

Statistical Comparison of Character Use Between English and Melanesian Languages

Shigeharu SUGITA*

Abstract

This paper simply presents several statistical data concerning character use in the text written in different languages. They are 1) frequency table of alphabet use, 2) distribution of word length, 3) distribution of sentence length measured by number of words used in the sentence, and 4) digram of alphabet connection.

Main object of this work is to make statistical comparison between English and Melanesian languages such as Pijin, Fijian, Bislama, and ARE'ARE.

Sample texts used for analysis are obtained from Fiji and Solomon islands when the second scientific survey group of the south pacific, organized by the Kagoshima University Research Center for the south pacific, visited to their island in 1982.

The solomon Pijin is not used for writing purpose but used for spoken purpose. Therefore character use has not so much meaningfull. But nowaday there are appearing newspapers and textbooks which are written in Pijin or Fijian etc., so that written from of their languages are becoming popular in Islands.

Sample texts

The textbooks inputted into computer for this purpose are listed below. The numerals indicated inside of parentheses means total number of words used in that text.

PIJIN	Stori Abaotem Sir Jacob Vouza	(2182 words)
	Samfala Kastom Stori an Kastom Wei	(5883 words)
	Kastom Stori long Narafala	(4100 words)
	Mista Aao an Mista Totel	(18100 words)
FIJIAN	Tavialolowai	(3460 words)
	Na I Tovo Vakapasifika	(43427 words)
	Dua Ga Na Bogi	(68548 words)
BISLAMA	Ripot blong rijinal konfrens long lan manejmen	(20760 words)
ARE'ARE	Six Stories of Rapu" anate	(8617 words)
ENGLISH	The golden bough	(411866 words)
	The chrysanthemum and sword	(89806 words)
	The Burman	(221147 words)

* National Museum of Ethnology, Expo Park, Senri, Suita, Japan

Statistics

3.1 Alphabet use

A simple frequency of alphabet use in each text is counted. Graph 1 shows distribution of each character in percent scale. In this graph, symbols such as space and hyphen etc are omitted.

3.2 Word length

Graph 2 shows distribution of word length measured by number of characters. Vertical unit of graph is percent (%).

3.3 Sentence length

Graph 3 shows distribution of sentence length measured by the number of words in the sentence. Horizontal unit of the graph is five words, that is, the first position indicates the percentage of total sentence of length from 1 to 5 words, the next position from 6 to 10, etc.

3.4 Digram ; frequency table of alphabet connection

Table 1 to 5 show the connection frequency of alphabets. The frequency is represented by the largeness of square. The blank position does not necessarily mean zero frequency, because small value can not be printed out in visual size.

In this table, if you want to see the frequency of alphabet connection, for example "BY", first look up the left most column for "B", and then look up the upper most row for "Y". The cross point of "B" row and "Y" column is the position of "BY" connection.

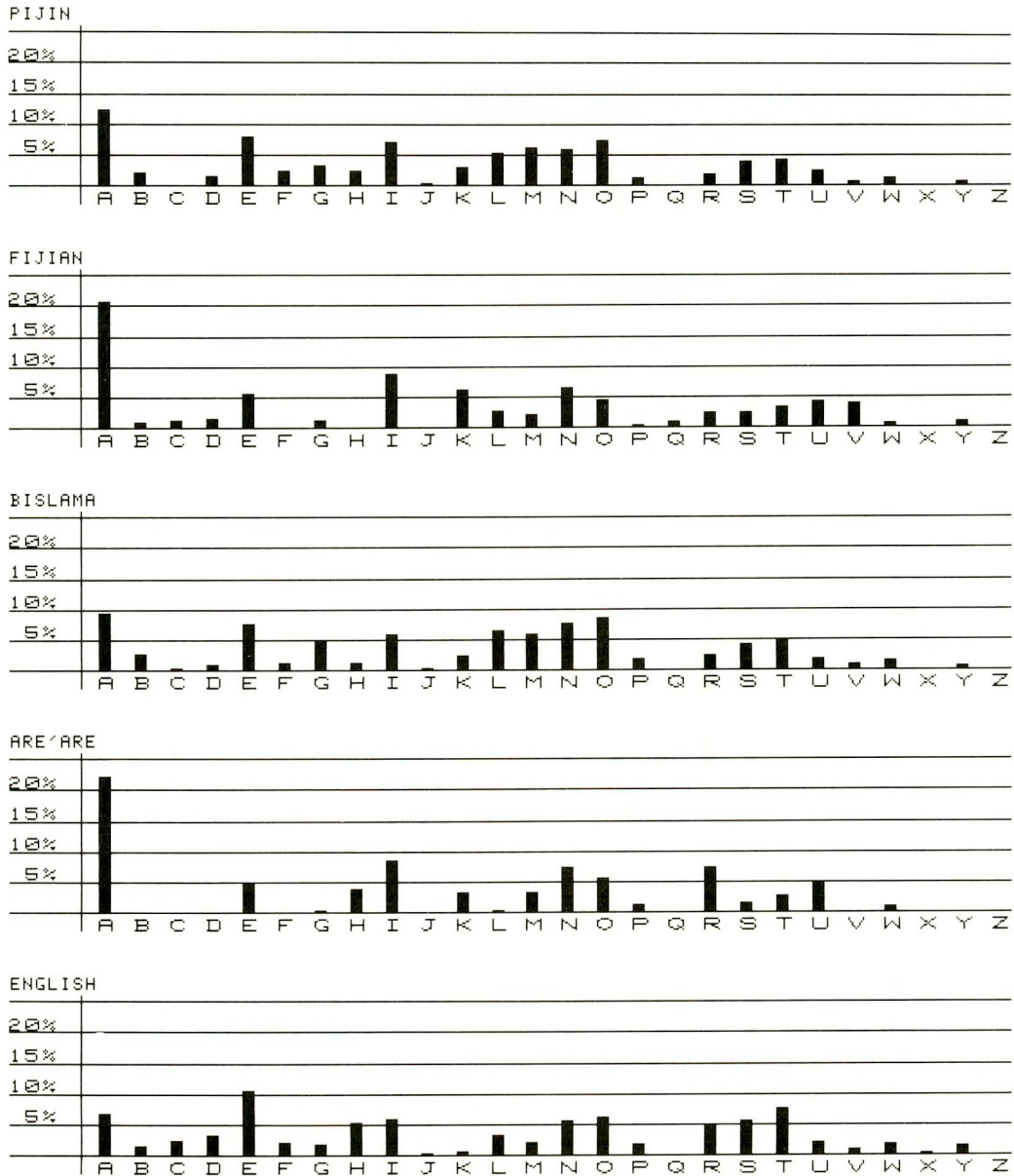
Aknowledgements

This work was supported by the special research grant of the ministry of education, science and culture, Japan (The second scientific survey of the south pacific, organized by the Kagoshima University Research Center for the South Pacific in 1982).

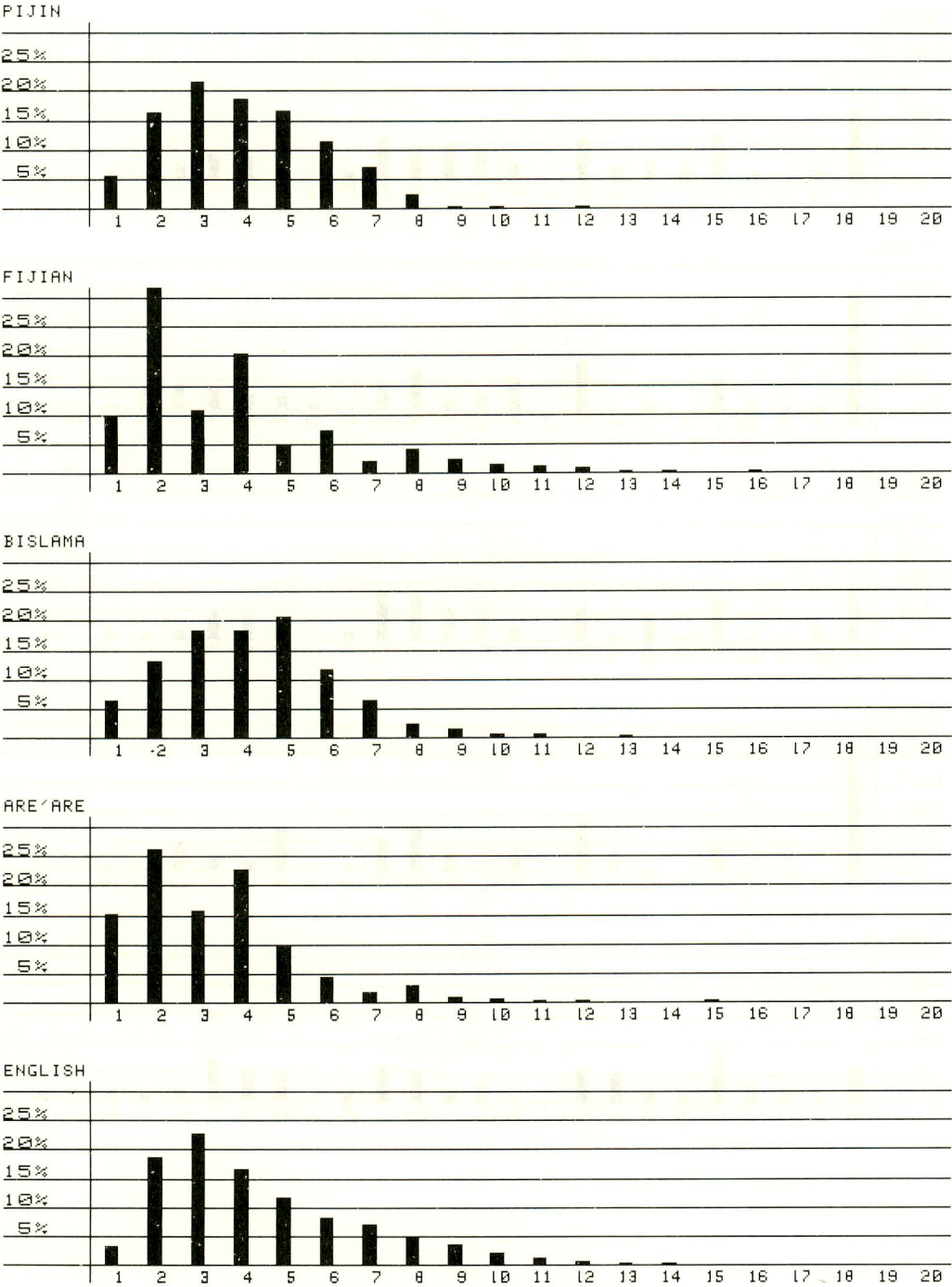
I would like to thank staffs of Research Center for the South Pacific, especially Professors Tadao Katayama, Hachiro Hirata, Shinichi Terashi, and Yoshiaki Matsuda.

I also thank Professor Yuuichi Wada, National Museum of Ethnology, who was a co-researcher of the language project and advised me to gather textbooks in the field.

Graph 1. Frequency of alphabet use in the text



Graph 2. Distribution of word length



Graph 3. Distribution of sentence length

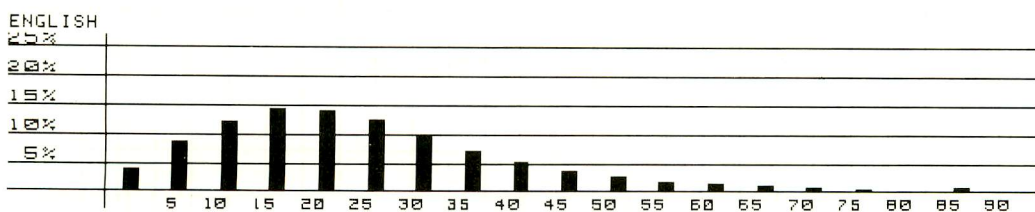
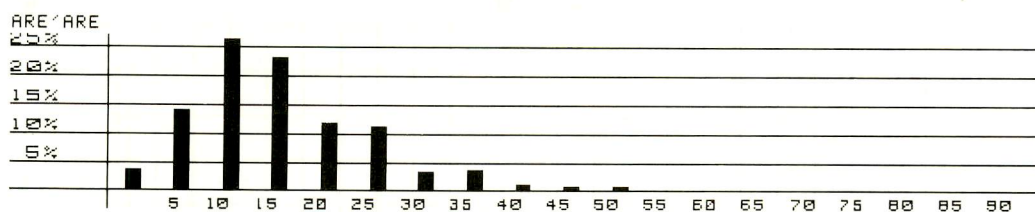
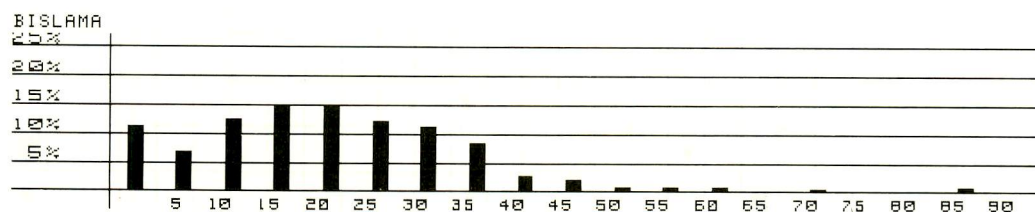
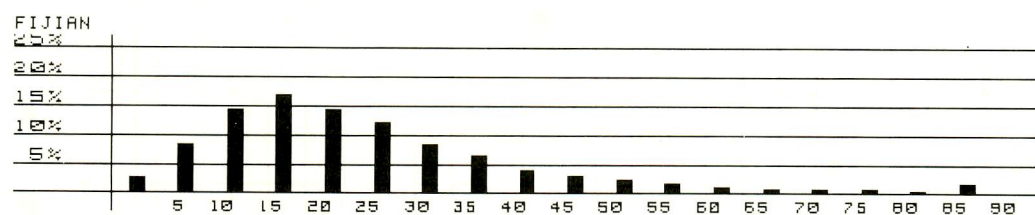
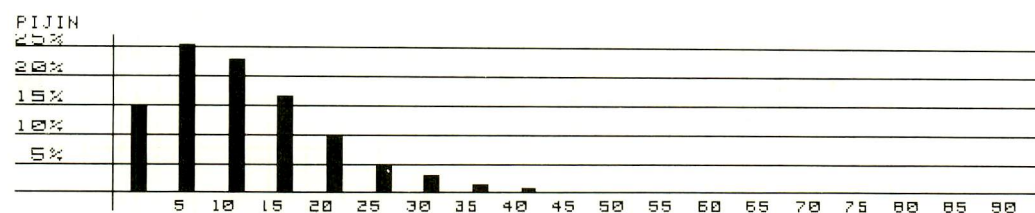


Table 1. Matrix of character combination in the text (The sign * Represents space or symbols)

PIJIN																												
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	*	
A					■				•	•	■	■	■	■	•			•	•	•		•						■
B	■								•			■			•													
C																												
D	•				■	•			•						•				•		•							•
E	•	•		•					•	•	•	■	•					•	•	■		•						■
F	■								•						•													
G															•						•							■
H	•				■										•													
I	■					•	•	•			•	•	•	■		•				•	•	•						■
J																												
K	■				■				•							•												•
L	■				•				•		■					■				•	•							•
M	■	•		•	•	•			■				•		•	•			•	•	•							■
N	■			•	•	•	■		•						•													■
O	•											■	•	■					•	•	•							■
P									•			•			•													•
Q																												
R	■				■				•							•						•						
S	■				■	•			•		•	•	•	•	•						■	•		•				■
T	■				•				■							•						■						■
U						•					•		•															•
V																												
W																												
X																												
Y																												
Z																												
*	■	■		■	•	■	■	■	■	■	•	■	■	■	■	■	•		•	■	■		■		•			

Table 2. Matrix of character combination in the text (The sign * represents space or symbols)

FIJIAN																												
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	*	
A		.	.	.					■		■	■	.	■		.		■	.	■	■		■
B	.																				.							
C	■								.					.														
D	
E		.	.			.			■						■	
F																												
G	■																											
H																												
I					■	.	.	■	.			.	.	■	.	.	.				■	
J																												
K	■				■				■						■						.							
L	■				■				■						.					.								
M	■	.			■				
N	■			.					■					.						.								
O											■	
P				.																								
Q	.														.						.							
R	■											
S	■				■				.						.					.								
T	■								■						.						■							
U								■	
V	■				■				.					.							■							
W	
X																												
Y	■																											
Z																												
*	.	.	.	■	■		.		■		■	■	■	■	■	■	.	■	.		.			

Table 4. Matrix of character combination in the text (The sign * represents space or symbols)

[illegible]

Table 5. Matrix of character combination in the text (The sign * represents space or symbols)

ENGLISH																												
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	*	
A					■	.	■		.		■	■	■		.			.		■	
B					■																							
C	.				■									
D	■	■	
E	■	.	.	■	■		.		■	■		■	
F									.																		■	
G				.				.																			■	
H	■				■				■											■	
I	■	■	■		.					.	
J																												
K																												
L	■			.	■				.			■				■	
M	■	.	.		■	■	
N	.	.	■	■			■		.					.	■				■		.	■	.	.			■	
O						■							.	■				■		.	■	.	.				■	
P										
Q																					.							
R	■		.	.	■				■			■	
S	.		.	.	■			■	.	.					■	
T	.			.				■	■	.					■			.	.								■	
U																		
V				.					.																			
W	
X																												
Y																											■	
Z																												
*	■	■	■	■	■	■	.	■	■	.	.	.	■	.	■	■		.	■	■	■	.	.	■	.			