Idiosyncratic Gap: A Tough Prolem to Structure-bound Machine Translation

Yoshihiko Nitta

Advanced Research Laboratory
Hitachi Ltd.
Kokubunji, Tokyo 185 Japan

ABSTRACT

Current practical machine translation systems (MT, in short), which are designed to deal with a huge amount of document, are generally structure-bound. That is, the translation process is done based on the analysis and transformation of the structure of source sentence, not on the understanding and paraphrasing of the meaning of that. But each language has its own syntactic and semantic idiosyncrasy, and on this account, without understanding the total meaning of source sentences it is often difficult for MT to bridge properly the idiosyncratic gap between source- and target- language. A somewhat new method called "Cross Translation Test (CTT, in short)" is presented that reveals the detail of idiosyncratic gap (IG, in short) together with the so-so satisfiable possibility of MT. It is also mentioned the usefulness of sublanguage approach to reducing the IG between source- and target- language.

1. Introduction

The majority of the current practical machine translation system (MT, in short) (See [Nagao 1985] and [Slocum 1985] for a good survey.) are structure-bound in the sense that all the target sentences (i.e. translated sentences) are composed only from the syntactic structure of the source sentences, not from the meaning understanding of those. Though almost all the MT are utilizing some semantic devices such as semantic feature agreement checkers, semantic filters and preference semantics (See [Wilks 1975] for example.) which are serving as syntactic structural disambiguation, they still remain in structure-bound approaches far from the total meaning understanding approaches.

The structure-bound MT has a lot of advantageous features among which the easiness of formalizing translation process, that is, writing translation rules and the uniformity of lexicon description are vital from the practical standpoint that it must transact a huge vocabulary and inumerable kinds of sentence patterns.

On the other hand, the structure-bound MT has the inevitable limitation on the treatment of linguistic idiosyncrasy originated from the different way of thinking.

In this paper, first of all, we will sketch out the typical language modeling techniques on which the structure-bound MT(= current practical machine translation systems) are constructed. Secondly, we will examine the difference between the principal mechanism of machine translation and that of human translation from the viewpoint of the language understanding ability. Thirdly, we will illustrate the structural idiosyncratic gap (IG, in short) by comparing the sample sentences in English and that in Japanese. These sentences are sharing the same meaning. This comparison will be made by a somewhat new method which we call "Cross Translation Test (CTT, in short)", which will eventually reveal the various IGs that have origins in the differences of culture, i.e., the way of thinking or the way of representing concepts. But at the same time, CTT will give some encouraging evidence that the principal technologies of today's not-yet-completed structure-bound MTs have the potential for producing barely acceptable translation, if the source language sentences are taken from the documents of less equivocations or are appropriately rewritten. Finally, we will briefly comment on the sublanguage to control or normalize source sentences as the promising and practical approaches to overcoming the IGs.

2. Modeling of Natural Language

Modeling natural language sentences is, needless to say, very essential to all kinds of natural language processing systems inclusive of machine translation systems. The aim of modeling is to reduce the superficial complexity and variety of the sentence form, so as to reveal the indwelling structure which is indispensable for computer systems to analyze, to transform or to generate sentential representations.

So far various modeling techniques are proposed (See for example [Winograd 1983].) among which the two, the dependency structure modeling (Figure 1) and the phrase structure modeling (Figure 2) are important. The former associated with semantic role labeling such as case marker assignment is indispensable to analyze and generate Japanese sentence structure (See for example [Nitta, et al. 1984].), and the latter associated with syntactic role labeling such as governor-dependent assignment, head-complement assignment, or mother-daughter assignment (See for example [Nitta, et al. 1982].) is essential to analyze and generate English sentences.

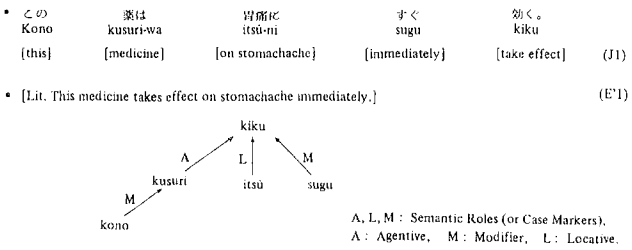


Figure 1. Example for Dependency Structure Modeling

"To what extent should (or can) we treat semantics of sentences?" is also very crucial to the decision for selecting or designing the linguistic model for

machine translation. But it might be fairly asserted that the majority of the current "practical" machine translation systems (MT, in short) are structure-bound or syntax-oriented, though almost all of them claim that they are semantics-directed. Semantics are used only for disambiguation and booster in various syntactic processes, but not used for the central engine for transformation, generation and of course not for paragraph understanding (See [Slocum 1985, pp. 14 ∿16] for a good survey and discussion on this problem; and see also [Nitta, et al. 1982] for the discussion on a typical (classical) structure-bound translation mechanism,i.e. local rearrangement method). Here "practical" means "of very large scale commercial systems" or "of the daily usage by open users", but neither "of small scale laboratory systems" nor "of the theory-oriented experimental systems". For structure-bound machine translation systems, both the dependency structure modeling and the phrase structure modeling are very fundamental technical tools.

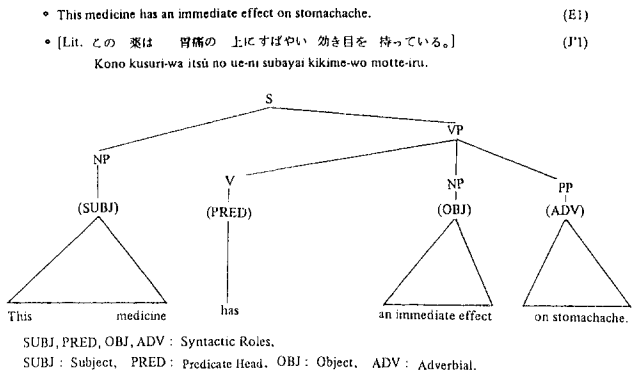


Figure 2. Example of Phrase Structure Modeling

The semantic network medeling, which is recently regarded as an essential tool for semantic processing for natural languages (See for examples [Simmons 1984].), might also be viewed as a variation of dependency modeling. However modeling problems are not discussed further here. Comparing Figure 1 and Figure 2, note that the dependency structure modeling is more semantics-oriented, logical and abstract, in the sense of having some distance from surface word sequences.

3. Machine Translation vs. Human Translation

Today's practical machine translation systems (MT, in short) (See for example [Nagao 1985] and [Slocum 1985].) are essentially structure-bound literal type. The reasons for this somewhat extreme judgement are as follows:

(1) The process of MT is always under the strong control of the structural information extracted from source sentences;
(2) In all the target sentences produced by MT, we can easily detect the traces of wording and phrasing of the source sentences;
(3) MT is quite indifferent to whether or not the output translation is preserving the proper meaning of the original sentence, and what is worse, MT is incapable of judging whether or not;
(4) MT is quite poor at the extra-sentential information such as situational information, world knowledge and common sense which give a very powerful command of language comprehension.

Now let us see Figure 3. This rather over-simplified figure illustrates the typical process of Japanese-English structure-bound machine translation. Here the analysis and transformation phase are based on the dependency structure modeling (cf. Figure 1) and the generation phase is based on the phrase structure modeling (cf. Figure 2) (For further details, see for example [Nitta, et al. 1984].). This figure reveals that all the process is bound by the grammatical structure of the source sentence, but not by the meaning of that.

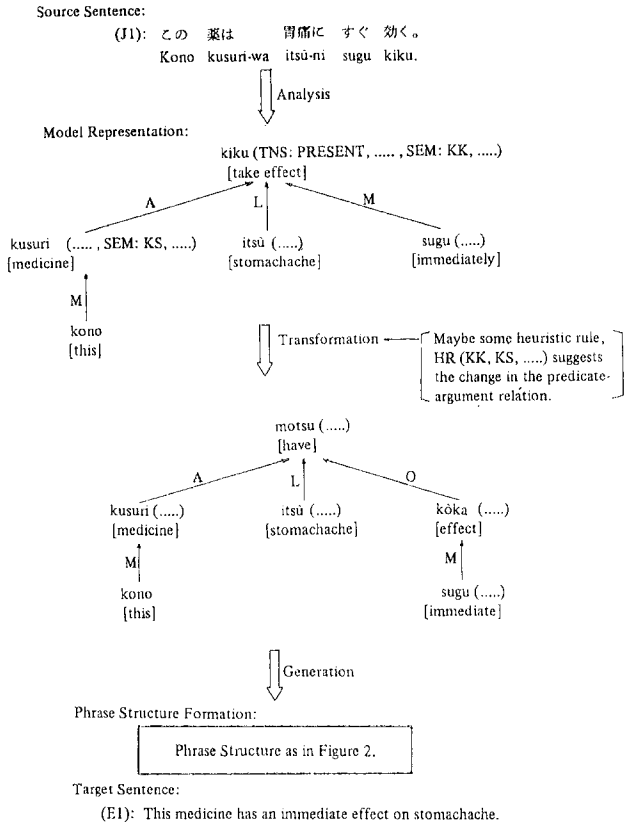


Figure 3. Simplified Sketch of Machine Translation Process

Thus, the MT can easily perform the literal syntax-directed translation such as 'from (J1) into (E'1)' (cf. Figure 1). But it is very very difficult for MT to produce natural translation which reflects the idiosyncrasy of target language, preserving the original meaning. (E1) is an example of a natural translation of (J1). In order for MT to produce this (E1) from (J1), it may have to invoke a somewhat sophisticated heuristic rule. In Figure 3, the heuristic rule, HR (KK, KS, ...), can sucessfully indicate the change of predicate which may improve the treatment for the idiosyncrasy of target sentence.

But generally, the treatment of idiosyncratic gap (IG, in short) such as 'that between (J1) and (E1)'

is very difficult for MT. It might be almost impossible to find universal grammatical rules to manipulate this kind of gaps, and what is worse, the appropriate heuristic rules are not always found successfully.

On the other hand, the human translation (HT, in short) is essentially semantics-oriented type or meaning understanding type. The reasons for this judgement are as follows:

(1) HT is free from the structure, wording and phrasing of a source sentence;
(2) HT can "create" (rather than "translate") freely a target sentence from something like an image diagram obtained from a source sentence (Figure 4); (Of course the exact structure of this image diagram is not yet known);
(3) HT often refers the extra-linguistic knowledge such as the common sense and the culture;
(4) Thus, HT can overcome the idiosyncratic gaps (IG) freely and unconsciously.

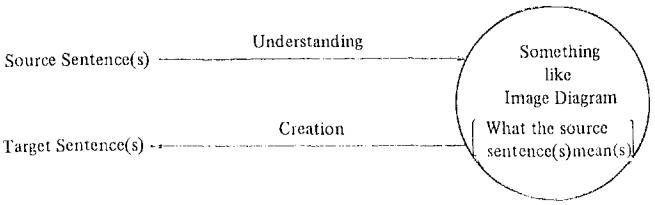


Figure 4. Human Translation Process

In order to simplify the arguments, let us assume that some kind of diagram is to be invoked from the understanding of the original sentence. This diagram may (or should) be completely free from the superficial structure such as wording, phrasing, subject-object relation and so on, and may be strengthened and modified by various extra-linguistic knowledge. It may be easy for human to compose the sentences such as (J2) and (E2) from this kind of image diagram invoked from (J1). But the sentences such as (J'1), (J'2), (E'1) and (E'2) will never be composed by human under the normal conditions.

・この	薬を	飲むと	胃の痛みが	すぐ	とれる。		(J2)
Kono	kusuri-wo	nomu-to	i-no-itami-ga	sugu	tore-ru.		
[this]	[medicine]	[if (you) take]	[stomachache]	[soon]	[deprived]		
・[Lit. If you take this medicine you will soon be deprived of a stomachache.]							#(E'2)

・This medicine will soon cure you of the stomachache. (E2)

・[Lit. この	薬は	あなたを	すぐに	胃痛から	救うだろう。]		#(J'2)
Kono	kusuri-wa	anata-wo	sugu-ni	itsū-kara	sukuu-daro.		
[this]	[medicine]	[you]	[soon]	[of the stomache]	[will cure]		

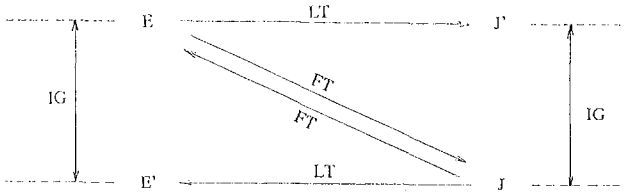
Now, note that there are big structural gaps between (J1) and (E1), and between (J2) and (E2), which are the natural reflections of linguistic idiosyncrasy orginated in the culture, i.e. the difference of the way of thinking. So far we have seen that MT is poor at the idiosyncrasy treatment and conversely HT is good at that. This difference between MT and HT depends on whether or not it has the ability of meaning understanding.

4. Idiosyncratic Gaps

In this section, let us examine the idiosyncratic gaps between the two sentences which share the same meaning but each of which belongs to different language. The reason for comparing the two sentences is that we cannot examine the linguistic idiosyncrasy itself. Because, currently, we cannot fix the one abstract neutral meaning without using something like the image diagram (cf. Figure 4) which is not yet elucidated.

In order to examine the idiosyncratic gap, we have devised the practical method named "Cross Translation Test (CTT, in short)." The outline of CTT is as follows:
First, take an appropriate well-written sample sentence written in one language, say English; Let E denote this sample sentence; Secondly, select or make the proper free translation of E in the other language, say Japanese; Let J denote this proper free translation; J must preserve the original meaning of E properly; At the same time, make a literal translation of E in the same language that J is written in; Let J' denote this literal translation; Lastly, make a literal translation of J in the same language that E is written in; Let E' denote this literal translation.

Here, the "literal" translation means the translation that is preserving the wording, phrasing and various sentential structure of the original (source) sentence as much as possible. Then, eventually we may be able to define (and examine) the idiosyncratic gap, IG, by Figure 5. In other words, we may be able to examine and grasp the idiosyncratic gap by comparing the structure of E and that of E', or by comparing that of J' and that of J.



IG: Idiosyncratic Gap
LT: Literal Translation
FT: Free Translation
E, E': Sentences Written in English
J, J': Sentences Written in Japanese
In this paper, we have assumed that:
LT ≒ MT and FT ≒ HT,
where, MT: Machine Translation, and HT: Human Translation.

Figure 5. Illustrative Definition of Idiosyncratic Gap

Now, note that we can assume the relationship,

LT ≒ MT,

and

FT ≒ HT,

where "≒" denotes "nearly equal" or "be almost equivalent to". Namely, we can assume that the literal translation, LT, which is preserving the wording, phrasing and structure of the source sentence, is almost equivalent to the idealized competence of today's practical structure-bound machine translation, MT. The rationale of this assumption has already been discussed in Section 3.

In this paper, the literal translation, LT (≒MT), is performed by tracing the procedural steps of a virtual machine translation system (VMTS) theoretically. Here, the VMTS is a certain hypothetical system

which never models itself upon any actually existing machine translation systems, but which models the general properties of today's practical structure-bound machine translation systems.

Now let us observe the gap, IG, by applying CTT to various sample sentences. First, let us take an example with large gaps.

• 国境の	長い	トンネルを	抜けると	雪国	であった。	(J3)
Kokkyō-no	nagai	tonneru-wo	nukeru-to	yuki-guni	de-atta.	
[of border]	[long]	[tunnel]	[after passing through]	[snow country]	[was]	

• [Lit. After passing through the long border tunnel, it was the snow country.] +(E'3)

• The train came out of the long tunnel into the snow country. (E3)

• [Lit. 列車は	長い	トンネルを	ぬけて	雪国に	出た。]	(J'3)
Ressha-wa	nagai	tonneru-wo	nuke-te	yuki-guni-ni	de-ta.	

(J3) is taken from the very famous novel "Yuki-guni" written by Yasunari Kawabata, and (E3) is taken from the also famous translation by Seidensticker. (E'3) is the slight modification of [Nakamura 1973, p.27] and (J'3) is taken from the same book. In (E3) the new word "the train [ressha]" is supplemented according to the situational understanding of the paragraph including (J3) which may, currently, be possible only for HT.

(J3) is a very typical Japanese sentence possessing the interesting idiosyncrasy, i.e., (J3) has no superficial subject. But in (J3) some definite subject is surely recognized, though unwritten. That is "the eyes of the storyteller", or rather "the eyes of the reader who has already joined the travel to the snow country by the train". So the actual meaning of (J3) can be explained as follows:

> After I (= the reader who is now experiencing the imaginary travel) passed through the long border tunnel by the train, it was the snow country that I encountered.

Thus (J3) is very successful in recalling the fresh and vivid impression of seeing (also feeling and smelling) suddenly the snow country to the readers. (J3) has a poetic feeling and a lyric appeal in its neat and concise style.

But the English sentence such as (E3) requires the concrete, clearly written subject, "the train [= ressha]" in this case, and this concrete subject requires the verb, "came", and again this verb requires the two locative adverbial phrases, "out of the long tunnel" and "into the snow country". Thus, the original phrase "yuki-guni de-atta. [= it was the snow country.]" in (J3) has completely disappeared in (E3), but the new adverbial phrase "into the snow country [= yuki-guni-ni]" appears instead. These drastic changes are made under the strong influence of linguistic idiosyncrasy, and, at the same time, with the effort to preserve the original poetic meaning as much as possible.

Consequently, these changes have invoked a large distant gap, IG between (J3) and (E3). But this gap is indispensable for this translation from (J3) into (E3),

$$\text{HT:}\quad (J3) \rightarrow (E3)$$
$$\text{where, } |(J3) - (E3)| \doteq |(E'3) - (E3)| \doteq \text{IG} = \text{large}.$$

One more comment. Note that as a result of this large gap, the literal translation from (J3) into (E'3),

$$\text{LT:}\quad (J3) \rightarrow (E'3)$$
$$\text{where, } |(J3) - (E'3)| \doteq |(E'3) - (E'3)| = 0$$

has failed to preserve the original meaning, i.e., (E'3) is an unacceptable translation which is misleading. Because (E'3) can be interpreted as:

> After something (=it) finished passing through the long border tunnel, something became (= changed into) the snow country.

However, it is not always the case with idiosyncratic gaps. Lastly, let us now observe the somewhat encouraging example favorable for structure-bound machine translation, MT (≐ LT). In the following quadruplet, the gap is not so small but the gapless translation, i.e., LT (≐ MT) is acceptable. The following sample sentence (E4), is the news line taken from [Newsweek, January 18, 1982, p.45].

• He may have saved the flight from a tragic
[kare] [kamo-shire-nai] [kyûjo-shi-ta] [sono] [teiki-bin] [kara] [higeki-teki]
repeat performance of the American Airlines DC-10 crash that killed 275
[hanpuku] [jikkô] [no] [tsuiraku] [koroshi-ta] [275 nin-no]
people in Chicago in 1979.
[hito-bito] [Chicago-de] [1979 nen-ni] (E4)

• Lit 彼は その 定期便を, 1979年に シカゴで 275人の 人々を 殺した
Kare-wa sono teiki-bin-wo, 1979 nen-ni Chicago-de 275 nin-no hito-bito-wo koroshi-ta
アメリカン 航空の DC−10の 墜落の 悲劇的 反復の 実行から
American-Kôkû-no DC-10-no tsuiraku-no higeki-teki hanpuku-no jikkô-kara
救助した かもしれない。
kyûjo-shi-ta kamo-shirenai. #(J'4)

• これによって この 機は, 死者 275名を 出した 1979年の シカゴ空港での
kore-ni-yotte kono ki-wa, shisha 275 mei-wo dashi-ta 1979 nen-no Chicago-kûkô-de-no
墜落事故の 悲劇の 二の舞を 避け得たと いえよう。
tsuiraku-jiko-no higeki-no ni-no-mai-wo sake-eta-to ie-yô. (J4)

• Lit. It may safely be said that, by this, this airplane could escape from
[to-ie-you] [kore-ni-yotte] [kono hikouki] [sake-eta] [kara]
tragic repetition of crash accident of American Airlines
[higeki-teki] [hanpuku, ni-no-mai] [no] [tsuiraku] [jiko] [no]
DC-10 in Chicago Airport in 1979 that produced 275 dead persons.
[Chicago-Kûkô-de-no] [1979 nen-ni] [dashi-ta] [shisha] #(E'4)

The free translation , (J4) is taken from [Eikyo 1982, p.203] with slight modifications. For the reason of space limitation we have omitted the comments to this example.

Let us see one more example sentence (E5) in order to confirm that the structure-bound MT, which lacks the ability to understand the meaning of source sentences, can produce the barely passable translation, and to try to search for the reason for this.

• The soldiers fired at the women and we saw several of them fall. (E 5)

• 兵士達は	女達に	発砲した	そして		(J' 5)
Heishi-tachi-wa	on-na-tachi-ni	happo-shi-ta	soshite		
[soldiers]	[at the woman]	[fired]	[and]		
我々は	彼等の	数人が	倒れるのを	見た。	
wareware-wa	kare-ra-no	sûnin-ga	taoreru-no-wo	mita	
[we]	[of them]	[several]	[fall]	[saw]	

(E5) is one of the sample sentences in [Wilks 1975] where anaphora and references are discussed as the important elements of sentence understanding. As is pointed out by Wilks, a certain extent of understanding is necessary to solve the anaphora and reference problem of the sentence (E5), that is, whether "them" refers "the soldiers" or "the women".

And actually, the structure-bound MT, which cannot understand the meaning of "fired at" and "fall", may translate "them" into "kare-ra" being indiffer-

ent to the anaphora and references. In Japanese "kare-ra" denotes the pronoun of [male, third person, plural], and "kanojo-ra" denotes the pronoun of [female, third person, plural], so (J'5) is somewhat misleading translation. Nevertheless, human (i.e. almost all the Japanese readers) can surely understand the sentence (J'5) correctly; that is, they can understand that "kare-ra" (= "them") is referring "on-na-tachi" (= "the women") not "heishi-tachi" (= "the soldiers"). The reason of this is that the human's brain can understand the meaning of the sentence (J'5) with the support of the common sense like:

X fires at Y → Y will severly wounded
→ Y will fall and die,

which functions as the compensator for the anaphora and references.

The above example shows that the lack of the anaphoric ability in structure-bound MT may sometimes be compensated by the human-side, which is the encouraging fact for MT.

So far the point we are trying to make clear is that even IG-neglecting MT (= structure-bound machine translation systems) can generate target sentences that convey the correct meaning of source sentences, when the latter are written in simple, logical structures.

5. Conclusions

This paper has dealt with the limitations and potentials of structure-bound machine translation (MT) from the standpoint of the idiosyncratic gaps (IG) that exist between Japanese and English. The commercial machine translation system (MT) currently on the market are inept at handling IG since they are still not capable of understanding the meaning of sentences like human translators can, and are thus bound by the syntactic structures of the source sentences. This was pointed out by applying the Cross Translation Test (CTT) to several sample sentences, which brought the performance limitations of structure-bound machine translation into sharp relief. But the CTT applications also showed that if the source language sentence is simple, logical and contains few ambiguities, today's IG-neglecting machine translation systems are capable of generating acceptable target sentences, sentences that preserve the meaning of the original (source) sentences and can be understood.

However, source sentences are not always simple, logical and unambiguous. Therefore, to improve the performance of machine translation systems it will be necessary to develop technology and techniques aimed at rewriting source sentences prior to inputting them into systems, and at formalizing (normalizing) and controlling source sentence preparation. One move in this direction in recent years has had to do with the source language itself. Research has been steadily advancing in the area of Sublanguage Theory. Sublanguages are more regulated and controlled than everyday human languages, and therefore make it easier to create simple, logical sentences that are relatively free of ambiguities. Some examples of sublanguage theories currently under study are "sublanguage" [Kittredge and Lehrberger 1982], "controlled language" [Nagao 1983] and "normalized language" [Yoshida 1984].

The aim of these sublanguage theories is to assign certain rules and restrictions to the everyday human languages we use to transmit and explain information, improving the accuracy of parsing operations necessary for machine processing, and enhancing human understanding. Some examples of the linguistic rules and restrictions envisioned by the sublanguage theories are rules governing the creation of lexicons [Kittredge and Lehrberger 1982], rules governing the use of function words related to the logical construction of sentences [Yoshida 1984] and rules governing the expression of sentential dependencies [Nagao 1983].

References

Eikyo [Nihon-Eigo-Kyôiku-Kyôkai] (eds.) (1982), '2 Kyû Jitsuyô Eigo Kyôhon' ('2nd Class Practical English Textbook'), Nihon-Eigo-Kyôiku-Kyôkai, Tokyo, 1982 pp.202-203 (in Japanese).

Kittredge, Richard and J. Lehrberger (eds.) (1982), 'Sublanguage: Studies of Language in Restricted Semantic Domains', Walter de Gruyter, Berlin, New York, 1982.

Nagao, Makoto (1983), 'Seigen-Gengo-no Kokoromi' ('A Trial in Controlled Language'), in Shizen-Gengo-Shori-Gijutsu Symposium Yokô-Shû, Information Processing Society of Japan, Tokyo, 1983 pp.91-99 (in Japanese).

Nagao, Makoto (1985), 'Kikai-Hon-yaku-wa Doko-made Kanô-ka' ('To What Extent Can Machine Translate?'), Kagaku, Iwanami, Tokyo, vol. 54 no.9, 1985, pp.99-107 (in Japanese).

Nakamura, Yasuo (1973), 'Hon-yaku-no Gijutsu' ('Techniques for Translation'), Chû-kô-Shinsho 345, Chûô-Kôron-Sha, Tokyo, 1973 (in Japanese).

Newsweek (1982), 'Newsweek' January, 18, 1982 p.45.

Nitta, Yoshihiko, et al. (1982), 'A Heuristic Approach to English-into-Japanese Machine Translation', in J. Horecky (ed). Proc. COLING 82 (at Prague) [Proceedings of the 9th International Conference on Computational Linguistics]. North-Holland Publishing Company, 1982, pp.283-288.

Nitta, Yoshihiko, et al. (1984), 'A Proper Treatment of Syntax and Semantics in Machine Translation', in Proc. COLING 84 (at Stanford) [Proceedings of the 10th International Conference on Computational Linguistics], Association for Computational Linguistics, 1984, pp.159-166.

Simmons, Robert F. (1984), 'Computations from the English', Prentice-Hall, Englewood Cliffs, New Jersey, 1984.

Slocum, Jonathan (1985), 'Machine Translation: Its History, Current Status and Future Prospects' Computational Linguistics, vol. 11 , no.1, 1984, pp.1-17.

Wilks, Yorick (1975), 'An Intelligent Analyzer and Understander of English', Communications of the ACM, vol.18, no.5, 1975, pp.264-274.

Winograd , Terry (1983), 'Language as a Cognitive Process: vol. I : Syntax', Addison-Wesley, Menlo Park, Calif. 1983.

Yoshida, Shô (1984), 'Nihongo-no Kikakuka-ni-kansuru Kisoteki Kenkyû,' ('Basic Study on the Normalization of Japanese Language'), Shôwa 58-nen-do Kagaku Kenkyû-Hi Hojokin Ippan-Kenkyû (B) Kenkyû-Seika Hôkoku-Sho (Research Result Report on the General Study (B) Sponsored by the Shôwa 58 Fund for Science Research) Kyushu University, Kyushu, 1984 (in Japanese).