Conjecture to Statistical Proximity with Tree of Language (?)

Report on Few Austronesian Languages of Indonesian Ethnics

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

Deni Khanafiah
<dk@students.bandungfe.net>
Scholar in Dept. Computational Sociology
Bandung Fe Institute

May 5th 2007

Abstract

We continue some steps showing the distinctions and proximities of languages over statistical facts as it has been pioneered previously [3]. In the paper, we construct the homology tree from the distance matrix yielded from the transformation of some statistical aspects of the empirical observations into binary sequences in order to conform to the concepts of memetics [2]. The resulting visualizations show interesting facts and possibly challenge some further steps for the advancement of our understanding to the discourse of languages and ethnicities.

Keywords: quantitative linguistics, ethnic languages, phylomemetic tree, evolution of language.

1. Statistical Language Proximity

The works in [3] and Situngkir [4] showed some directions that statistical perspective on language based on the Zipfian plot might reveal the differences between languages. Apparently, some observation could be brought in order to see that some languages are different (as emerged from the particular linguistic structures as yielded from the adaptability of human developing the languages). This short report reflects a slightly different motive: the conjecture algorithmic works to see the proximity of languages as distinguished via the statistical properties.

The previous work has confirmed the existence of three regimes (*Mandelbrot* → *Zipf* → *Cancho-Solé-Montemurro*) as we plot the ranked words used in a corpus/corpora regarding to their occurrences (see Appendix 1). We use the most sophisticated method to fit the exhibited power law of the specific language,

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

with β as a variable attached to let the crossover over regimes, and from the table in appendix 1, it is clear that variable ϑ becomes the significant variable that apparently distinguishes different texts over different languages. Furthermore, we also see that the ratio between the used words in the text and the text length (θ) also become one variable depicting different texts that are observed.

$$\theta = \frac{\text{text length}}{\text{number of used words}} \qquad (2)$$

In advance the two variables could be interestingly used to show the proximities of the languages over the same text.

We use the algorithmic computation introduced in [1] to build the phylomemetic tree. The evolutionary innovative artifact tree is built upon the the homology between different products which feature characteristics depicted in binary matrix (called the binary matrix of memeplexes). This main concept of the imported evolutionary theory and memetics are discussed in [2].

In our case, the two variables ϑ and θ, as empirically become variables distinguishing distinct languages in our statistical observation, are thus transformed into the binary matrix and hence it easy for us to have the homology matrix reflecting the similarities between languages – of course, statistically speaking or more specifically the Zipfian analysis of textual artifacts. In order to visualizing the differences and similarities between the languages fitted by the equation [1], we use the notion of Hamming distance. This distance shows how much changes occurred upon two binary sequences. The Hamming Distance between two sequences x and y with the same length is the number of different state for

each same position for the two sequences. We can write it as $H(x, y) = \left|\{i : x_i \neq y_i\}\right|$. Normalized Hamming distance is another word for Manhattan distance, which have mutation value = 1, as we have previously explained. In advance, we can normalize the distance matrix by dividing the value with the number of memes constituting its memeplex. The shortest distance of a tree is defined as minimum number of Hamming distance from the sequences constituting the tree.

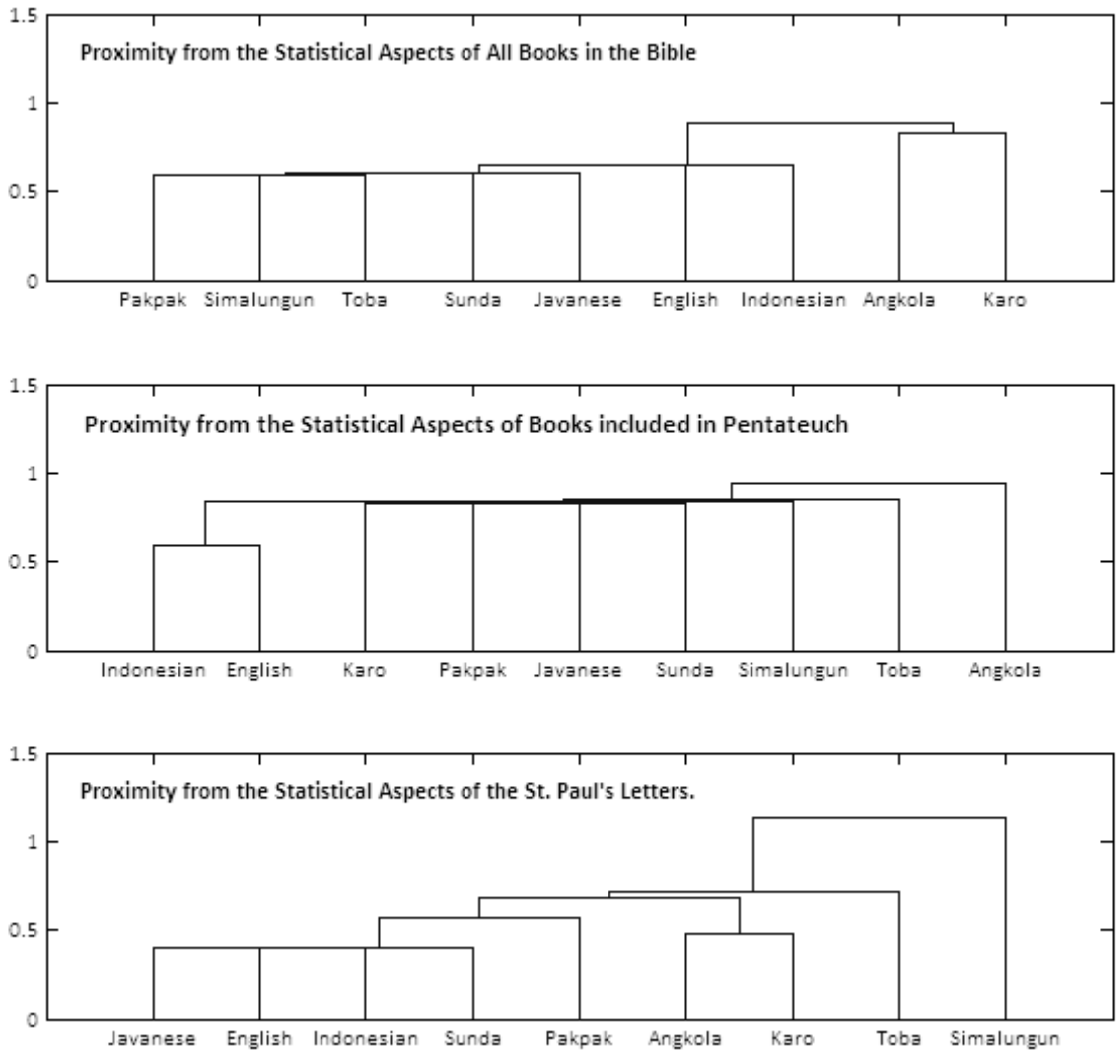

Figure 1. The proximities among languages showed in the shortest spanning tree, the y-axis are the distance between languages found in each text.

Thus, after we have the distance matrix, it would be easy for us to construct the tree of language by using the 'linkage' algorithm that uses the smallest distance between objects in the two clusters,

$$d(r, s) = \min(dist(x_{ri}, x_{sj})),\ i \in (i, ..., n_r),\ j \in (1, ..., n_s) \qquad (3)$$

where x_{ri} is the *i*-th object in cluster r, n_r is the number of objects in cluster r, and n_s is the number of objects in cluster s. From our statistical observations, we have three homology matrices from three sources of corpora, the book of Pentateuch (first five books of Old Testament Bible), the book of St. Paul's letters (from the New Testament), and the whole Holy Bible (both the Old and New Testament). The finding is very interesting and the result is shown in figure [1].

2. Discussions

In a glance view, the three figures shown in figure [1] are slightly different, especially as we see the leaves and the absolute positions of the nodes. However, the clustering of the nodes are interestingly, in some cases, are persistent over texts. For instance, we could see that anthropologically and culturally different ethnics of Javanese and Sundanese are always placed in different cluster with those of the Sumatera languages (Pakpak, Toba, Angkola, Simalungun, and Karo) as well as the two languages we used as reference, i.e.: Indonesian and English.

The clustered languages, i.e.: Javanese and Sundanese as well as the Indonesian and English languages are always seen in our observation to the three corpora. This fact is actually very interesting since the twos are culturally and relatively very different with the other Sumatera ethnics from the characteristics of customs, cultural artifact, traditions, etc.

Scrutinizing the Sumatera languages in the three homology tree, we could see that qualitatively speaking the clustering conform to some common understanding of the ethnics. As Toba language are one of the most used Batak languages in North Sumatera, we do not see unique clustered languages with the Toba (see the map in [3]). Throughout the three trees, we can see that Batak Toba may be showed up in some branches but never be with those of Javanese and Sundanese. Angkolanese and Karonese are shown twice at the same unique cluster; this might reflect some facts of similarity that is possible beyond merely languages. Some relative proximity over languages are also possibly visible in the language homology trees. For instance, from the sub-figure showing the tree retrieved from the Pentateuch, we could clearly see that Simalungun, Karo, and Pakpak are relatively closer to the languages of Javanese and Sundanese than Toba and Angkola. However, the latest two claims must be confirmed by anthropological and more casuistic linguistic cases.

In some cases, the outlines of our observations confirm the robustness of the yielded homology tree. However, we must realize the limited languages we employ in our analytical observations must be put into consideration for further step of more specific cases of quantitative linguistics. The paper is however motivated solely showing the possibility to visually demonstrate the distinctions and proximities of languages through statistical approach, or in general quantitative linguistics.

Acknowdgement

Both authors thank Surya Research International for support during the research period in which the paper is written.

Works Cited

[1] Khanafiah, D. & Situngkir, H. (2005) "Visualizing the Phylomemetic Tree: Innovation as Evolutionary Process". *Journal of Social Complexity* 2 (2): 20-30.

[2] Situngkir, H. (2005). "On Selfish Memes: Culture as Complex Adaptive System". *Journal of Social Complexity* 2 (1): 20-32.

[3] Situngkir, H. (2007a). "An Observational Framework to The Zipfian Analysis Among Different Languages: Studies to Indonesian Ethnic Biblical Texts". *Working Paper Series WPA2007*. Bandung Fe Institute.

[4] Situngkir, H. (2007). "Regimes in Babel are Confirmed: Report on Findings in Several Indonesian Ethnic Biblical Texts". *Working Paper Series WPC2007.* Bandung Fe Institute.

APPENDIX 1

The fit parameters by incorporating the three regimes of Zipfian Plot.

BIBLE		*General Fit Parameters*				*text length*	*used words*
		λ	μ	β	R		
ALL	***ANGKOLA***	5.26	0.0001	1.1061	0.9982	22375	802508
	KARO	9.5503	0.0002	0.9812	0.9973	17989	666634
	TOBA	5.8293	0.0001	1.0855	0.9974	23968	761282
	PAKPAK	7.4217	0.0001	1.0338	0.9986	23881	699861
	SIMALUNGUN	5.8798	0.0001	1.0806	0.9978	22300	705830
	JAWA	13.572	0.0002	0.93	0.9971	22200	639806
	SUNDA	15.715	0.0001	0.9127	0.9976	25022	609191
	INDONESIA	12.067	0.0002	0.9426	0.9977	18123	659943
	ENGLISH	6.7218	0.0003	1.0405	0.9981	12688	790664
PENTATEUCH	***ANGKOLA***	6.811	0.0003	1.0352	0.9978	7725	135683
	KARO	11.292	0.0005	0.9312	0.9969	6509	121549
	TOBA	6.6782	0.0003	1.0419	0.9968	8109	149432
	PAKPAK	8.2499	0.0003	0.9957	0.9982	7928	126641
	SIMALUNGUN	6.7033	0.0003	1.0396	0.9977	8027	144210
	JAWA	17.624	0.0004	0.8734	0.9967	8200	112435
	SUNDA	20.264	0.0003	0.8628	0.9968	9450	106949
	INDONESIA	15.089	0.0005	0.8849	0.9976	6771	128962
	ENGLISH	7.5946	0.0007	0.9945	0.9975	4770	157823

ST. PAUL'S LETTER	***ANGKOLA***	5.9751	0.0004	1.0604	0.9959	66079	4477
	KARO	11.627	0.0007	0.9227	0.9931	56417	3996
	TOBA	5.8249	0.0002	1.0679	0.9971	48485	4521
	PAKPAK	8.6471	0.0004	0.9795	0.9955	55845	4903
	SIMALUNGUN	7.1552	0.0004	1.0175	0.9963	49980	4345
	JAWA	12.466	0.0006	0.914	0.9931	56324	4600
	SUNDA	13.237	0.0004	0.9096	0.9955	51463	5333
	INDONESIA	14.336	0.0007	0.8863	0.9943	47700	4008
	ENGLISH	9.5942	0.0008	0.9467	0.9959	50233	3476