PSYCHO-LINGUISTIC STUDIES OF MECHANICAL TRANSLATION IN JAPAN*

Seiichiro OHNISHI

1. INTRODUCTION

Psycho-linguistic studies of mechanical translation in Japan can't be said to have been made so vigorously as the studies in the field of computer engineering. The chief reason is that any organization to study the problems of mechanical translation in connection with other fields of science and to promote the inquiry in cooperation with other scholars has not been established in the academic world in Japan.

But, under these circumstances, a research worker has long since been interested in the problems. He was Prof. T. Obonai (Tokyo Education University). He made many experiments (in 1957) with his collaborators, on the formula of the mechanical translation of language. Furthermore, the mechanical translation was already a concrete problem among the engineers at that time. In 1958 and further in 1960, his group members, Obonai being the chief of the group, receiving research-aid funds of the Ministry of Education, advanced the study lively, and promoted a comprehensive study of computer engineering, science of language, English, logic and psychology. Through their study psycho-linguistic study began to have a close connection with other fields of science, and as the Society for the Study of Mechanical Tranalation was organized, (in 1961, headed by Dr. Y. Niwa) the connection became much closer.

Here we were going to report the outline of the studies conducted by this group and others under the following two headings:

1. The studies of the Japanese language based on the information theory.
2. Psycho-linguistic studies connected with the mechanical translation.

2. THE STUDIES OF THE JAPANESE LANGUAGE BASED ON THE INFORMATION THEORY

The studies of the Japanese language based on the information theory can be divided into two parts: one is the measurement of the amount of information on the basis of letter as the element, the other the measurement of the same on the basis of word as the element.

2. 1. First, on the amount of information on the basis of letter as the element, H. Gamoo,

---

* 備　　考

1964年4月20日から24日まで5日間にわたって東京において U.S.A.-Japan Seminar on Mechanical Translationが開催された。これは，日米科学協力委員会の事業の一環として行なわれたものであって，機械飜訳に関する両国の情報交換と，今後の協力のありかたについて協議が行なわれた。

日本側からの参加者は，丹羽保次郎のほか12名，アメリカ側はF.L.Alt のほか7名であり，機械工学，数学，言語学，心理学，社会学の専門分野にわたっている。たまたま大西はその一員としてゼミナーに参加し，本問題について報告を行なったのである。

and T. Tamachi and others published their studies. Gamoo measured F0, F2 of the Kana-letters and the Roman letters. Tamachi and others measured F0, F20 and F40, F60, F80, F100 by means of the so-called Shannon's guessing method. Here, the value of F100 inferable from the actual survey is about 2 bits per letter, and the entropy of Japanese (F-H) is 1.8 bits, so that the redundancy is estimated to be about 0.74 (r=1−1.8/6.93=0.74). In a rough calculation of the actual survey that Shannon measured on the basis of 100 English sentences, entropy was 0.9 bits and redundancy was 0.75. So Japanese has twice as much entropy as English, and the redundancy is equal in the two languages. If one Kana-letter can be indicated by 2 Roman-letters (for example; か=ka, さ=sa, す=su, etc.), English and Roman letters have equal entropy and redundancy.

Furthermore, it is very interesting from the point of view of the sentence structure in Japanese to notice the fact that the amount of information at F3 and F4 is low and that the average number of letters in Japanese words is about 3. 2.

2. 2. H. Izawa slso measured the amount of information of F0, F3. He selected 2000 trigrams (diagrams as to Hiragana sentences and Kana-letter sentences) from comments, essays, and novels.

(1) As the result, it was clarified that the value of the FN showed a monotonous decrease and a sudden decrease at about the points of F2 and F3. From those points the value of FN kept on decreasing gradually, and as the value of N became larger, the value of FN showed a constant one. These results accord with those of Tamachi, and show the characteristics of the Japanese sentence structure.

(2) Then, Izawa compared the amount of information of various ways of expression. The findings are as follows: (Table 1)

(Table 1) Kind of Sentences	Number of elements	F0	F1	F2	F3 ····· F10		F20
Writings in Hiragana	72	6.17	5.46	3.91		3.09	2.41
〃 in Kana Letters	77	6.27	5.11	3.51		2.71	2.39
〃 in Roman Letters	25	4.64	3.85	2.89	2.05	2.15	1.58
Shannon	27	4.76	4.03	3.32	3.10	2.10	2.10 (F15)

Accordingly the redundancies R1, R2, R10, R20 are respectively as follows: (Table 2)

(Table 2)	R1	R2	R10	R20
Writings in Hiragana Letters	0.11	0.225	0.499	0.609
〃 in Kana Letters	0.182	0.285	0.548	0.627
〃 in Roman Letters	0.297	0.358	0.530	0.659

In Izawa's study the sentence is thought to be a probability process of the letters and the spaces, but in the following Ohnishi's study the amount of information is measured on the basis of the letter in the word as the element.

1. 3. S. Ohnishi measured the amount of information in Japanese words selected from the book "Vocabulary in the all-round magazine". (Report No. 12 of the National Language Research Institute.) The total number of the vocabulary was 195,000 consisting of 4,200 different words. From this number Ohnishi selected 3,000 words of 3 letters and more and another 3,000 words of 5 letters and more, and measured the amount of information F1, F2, of the former and F1—F4 of the latter.

The findings are as follows: (Table 3)

(Table 3)	F0	F1	F2	F3	F4
Word of 3 letters and more	6.15	5.26	2.83		
Word of 5 letters and more	6.15	4.63	1.15	0.92	0.48

These results show that the values of F2 and less are pretty low, especially the values of the words of 5 letters and more are lower than those of the words of 3 letters and more. It is also made clear that in Japanese words a letter is strongly affected by the preceding letters.

3. PSYCHO-LINGUISTIC STUDIES CONNECTED WITH THE MECHANICAL TRANSLATION

Psycho-linguistic studies connected with the mechanical translation include the following studies: (1) statistical studies of the language, (2) studies of sentence patterns, and (3) studies of the verbs and the prepositions connected with mechanical translation.

3. 1. First, T. Obonai and T. Kaneko attempted to make a framework necessary for a mechanical word dictionary from the statistical study of the language. As the sample source, they used two kinds of texts, (1) Oswald Vehlen and others: The Foundation of Differential Geometry and its Japanese translation, and (2) some essays in the Scientific American and their Japanese translations. The analysis on the Japanese translations is not proceeding, the samples were selected from the original works. The samples selected from Geometry were 740 sentences with the total number of 15,000 words, and from the Scientific American 520 sentences with the total number of 13,000 words.

(1) According to the investigation of the words, there were about 1,170 different words in Geometry, and about 2,270 different words in the Scientific American.

Between these two groups of samples, the proportion of the total number of words was 100 to 87, but that of different words were 100 to 228. In Geometry there was a very small number of different words. Though the number of different words will become larger as the total number of words increase, the rate of increase in different words will become acceleratedly smaller. From these results we can guess that the different words necessary will be 1,300 words at the maximum.

(2) Secondly, if we see the relation between the order of frequency in the use of words and the coverage by those words, 150 words covered 80% of the whole Geometry, while the same number

of words could cover only 60% of the Scientific American. In other words, in Geometry relatively few words can cover the whole contents but in Scientific American more different words are required.

(3) Thirdly, the law of Zipf that the order and the frequency of words have a straight relationship on the logarithmic scale, has been shown almost valid by these data, too.

In Geometry, there were tendencies showing that the gradient of straight line was large and the frequency of the use of a few special words was large, but there were not any significant differences between them. Accordingly it was shown that the coefficient of Zipf was constant to some extent.

3. 2. T. Kaneko made a study on the multiple meaning of the words and on the verb patterns using the materials selected from the above-mentioned Geometry and the Scientific American.

(1) First, on the multiple meaning he investigatad the two words, "linear" and "set". He investigated in what manner these words were translated into Japanese equivalents. According to this study it was shown that "linear" was translated in the sense of "senjo" and "ichi-ji".

Of thirty phrases, "linear" was translated into "senjo" and in those cases "linear" followed the word "space". On the other hand, of thirty-four phrases, the same word was translated into "ichi-ji", and in those cases the words following it were the next 5 words: transformation, equation, translation, dependence and combination.

(Table 4)	Japanese equivalent(f)		Words followed it or words combined with it (f)			
linear	senjo	(33)	* space	(33)		
	ichi-ji	(34)	* transformation	(19)	equation	(9)
			translation	(1)	dependence	(4)
			combination	(2)		
set	kumi	(15)	numbers	(2)	points	(11)
			equations	(1)	values	(1)
	ren ritsu	(11)	equation	(11)		
	gun	(10)	axioms	(7)	objects	(1)
			vilations	(1)	statements	(1)
	kei	(1)	points	(1)		
	shūgo	(36)	* points	(21)	transformations	(3)
			permulations	(3)	objects	(3)
			dependent	(1)	elements	(1)
			automorphism	(1)	translation	(1)
			displacements	(1)	symmetries	(1)
			parametrization	(1)		

* For example: a linear k-space
one linear transformations
a set of points

The Table 4 shows how these words "linear" and "set" were translated into Japanese equivalents, and what words followed them and what words were combined with it.

(2) Next, Kaneko tried to apply Hornby's verb patterns to these materials. 1) How much can Hornby's verb patterns cover these sentences? 2) Which pattern is most frequently used? 3) What differences are there between these two materials in terms of the frequency of the verb patterns?

1) According to the results of those researhes, it was shown that, almost every sentence conld be coped with by these patterns if some routine ways are found to change passive, negative, interrogative, and inversion sentences automatically into affirmative descriptive sentences. It was also shown that dividing verbs into 25 subgroups would be generally appropriate.

2) The verb pattern 1 (VP1) (for example, He cut his fingers.) and the verb pattern 22 (VP22) (for example, The weather has become warmer.) were used frequently. In Geometry, the frequency of the VP1 amounts to 29%, and the VP22 amounts to 38%, and in the Scientific American, the VP1 and VP22 amount to 32%, 24% respectively.

There were few differences between these two materials in terms of the frequency of each sentence pattern.

The assumption that certain verbs will take certain nouns as their objects more often than other nouns, namely nouns will be grouped under certain verbs did not prove always valid.

3) Further, in this study the passives are all changed into the actives; if we presume that the passive sentence forms another patterns, it will be necessary to increase the number of the sentence patterns.

Since the correspondence of these patterns with Japanese is not examined in this study, some more considerations will be necessary in order to apply these findings to translation.

3. 3. S. Ohnishi selected a thousand 5-word-series from twenty primary school readers approved by the Ministry of Education. Each 5-word-series was classified into ten parts of speech and was analyzed in terms of (1) frequency of occurrence of these parts of speech, (2) the way of combination of words. and (3) the frequency of occurrence of nouns and particles.

(1) The amounts of information of 5-word-series, measured on the basis of the parts of speech as the element, were as follows.

Fo=3.32 bits	F1=1.81 bits	F2=0.98 bits
F3=1.47 〃	F4=1.30 〃	F5=0.58 〃

These results show that the value of F2 is pretty low, while the value of F3 and F4 are high. From these points we can say that in Japanese sentences the second word can be more easily guessed than the third one. This is a characteristic trait in Japanese senteces.

(2) From the point of frequency of the use, it is evident that the noun is used most frequently, and the particle is the second; the former amounts to 36.7% and the latter 34.3%, and so both amount to about 70%.

In contrast with this, the least used are interjections (0.8%) and verb-modifiers (1.1%). The rate of the use of these parts of speech varies according to their positions in the sentence. In the first position, noun is used most frequently (63.5%). On the contrary, the particles, though

used most frequently in a whole sentence are never used in the first position. It is reasonable as we consider that a particle as an attached word has no meaning of its own. Sixty percent of the words in the second position are particles.

Ninety two percent of all the particles follow nouns, and few of them follow verbs, and auxiliary verbs. These are the reasons why the value of F2 becomes suddenly low while the value of F3 remains rather high.

(3) Next, we will explain the way of combination of the 5-word-series.

As the first word, any of the eight parts of speech except particles and auxiliary verbs is used. What kind of combinations do these 5-word-series have as the second and the following words? Among the results let's pick up those cases where the first word is adjective.

There are 26 cases in which adjectives are used as the first word, but the second words which follow them are limited to the five kinds of words, i.e., noun, "rentai-shi", verb, adjective, and particle.

The 26 5-word-series can be divided into 15 kinds of combinations at the fifth word.

(4) In Table 5 are shown the combinations of the words around the first word.

(Table 5) Parts of speech of the 1st word	1st word	2nd 〃	3rd 〃	4th 〃	5th 〃
noun	1	6	22	61	113
adverb	1	7	21	41	62
"rentai-shi"	1	2	6	11	20
conjunction	1	6	11	19	34
interjection	1	6	12	17	21
verb	1	4	10	13	17
adjective	1	6	8	14	15
adjective verb	1	4	7	10	12
total	8	41	97	186	294
Nos. of possible kinds of combination	10	100	1,000	10,000	100,000

As is shown in this Table, combinations of the words in the sentence ars limited to a narrow range. A sentence beginning with a noun will have 6 parts of speech as the second word, and will be divided into 22 kinds of patterns at the 3rd word, and into 113 kinds of patterns at the fifth word.

Similarly a sentence beginning with an adverb will take one of 7 parts of speech as the second word, and will be divided into 61 kinds of pattern at the fifth word.

As mentioned above, 8 out of 10 parts of speech are used as the first word, and as the second word 41 kinds of combinations actually appear out of possible 100 kinds. And after that there appear 97, 186, 294 kinds of combinations out of possible 1,000, 10,000, 100,000 kinds.

(5) As to the particles, "no, ga, wa, o, ni" are most frequently used. They amount to 77 percent of all the particles. Furthermore, if the other five particles, "to, de, mo, kara, e" are added, they amount to 90 percent of all the particles. As compared with the Fries's results, which reported that the prepositions, "at, by, for, from, in, of, on, to," amount to 62.6 percent of all

the prepositions, our results are very interesting.

3. 4. T. Taguchi made a study of Japanese sentence structure. He selected 1,328 words from a Japanese language dictionary, and classified them according to parts of speech.

(1) As the results it was shown that about 86 percent of all were nouns and 5.3 percent verbs. Though these percentages were high as compared with the results of other studies, it was perhaps caused by his using a dictionary as the source.

(2) Then, Taguchi investigated in what sentence pattern the verbs were used. He used novels and essays in the newspaper Asahi as the materials. The number of sentences were 883, and there were 1,630 verbs in these materials. He analyzed these verbs from various view points.

First, in those 1,630 verbs, 1,030(63.2%)were used in the form of modifier to verbs("renyo-kei"). adjectives or advrerbs. The number of the verbs accompanied by "formal" verbs was 315(30.6%) (for exampe; kaite+iru=writing on, kaite+aru=it is said (in the letter) that, etc.)

(3) Taguchi also investigated from what words the S series irregular conjugation verbs had developed. In Japanese there are many words which function as verbs by adding the verb "suru" after them. For instance, noun+suru (denwa+suru=to telephone, hayagawari + suru=to make a quick change, deiri+suru=to go in and out, annai+suru=to guide, kinshi+suru=to prohibit, kyoryoku+suru=to cooperate with), and adjective+suru (utsukushiku+suru=to make beautiful, kuraku+suru=to make dim), adverb+suru, (nikoniko+suru=to smile, sappari+suru=to feel refreshed), and the like. There were 170 verbs belonging to these kinds.

(4) Further, there were some words which function as verbs by adding a term (a particle) before the word "suru". For instance, (bengoshi+o+suru=to set up as a lawyer, gochiso+o+suru =to entertain with a feast, hai+ni+suru=to burn to ashes, kanji+ga+suru=to feel, toku+to+suru =to begrateful to a person for something, ki+mo+suru=to feel also, denwa+nado+suru=to make a telephone call and the like), and the like, and there were 92 words belonging to these kinds.

(5) Since the classification of verbs is made according to grammar, Taguchi's study may be said to be based on the traditional method, and how to deal with the "Renyo-kei" is certainly one of the most difficult problems in mechanical trsnslation.

Seventy verbs (5.3%) were used in these materials, and when we consider the translation of other languages into Japanese, it would be most efficient if we could limit the number of words within a narrow range.

3. 5. Next, H. Hatori investigated the prepositions and Japanese equivalents. He investigated the Japanese which can be thought to be the Japanese equivalents, and what kinds of words and to what degree are used in the translation. The prepotions selected as the objects of investigation were the next 10 words, "as, at, by, for, from, in, of, on, to and with".

As the sources of the materials were used 10 kinds of sentences, books of mathematics, chemistry, biology, psychology, history, commercial correspondences, and editorial articles of the newspaper, ordinary wrtings, novels, and cenarios.

The prepositions selected were shown in Table 7.

From the results of these investigations, it was shown that the prepositions have considerable

(Table7)	as	at	by	for	from	in	of	on	to	with
f	575	363	466	544	402	1607	2960	401	1940	456
Most used way	to-shite	ni	ni-yotte	no-ta-meni	kara	no naka-ni	no	no-ue-ni	suru-tame-ni	to-tomo-ni
f	253	254	433	298	246	971	1183	127	1070	165
Nus. of kinds most used	7	4	7	8	3	7	9	11	8	6

differences between the frequencies of the use. The findings are as follows:

1) Through the kinds of Japanese equivalents were divided into some groups, they may probably be rearranged into more reasonable categories and it will be a problem to be solved in future.

2) As there are many compound-words which include prepositions, it was suggested to be important to study the compound-words more inclusively.

3) In many kinds of sentences, there were not remarkable differences between the frequencies of the use of the prepositions. That will be reasonable when we think the characteristics of prepositions as the colourless words.

3. 6. And lastly T. Kaneko and K. Ohtsubo investigated the prepositions selected from the secondary school readers. When the prepositions were used to modify the preceeeding nouns, the prepositions could be translated as the particle "no" irrespective of the kinds of these prepositions.

(1) According to Fries, 9 kinds of prepositions counted for 92.6% of all prepositions (3,448) found in his data. These kinds of 9 prepositions in these secondary school readers used in the form of NI PN2* were 488. (Table 8)

* NI PN2 NI is a modified noun P is a noun

(Table 8)	at	by	for	from	in	of	on	to	with	others	Total	%
Adjective phrase	1	0	12	6	37	145	14	7	14	32	268	55
Adverbial phrase	13	6	23	13	65	1	16	23	16	44	220	45

The prepositions picked np in this Table are only those belonging to the special word group in the "NPN". Among them, the numbers of the prepositions used in adjective phrases and in adverbial phrases were 268 (55%) and 220 (45%) respectively. Among them "of" was used most frequently, and it was used most often in adjective phrases. It could be translated in the sense of "no" as in the pattern of "N no N" in Japanese.

As compared with Fries's results, there are a few differences between the numbers of prepositions used in adjective phrases and in adverb phrases. It is not evident whether this is caused by different standards of classification between Kaneko and Fries, or else by an inherent bias

of the materials used.

(2) The Table 9 shows the degree in which prepositional phrases can be translated into "no", when they are used as adjectives.

(Table 9)	at	by	for	from	in	of	on	to	with	total
Can be translated into "no"	0	0	3	3	37	145	13	0	1	202
Need some Japanese equivalents other than "no"	1	0	9	9	0	0	1	7	13	34

The materials are too few for us to draw a conclusion, but we can say tentatively as follows: 1) When prepositions are found in adjective phrases, the prepositions which can be translated into "no" will be five times as many as those which need other Japanese equivalents. 2) When phrases containing "of" "in" "on" are found to be adjective phrases, we can translate them into "no" in almost cases.

(3) Furthermore, Kaneko and others investigated the roles of the prepositional phrase. In English there are cases in which it is ambiguous which words are modified by the prepositional phrases. For example, "in London" can be regarded either as an adjective phrase or as an adverb phrase according to the context.

"He is in London". (Adverb phrase)

"A college in London is good." (Adjective phrase)

It is necessary to distinguish adjective phrase from abverb phrase in order to translate English into normal Japanese. "In London" can be translated into Japanese, only when we understand its role in the sentence.

(4) Difficulties arise when the words appear in the order of noun, preposition and noun (or NPN).

"I saw a college in London."

"Watashi wa London de daigaku o mita."

or "Watashi wa London no daigaku o mita."

In these cases, we are completely unable to distinguish the correct translation from the wrong one. More infomation may be needed than this one sentence.

In order to clarify these problems Kaneko and others classified the word-groups appearing in the order of NPN into three categories.

1) NP) N　　2) N(PN,　　3) V(or A)NPN

1) The followings are the examples of the NP).N type: "Two fnndamental operations in vector algebra are multiplication by a number and addition." "a single valued transformation of x into y."

In the case of the NP)N type, P-gronp is almost always uniquely adjective phrase and correctly translated as such.

2) The followings are the samples of the N)PN: "Similarly, if x and y are any points in the flat K-space……" "The set of all arithmetic points, for a given value n,……"

In the case of N)PN type, some knowledge about the subject is necessary for the human translator to do correctly, and for the machine translator it is impossible to distinguish whether the P-group is an adjective phrase or adverb phrase, on the present stage.

3) The followings are the examples of V(orA)NPN type: "a different object from the same set of points". "We shall sometimes distinguish this space from the arithmetic space……".

In the case of the V(or A) NPN type, P-group is almost always uniquely the adverb phrase, and translated correctly as such.

(5) It will be desirable that mechanical translation will make normal Japanese but that is not sufficient. Though it will be desirable that the contents written in foreign languages take a form easy to read, they must not lack clarity. For example, English prepositions may have less functions than Japanese prepositions from the viewpoint of syntax, but they may be better functioners in that they can show clearer relationship between words. We must endeavor not only to make a translation natural but also to make it clear.

4. CONCLUSION

We have reported the outline of the psycho-linguistic studies of the mechanical translation in Japanese. One was the studies of the Japanese language based on the information theory, and the other was the psycho-linguistic studies connected with the mechanical translation.

The studies of the Japanese language based on the information theory was divided into two parts: one is the measurement of the amount of information based on letter as the element, the other is the measurement of the same on the basis of word as the element.

From these studies, some characteristics have been seen in the Japanese sentence structure.

Next, on the psycho-linguistic studies connected with the mechanical translation, we reported (1) the studies of the statistical studies of the language, (2) sentence patterns, (3) and the studies of the verbs and the prepositions connected with mechanical translation. In these studies, not only Japanese but also English books were used as the material sources.

From these studies have been clarified some of the characteristics of the usage of words, and the characteristics of the verbs and the prepositions in Japanese.

Though we can't say that all these studies have been closely connected with the mechanical translation, we wish to contribute these works to the promotion of the studies on the mechanical translation.