MACHINE TRANSLATION: A PROBLEM OF UNDERSTANDING

Yoshihiko Nitta
Senior Researcher
Advanced Research Laboratory
Hitachi Ltd.

INTRODUCTION

Expectations surrounding the development and use of commercial machine translation systems are running extremely high today. However, quite a few problems remain to be solved before systems capable of "automated" translation can be realized. One of the most formidable problems facing machine translation at this stage of the technology is how to overcome differences in the semantic structure of source and target language sentences resulting from the distinctive ways of thinking endemic to the different cultures upon which those languages are based. To be more precise, the biggest problem facing automatic translation today is the inability to effectively deal with the idiosyncratic gaps inherent between different languages, i.e. machine translation systems are still incapable of "understanding." Therefore, even the large-scale commercial machine translation systems on the market today are still, for the most part, in the developmental stage.

The first section of this paper is devoted to language modelling, the techniques for which form the basis of presentday machine translation systems. This is followed by a brief discussion of the differences between human translation and machine translation from the standpoint of language understanding capabilities. The third section of this paper describes certain idiosyncratic structural differences that exist between the Japanese and English languages, and presents examples of these differences brought out by comparing English and Japanese sentences with the same meanings. These comparisons are made using a new method called the "Cross Translation Test" (CTT), which reveals the deep gaps that exist in the sentence structures of English and Japanese as a result of the distinctive cultures, i.e. the peculiar ways of thinking and expressing concepts, in which these languages have their respective origins. By showing the extent of these idiosyncratic structural differences, or "gaps" as they are called here, CTT points out the futility of trying to deal with them via formal processing methods. At the same time, however, CTT also provides encouraging evidence pointing to the fact that even the structure-bound machine translation systems that comprise the brunt of today's commercial systems possess the ability, or at least latent capability, to produce translations that can be judged "acceptable." And finally, this paper concludes with a brief summation of the main points made throughout the argument, plus a short discussion of metalanguages, such as sublanguages and normalized language, which are presently being studied as possible means of augmenting the capabilities of today's still unperfected machine translation systems.

NATURAL LANGUAGE MODELING

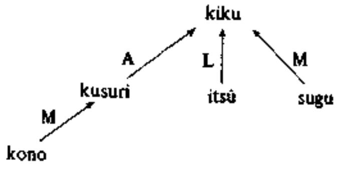
Before a natural language can be processed by a computer, it must first undergo something called language modeling to remove the superficial complexities that abound in linguistic expressions, accurately extracting and formalizing only that intrinsic language data required for computer processing.

There are currently a number of techniques being used in the modeling of natural languages, including phrase structure models, network models, boolean models and a variety of graphic models.[1] [2] For our purposes, here, however, we will limit our discussion to two very fundamental techniques, the dependency structure model (See Figure 1) and the phrase structure model (See Figure 2).

Dependency structure modeling makes it possible to build a very basic semantic representation that labels the semantic roles of the words in the source sentence by classifying the dependency relations that exist among them and by assigning them case markers. Dependency structure modeling is indispensable for the analysis and generation of Japanese sentence sturctures.[3] Phrase structure modeling labels the syntactic roles of words by determining the governor-dependent relation, the head-complement relation or the mother-daugher relation. Phrase structure modeling is very effec-

•	この Kono	薬は kusuri-wa	胃痛に itsû-ni	すぐ sugu	効く。 kiku	
	[this]	[medicine]	[on stomachache]	[immediately]	[take effect]	(J1)

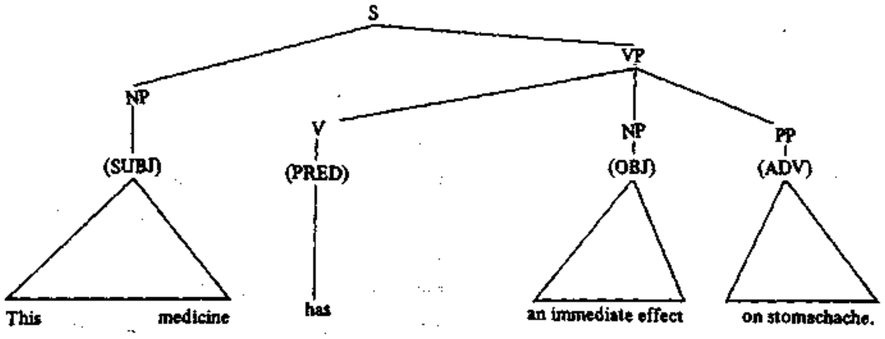
• [Lit. This medicine takes effect on stomachache immediately.] (E'1)



A, L, M : Semantic Roles (or Case Markers),
A : Agentive, M : Modifier, L : Locative.

Figure 1. An Example for Dependency Structure Modeling

- This medicine has an immediate effect on stomachache. (E1)
- [Lit. この 薬は 胃痛の 上にすばやい 効き目を 持っている。] (J'1)
  Kono kusuri-wa itsū no ue-ni subayai kikime-wo motte-iru.



SUBJ, PRED, OBJ, ADV : Syntactic Roles,
SUBJ : Subject, PRED : Predicate, OBJ : Object, ADV : Adverbial.

Figure 2. Example of Phrase Structure Model

tive for analyzing and generating English sentence structures.[4]

Two major questions of great interest, not only with regards to the field of machine translation, but for all fields related to natural language processing, are the extent to which semantic information should be incorporated in the language model, and the degree to which the semantics (meaning) of the source language sentence is related to the model representation. Unfortunately, the scope of this paper does not permit a discussion of these points. But it should at least be pointed out that semantic network modeling,[5] which is once again being looked upon as an effective and practical approach to representing the semantic structures of source language sentences, can be considered as a variation or extension of dependency structure modeling.

MACHINE TRANSLATION vs. HUMAN TRANSLATION

Most of the machine translation systems on the market today[6] [7] are generally "structure-bound" systems. That is, they are designed to translate source sentences in strict adherence to the syntax structures of those sentences, i.e. they perform literal, word-for-word translations. The principal reasons for this are as follows:

1) The machine translation process is controlled by structural information extracted from the source language sentence (The term structural information here refers to syntactic data, i.e. information concerning the way words are put together to form phrases, clauses and sentences.):
2) Therefore, all target sentences pro-

duced by machine translation systems are literal translations of source sentences, and as such closely resemble the source sentences in terms of wording and structure;

3) Machine translation systems are incapable of judging whether or not the meanings of the target sentences they produce accurately convey the meanings of the source sentences upon which they are based; and

4) Machine translation technology is still unable to make adequate use of information other than the linguistic information contained in source sentences to produce target sentences that convey the full meaning of the original text. The non-linguistic information referred to here consists of data concerning the situation surrounding the original text, the context within which a sentence or passage is used, plus common sense factors and world knowledge possessed by both the authors and readers alike, i.e. information that can shed light on the meaning of a sentence.

Figure 3 presents a simplified overview of a typical Japanese to English translation process, based on the dependency structure and phrase structure models discussed earlier. This diagram clearly indicates the strong control over the language transformation and generation phases of the machine translation process exerted by the structural information obtained from the source sentence.

Another point brought out in this diagram is the characteristic differences that exist between predicate verbs in Japanese and English. As you can see, the Japanese verb "KIKU," which literally means "to take effect," has been transformed to the English verb "to have," the literal Japanese translation of which is usually "MOTSU," plus the noun "effect," "KŌKA" in Japanese, which serves as the object of the verb. This is a typical example of a 'structural idiosyncratic gap' (SIG) something present-day commercial machine translation systems still find extremely difficult to handle. This is because SIGs require precise, formal rules before they can be processed by a machine translation system, and formulating the universal grammatical rules needed to manipulate these SIGs correctly is an exceedingly difficult task. At present, SIGs are being dealt with to a certain degree by heuristic (trial and error) rules put together based on experience obtained via in-depth analysis of the two languages. But appropriate heuristic rules are not easy to come by, and therefore are not always available when the system needs them.

Compared to machine translation, human translation, i.e. the translation of a source language to a target language by a human translator, is essentially semantic oriented, that is, it places much more emphasis on comprehending the meaning of the source language and preserving that meaning in the target language. Of course, human translators also must be concerned with sentence structure (syntax), and have difficulty understanding source language sentences that feature structural abnormalities. But while there is no

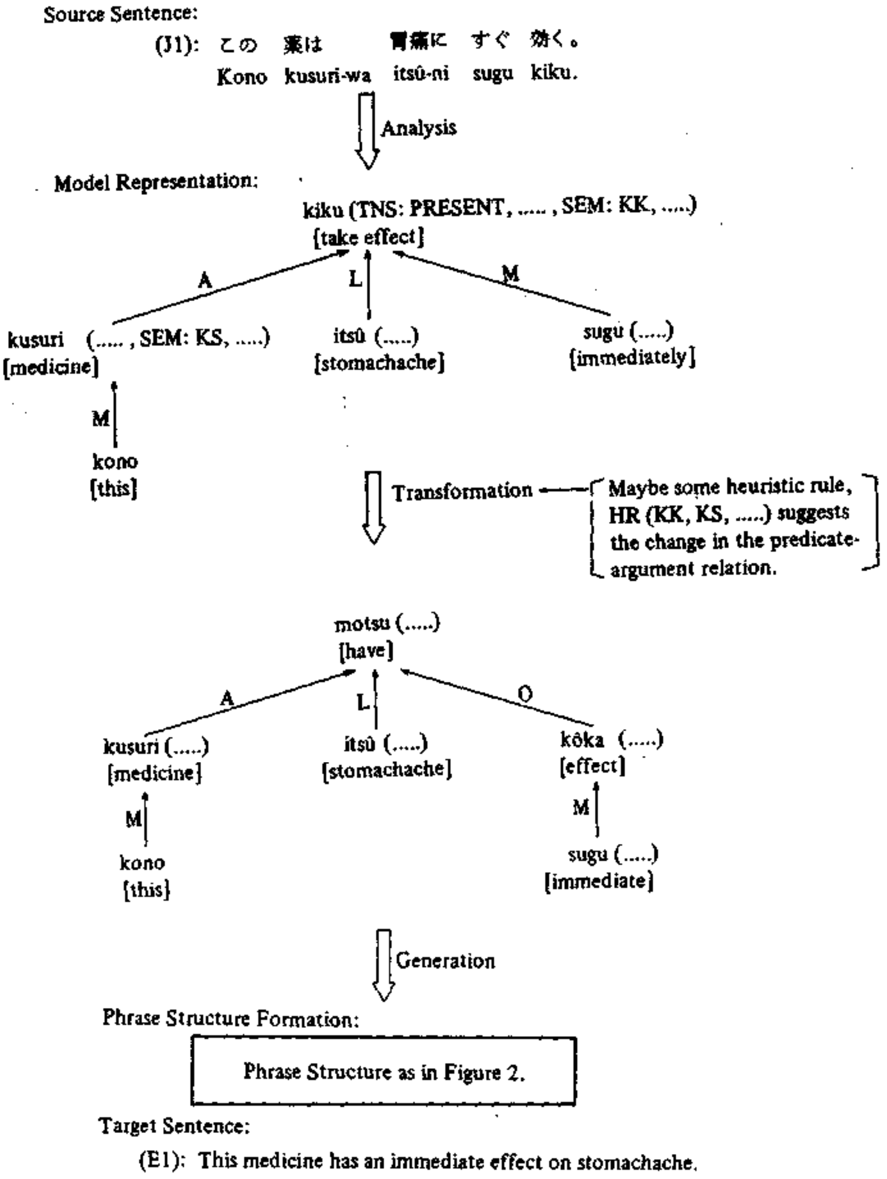

Source Sentence:
(J1): この 薬は 胃痛に すぐ 効く。
Kono kusuri-wa itsû-ni sugu kiku.
Analysis
Model Representation:
kiku (TNS: PRESENT, ….. , SEM: KK, …..)
[take effect]
A
L
M
kusuri (….. , SEM: KS, …..)
[medicine]
itsû (…..)
[stomachache]
sugu (…..)
[immediately]
M
kono
[this]
Transformation
Maybe some heuristic rule, HR (KK, KS, …..) suggests the change in the predicate-argument relation.
motsu (…..)
[have]
A
L
O
kusuri (…..)
[medicine]
itsû (…..)
[stomachache]
kôka (…..)
[effect]
M
M
kono
[this]
sugu (…..)
[immediate]
Generation
Phrase Structure Formation:
Phrase Structure as in Figure 2.
Target Sentence:
(E1): This medicine has an immediate effect on stomachache.

Figure 3. Simplified Sketch of Machine Translation Process

doubt as to the vital role played by structural information derived from the source sentence in the human understanding process (at least in the initial stage of that process), in the final stage much more emphasis is placed on semantics, on grasping the meaning of a sentence in order to create adequate translations. The reasoning behind this statement can be summed up as follows:

1) Human translators, unlike their machine counterparts, are capable of understanding, and can therefore comprehend the meaning, both explicit and implicit, of a source language sentence, completely free of the restraints inherent in the structure, wording and phrasing of that source sentence;
2) Human translators are therefore capable of 'creating' a target language sentence based on a 'mental model' or 'image-like mental diagram' they construct in their minds predicated on their understanding of the meaning of the source language sentence (See Figure 4);
3) And in his/her efforts to understand the meaning of a source language sentence, the human translator can also rely on non-linguistic information (sometimes referred to as extra-linguistic information) not contained in that sentence, such as data concerning the situation surrounding the original text, the context within which the sentence is used, and common sense factors and world knowledge which the author of the source sentence and the translator share in common; and
4) Thus, the human translator is able to overcome SIGs that occur between the source and target language sentences quite easily, even unconsciously.

We still have no idea whatsoever what the mental model or image-like mental diagram of a source language sentence formed in the mind of a human translator is, or how it is constructed. But a simple experiment can be used to show that such a mental model must really exist.

For instance, if we were to show a person the sample Japanese sentence (J1) or English sentence (E1) given in Figures

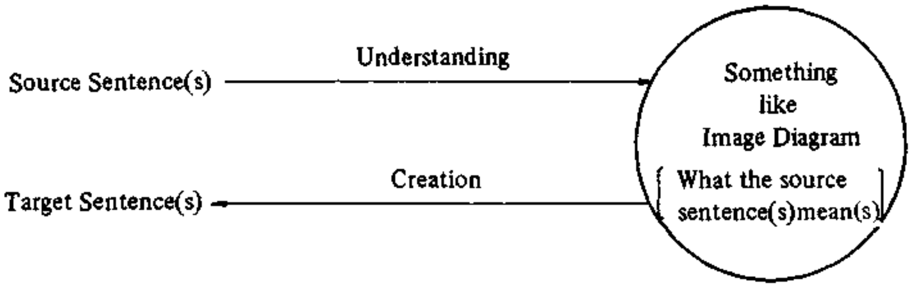

Source Sentence(s)
Understanding
Something like Image Diagram
What the source sentence(s)mean(s)
Creation
Target Sentence(s)

Figure 4. Human Translation Process

3 and 4, and ask him/her to draw a diagram using key words and relational links to illustrate what he/she understands that sentence to mean, we might wind up with a diagram such as that shown in Figure 5 (apart from the slight differences on linking topology and key words). Conversely, if we were to ask a different person to look at the diagram in Figure 5, and then based on his/her understanding of that diagram, to render that meaning into either Japanese or English, we might end up with sentences such as J1 or E1, or perhaps like J2 or E2, shown below (apart from the delicate differences on determiner, tense, aspect and mood). For example, E'1 is a literal translation of J1 (See Figure 1) and J'1 is a literal translation of E1 (See Figure 2). Also, the pairs 'J1 and E1' in Figures 1 and 2, respectively, and 'J2 and E2' shown below, are free translations of each other.

To get back to the point of this example, for a person (read human translator) who is capable of constructing a mental model or diagram like the one shown in Figure 5 in his/her mind based on an understanding of the meaning of a sentence, it is no problem whatsoever for him/her to create free translations of that source sentence, translations that deftly manipulate any SIGs that might exist between the source language and target language sentences.

Figure 6 illustrates the difference between machine translation and human translation, which briefly put, is the ability to grasp and understand the meaning of a sentence. The human translator is capable of understanding the meaning of a source sentence, and is therefore able to manipulate SIGs quite easily; commercial machine translation systems on the market today do not possess these capabilities.

DIAGNOSING SIGs

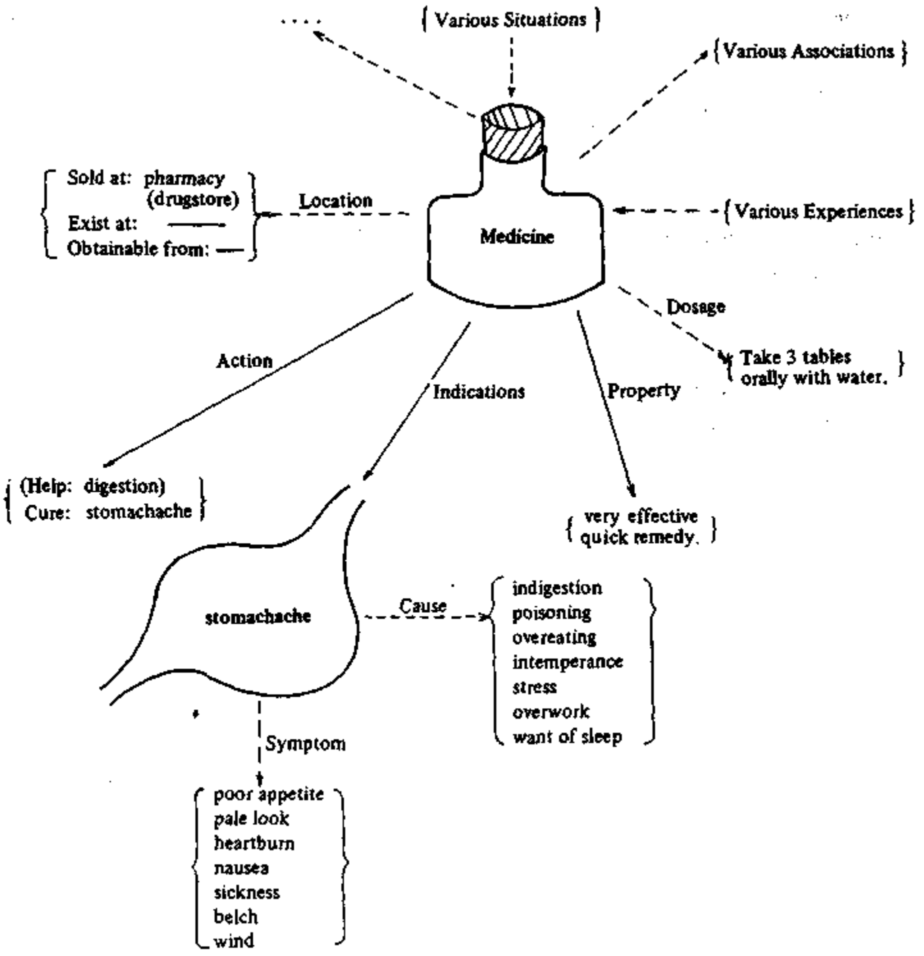
The best method we can think of for analyzing concrete examples of SIGs would be to construct and fix the semantic information in a single, abstract, universal semantic representation such as a mental model/diagram, and then compose sentence structures possible in a variety of target languages based on this fixed semantic information in order to clarify the differences inherent in these sentences by comparing them with one

• 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	itsû-kara	sukuu-darô.	
	[this]	[medicine]	[you]	[soon]	[of the stomache]	[will cure]	



——→ : Information directly obtained from a source sentence.

- - -→ : Information obtained from outside of a source sentence; (including common sense, world knowledge, and so on);

Source Sentence:

(J1):	この	薬は	胃痛に	すぐ	効く。
	Kono	kusuri-wa	itsū-ni	sugu	kiku.
	{this]	[medicine]	{stomachache]	{immediately]	{take effect]

[Lit. (E'1) : This medicine takes effect on stomachache immediately.].

Figure 5. Image Diagram for Sentence Understanding by Humans

another. Unfortunately, at present we have no way of representing the mental models constructed in the heads of human translators. Therefore, we have devised the second best method to analyzing SIGs. This is the CTT method mentioned earlier in this paper.

An intuitive understanding of the CTT method can be obtained from the following CTT test procedure. However, space does not permit a complete description of the analysis code system or SIG measurement methods used with CTT.

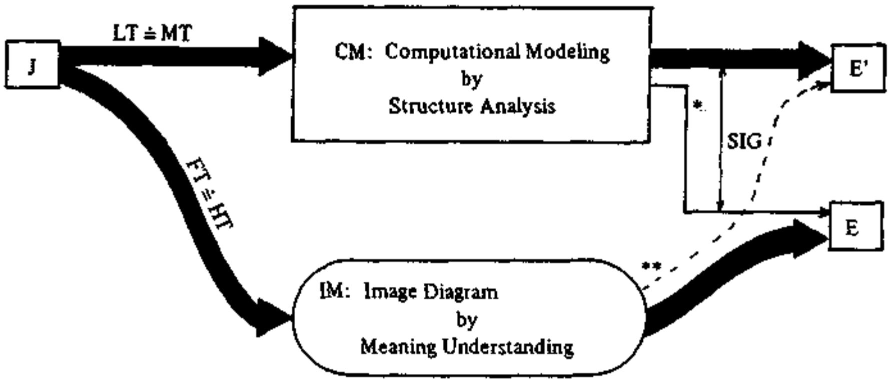
CTT involves the following steps:

1) Select or produce a sample sentence correctly written in one language (for example, in English) and label it E;

2) Select or produce a free translation of sample sentence E in the other language (for example, in Japanese), and label this translation J;

3) Finally produce a literal translation of E in the same language as J and label it J', and produce a literal translation of J in the same language as E and label it E'.

The term free translation as used here refers to translations that correctly convey the meaning of the source sentence,



LT: Literal Translation,
MT: Machine Translation,
FT: Free Translation,
HT: Human Translation,

SIG: Structural Idiosyncratic Gap
*: It is very difficult for MT (≐LT) to take this root because of the gap G.
**: Mediocre HT often drifts into this root;

J: Source Language Sentence,
E: Target Language Sentence,
E': (Awkward) Literal Translation.

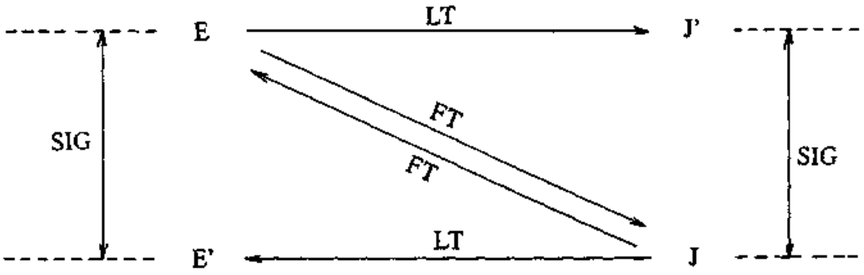
Figure 6. Machine Translation, Human Translation and Structural Idiosyncratic Gap

while the term literal translation refers to translations that preserve the wording, phrasing and syntactic structure of the source sentence. For our purposes here, we can assume that machine translation is almost equivalent to literal translation; and human translation is almost equivalent to free translation. In other words, the term literal translation can be used to represent the "competence" or capabilities of the structure-bound commercial machine translation systems on the market today, while the term free translation can be assumed to represent the semantics-oriented translation capabilities of human translators.

This being the case, the SIGs that exist between English and Japanese sentences that possess the same meaning can be measured. For example, the English and Japanese sentences selected or produced for a CTT test (E, E', J and J') can be used to express SIGs as shown in Figure 7, where $|E - J| = G, |J - J'| = G$ and $|E - E'| = G$.

Although space does not permit a complete description of CTT here, the intuitive meaning of the CTT is inspecting and analyzing the SIGs by means of comparing the four sets of sentences (E, J', J, E') as is indicated in Figure 7.

The utility of CTT does not reside simply in accurately measuring the structural gaps that appear between source and target sentences at translation time. Rather, the results of CTT can be used to derive clues to the formulation of heuristic rules for dealing with these gaps whenever they crop up. CTT results can



SIG: Structural 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:
$LT \doteq MT$ and $FT \doteq HT$,
where, MT: Machine Translation, and HT: Human Translation.

Figure 7. Illustrative Definition of Idiosyncratic Gap

also be used to devise and/or select metalinguistic methods aimed at reducing SIGs, such as sublanguages,[8] controlled language[9] and normalized language.[10]

The remainder of this section is devoted to analyzing a few actual examples of SIGs using CTT.

The Japanese sentence above marked (J3) is the opening sentence in a famous novel, titled "Yukiguni," written by one of Japan's literary greats, Yasunari Kawabata. (E3) is the English rendering of that sentence produced by Edward G. Seidensticker in his equally famous translation of that novel, "Snow Country". This example was selected because of the major SIGs that exist between the Japanese and English sentences.

The first, and most obvious SIG is the lack of a subject in the original Japanese sentence:

Φ1-ga	...... wo	nukeru-to
[Subject]	[Object]	[Predicate + Conjunction]

Φ2-wa	yuki-guni	de-atta.
[Subject]	[Complement]	[Predicate]

Where Φ is the empty mark denoting the omitted term. (J3) is a poetic and lyrical Japanese sentence. But it is also quite common in ordinary Japanese to omit the subject of a sentence when that subject is "understood," i.e. when the writer or speaker feels reasonably certain that the reader or listener will understand what he is writing/speaking about from the context. The omission of the subject in Japanese is considered to give the sentence a certain lyric appeal, to make it poetic and concise in nature. This device is so common in Japanese that it is even used quite frequently in technical papers and/or business reports. And the Japanese people are so accustomed to not using subjects in their sentences, that they don't feel the least unnatural with this type of construction, and their understanding of the meaning of such sentences isn't impaired in the least. This is a major idiosyncracy of the Japanese language.

However, in English, the subject of the sentence must be clearly stated. Therefore, in his translation (E3) of the original (J3) sentence, Seidensticker supplied the missing subject, "train (=RESSHA)," the Φ1 of the predicate "NUKERU (='pass through' or 'come out')," based on his understanding of the situation surrounding the sentence (J3). But then, to preserve the simplicity and lyric appeal of the original Japanese as much as possible, Seidensticker chose not to provide a

• 国境の	長い	トンネルを	抜けると	雪国	であった。	(J3)
Kokkyô-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.	

subject Φ2 for the second predicate "YUKI-GUNI DE-ATTA (=was the snow country)," instead opting to use the locative adverbial phrase "into snow country" as a goal of the first predicate "NUKERU." Thus, while the poetic style of the original sentence written by Kawabata was rendered into somewhat more ordinary English prose by Seidensticker, the translation still managed to preserve a good deal of the original simplicity.

But what kind of mental model/diagram did Seidensticker form in his mind based on his understanding of the meaning of the source sentence? Although any number of explanations are possible, without going into too much detail, the following sentence can be given as an example of how he might have filled in the blanks (the missing subjects Φ1 and Φ2 of the sentence (J3)) in his head prior to translating the sentence.

Ressha-ga kokkyo-no nagai tonneru-wo
[train] Φ1

nukeru-to ressha-ni notte-iru watashi-
[what I encountered was]

no (=of Shimamura) me-ni tobikonde-
kita mono-wa yukiguni de-atta.
Φ2

This kind of Japanese sentence is much too verbose, too wordy, and is therefore unnatural. It completely lacks a sense of literary appeal. The omission of subjects from Japanese sentences is an idiosyncracy that is firmly rooted in our history and culture, and as such, is not likely to change in the near future. But if we try to omit subjects from English sentences in the same way, then we wind up with very awkward, ill-formed and hard to understand sentences which exhibit extremely large SIGs.

A literal translation of the source sentence (J3), such as that given as (E'3), is totally unacceptable. This is because (E'3) preserves the structure of the original, and only provides the subject "it." This usage of "it" is unacceptable because it can be interpreted to mean that "it became the snow country, only after the train passed through the long border tunnel," when in fact it had always been snow country.[11] Therefore, a machine translation of (J3) would most likely result in a sentence similar to (E'3), and would be considered unacceptable for its inability to overcome the major SIG present in this sentence.

• One thing is certain. (E4)
[hitotsu-no] [koto] [de-aru] [tashika-na]

• [Lit. 一つの 事が 確か である。] #(J'4)
Hitotsu-no koto-ga tashika de-aru.

• 確かな 事が 一つ ある。 (J4)
Tashika-na koto-ga hitotsu aru.
[certain] [thing] [one] [exist]

• [Lit. A certain thing exists by one.] #(E'4)
[tashika-na] [koto] [aru] [hitotsu-dake]

```
• What        a  great     artist              dies     with        me !                    (E5)
  [nanto-iu]     [idai-na] [geijutsuka]        [shinu]  [to-tomoni] [watashi]

• [Lit  何という  偉大な    芸術家が          私と共に       死ぬ  の  だろうか。        +(J'5)
     Nanto-iu  idaina    geijutsuka-ga    watashi-to-tomoni  shinu no-darō-ka.

• 私が       死ねば  何という  偉大な   芸術家が  一人   この世から  去ることに  なるのだろうか。 (J5)
  Watashi-ga shine-ba nanto-iu idai-na geijutsuka-ga hitori konoyo-kara saru koto-ni naru-no-darō-ka.
  [I]        [if die] [what]  [great] [artist]      [one] [from this world]|[that will imply]
                                                                 [disappear]

• [Lit. If  I  died, it would imply that one  great  artist    disappeared from this world.]   (E'5)
  [Moshi] [shinu] [koto-naru-darō]     [idaina]     [saru]           [konoyo-kara]
     [watashi]                     [hitori-no]  [geijutsuka]
```

Let us just look at a couple more examples of major SIGs that exist between Japanese and English.[12]

In the above example, a human translation of the English sentence (E4) is shown as Japanese sentence (J4). In this translation, the adjective "certain" in the original sentence has been changed from a predicative (TASHIKA DE-ARU) to an attributive (TASHIKA-NA KOTO-GA). This is another example of a SIG that machine translation systems find very difficult to handle. The literal translation shown as (J'4), while understandable, is unnatural. Therefore this machine translation cannot be considered successful. Similarly, the machine translation of (J4) back into English (E'4) might be capable of being understood by some readers, but it is not a natural English construction by any means.

The human translation of English sentence (E5) into the Japanese sentence identified as (J5) obviously made use of knowledge concerning the situation, information not directly perceivable from the linguistic data available. Otherwise, the prepositional phrase "with me" would most likely have been translated as "WATASHI-TO-TOMO-NI," which would indicate that more than one person is dying; or rather, 'I (=the speaker)' and 'the great artist' are going to commit a double suicide. Since a machine translation system is not capable of drawing on this type of situational knowledge, i.e. the speaker and the great artist are one in the same person, it would produce a literal translation such as (J'5), and would thereby distort the meaning of the source

(NOTE)

The symbols "*", "#" and "+" which appear to the left of the sample sentence identifiers possess the following meanings:

*: Ungrammatical sentence that cannot be understood;

#: A little unnatural, but grammatically accurate, and therefore capable of being understood with some effort;

+: No grammatical errors, and therefore understandable, but does not preserve the meaning of the original sentence; and

No symbol: Grammatically and semantically accurate and understandable.

• 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] (E6)

• 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
Karewa 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'6)

• 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. (J6)

• 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'6)

sentence.

Not all SIGs lead to improper machine translations, however. The above source and target sentences are good examples of cases where a machine translation system, although incapable of processing the SIGs, has produced an acceptable translation of the original sentence (E6) by ignoring them. The sample sentence (E6) was taken from an article in the January 18, 1982 edition of Newsweek,[13] and the free (human translation (J6) was lifted from an English-language textbook used in Japan.[14] (J'6) is the literal translation of (E6) and thus is the most conceivable output of the machine translation.

Although (J'6) is translated in very faltering Japanese, it adequately conveys the meaning of the original English sentence. If the machine translation system had been capable of anlyzing the function of the model auxiliary verb "may have" more accurately, the final translation of the source sentence might have ended "Kyûjo-shita-to itte-yoi de-arô" instead of "kyûjo-shita kamo-shirenai," the former being the better translation.

In the case of (J'6), the machine translation system did a good job rendering the word "crash" in (E6) as "TSUIRAKU [a sudden fall] in sentence (J'6), selecting the correct translation for a word that has multiple meanings. More often than not, however, a machine translation system will fail to come up with the correct word when it has more than one meaning to choose from. This is because the machine translation system is not able to deduce the correct meaning from the context of the sentence and/or by making a judgement based on its understanding of the situation involved.

According to a widely-used and respected English-language dictionary,[15] "crash" has at least three different meanings in Japanese, all three of which could be applicable in the case of (E6). The meanings given are:

1) "SHÔTOTSU" [= the violent striking of one solid object against another];
2) "DAI-ONKYÔ" [= a sudden loud noise]; and
3) "TSUIRAKU" [= the sudden, accidental fall of an aeroplane].

As you can see, even if we limit ourselves to the field of aeronautics, any one of these three translations of "crash" is still possible. The fact that this one English word, "crash," represents "one" concept that has at least three Japanese translations, can be said to be a kind of idiosyncrasy of the English language (as far as that language is related to Japanese).

The literal translation (E'6) is another example of where the machine translation system used some pretty awful expressions, such as "produced 275 dead persons," but nevertheless did an acceptable translation.

To get a better idea of where the SIG lies between the source sentence (E6) and the free translation (J6), let us break the two down into their kernel sentences and label them using case markers.

(E6): Agentive + Predicate-Head
+ Objective + Locative

(J6): Instrumental + Agentive + Objective + Predicate -Head

The above two case patterns indicate that the agentives denote a different entity in the two sentences. In (E6) the agentive is "he" ("KARE"), whereas in (J6) it is "KONO KI" ("this airplane"). Thus we can say that there are many SIGs which are more complicated than simple word-ordering differences. In (J6) the case marker "Instrumental" denotes the phrase "KORE-NI-YOTTE" in Japanese (which means "by this" in English). Since this expression is not found in the original source sentence (E6), it must have been introduced by the translator as a result of his or her understanding of the situation surrounding the contents of the sentence.

The point we are trying to make clear so far with the examples is that even SIG-neglecting machine translation can generate target sentences that convey the correct meaning of source sentences, when the latter are written using simple, logical structures.

CONCLUSIONS

This paper has dealt with the limitations and potentials of machine translation from the standpoint of the SIGs that exist between Japanese and English. The commercial machine translation systems currently on the market are inept at handling SIGs 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 they translate. This was pointed out by applying the Cross Translation Test (CTT) to several sample sentences, to bring 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 SIG-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 sublanguages currently under study are controlled language and normalized language.[16]

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 sublanguage theory are rules governing the creation of lexicons, rules governing the use of function words related to the logical construction of sentences and rules for the expression of sentential dependency patterns.

Quite naturally, there are many people who exhibit an almost allergic reaction to the mere mention of ideas aimed at "assigning restrictions to everyday human languages." Nevertheless, in specialty fields such as medicine, pharmacology, military and computer applications,[17] this assignment of restrictions and rules to the everyday languages used in these fields has already resulted in a number of sublanguages that are being used daily by experts in these fields to communicate with one another. These sublanguages are full-fledged languages in both the lexical and semantic, as well as the discourse sense of the word, and have enabled specialists in these fields to exchange information with one another more accurately and economically than was possible before these sublanguages were developed. It may be time for those people still "allergic" to the development of such sublanguages to reconsider the issue in light of current realities.

This paper is a revised version of one presented by the author at the International Symposium of Machine Translation, 1985, held at the Keidanren Kaikan Hall in Tokyo on October 14, 1985 under the co-sponsorship of

JIPDEC and the Ministry of International Trade and Industry (MITI). In closing, I would like to express my sincere appreciation to JIPDEC's Yoshimitsu Hirai for providing me with the opportunity to write and present this paper, and for his valuable advice and encouragement throughout its preparation.

Footnotes

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

2) Ozeki, Masanori (supervisory ed.) & Y. Aoyama (ed.) (1985), 'OA-no Software' ("Software for Office Automation'), Ohm-sha, Tokyo, 1985 chap. 5, pp. 113-154 (in Japanese).

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

4) 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.

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

6) Nagao, Makoto (1985), 'Kikai-Hon-yaku-wa Doko-made Kanô-ka' ('To What Extent Can Machines Translate?'), Kagaku, Iwanami, Tokyo, vol. 54 no. 9, 1984, pp. 524-532 (in Japanese).

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

8) Kittredge, Richard (chair.) (1982), 'Sublanguages', American Journal of Computational Linguistics, vol. 8, no. 2, 1982 pp. 79-84.

9) 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.

10) 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).

11) Nakamura, Yasuo (1973), 'Hon-yaku-no Gijutsu' ('Techniques for Translation'), Chû-kô-Shinsho 345, Chûô-Kôron-Sha, Tokyo, 1973 p. 27 (in Japanese).

12) Nakamura, Yasuo (1973), 'Hon-yaku-no Gijutsu' ('Techniques for Translation'), Chû-kô-Shinsho 345, Chûô-Kôron-Sha, Tokyo, 1973 pp. 111-112 (in Japanese).

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

14) Eikyô [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).

15) Kenkyusha-Jisho-Henshû-Bu (eds.) (1973), 'New Collegiate Dictionary of the English Language' ('Shin-Ei-Ei-Jiten'), Kenkyusha Ltd. 1973 p. 244.

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

17) Kittredge, Richard (chair.) (1982), 'Sublanguages', American Journal of Computational Linguistics, vol. 8, no. 2, 1982 pp. 79-84.