Exploring Letter's Differences between Partial Indonesian Branch Language and English

Nankai Lin, Sihui Fu, Jiawen Huang and Shengyi Jiang✉
School of Information Science and Technology
Guangdong University of Foreign Studies
Guangzhou, China
neakail@outlook.com,jiangshengyi@163.com

Abstract—**Differences of letter usage are the most basic differences between different languages, which can reflect the most essential diversity. Many linguists study the letter differences between common languages, but seldom research those between non-common languages. This paper selects three representative languages from the Indonesian branch of the Austronesian language family, namely Malay, Indonesian and Filipino. To study the letter differences between these three languages and English, we concentrate on word length distribution, letter frequency distribution, commonly used letter pairs, commonly used letter trigrams, and ranked letter frequency distribution. The results show that great differences do exist between three Indonesian-branch languages and English, and the differences between Malay and Indonesian are the smallest.**

Keywords-Differences of letter usage; Austronesian family language; word length distribution; ranked letter frequency distribution

I. Introduction

Although morphemes, not letters, are usually considered to be the smallest linguistic unit, studying statistics of letter usage has its own merit [1]. For example, information on letter frequency is essential in cryptography for deciphering a substitution code [2] . Moreover, in Morse code, the more commonly used letters, the shorter the coding symbol. There are also some similar methods in data compression techniques, such as Huffman coding, in which a source letter is coded based on the probability of its occurrence. Therefore, it is of great significance to carry out letter usage research.

There are many differences in letter usage between different languages. The study of them promotes the related research in cryptology, natural language processing, linguistics and so on. However, relevant research at this stage mainly focuses on common languages. As to non-common languages such as Malay and Indonesian, the focus is on lexical differences. Lin et al. studied the differences in the word frequency distribution between Indonesian and English using Zipf's Law [3] . They also explored lexical differences between Indonesian and Malay [4] .

Indonesian, Malay and Filipino are all members of the Indonesian branch of the Austronesian (Malayo-Polynesian) language family [1] and therefore could somehow reveal the characteristics of this language branch. In this paper, we choose these three languages to conduct a preliminary survey of the letter differences between the Indonesian language branch and English. On the other hand, all these four languages adopt the Latin alphabet as their writing systems, which makes the comparison much easier. We target the news texts in the above three languages and English as the subjects of the research. Four aspects of letter differences among these four languages are studied, including word length, letter frequency distribution, common letter pairs and common letter trigrams. What's more, we fit the ranked letter frequency distributions with ten letter frequency distribution models.

TABLE I. The Data Distribution of Our Crawled Data

Language	Website	Number of words
Filipino	Bandera [2]	18521278
Indonesian	Antara News[3]	183210621
Malay	Bharian[4]	31795086
English	Guardian [5]	255040052

The remaining part of this paper is organized as follows: Section 2 provides information of our data; Section 3 demonstrates how we analyze the difference of word length; Section 4 introduces the letter frequency distribution analysis; Section 5 discusses distributions of the letter pair frequency and the letter trigrams frequency; Section 6 demonstrates how we fit the ranked letter frequency distributions; Section 7 concludes our work.

II. Data Description

To obtain news texts, we crawl four influential news websites to represent these four languages' news texts and segment the articles into sentences and then words. The data distribution concerning these websites is shown in Table 1 (note that the punctuations have been excluded).

III. Word Length Frequency Distribution

The literature on word-length frequency distributions is in abundance within quantitative linguistics. Most research nevertheless mainly focus on common languages such as English, German, etc. [5][6][7][8][9] . Word length frequency typically investigates the frequency of words of different lengths in syllables [10] . We count the length frequency to get the frequency distribution of different languages. Table 2 presents the relative frequency distribution of word length for four languages. We could see that compared with the other three languages, Filipino is more likely to use shorter words.

[1] https://www.britannica.com/topic/Malay-language

[2] http://bandera.inquirer.net/

[3] http://www.antaranews.com/

[4] http://www.bharian.com.my/

[5] http://www.theguardian.com/uk

978-1-7281-5014-7/19/$31.00 ©2019 IEEE

We also calculate the average word length for each language. The average word lengths for English and Filipino are close, 4.69 and 4.77 respectively, while the ones for Malay and Indonesian are relatively longer, which are 6.07 and 6.11. Although Indonesian, Malay and Filipino all belong to the Indonesian language branch, the average word length of Filipino is shorter than Malay and Indonesian's. This is because many words in Indonesian and Malay are formed by adding affixes.

IV. Letter Frequency Distribution Analysis

Studies of letter statistics for many languages have been carried out, such as English, German, Spanish, Esperanto, Russian and Malay [11][12][13][14][15][16][17] . Most studies simply direct towards the distribution of a single language, but rarely compare it with those of other languages. We separately calculate the letter frequency distributions of the four languages in question. Detailed results are shown in Table 3. Note that although C, F, J, Q, V, X and Z are not indigenous Filipino letters, with the introduction of foreign words, these letters gradually appear in Filipino, so we also suppose Filipino has 26 letters, and we could see these letters have lower frequency. The five most commonly used letters in Filipino are A, N, I, G and S. The five most commonly used letters in Indonesian and Malay are the same, which are A, N, E, I and R. As for English, the experimental results we have are consistent with those of other previous studies，which are E, T, A, O and I. In addition, we also count the first, last and middle letter frequency distribution of these four languages. The results are presented in Fig. 2-4. Each chart suggests that the frequency distribution curve of Indonesian is identical to that of Malay. N is the most commonly used first letter of Filipino, with A the middle letter and G the last letter. Of English, they are T, E and E.

TABLE II. WORD LENGTH FREQUENCY DISTRIBUTION

Length	Language			
	Filipino	Indonesian	Malay	English
1	0.0187	0.0040	0.0020	0.0490
2	0.2277	0.0461	0.0427	0.1656
3	0.1426	0.0801	0.0895	0.1918
4	0.1560	0.1708	0.1592	0.1626
5	0.1205	0.1750	0.1752	0.1106
6	0.0977	0.1356	0.1372	0.0881
7	0.0773	0.1182	0.1337	0.0787
8	0.0600	0.0908	0.1024	0.0544
9	0.0380	0.0716	0.0623	0.0406
10	0.0263	0.0495	0.0428	0.0268
11	0.0153	0.0284	0.0284	0.0149
12	0.0085	0.0137	0.0140	0.0081
13	0.0056	0.0084	0.0055	0.0049
14	0.0027	0.0030	0.0023	0.0020
15	0.0015	0.0020	0.0008	0.0010
16	0.0007	0.0008	0.0005	0.0005
17	0.0005	0.0005	0.0003	0.0003
18	0.0002	0.0006	0.0003	0.0001
19	0.0001	0.0002	0.0001	0.0001
20	0.0001	0.0002	0.0001	0.0001

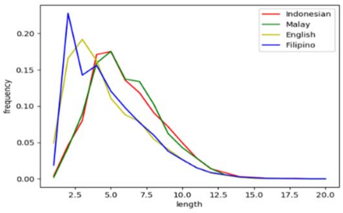

Figure 1. Word length frequency distribution.

TABLE III. LETTER FREQUENCY OF EACH LANGUAGE

Letter	Language			
	Filipino	Indonesian	Malay	English
A	0.1887	0.1908	0.2013	0.0835
B	0.0191	0.0247	0.0282	0.0167
C	0.0141	0.0058	0.0054	0.0310
D	0.0235	0.0393	0.0398	0.0377
E	0.0537	0.0815	0.0840	0.1212
F	0.0072	0.0032	0.0031	0.0217
G	0.0637	0.0361	0.0345	0.0216
H	0.0230	0.0208	0.0240	0.0494
I	0.0805	0.0770	0.0754	0.0741
J	0.0023	0.0108	0.0106	0.0019
K	0.0269	0.0485	0.0505	0.0080
L	0.0402	0.0317	0.0349	0.0423
M	0.0340	0.0430	0.0479	0.0252
N	0.1194	0.1006	0.0981	0.0712
O	0.0545	0.0224	0.0157	0.0745
P	0.0281	0.0340	0.0326	0.0208
Q	0.0008	0.0002	0.0002	0.0009
R	0.0381	0.0562	0.0515	0.0622
S	0.0585	0.0463	0.0420	0.0678
T	0.0526	0.0545	0.0487	0.0900
U	0.0263	0.0480	0.0487	0.0280
V	0.0050	0.0016	0.0013	0.0108
W	0.0112	0.0064	0.0049	0.0190
X	0.0009	0.0003	0.0002	0.0020
Y	0.0262	0.0154	0.0146	0.0179
Z	0.0014	0.0007	0.0017	0.0008

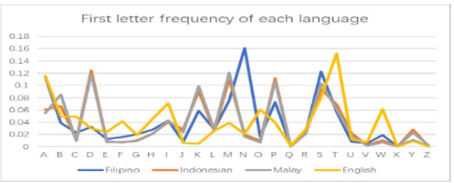

Figure 2. Frequencies of the first letter in each language.

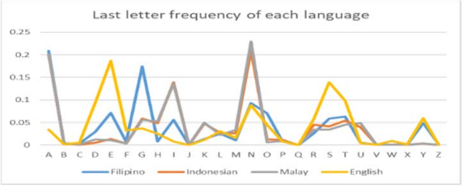

Figure 3. Frequencies of the last letter in each language.

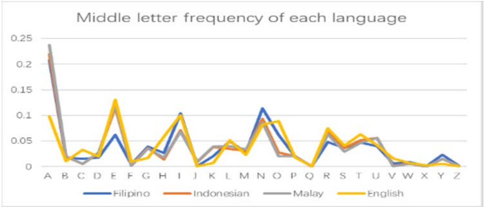

Figure 4. Frequencies of the middle letter in each language.

TABLE IV. TWENTY MOST COMMONLY USED LETTER PAIRS IN EACH LANGUAGE

Filipino		Indonesian		Malay		English	
Pair	Freq	Pair	Freq	Pair	Freq	Pair	Freq
NG	0.0513	AN	0.0603	AN	0.0644	TH	0.0326
AN	0.0500	NG	0.0275	ER	0.0274	HE	0.0281
NA	0.0316	KA	0.0231	NG	0.0267	IN	0.0256
IN	0.0263	ER	0.0229	KA	0.0228	ER	0.0200
SA	0.0217	EN	0.0226	DA	0.0220	AN	0.0194
MA	0.0186	TA	0.0205	EN	0.0211	RE	0.0178
LA	0.0182	AR	0.0203	LA	0.0192	ON	0.0158
KA	0.0160	DA	0.0194	TA	0.0177	EN	0.0136
AL	0.0156	ME	0.0174	ME	0.0169	AT	0.0135
AT	0.0151	LA	0.0161	AT	0.0162	ND	0.0127
PA	0.0147	RA	0.0161	RA	0.0160	OR	0.0123
ON	0.0145	AT	0.0157	AR	0.0150	ES	0.0121
AG	0.0141	DI	0.0150	MA	0.0147	NG	0.0119
GA	0.0125	YA	0.0145	PE	0.0142	IT	0.0117
SI	0.0122	GA	0.0144	DI	0.0140	TO	0.0115
NI	0.0120	AK	0.0139	YA	0.0135	AR	0.0114
AY	0.0116	IN	0.0138	IN	0.0132	ST	0.0113
YA	0.0110	PE	0.0133	GA	0.0130	ED	0.0111
TA	0.0109	SE	0.0126	SE	0.0128	IS	0.0111
ER	0.0107	AS	0.0113	AL	0.0128	TE	0.0110

TABLE V. TEN MOST COMMONLY USED LETTER TRIGRAMS IN EACH LANGUAGE

Filipino		Indonesian		Malay		English	
Pair	Freq	Pair	Freq	Pair	Freq	Pair	Freq
ANG	0.0349	ANG	0.0169	ANG	0.0171	THE	0.0287
ING	0.0097	KAN	0.0138	KAN	0.0139	ING	0.0132
ALA	0.0088	MEN	0.0126	MEN	0.0118	AND	0.0116
INA	0.0070	ARA	0.0087	BER	0.0102	ION	0.0066
IYA	0.0068	ENG	0.0087	ALA	0.0091	ENT	0.0062
LAN	0.0069	NGA	0.0085	ENG	0.0082	FOR	0.0055
ONG	0.0064	NYA	0.0076	ATA	0.0080	HAT	00052
AMA	0.0060	ATA	0.0075	ADA	0.0078	THA	0.0052
ILA	0.0058	AKA	0.0072	DAN	0.0078	TIO	0.0050
MAN	0.0057	DAN	0.0072	NGA	0.0077	HER	0.0050
AKA	0.0051	TER	0.007	PER	0.0075	TER	0.0046
THE	0.0050	PER	0.0067	ARA	0.0069	VER	0.0039
YAN	0.0050	YAN	0.0064	TAN	0.0069	ERE	0.0038
ARA	0.0049	GAN	0.0062	NYA	0.0065	ATI	0.0038
AGA	0.0044	BER	0.0061	GAN	0.0064	ALL	0.0036
NAG	0.0044	ALA	0.0056	TER	0.0059	ERS	0.0036
AND	0.0042	ADA	0.0055	ERA	0.0058	HIS	0.0036
HIN	0.0041	ELA	0.0055	RAN	0.0055	ATE	0.0034
MGA	0.0040	TAN	0.0054	YAN	0.0055	ITH	0.0032
PAG	0.0040	NTA	0.0049	AKA	0.0052	WIT	0.0031

TABLE VI. RESULT OF THE JACCARD SIMILARITY COEFFICIENT CALCULATION

Language	Number of common letter pairs		Jaccard similarity coefficient	
	bigram	trigram	bigram	trigram
Filipino, Malay	12	5	0.42	0.14
Filipino, Indonesian	10	5	0.33	0.14
Filipino, English	5	3	0.14	0.08
Indonesian, Malay	18	18	0.82	0.82
Indonesian, English	7	1	0.21	0.03
Malay, English	7	1	0.21	0.03

A and N are the most commonly used middle and last letters in both Malay and Indonesian. M is the most common first letter in Malay and D in Indonesian.

V. LETTER PAIR DIFFERENCE ANALYSIS

Dahlqvist described some of the characteristics of words and single character distribution as well as the distribution of character bigrams and trigrams in the Uppsala Newspaper Corpus [18] . Jones and Mewhort studied case-sensitive letter and character bigrams frequency counts from large-scale English corpora [19] . In this work, we count all the letter pairs' frequency and present the twenty most commonly used letter pairs in each language. The results are shown in Table 4. And then we use the Jaccard similarity coefficient to calculate the similarity between any two languages. The Jaccard similarity coefficient calculation formula is as follows:

$$\mathrm{J}\,(\mathrm{A},\mathrm{B}) = \frac{|A \cap B|}{|A \cup B|}$$

where $|A \cup B|$ is the union of the twenty most commonly used letter pairs in two languages (i.e. the number of letter pairs that occur in any one of the languages), and $|A \cap B|$ is the intersection of the twenty most commonly used letter pairs in two languages (i.e. the number of letter pairs that both languages share). The result is shown in Table 6.

From the perspective of letter pairs, we could see that Malay and Indonesian have the highest similarity, followed by Filipino and Malay, and then by Filipino and Indonesian. It suggests that there is a certain commonality in the use of letter pairs between languages of the Indonesian branch.

We then use the same strategy to study the differences of letter trigrams among languages. Table 5 and Table 6 are the experimental results. For any two languages, except the Indonesian-Malay pair, there is a small overlap between their most frequently used letter trigrams. Most of these commonly used letter trigrams are typical affixes in these languages. For instance, ING is one of the English affixes, and KAN, MEN, BER, NYA etc. are common affixes in Indonesian and Malay. What's more, the frequencies of these letter trigrams in Indonesian approximates to those in Malay.

VI. FITTING RANKED LETTER FREQUENCY DISTRIBUTIONS

The limited range in its abscissa of ranked letter frequency distributions causes multiple functions to fit the observed distribution reasonably well. In order to critically compare various functions, Li applied the statistical model selections to some functions, using the texts of U.S. and Mexican presidential speeches in the last 1-2 centuries [1] .

In our work, we use ten different functions to fit the ranked letter frequency distribution. The following is a list of these functions (y denotes the normalized letter frequency, and x the rank of the frequency; x = 1 for most frequent letter and x = 26 for the rarest letter, and n = 26 is the maximum rank value).

Gusein-Zade function is based on the study of Gusein et al. They studied the frequency distribution of letters for

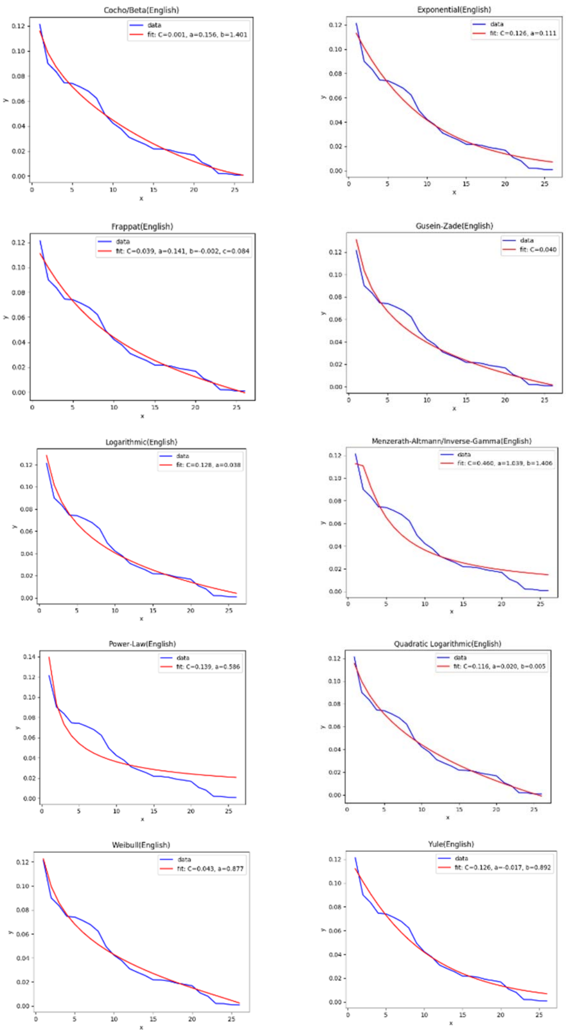

Figure 5. Fitting results for English data.

Russian and fitted this function to the distribution [16][20][21] . Gusein-Zade function is as follows:

$$y = \mathrm{C} \log \frac{n+1}{x}$$

Power-Law is a simple but effective function to fit all kinds of distribution. Clauset presented a principled statistical framework for discerning and quantifying power-law behavior in empirical data [22]. Li used this function to fit their data's letter frequency distribution [1]. The function is as below:

$$y = \frac{C}{x^a}$$

Exponential is also a simple but effective function to fit all kinds of distribution. And it is the basis of other functions. Exponential formula is as follows:

$$y = \mathrm{C}\, e^{-ax}$$

Logarithmic is an extension of the Gusein-Zade function by allowing the coefficient of log(x) term to be independently fitted [23][24] . The function is as below:

$$y = \mathrm{C} - \mathrm{a} \log(\mathrm{x})$$

Weibull is a statistical distribution function proposed by Weibull. It could be used in many situations [25] . It corresponds to the stretched exponential cumulative and many people apply it to fit letter frequency distribution [26][27] . Weibull distribution is as follows:

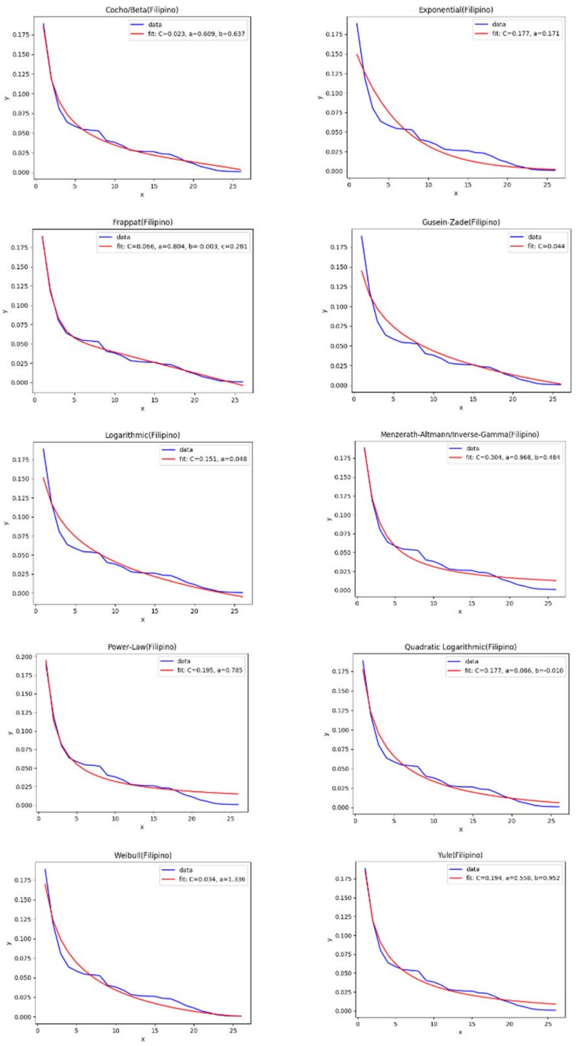

Figure 6. Fitting results for Filipino data.

$$y = \mathrm{C}\, (\log \frac{n+1}{x})^a$$

Quadratic Logarithmic is an extension of the logarithmic function by adding one extra term [23][24] . Quadratic Logarithmic formula is as follows:

$$y = \mathrm{C} - \mathrm{a} \log x - b\, (\log x)^2$$

Proposed by Frappat, Frappat adds a linear trend over the exponential function [34][35] . The function is as follows:

$$y = \mathrm{C} + \mathrm{bx} + ce^{-ax}$$

Proposed by Yule et al., Yule uses an exponential function. This function was put forward based on the conclusions of Dr. J. C. Willis, F.R.S. [27] . Li et al. used it to fit the letter frequency distribution and Martindale used it to fit graphemes and phonemes frequency distribution [28][29] . The function is as follows:

$$y = \mathrm{C}\, \frac{b^x}{x^a}$$

Menzerath-Altmann/Inverse-Gamma, concerning the relationship between the length of two linguistic units, uses an exponential function of the inverse of rank [27][30] . Menzerath-Altmann/Inverse-Gamma is as below:

$$y = \mathrm{C}\, \frac{e^{\frac{-b}{x}}}{x^a}$$

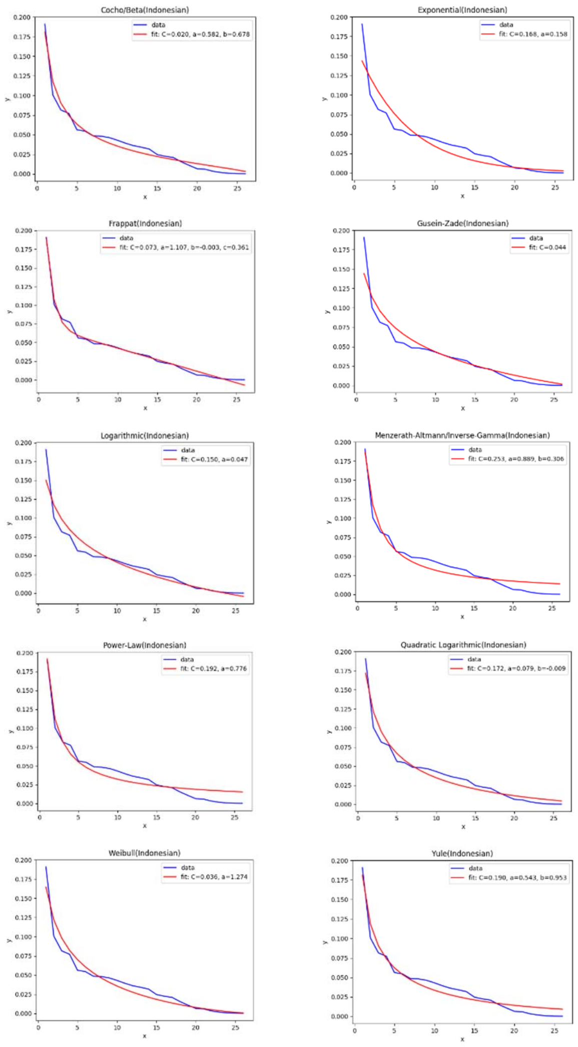

Figure 7. Fitting results for Indonesian data.

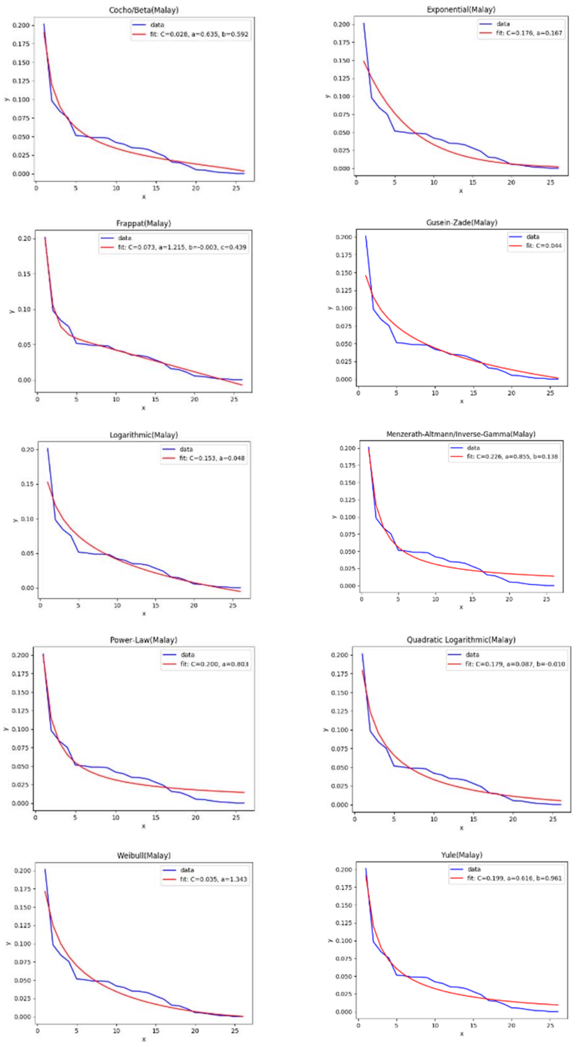

Figure 8. Fitting results for Malay data.

Cocho/Beta attempts to fit the two ends of a rank-frequency distribution by power-laws with different exponents [27][31][32][33] . It can be expressed as:

$$\mathrm{y} = \mathrm{C}\,\frac{(n+1-x)^b}{x^a}$$

We use these functions to fit data in four languages, and the results are shown in Figures 5-8. How well a function y fits the data can be measured by the sum of squared errors (residuals) SSE:

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - y_i')^2$$

where the loss of each model is measured by above function. Table 8 lists the SSE of each function.

The best function for Filipino, Indonesian and Malay, selected by SSE, is the Frappat function. For English, the best functions are Cocho/Beta and Quadratic Logarithmic. And Cocho/Beta function is also the second best function for Filipino, Indonesian and Malay. The SSE values of four languages' best fitting functions are all less than 0.001. The results show that the fitting of ranked letter frequency distribution of Filipino is the best, with a SSE value of 0.002, using Frappat as the fitting function. We could see that the statistics for the three Indonesian branch languages are quite similar.

TABLE VII. THE SSE OF EACH FUNCTION

Function	Language			
	Filipino	Indonesian	Malay	English
Gusein-Zade	0.0033	0.0034	0.0048	0.0010
Power-Law	0.0018	0.0025	0.0027	0.0049
Exponential	0.0043	0.0049	0.0064	0.0008
Logarithmic	0.0029	0.0030	0.0043	0.0009
Cocho/Beta	0.0006	0.0011	0.0016	**0.0005**
Weibull	0.0018	0.0023	0.0032	0.0007
Quadratic Logarithmic	0.0011	0.0018	0.0025	**0.0005**
Yule	0.0010	0.0017	0.0021	0.0008
Menzerath-Altmann/Inverse-Gamma	0.0015	0.0024	0.0026	0.0026
Frappat	**0.0002**	**0.0004**	**0.0005**	0.0006

VII. Conclusion

In this paper, we study the differences of letter usage among four languages in terms of vocabulary length, letter frequency distribution, common letter pairs and common letter trigrams. What's more, we fit the ranked letter frequency distributions with ten letter frequency distribution models. Our results show that there are considerable differences between three Indonesian branch languages and English. In addition, although Filipino, Indonesian and Malay belong to the same language branch, the differences between Filipino and the other two languages could not be ignored. We could also see that Indonesian and Malay are indeed very similar, which further validates that they share the same origin.

In future, we will try to study the differences of letter usage among these languages from other perspectives. And we also consider processing other languages in the Indonesian branch, to verify the results present in this work.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 61572145) and the Major Projects of Guangdong Education Department for Foundation Research and Applied Research (No. 2017KZDXM031). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

References

[1] W. Li and P. Miramontes, "Fitting Ranked English and Spanish Letter Frequency Distribution in U.S. and Mexican Presidential Speeches," J. Quant. Linguist. - JQL, vol. 18, 2011.

[2] W. F. Friedman, Elements of Cryptanalysis. Government Printing Office, Washington, 1976.

[3] N. Lin, S. Fu, S. Jiang, C. Chen, L. Xiao, and G. Zhu, "Learning Indonesian Frequently Used Vocabulary from Large-Scale News," in 2018 International Conference on Asian Language Processing (IALP), 2018, pp. 234–239.

[4] N. Lin, S. Fu, S. Jiang, G. Zhu, and Y. Hou, "Exploring Lexical Differences Between Indonesian and Malay," in 2018 International Conference on Asian Language Processing (IALP), 2018, pp. 178–183.

[5] G. A. Miller, E. B. Newman, and E. A. Friedman, "Length-frequency statistics for written English," Inf. Control, vol. 1, no. 4, pp. 370–389, 1958.

[6] F. Golcher, "Wiederholungen in Texten," Humboldt-Universität zu Berlin, Philosophische Fakultät II, 2013.

[7] R. H. Baayen, Word Frequency Distributions. Springer Netherlands, 2012.

[8] K. Best, "Results and perspectives of the Göttingen project on quantitative linguistics," J. Quant. Linguist., vol. 5, no. 3, pp. 155–162, 1998.

[9] K.-H. Best, "Häufigkeitsverteilungen in Texten. By Simone Andersen," Glottometrics, vol. 2, 2002.

[10] R. D. Smith, "Distinct word length frequencies: distributions and symbol entropies," CoRR, vol. abs/1207.2334, 2012.

[11] D. R Ridley and M. Lively, "English letter frequencies and their applications: Part I," Percept. Mot. Skills, vol. 96, pp. 545–548, 2003.

[12] R. L. Solso and J. F. King, "Frequency and versatility of letters in the English language," Behav. Res. Methods, vol. 8, pp. 283–286, 1976.

[13] A. Beutelspacher, Kryptologie: eine Einführung in die Wissenschaft vom Verschlüsseln, Verbergen und Verheimlichen ; ohne alle Geheimniskrämerei, aber nicht ohne hinterlistigen Schalk, dargestellt zum Nutzen und Ergötzen des allgemeinen Publikums. Vieweg+Teubner Verlag, 2007.

[14] F. Pratt, Secret and Urgent: The Story of Codes and Ciphers. Aegean Park Press, 1939.

[15] D. G. Simpson, "La Oftecoj de la Esperantaj Literoj," 2007.

[16] S. M. Gusein-Zade, "Frequency distribution of letters in the Russian language," vol. 24, pp. 338–342, 1989.

[17] A. Shah, A. Z. Saidin, I. F. Taha, and A. M. Zeki, "Frequencies Determination of Characters for Bahasa Melayu: Results of Preliminary Investigations," Procedia - Soc. Behav. Sci., vol. 27, pp. 233–240, 2011.

[18] B. Dahlqvist, "The Distribution of Characters, Bi- and Trigrams in the Uppsala 70 Million Words Swedish Newspaper Corpus," 1999.

[19] M. N. Jones and D. J. K. Mewhort, "Case-sensitive letter and bigram frequency counts from large-scale English corpora," Behav. Res. Methods, Instruments, Comput., vol. 36, no. 3, pp. 388–396, 2004.

[20] S. M. Gusein-Zade, "On the frequency of meeting of key words and on other ranked series," Sci. Information, Ser. 2 Inf. Process. Syst., vol. 1, pp. 28–32, 1987.

[21] M. Y. Borodovsky and S. M. Gusein-Zade, "A General Rule for Ranged Series of Codon Frequencies in Different Genomes," J. Biomol. Struct. Dyn., vol. 6, no. 5, pp. 1001–1012, 1989.

[22] A. Clauset, C. Shalizi, and M. Newman, "Power-Law Distributions in Empirical Data," SIAM Rev., vol. 51, no. 4, pp. 661–703, 2009.

[23] I. Kanter and D. A. Kessler, "Markov Processes: Linguistics and Zipf's Law," Phys. Rev. Lett., vol. 74, no. 22, pp. 4559–4562, 1995.

[24] A. Vlad, A. Mitrea, and M. Mitrea, "Two frequency-rank law for letters in printed Romanian," Proces. del Leng. Nat., vol. 26, pp. 153–160, 2000.

[25] W. Weibull, "A Statistical Distribution Function of Wide Applicability," ASME Journal of Applied Mechanics, Vol. 18, pp. 293-297, 1951.

[26] T. Nabeshima and Y.-P. Gunji, "Zipf's law in phonograms and Weibull distribution in ideograms: Comparison of English with Japanese," Biosystems., vol. 73, no. 2, pp. 131–139, 2004.

[27] W. Li, P. Miramontes, and G. Cocho, "Fitting ranked linguistic data with two-parameter functions," Entropy, vol. 12, no. 7, pp. 1743–1764, Jul. 2010.

[28] F. Y. E. and G. U. Yule, "A Mathematical Theory of Evolution Based on the Conclusions of Dr. J. C. Willis, F.R.S.," J. R. Stat. Soc., 2006.

[29] C. Martindale, S. M. Gusein-Zade, D. McKenzie, and M. Y. Borodovsky, "Comparison of equations describing the ranked frequency distributions of graphemes and phonemes," J. Quant. Linguist., 1996.

[30] G. Altmann, "Prolegomena to Menzerath's law," Glottometrika, vol. 2, pp. 1–10, 1980.

[31] R. Mansilla, E. Köppen, G. Cocho, and P. Miramontes, "On the behavior of journal impact factor rank-order distribution," J. Informetr., 2007.

[32] G. G. Naumis and G. Cocho, "Tail universalities in rank distributions as an algebraic problem: The beta-like function," Phys. A Stat. Mech. its Appl., 2008.

[33] G. Martínez-Mekler, R. A. Martínez, M. B. del Río, R. Mansilla, P. Miramontes, and G. Cocho, "Universality of rank-ordering distributions in the arts and sciences," PLoS One, 2009.

[34] L. Frappat, C. Minichini, A. Sciarrino, and P. Sorba, "Universality and Shannon entropy of codon usage," Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top., 2003.

[35] L. Frappat and A. Sciarrino, "Conspiracy in bacterial genomes," Phys. A Stat. Mech. its Appl., 2006.