STATISTICAL ANALYSIS OF JAPANESE CHARACTERS

by
Takushi Tanaka
The National Language Research Institute
3-9-14 Nishigaoka Kita-ku, Tokyo

Summary

The purpose of this study is to analyze the statistical property of Japanese characters for computer processing. Sentences in high school textbooks and newspapers have been investigated in this study. This paper contains the following points : the number of different words written in each character, position of characters in a word, relation between word boundaries and character strings, relation between parts of speech and patterns of character strings, relation between parts of speech and each character.

The results of these investigations can be applied to the processing of written Japanese for practical purpose.

1. Introduction

There are several different aspects between English and Japanese in the information processing of natural language. The first concerns the number of characters. In order to write Japanese more than 2,000 characters are used. The second concerns the way of writing. A Japanese sentence consists of a continuous character string without any space between words. The third concerns word order and other syntactic features. Among these aspects, the second and third features are closely related to the characters.

Japanese characters consist of three kinds. A KANJI (Chinese character) is used to write nouns and the principal part of a predicate, and expresses the concepts contained in the sentence. A HIRAGANA (traditional Japanese character) is used to write conjunctions, adverbs, JODOSHI (mainly expresses many modalities of a predicate) and JOSHI (post-position, mainly expresses case relations). A KATAKANA (traditional Japanese character) is used mainly as phonetic signs to write foreign words.

Accordingly, Japanese characters are regarded as elements of words, at the same time, they function to characterize the syntactic or semantic classes of words and express word boundaries in a character string.

The following Japanese character strings, (A) to (D), are the same sentences written by using KANJI to different degrees.
(D) is quoted from a high school textbook (world history).
While (A), (B) and (C) are transliterated from (D) by computer.[1,2]

(Example of Japanese sentence)

(A)

ヨーロッパれっきょうはしほんしゅぎがいちじるしくはったつすると，せいさんぶつのはんろをかいがいにひろげるために，くんびをかくちょうしてじつりょくてしょくみんちをもとめるきょうそうをおこなった．こうして２０せいきをむかえるとていこくしゅぎのじだいとなり，そのなみはとうようにおよんて，れっきょうのちゅうごくぶんかつがはじまった．

(B)

ヨーロッパれっ強はし本主ぎがいちじるしく発たつすると，生産物のはんろを海外にひろげるために，軍びをかくちょうして実力てしょく民地をもとめるきょうそうをおこなった．こうして２０世きをむかえるとてい国主ぎの時代となり，そのなみは東ようにおよんて，れっ強の中国分かつがはじまった．

(C)

ヨーロッパ列強は資本主義が著しく発達すると，生産物のはん路を海外にひろげるために，軍備を拡張して実力で植民地を求める競争をおこなった．こうして２０世紀をむかえるとてい国主義の時代となり，その波は東洋におよんて，列強の中国分割がはじまった．

(D)

ヨーロッパ列強は資本主義が著しく発達すると，生産物の販路を海外にひろげるために，軍備を拡張して実力て植民地を求める競争をおこなった．こうして２０世紀をむかえると帝国主義の時代となり，その波は東洋におよんて，列強の中国分割がはじまった．

(A) is written in KATAKANA (only for 'ヨーロッパ') and HIRAGANA (the rests) without using KANJI.

(B) is written in HIRAGANA, KATAKANA and 200 KANJI of high frequency in Japanese writing.

(C) is written in HIRAGANA, KATAKANA and the so-called educational KANJI (996 characters).

Low graders in elementary school tend to write sentences like (A). The older they get the more KANJI they learn and they begin to write sentences like (D) in high school. When we read sentences like (A), we realize it is very difficult to read them, because we cannot find word boundaries easily. On the other hand, in (B), (C) and (D) we find less difficulty in this order. Because we can easily find out word boundaries by means of KANJI in a character string. Boundaries between a HIRAGANA part and a KANJI part play a role to indicate word boundaries in many cases. We can also grasp main concepts in a sentence by focusing our attention to the KANJI parts of the sentence.

Therefore, it is very important to use HIRAGANA and KANJI appropriately in a character string. It is, however, hard to say the rules for the appropriate use of HIRAGANA and KANJI have been established. Due to the fact, it is necessary for us to study more about the actual use of Japanese characters. Because, explication of rules for the appropriate use of the characters is a prerequisite for information processing in commonly written Japanese.

2. Outline of Japanese characters

Fig.1 illustrates the rate of total characters contained in the high school textbooks (9 subjects × 1/20 sampling). The data contains 48,096 characters in total.[3] HIRAGANA occupies the first place accounting for 47.1%. According to the result of Nakano's study which will be presented here, KANJI takes the first place in the newspaper, because they have TV-programs and mini advertisement which are both written mainly in KANJI.[4]

Fig.2 illustrates the rate of different characters in the data of textbooks. The data contains 1,525 different characters. KATAKANA and HIRAGANA are composed of basic 47 characters respectively, however the data also contains variations like small letters and letters with special symbols, and both kind of KANA exceed 70. Most of HIRAGANA and KATAKANA were appeared in the data of textbooks. The data contains 1,312 different KANJI. The more data is investigated the more KANJI appear, and the rate of KANJI increases in the graph.

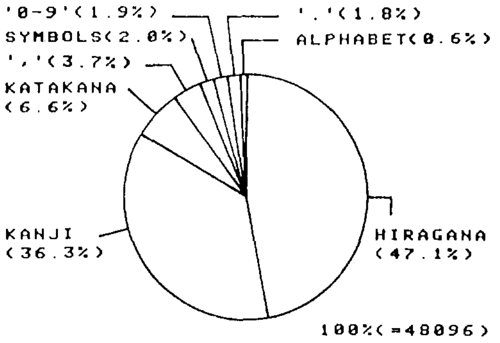


Fig.1 Rate of total characters

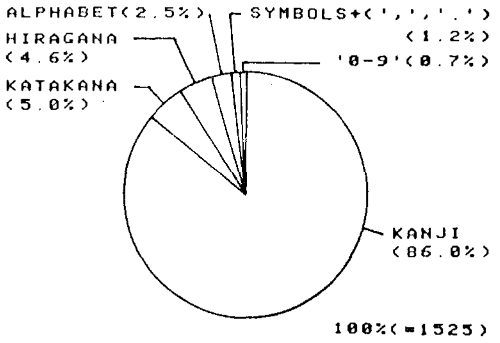


Fig.2 Rate of different characters

According to the investigation of Nomura 3,213 KANJI were found in the newspaper.[5] The largest Japanese KANJI dictionary (edited by Morohashi) contains about 50,000 characters.[6]

Fig.3 shows relation between frequency and order of frequency in every kind of characters. From Fig.3 we see that a few HIRAGANA have high frequency. They play an important role in writing grammatical elements in a sentence as JOSHI and JODOSHI.

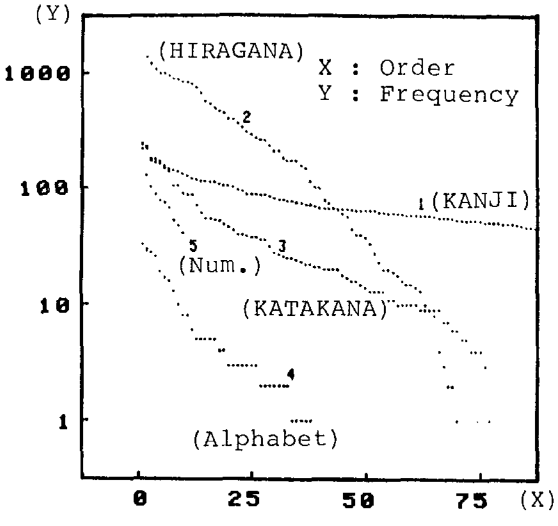


Fig.3 Frequency and Their order

Fig.4 shows the relation between order of frequency and total number of characters up to their order. In this graph, we see about twelve different HIRAGANA occupy 50% of total HIRAGANA. About 120 different KANJI occupy 50% of total KANJI.

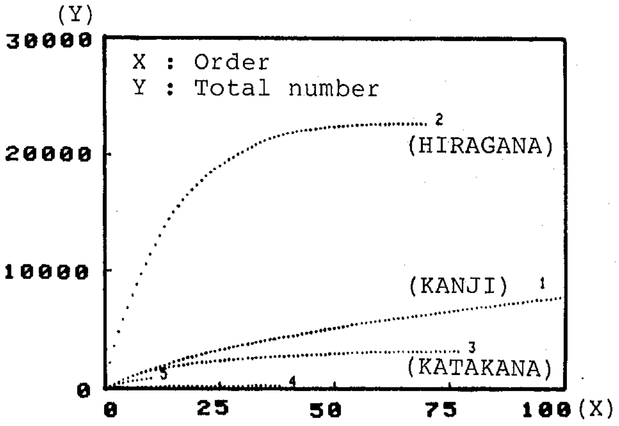


Fig.4 Order and Total up to the order

3. Number of different words written in each character

As we have more than 50,000 characters, it is necessary to decide the degree of importace of them. In order to decide the degrees two criteria are assumed here. One is the frequency of the characters. The other one is the number of different words in which the same character is used. The similar concept has been proposed by A. Tanaka.[7]

In Fig.5, axis X represents the frequency of the character as first criterion. Axis Y represents the number of different words in which the same character is used. The graph shows the distribution of characters in the textbooks except KANJI. Each character on Y=1 is used for only one word. For instance, HIRAGANA 'を'(o) on Y=1 is used for only one word (one of case-JOSHI, indicating accusative case) exclusively. Each character on Y=X is used for a new word in every occurrence of the character.

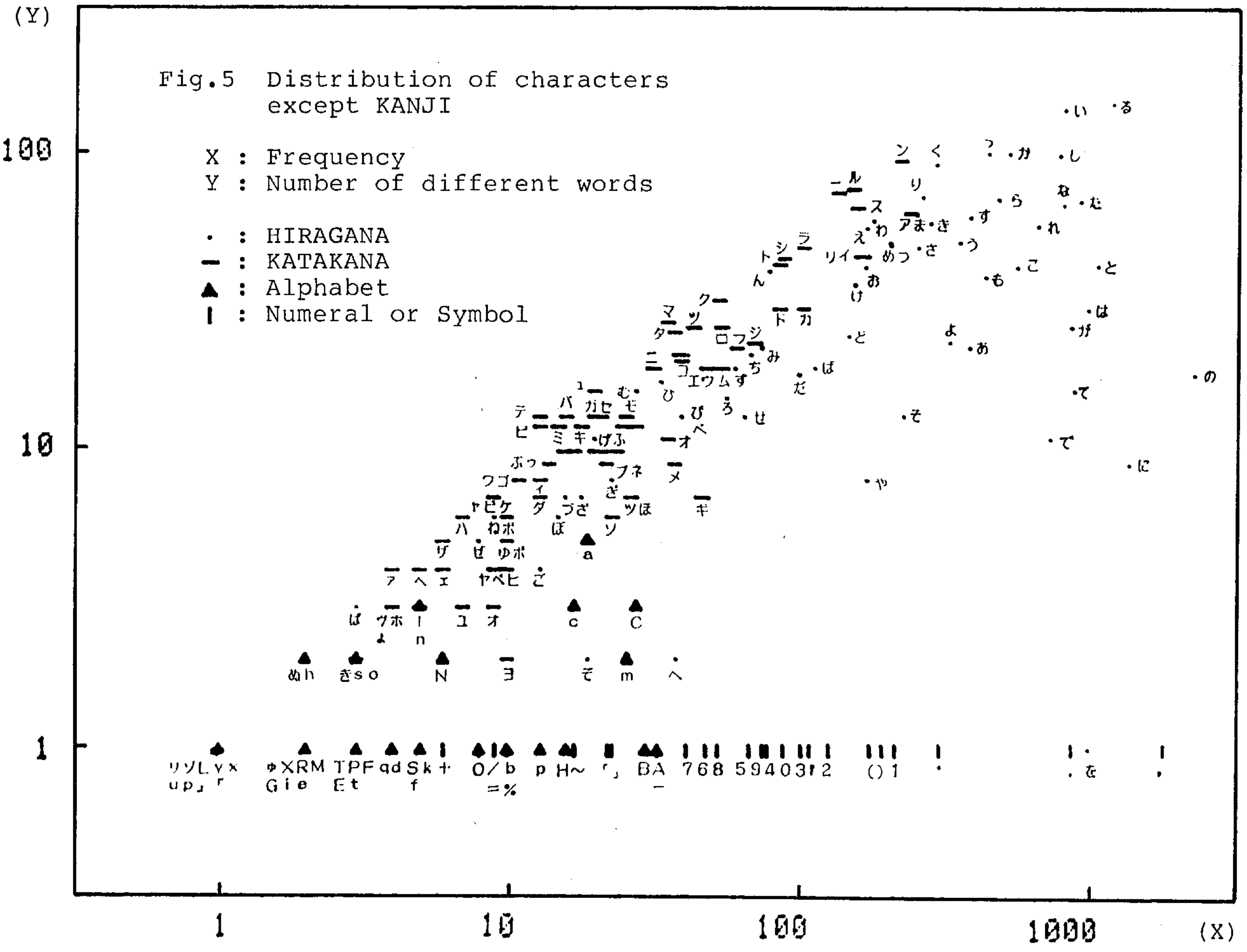


Fig.5 Distribution of characters except KANJI

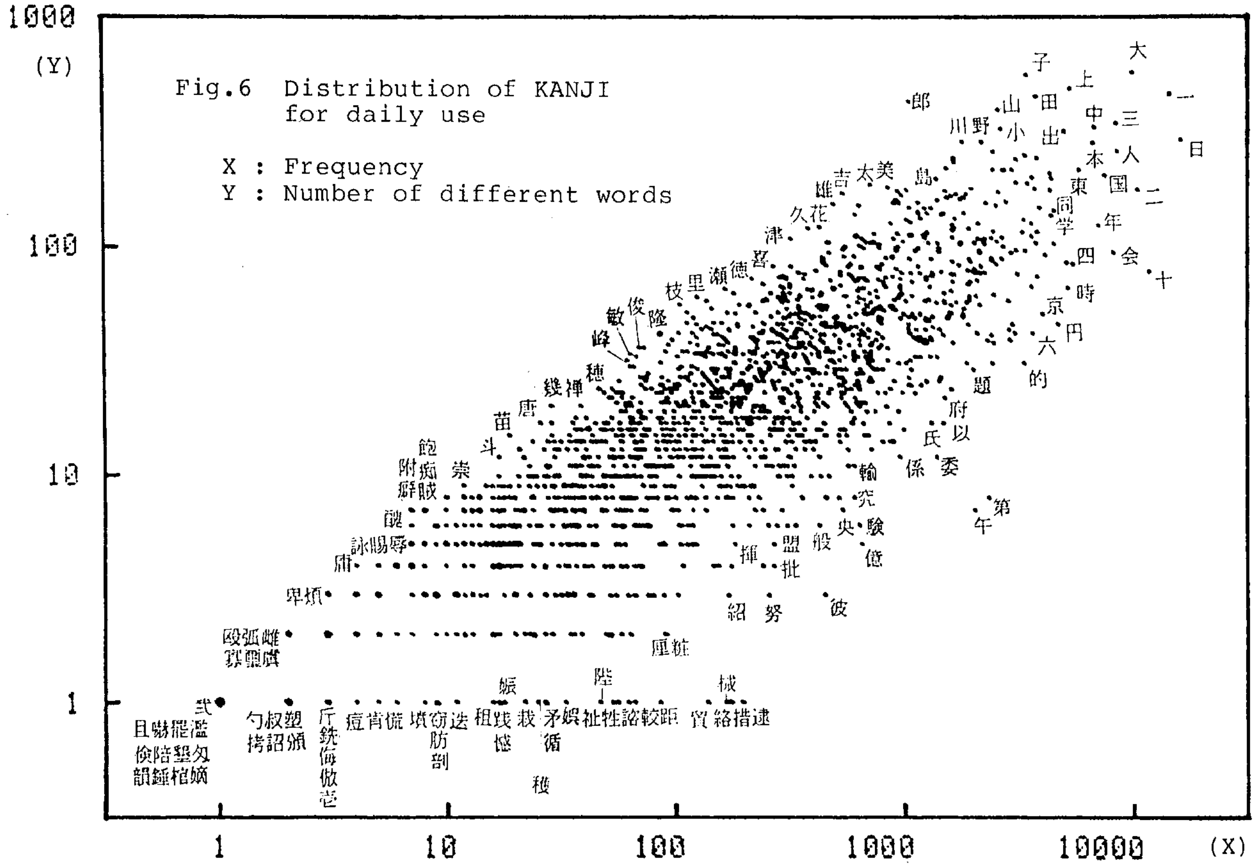


Fig.6 Distribution of KANJI for daily use

X : Frequency
Y : Number of different words

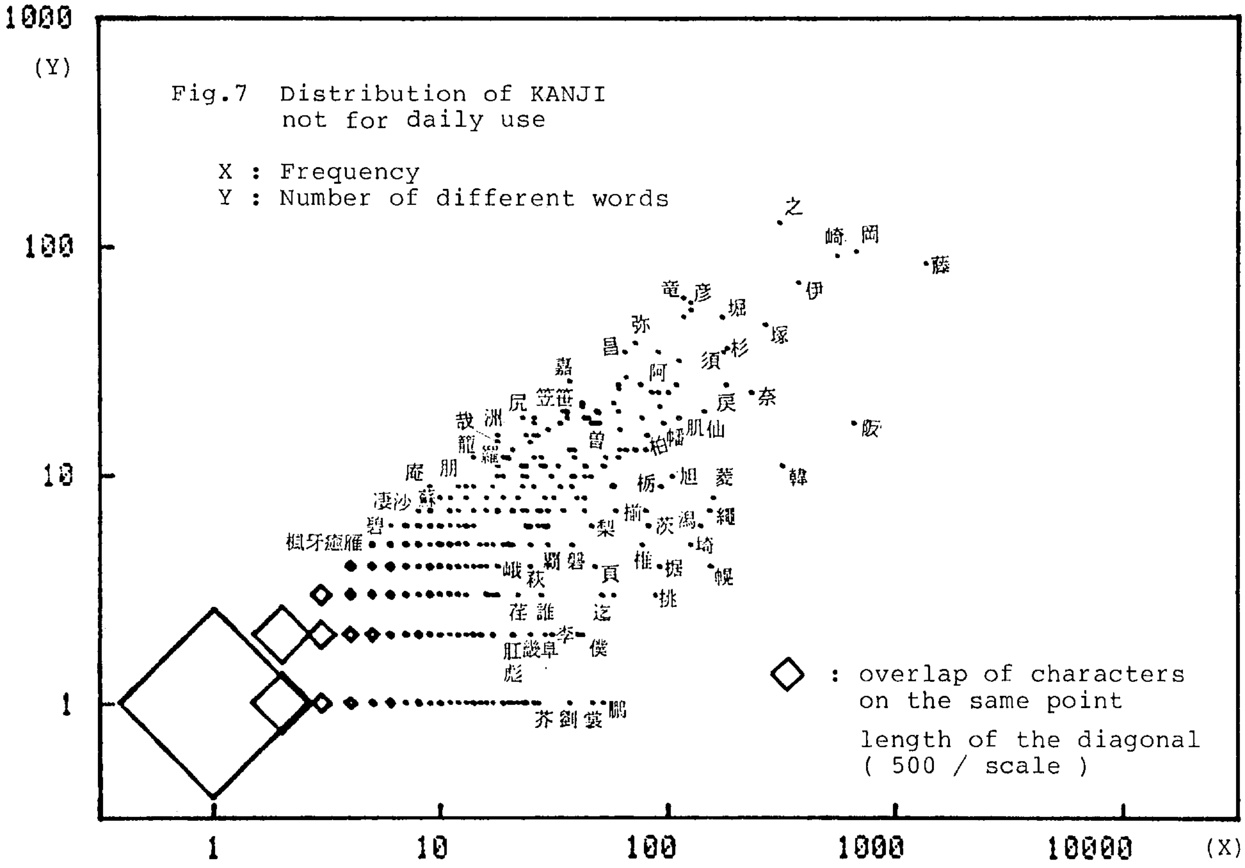


Fig.7 Distribution of KANJI not for daily use

X : Frequency
Y : Number of different words

◇ : overlap of characters on the same point
length of the diagonal (500 / scale)

KATAKANA appear near Y=X, because KATAKANA are mainly used for writing proper nouns of foreign words. The same words of such a category do not appear frequentry.

HIRAGANA,'る'(ru),'い'(i),'し'(shi), 'っ'(tsu), 'か'(ka) and 'く'(ku) are localized on the upper right side. These are often used for writing some parts of inflectional forms of verbs (e.g. 'い' for '書いた', 'し' for 'しない', 'か' for '行かない'). 'い'(i), 'か'(ka) and 'く' (ku) are also often used for writing some parts of inflectional forms of adjectives. 'の'(no), 'に'(ni), 'を'(o), 'は'(wa), 'と'(to), 'が'(ga) and 'で'(de) on the right side are frequently used for JOSHI (post-position, expressing-case relations or other grammatical relations). 'た'(ta) on the upper right side is often used for JODOSHI of the past tense. 'な'(na) on the upper right side is often used for the initial syllable of JODOSHI of negative.

Fig.6 and Fig.7 show the same investigation into the KANJI of newspapers (the original work was carried out by Nomura).5

Fig.6 shows the distribution of the so-called "TOYOKANJI" selected by the Japanese government for daily use in 1946. The upper right area on the graph is occupied by the so-called educational KANJI. Each KANJI on Y=1 is used only for one word (e.g. '逮'(tai) for '逮捕' (taiho : arrest), '貿'(bou) for '貿易' (boueki : trade), '械'(kai) for '機械' (kikai : machine)). The same as Fig.5, characters used for persons' names are localized near Y=X.

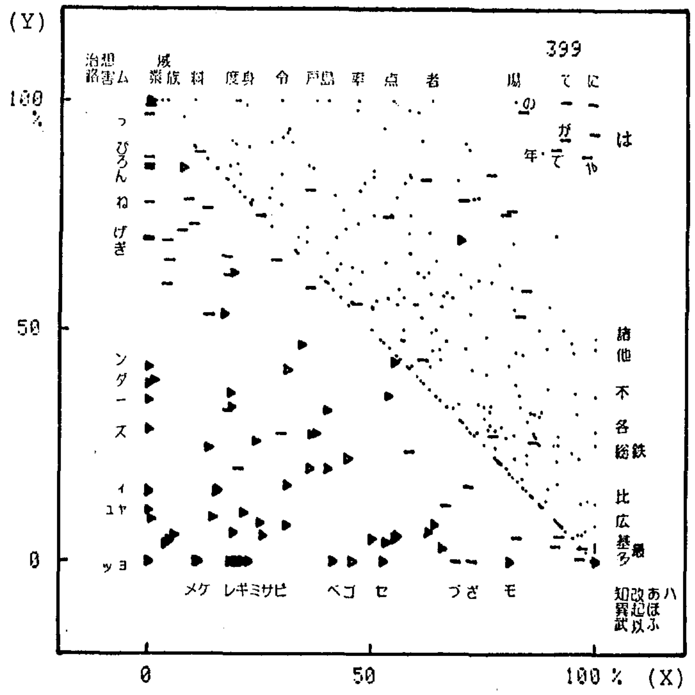
Fig.7 shows the distribution of KANJI other than TOYOKANJI. The most of characters in upper right part of the graph are the ones which are used for persons' names or for place names. (e.g. '藤' and '崎' for '藤崎'(Fujisaki:person) '岡' for '福岡'(Fukuoka:place).

4. Position of characters in a word

For the information processing of Japanese sentences, at first, it is important to find out word boundaries in a continuous character string. If there are some characters which always come to the initial position or the final position of a word, these characters are available to find the boundaries.

Fig.8 shows the position of characters in words. In the data of textbooks, there are 399 characters which are used for more than 6 kinds of different words. The characters on X=100 always come to the initial position of a word. The characters on X=0 are never used at the initial position. The characters on Y=100 always come to the final position of a word. The characters on Y=0 are never used at the final position.

KANJI, represented with dots, spread over the area of Y≥-X+100. Namely, the value of X+Y are always greater than or equal to 100. In other words, rates of the initial position plus final position are always greater than or equal to 100%. It means that all KANJI have a tendency to be used for the initial position or the final position or both position (as a word of one character) of a word (short unit *). Most KANJI on Y = -X+100 form only words of two KANJI. The tendency originates in the composition of words written by KANJI. This matter will be observed in section 6. The group of HIRAGANA in the upper right area has a tendency to be used for JOSHI. KATAKANA represented by '▶' appear around the under left area on the graph. Words written in KATAKANA have relatively long length (See section 6). Therefore, the rates of the initial position and the final position are relativery decreased.



X : Rate of initial position
Y : Rate of final position

Fig.8 Position of character in a word

* word (long unit) : 国立国語研究所 (National-language-research-institute)

word (short unit) : 国立, 国語, 研究, 所 (National,Language,Research,Institute)

5. Relation between word boundaries and character strings

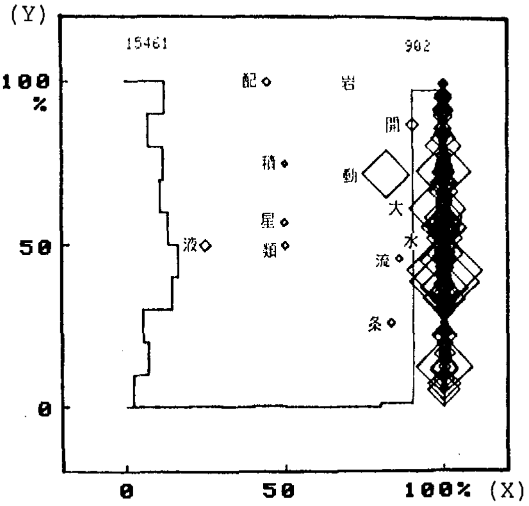
(Simple Japanese grammar)

$$N_1 J_1 N_2 J_2 \cdots V. \quad \text{———} \quad (1)$$

N_i : Noun
J_i : Case-JOSHI for N_i
V : Verb

A Japanese sentence fundamentally belongs to pattern (1). Many nouns (N_i) tend to be written in KANJI (See next section). All the case-JOSHI are written in HIRAGANA. Stems of verbs are often written in KANJI and their inflectional parts in HIRAGANA. So both a phrase of $N_i J_i$ and V have such a pattern that the initial position is occupied by a KANJI and the final position is occupied by a HIRAGANA. Therefore, the changing point from HIRAGANA to KANJI in a character string is always regarded as a word boundary. On the other hand, a word boundary is not always a changing point from HIRAGANA to KANJI. One of the exception is Japanese nouns (long unit) which are composed of some concatenation of nouns (short unit). (See page 5 *)

Fig.9 shows one of the relations between word boundaries and character strings. The graph contains 902 KANJI (total : 1,546) in the textbooks. The axis X represents the rate that the changing points from HIRAGANA to KANJI correspond to word boundaries. Each KANJI on X=100 is considered as the initial character of a word if it is preceeded by a HIRAGANA. The axis Y represents the rate that the word boundaries correspond to changing points from HIRAGANA to KANJI. The symbol of '◇' represents a KANJI. The length of diagonal of '◇' is proportionate to the frequency of the KANJI. In the graph, the length of 10% of axis is equal to 100 times of the frequency.



X : Rate of word boundary
Y : Rate of H-K boundary

Fig.9 Character string and boundary

6. Parts of speech and patterns of character strings

In the investigation of newspapers, 20 parts of speech were assumed.[8] Each part of speech has a particular pattern of character strings. It is possible to decide the part of speech of a word based on the knowledge of such patterns in computer processing of Japanese sentences.

In Fig.10, 'K' in the column of pattern represents a KANJI, 'H' represents a HIRAGANA, and 'I' represents a KATAKANA. The left side of the bar chart shows the rate of total words. The right side of the bar chart shows the rate of different words.

Fig.10-(1) shows the pattern of common nouns. The left side of the bar chart shows that KK-pattern accounts for 68.0% of total common nouns in the newspapers. The right side of the bar chart shows that KK-pattern accounts for 68.5% of different common nouns in the newspapers.

Fig.10-(2) shows the pattern of proper nouns. Most of the proper nouns also have KANJI strings. The rest of proper nouns have KATAKANA strings expressing foreign words.

Fig.10-(3) shows the pattern of verbal nouns which change to verbs with succeeding characters 'せ'(se), 'さ'(sa) 'し'(shi), 'す'(su), 'する'(suru), 'すれ'(sure), 'せよ'(seyo). The verbal nouns consist of KK-pattern up to 97.1% of total. If KK-pattern and succeeding characters 'せ'(se),'さ'(sa), 'し'(shi) ...are found, such a character string can be treated as a form of this kind.

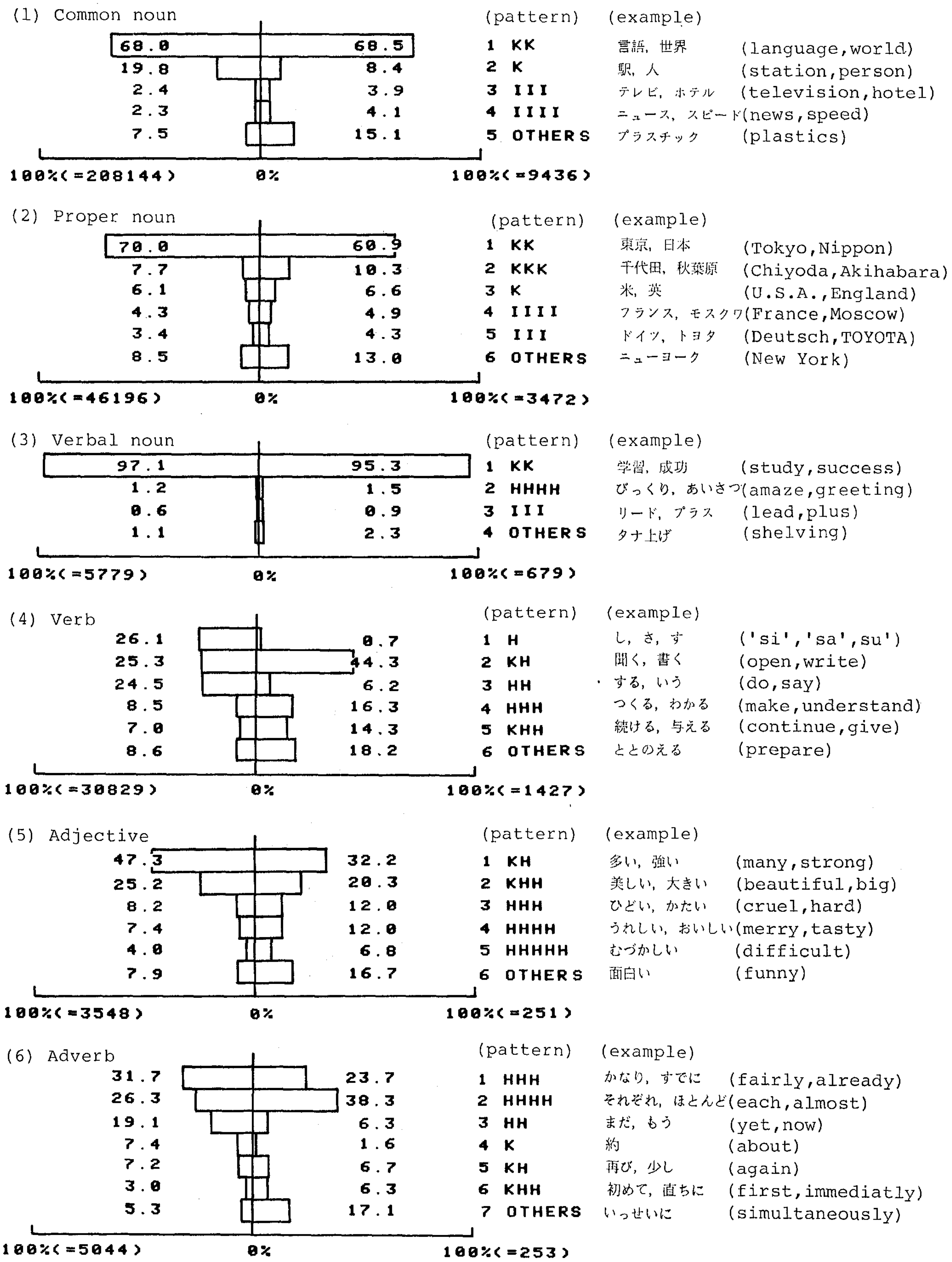
Fig.10-(4) shows the pattern of verbs. The verb of H-pattern is often used with preceding verbal nouns. Most different verbs have KH-pattern.

Fig.10-(5) shows the pattern of adjective. Most of the adjectives are written with KH-pattern or KHH-pattern.

Fig.10-(6) shows the pattern of adverbs. Most of the adverbs are written with HHH-pattern or HHHH-pattern. Namely they are written in HIRAGANA.

7. Relation between each character and part of speech

We have assumed patterns of character strings and the patterns are basically available for classifing part of speech in actual data. However, the patterns do not provide sufficient criteria for the classification. For example, the



(1) Common noun

		(pattern)	(example)	
68.0	68.5	1 KK	言語, 世界	(language,world)
19.8	8.4	2 K	駅, 人	(station,person)
2.4	3.9	3 III	テレビ, ホテル	(television,hotel)
2.3	4.1	4 IIII	ニュース, スピード	(news,speed)
7.5	15.1	5 OTHERS	プラスチック	(plastics)

100%(=208144) 0% 100%(=9436)

(2) Proper noun

		(pattern)	(example)	
70.0	60.9	1 KK	東京, 日本	(Tokyo,Nippon)
7.7	10.3	2 KKK	千代田, 秋葉原	(Chiyoda,Akihabara)
6.1	6.6	3 K	米, 英	(U.S.A.,England)
4.3	4.9	4 IIII	フランス, モスクワ	(France,Moscow)
3.4	4.3	5 III	ドイツ, トヨタ	(Deutsch,TOYOTA)
8.5	13.0	6 OTHERS	ニューヨーク	(New York)

100%(=46196) 0% 100%(=3472)

(3) Verbal noun

		(pattern)	(example)	
97.1	95.3	1 KK	学習, 成功	(study,success)
1.2	1.5	2 HHHH	びっくり, あいさつ	(amaze,greeting)
0.6	0.9	3 III	リード, プラス	(lead,plus)
1.1	2.3	4 OTHERS	タナ上げ	(shelving)

100%(=5779) 0% 100%(=679)

(4) Verb

		(pattern)	(example)	
26.1	0.7	1 H	し, さ, す	('si','sa',su')
25.3	44.3	2 KH	開く, 書く	(open,write)
24.5	6.2	3 HH	する, いう	(do,say)
8.5	16.3	4 HHH	つくる, わかる	(make,understand)
7.0	14.3	5 KHH	続ける, 与える	(continue,give)
8.6	18.2	6 OTHERS	ととのえる	(prepare)

100%(=30829) 0% 100%(=1427)

(5) Adjective

		(pattern)	(example)	
47.3	32.2	1 KH	多い, 強い	(many,strong)
25.2	20.3	2 KHH	美しい, 大きい	(beautiful,big)
8.2	12.0	3 HHH	ひどい, かたい	(cruel,hard)
7.4	12.0	4 HHHH	うれしい, おいしい	(merry,tasty)
4.0	6.8	5 HHHHH	むづかしい	(difficult)
7.9	16.7	6 OTHERS	面白い	(funny)

100%(=3548) 0% 100%(=251)

(6) Adverb

		(pattern)	(example)	
31.7	23.7	1 HHH	かなり, すでに	(fairly,already)
26.3	38.3	2 HHHH	それぞれ, ほとんど	(each,almost)
19.1	6.3	3 HH	まだ, もう	(yet,now)
7.4	1.6	4 K	約	(about)
7.2	6.7	5 KH	再び, 少し	(again)
3.0	6.3	6 KHH	初めて, 直ちに	(first,immediatly)
5.3	17.1	7 OTHERS	いっせいに	(simultaneously)

100%(=5044) 0% 100%(=253)

Fig.10 Pattern of character string of word

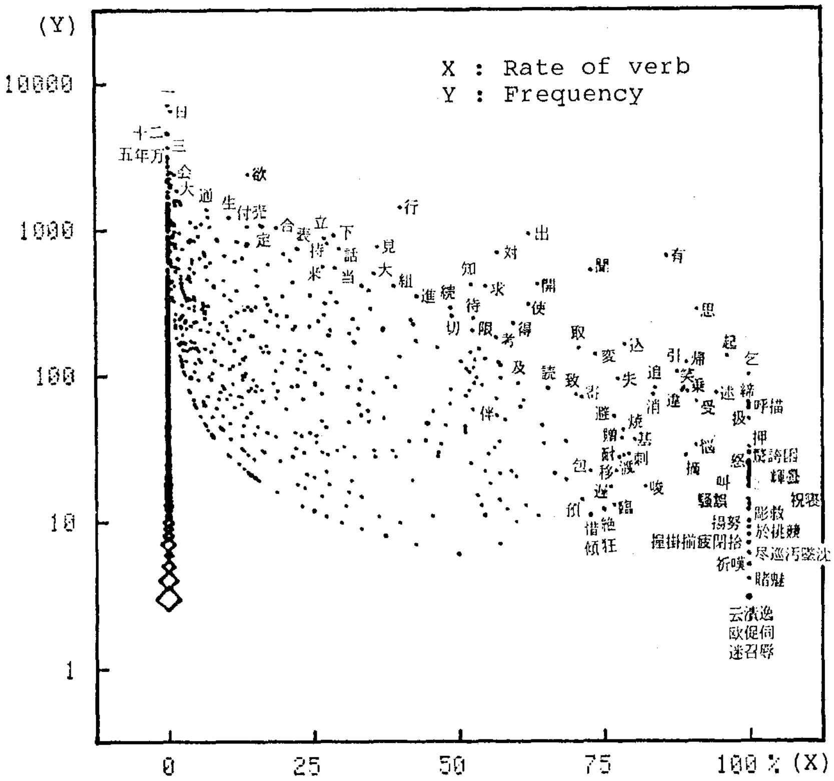


Fig.11 KANJI for verbs

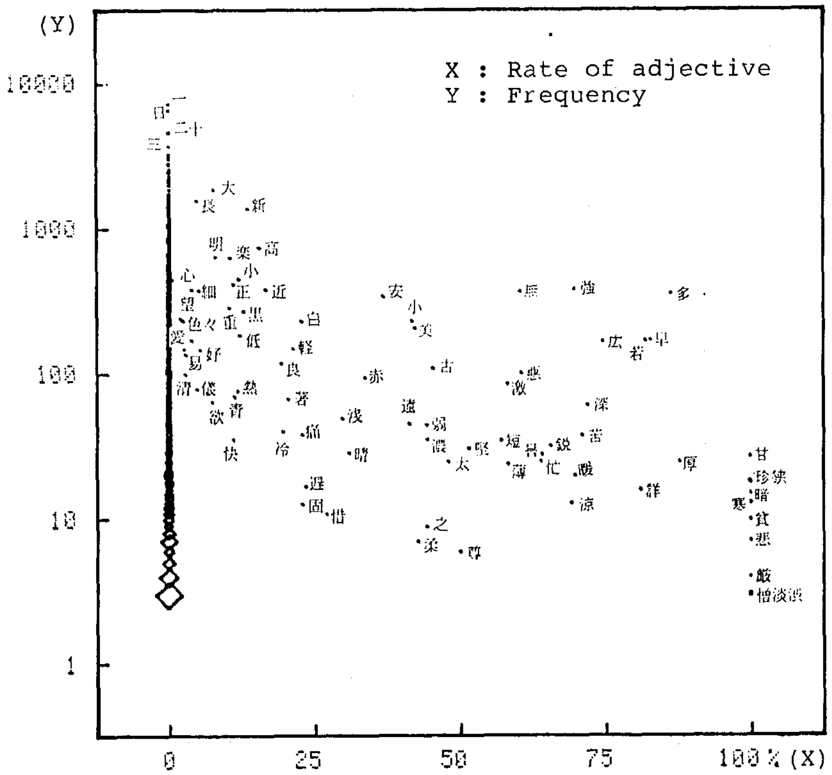


Fig.12 KANJI for adjectives

same pattern was found among different parts of speech. In order to obtain more accurate results, we analyzed relations between each character and each part of speech in the data of newspaper (restriction : word-frequency ≥ 3).

In Fig.11 the axis Y represents total number of the last KANJI in a word.

e.g. KKHH
└─last KANJI in a word

The axis X shows the rate of KANJI used for verbs. KANJI on X=100 are used for verbs in the all occurrence of the last KANJI in a word. The reliability of axis X increases according to the value of axis Y. In the lower area of the graph, the value on axis X seems to be discrete because of shortness of the data.

8. Conclusion

These analyses are preliminary works to make character dictionary having statistical data. We plan to use the dictionary for computer processing of various written Japanese.

References

[1] T.Tanaka, "A similation system for transliteration of writing form of Japanese",Mathematical Linguistics, Vol.11,No.15,1978

[2] T.Tanaka, "Transliteration of Japanese writing", bit,Vol.10, No.15, 1978

[3] T.Tanaka, "Statistics of Japanese characters", Studies in computational linguistics, Vol.X, (National Language Research Inst. Report-67), 1980

[4] H.Nakano et al., "An automatic processing of the natural language in the word count system", (in this proceeding)

[5] M.Nomura et al., "A study of Chinese characters in modern newspapers", N.L.R. Inst. Report-56, 1976

[6] T.Morohashi,"DAIKANWA dictionary", Taishu-kan Book Co. 1971

[7] A.Tanaka, "A statistical measurement on survey of KANJI", Studies in computational linguistics, Vol.Ⅷ,(N.L.R.Inst. Report-59, 1976

[8] T.Ishiwata, A.Tanaka, H.Nakano et al., "Studies on the vocabulary of modern newspapers", Vol.1, Vol.2, N.L.R. Inst. Report-37,38, 1970,1971