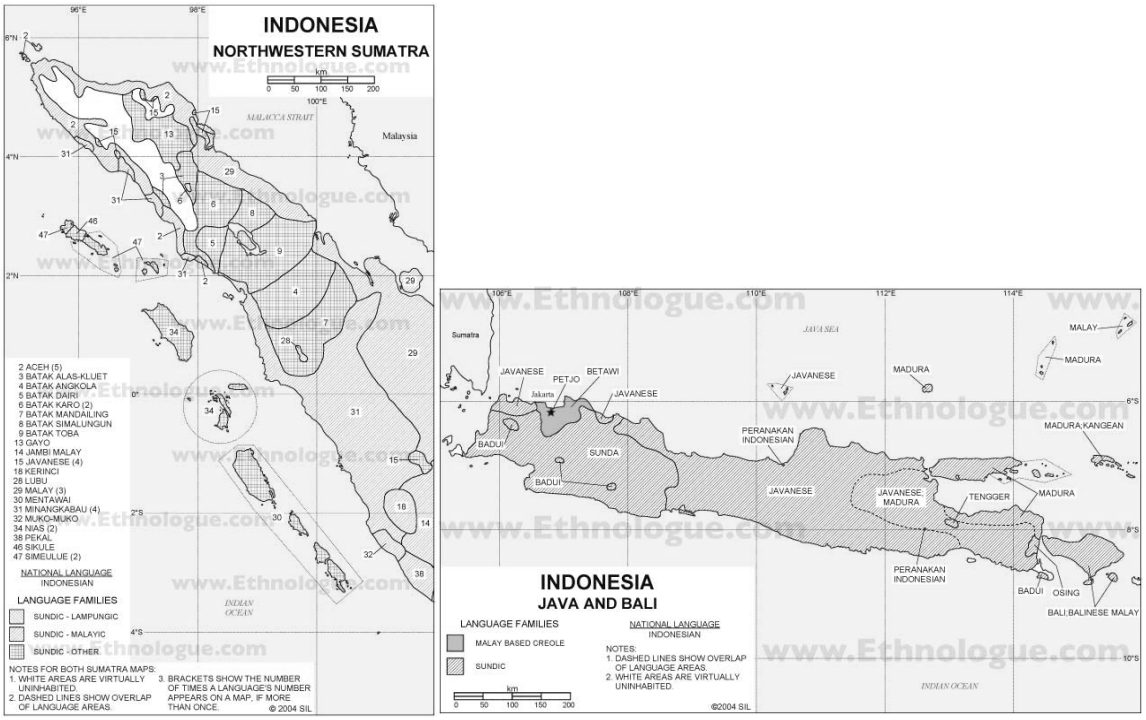

Figure 1.
The ethnic population adopted languages analyzed in the paper (reproduced from Gordon, 2005).

We use the bible texts from various 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. It is certain that those language cannot represent the whole language used among Indonesian ethnic groups but yet what we are trying to do is to see any hypothetical similarities and distinctions as well as the persistence of Zipf's law among them.

From the lists of the language we use, the first five ethnic ones come from the population in Northern Sumatera, the most populated province in the island of Sumatera while the next two ones are respectively the western and eastern Java Island. The cultural systems in which both Sumatera and Java populated are not very typical to one another although among the world's languages they are categorized at the same spot of Austronesian Malayo-Polynesian (Gordon, 2005). Thus the Batak Toba, Angkola, Simalungun, Karo, Pakpak Dairi, Java and Sunda were coming from the same proto-language while the modern populations are now using the Bahasa Indonesia as the national language, a form of language developed from the Malaya language, a *lingua franca* widely used before the country gained independence.

Even though the languages observed in this paper are quite closely from the historical perspective of the evolution of language, practically, there are quite a lot difference among them respect to grammars or structures, lexicons or vocabularies, and of course reflecting different dialect when each population speak by using the national language. In these not-so-different languages lies the main motivation behind the paper. We use the same text/corpus across different languages and observe the statistical properties of each. Since the "meanings" of the text have same intention showing what the particular biblical texts told the readers about, then any difference that we have from the statistical analysis would be deviated by the nature of the structure of the language. This, however, comes from our consideration that the respective institution incorporating not just theologians but also those with broad linguistic understanding would carry the translation of bible from one language to another very delicately.

Henceforth, the simple statistical analysis presented by the paper would be merely a first step while the future works are waiting forward to be conducted within other quantitative and sophisticated statistical models.

3. Zipf and Mandelbrot-Zipf

Zipf's plot can be viewed as follows: we count the frequency of the use of all words in a large corpus and then rank the words in order of their frequency of occurrence, we can see the relationship between the frequencies of words, say *f(r)*, and its position in the ranking plot, say *r*. Then a very simple mathematical relationship between them fulfils:

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

and we can write it in logarithmic form,

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

where α is the Zipf exponent that should close to unity. In the Zipf's plot, we draw the equation as a log-log plot in order to have the Zipf exponent as the slope of the curve in logarithmic coordinates while $\log A$ becomes the intercept. Here we also have the normalization constant *A* as the number of times a particular word appears in our corpus/text divided by the total number of words there are, say *N*.

Despite its simple relation, as we have discussed before, this sort of relation is discovered in a wide range of phenomena. There have been several explanations introduced for this seemingly 'universality' beyond various texts. Some attempts have been laid upon the information theory and statistics with some advanced mathematical models of random symbolic sequences to explain the emergence of the law. The first are set up by Herbert Simon (1955) that simulated the dynamics of text generation as a multiplicative process (later also known as Yule process) that discovered leading to Zipf's law for asymptotically long texts (see the detail in Newman, 2005). The simpler yet interesting attempt to explain the presence of Zipf's law in texts by random multiplicative models is also shown recently by Li (1992). The other way to understand this phenomenon is proposed by Benoit Mandelbrot (1983) that is built on the special properties of the hierarchical structure of natural languages. Here, an interesting framework of fractal model is used with assumptions of the existence of self-similarity to explain the power-law of the involved distributions of texts. In

advance, the work of Mandelbrot has brought into a well-known improvement to the fitting process of the power-law over the frequency rank.

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

or in the logarithmic form of,

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

where C is a second parameter that needs to be adjusted to fit the data. In this equation, we now have three variables used depicting the richness of the text's used of words. Here, as we have $B = 1$, we could see that we are now back at the original form of the Zipf's law. It is obvious that the existence of $B \neq 1$ exhibits a more hyperbolic curve at the upper data (lowest rank words) of the Zipf's plot. Regarding to Mandelbrot (1983), this might be related to hyperbolic functions – rather than Gaussian function – discovered in a lot of stylized statistical properties of nature.

However, a more converged analysis was proposed by Zannette & Montemurro (2005) that is grounded on the statistical pattern found in written human language for the reason that of multiplicative processes in human language generative phenomena. This latter explanation becomes interesting since it emphasized of the analytical observation on human written texts in which most empirical works discovered the persistence of Zipf's law. However, the paper presented here is not an attempt to give alternative view on the generative explanation of texts exhibiting Zipf's law or trying to explain the reason behind its presence. The paper wants to report a direct observation on what we can see and learn from the Zipf's law as different languages of the same content of corpus (or corpora).

4. Discussions: Zipf Plot over Languages

Talking about the bible, it is in fact coalesced by 66 books, 39 known as the Old Testament and 27 as the New Testament. Those books are considered written by different persons except, for example, the first five books of the Old Testament, literally known as the Pentateuch or the majority books in the new testament recognized written by St. Paul. Regarding to the sizes of the books and understanding the nature of Zipf's law as "macro" properties with some requirements of (statistically) large enough data, we realize that it is obvious that Zipf's law is deviated for some short books compiled in the bible. We do the observations of the Zipf and Zipf-Mandelbrot fit and analysis; however, some of them are displayed in the paper to ease us to see the emerging patterns.

As an exemplification, we display the Zipf's plot in figure 2 and it showed that all of the texts exhibit the power-law with the exponent close to unity. However, from the figure 1, we can also see that there are some slight differences among them – especially when we compare the English version with the Indonesian and Indonesian ethnic languages. This interesting feature can be observed even more clearly in figure 3 (and of course quantitatively in table 1). Throughout our observation, the data of national Indonesian language will "slightly deviated" more to the English language while it will just "close enough" to the ethnic ones. This interesting discussion surely could bring us to the conjecture that there is some different statistical properties that can be differed among multiple language corpus.

Our observation is now forwarding to see the similar methodology applied to the Zipf-Mandelbrot fit. Here, we have two fitting parameters to test among languages that in return promise us for the possible better macro distinctions among languages. As shown in table 2, we can see the values of the yielding of both variables we used in the Zipf-Mandelbrot fit. In general, it is obvious that the resulting error is smaller by using the Zipf-Mandelbrot since it incorporates the possible hyperbolic pattern at the highest ranks of the list.

Table 1

The Fit Parameters with Zipf's Law

	Genesis			Psalm			Luke			Rome		
	A	***α***	***R***	***A***	***α***	***R***	***A***	***α***	***R***	***A***	***α***	***R***
ANGKOLA	0.24724	1.1645	0.98704	0.17702	1.0996	0.98556	0.20661	1.1344	0.98315	0.20434	1.1151	0.98063
KARO	0.35927	1.1886	0.98754	0.28079	1.1474	0.98715	0.27406	1.1496	0.98708	0.24669	1.1277	0.98439
TOBA	0.21179	1.1394	0.98674	0.12414	1.047	0.98301	0.11735	1.0299	0.97981	0.12409	1.0104	0.97404
PAKPAK	0.25126	1.1444	0.98711	0.17569	1.0758	0.98575	0.17865	1.0816	0.98481	0.169	1.0534	0.9798
SIMALUNGUN	0.26596	1.1683	0.98704	0.20522	1.1178	0.98652	0.15878	1.0717	0.98436	0.16288	1.0554	0.97768
JAWA	0.24193	1.0959	0.98587	0.20456	1.0908	0.98559	0.19564	1.082	0.98627	0.19677	1.0793	0.98405
SUNDA	0.16832	1.0364	0.98474	0.15526	1.0381	0.98486	0.13553	1.0005	0.98572	0.16233	1.0326	0.98056
INDONESIA	0.39601	1.1925	0.987	0.25608	1.1196	0.98676	0.22326	1.1038	0.98648	0.22406	1.0937	0.98436
KJV	0.69169	1.3285	0.98801	0.60662	1.2993	0.98905	0.38686	1.2207	0.98803	0.25568	1.1291	0.98303

Table 2

The fit parameters with Zipf-Mandelbrot

	Genesis				Psalm				Luke				Rome			
	A	***α***	***B***	***R***	***A***	***α***	***B***	***R***	***A***	***α***	***B***	***R***	***A***	***α***	***B***	***R***
ANGKOLA	0.185	1.176	0.727	0.987	0.268	1.105	1.406	0.986	0.323	1.141	1.416	0.983	0.282	1.128	1.233	0.981
KARO	0.038	1.247	0.116	0.990	0.045	1.184	0.169	0.988	0.067	1.184	0.247	0.988	0.092	1.168	0.343	0.986
TOBA	0.272	1.147	1.188	0.987	8.352	1.047	55.676	0.983	32.818	1.030	238.300	0.980	35.792	1.010	272.110	0.974
PAKPAK	0.078	1.167	0.316	0.988	0.110	1.086	0.602	0.986	0.161	1.092	0.849	0.985	0.146	1.070	0.786	0.980
SIMALUNGUN	0.123	1.186	0.467	0.987	0.151	1.127	0.712	0.987	0.395	1.076	2.263	0.984	0.366	1.063	2.039	0.978
JAWA	0.023	1.150	0.091	0.988	0.054	1.112	0.260	0.986	0.069	1.103	0.336	0.987	0.103	1.107	0.470	0.985
SUNDA	0.031	1.066	0.164	0.986	0.057	1.053	0.348	0.985	0.050	1.019	0.326	0.986	0.088	1.057	0.479	0.981
INDONESIA	0.024	1.278	0.066	0.990	0.030	1.164	0.117	0.988	0.063	1.133	0.272	0.987	0.066	1.145	0.256	0.986
KJV	0.027	1.464	0.055	0.992	0.030	1.400	0.069	0.992	0.054	1.279	0.154	0.990	0.081	1.179	0.284	0.985

An interesting thing emerges in the table 2, where we could see that one language, the Batak Toba language exhibits the same error for both Zipf and Zipf-Mandelbrot for the book of Psalm, the Gospel of Luke and St. Paul's letter for the people of Rome. This opens a broad possible discussions as this possibly came from the number of the words at those corpus are still not yet significantly fit best with the Zipfian power law. Since the Mandelbrot's modification to the Zipf's law is sensitive to words with lowest ranks ($r < 200$) in order to have the hyperbolic shape before the plateau of the straight line of the power law, this could be an issue that also, modestly saying, characterize a language relative to others.

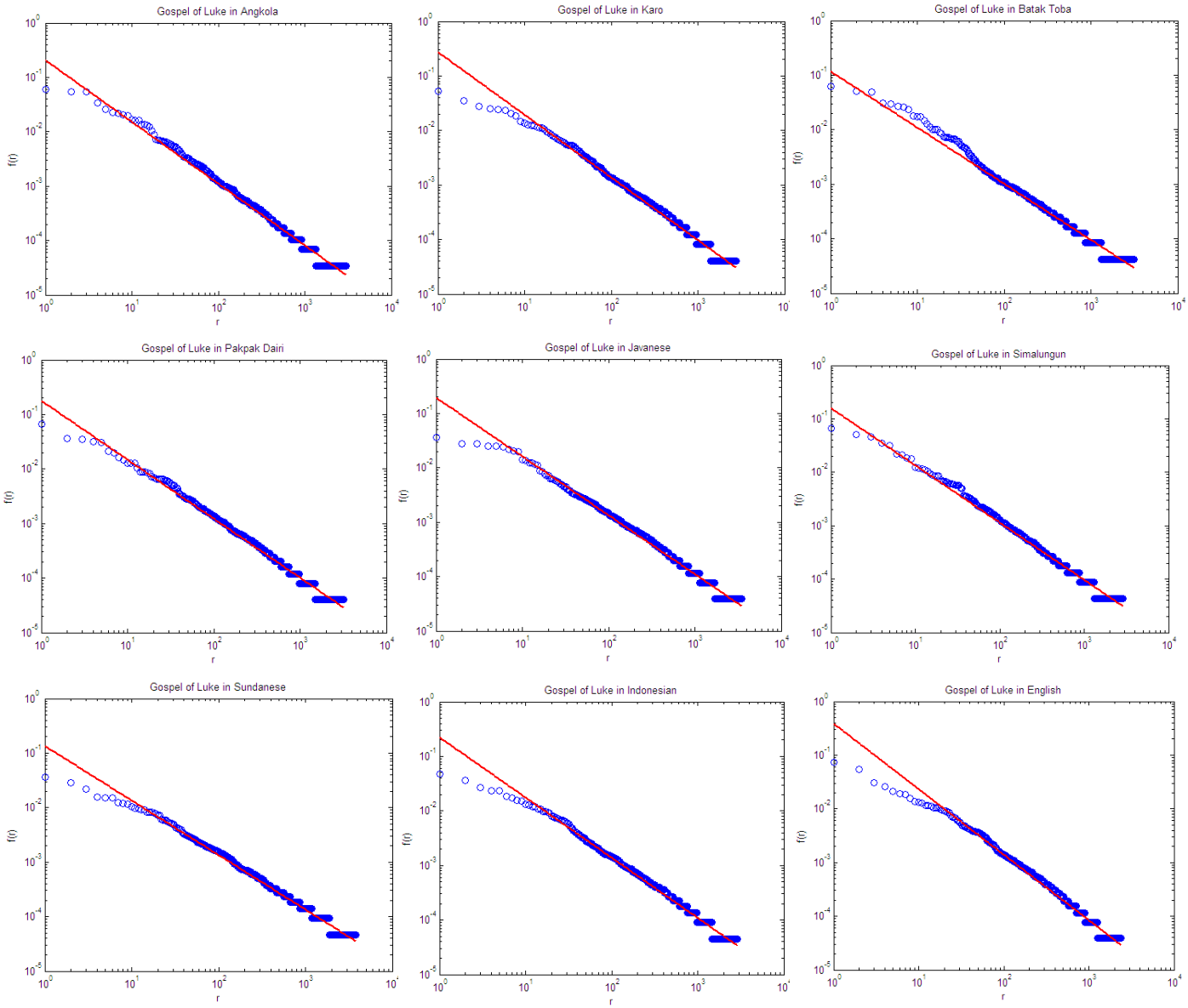

Figure 2.
Zipf's Law over languages for Gospel of Luke

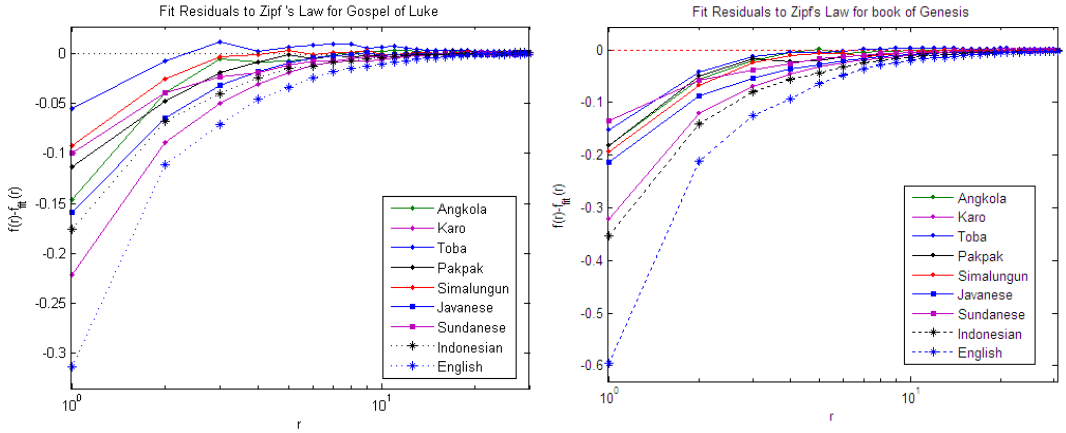

Figure 3.
Fit Residuals to Zipf's Law

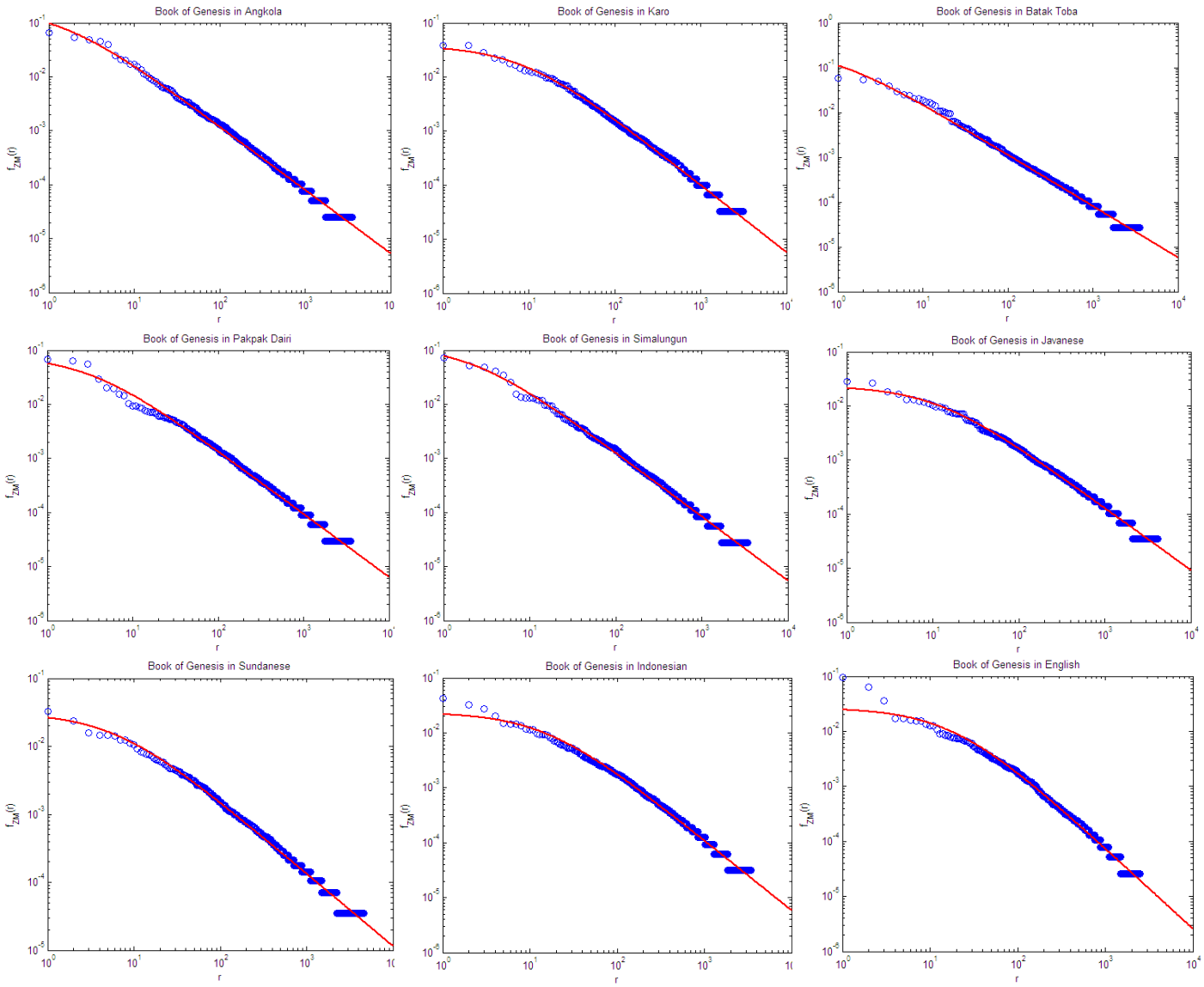

Figure 4.
Zipf-Mandelbrot signatures over languages for the book of Genesis

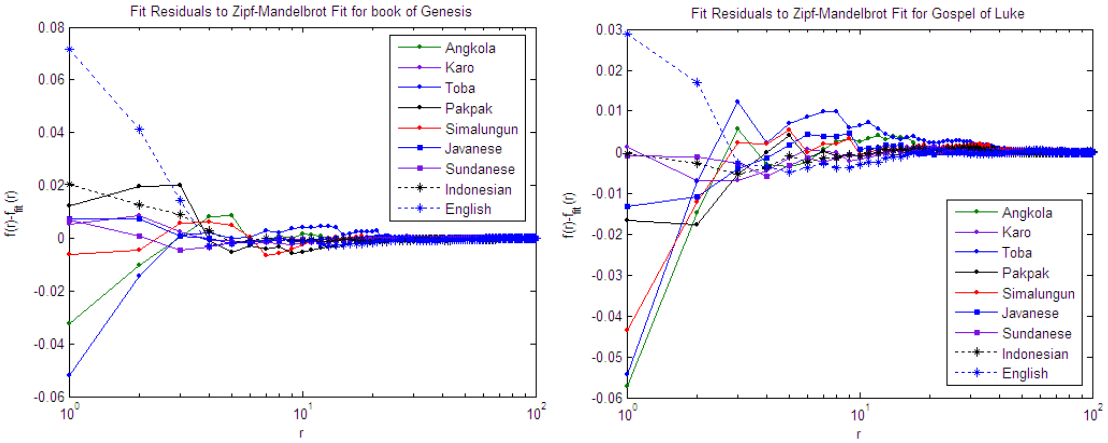

Figure 5.
The fit residuals to Zipf-Mandelbrot fit for book of Genesis and the Gospel of Luke

This hyperbolic signature is showed apparently in figure 4 for words with lowest ranks, while the straight power law line are still fulfilled by the words with higher ranks. We should put into account that Zipf's law of the word frequency has a little difference with the one with many applications to other systems that are also exhibiting the power law – while we regard the rank size as a cumulative distribution function (see for instance Newman, 2005 or an empirical works on sizes of cities Moura & Ribeiro, 2006). As a cumulative distribution function, $F(x)$, of such probability distribution function, $f(x)$,

$$F(x) = \int_{x_{min}}^{\infty} f(x)\, dx$$

which is in roughly speaking the random variable, x, equals to or greater than x_{min}, and the power law of $f(x) = Cx^{-\alpha}$ can also be written in power-law form of the

$$F(x) = \frac{C}{(\alpha-1)x^{(\alpha-1)}}$$

thus, obviously the $p(x)$ would diverges from the power law for $\forall\ \alpha > 0$ as $x \to 0$, and the distribution should deviate from power law below the minimum value of x_{min}. In linguistics, the variable x is the integers $(x = \{1,2,3, \dots\})$ representing the sequence of the rank list while the distortion from the power law of the lower rank words is then fulfilled by the Mandelbrot's modification.

As clearly shown in figure 5 the better fit is brought by the Zipf-Mandelbrot and still as we focus on the statistical distinction of languages, the differences between languages are persist. There is some kind of robustness here as we can still figure that the English version of the biblical books is still distinctive with the one from Indonesia national and ethnic languages. The English version is placed higher in the fit residuals plot and this accentuate what we see before. Moreover, in the Zipf-Mandelbrot plot we have now two parameters that possibly brought us to more statistical variations of languages.

Chomsky (1957) defined grammar (or we could roughly said the observation to the structure of languages) as "a theory of the set of sentences constituting the language", i.e. with an explicit ontological commitment to language as sentences. Thus, there is something in the evolution of languages with the evolution of human mind and for this understanding, to the structure of language; we outline at least three things for further conjectures and advancement of scientific – especially statistical approach – of the natural language of human, i.e.:

- The sophistication to the statistical model used to see cope with how different language utter the same things henceforth we can discern better the differences between language. However this is a core point lack in conventional linguistics but modestly promising. Our elaboration in this paper is based on observation with some technical things are omitted since we discount the use of punctuations even though we realize well enough that punctuations are very important in any human corpora sensitive to languages. It is becoming a challenging problem how to put into account the punctuations in the statistical models of the natural languages.
- The understanding of the statistical approaches will nicely bring us to the betterment of our understanding to how human mind generate thinking and here is the core point of the complexity of the cognitive processing. This kind of explanation can be

brought by the computational generative processes as it has even been started since the classical works of information theory.

- The twos are expected for further understanding to see what happened to our language in its evolution as it occurred to the social diversification based on ethnicity and even the social cultures.

5. Concluding Remarks

The paper reports the statistical observation of Zipf's law to different human languages while the approached corpus is being telling the same things. This is expected to reduce the possible sensitivity to the meaning of the texts and the different stylized statistics are closer to what emerging from the respective structure of language, whether it grammatical or lexical. Interestingly, it has also been showed that Zipfian statistics is robust throughout those raw corpuses we analyzed.

From the comparative analysis, we point out that there are some possible conjectures of statistical distinctions between languages and this may open a good challenge in our understanding to the broader sense of our scientific recognition to nature of human language. Since different languages are sensitive to the evolution of human species with very tight relationship with cultural phenomena, social environments, and even geographical and natural system in which the language widely used and developed, this can bring us further to the evolution of language and optimistically reveal its evolutionary life history. We modestly point out here that human utterances can be observed by its quantitative nature by using better and better statistical tools.

Acknowledgement

Author thanks Surya Research International for financial support during the period of the research.

Works Cited

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

Chomsky, N. (1957). *Syntactic Structures*. The Hague.

Gaffeo, E., Gallegati, M., Giulioni, G., Palestrini, A. (2003). "Power Laws and Macroeconomic Fluctuations". *Physica A* 324:408-416.

Gordon, Raymond G., Jr. (ed.), 2005. *Ethnologue: Languages of the World, 15th edition*. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/.

Kennedy, J. (1971). "A History of Malaya". *The Journal of Asian Studies* 30 (3):736-7.

Kosmidis, K., Kalampokis, A., Argyrakis, P. (2006). "Statistical Mechanical Approach to Human Language. *Physica A* 366:495-502.

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*.

Li, W. (1992). “Random Texts Exhibit Zipf’s-Law-like Word Frequency Distribution”. *IEEE Transaction Information Theory* 38 (6):1842-45.

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

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

Mantegna, R. N., Buldyrev, S., Goldberger, A. L., Havlin, S., Peng, C-K., Simons, M. (1994). “Linguistic Features of Noncoding DNA Sequences. *Physical Review Letter* 73:3169-72.

Mulianta, I., SItungkir, H., Surya, Y. (2004). “Power Law Signature in Indonesian Population: Empirical Studies of Kabupaten and Kotamadya Population in Indonesia”. *BFI Working Paper Series* WPT2004. Bandung Fe Institute.

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

Powers, D. M. W. (1998). “Applications and Explanations of Zipf’s Law”. In D. M. W. Powers (ed.). *New Methods in Language Processing and Computational Natural Language Processing*. ACL.

Primo, C., Galvàn, A., Sordo, C., Gutiérrez, J. M. (2007). “Statistical Linguistic Characterization of Variability in Observed and Synthetic Daily Precipitation Series”. *Physica A* 374:389-402.

Simon, H. A. (1955). "On a Class of Skew Distribution Functions". *Biometrika* 42: 425-40.

Situngkir, H., Surya, Y. (2003). “Dari Transisi Fasa ke Sistem Keuangan: Distribusi Statistika pada Sistem Kompleks”. *BFI Working Paper Series* WPQ2003. Bandung Fe Institute.

Situngkir, H., Surya, Y. (2004). “Democracy: Order out of Chaos – Understanding Power-Law in Indonesian Elections”. *BFI Working Paper Series* WPO2004. Bandung Fe Institute.

Situngkir, H., Surya, Y. (2005). “What can We See from Investment Simulation based on Generalized (m,2)-Zipf Law”. *BFI Working Paper Series* WPO2005. Bandung Fe Institute.

Zanette, D. H., Montemurro, M. A. (2005). “Dynamics of Text Generation with Realistic Zipf Distribution”. *Journal of Quantitative Linguistics* 12(1): 29-40. Routledge.

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