UDC 681.39-181.48:651.926.011.56:[802.0:809.56]

Japanese-English Machine Translation System "PENSEE"

Masashi SAKAMOTO* and Tsutomu SHIINO**

Abstract

This paper outlines a machine translation system using full semantic analysis based on case grammar. This system which could previously be implemented only on a mainframe, has been developed for the if1000 engineering workstation. Due to improvement of the algorithm for analysis, concise expression of inner expressions, a compact dictionary, reduction of required memory and speed-up of the process have been achieved, resulting in a practical system which has been implemented on the workstation.

In addition, semantic information, case frame, and advanced grammatical rules have made it possible to translate long sentences or complicated ones using lots of modifiers, which even mainframes could not translate properly.

1. Introduction

It has been our dream for a long time that we could design a computer to understand and generate natural language. Effective machine translation using symbolic and integrated techniques has been shown to be unfeasible in the famous A.L.P.A.C. report[1] (1966). Many researchers, however, have continued to make consistent efforts to make their dreams into reality. In addition to the remarkable progress that has been made in hardware and software, research on natural language itself and developments of in the study of artificial intelligence technology[2] have contributed to this advancement.

Whether machine translation can become practical depends on how accurately it can translate sentences written without restrictions by humans and how much post-edit it requires.

In early machine translation systems with simple phrase structure grammar[3], an input sentence could not be analyzed definitely, because of many interpretation for it. Selected words often had a lack of generally applicable meanings. For example, a word appropriate in one specific case can often be inappropriate in other cases.

Case grammar[4] which is generally used today has solved a considerable number of these difficulties, and make it possible to translate accurately even relatively long complicated sentences. However, even if case grammar is used in principle, the width of the translatable domain and the accuracy and naturalness of the output are a function of several factors. These factors include preservation of case frame to declinable words, assignment of semantic information to each word, composition method of grammatical rules, etc. Differences in processing speed ranging from several to dozens of times as fast can also be achieved by using a middle language method, choosing a more efficient process algorithm and so on. In the past mainframes were employed for machine translation based on case grammar because of their huge memory capacity for dictionary and processing and the great number of dynamic steps in the program.

However, the machine translation system "PENSEE" which is based on case grammar, using the engineering workstation if1000/10M, has achieved a wider translatable domain and higher quality than mainframe systems. Using our system, remarkably high speed processing for a workstation, 4000 words/hour (CPU time), has been achieved. This paper describes the features of this machine translation system called "PENSEE".

2. Outline of system

The hardware module configuration of this system is shown in Figure 1, and the software module configurated in Figure 2. The engineering workstation if1000/10M (CPU:M68010), with 8M byte of main memory and an 80M byte hard disk make up the hardware system. UNIPLUS +

* Natural Language Processing Section, Office Systems R & D Department, Systems Laboratory
** General Manager, Office Systems R & D Department, Systems Laboratory

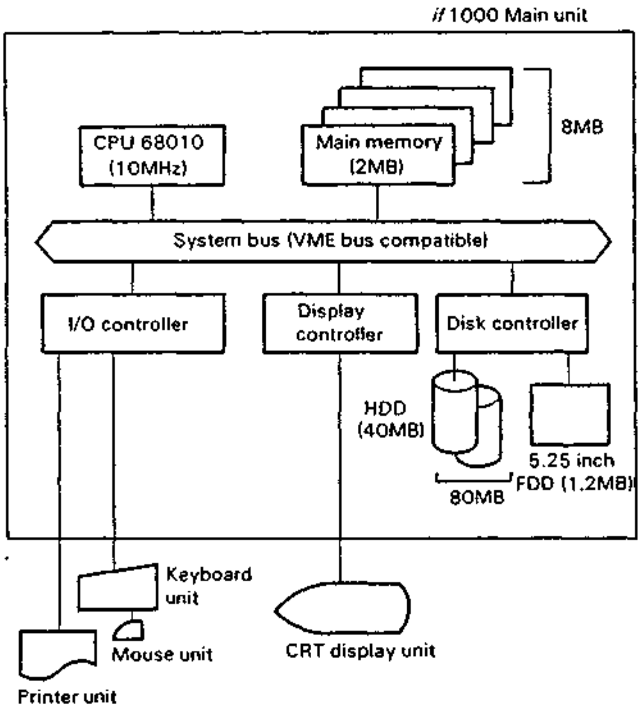


Figure 1 Hardware Module Configuration

(UNIX System V base) is used as the OS, and C is the programming language. The total management system includes a bilingual editing system for Japanese and English and also a dictionary editing system arranged for the registration of a private vocabulary dictionary.

The translation process is made up of Japanese morphological analysis, syntactic analysis and English morphological generation. The rules for syntactic analysis are described in a special language for rule description, which is translated into the C language and stored in the system knowledge base.

An analysis dictionary is used for analyzing Japanese, and a generation dictionary is used for generating English. The basic dictionary contains approximately 50,000 words to begin with. Up to 20,000 words can be added in a given technical field, and also private vocablulary can be registered within this limitation. Access to each dictionary is carried out using ISAM (indexed sequential access method). Figure 3 shows the flow chart of the translation process.

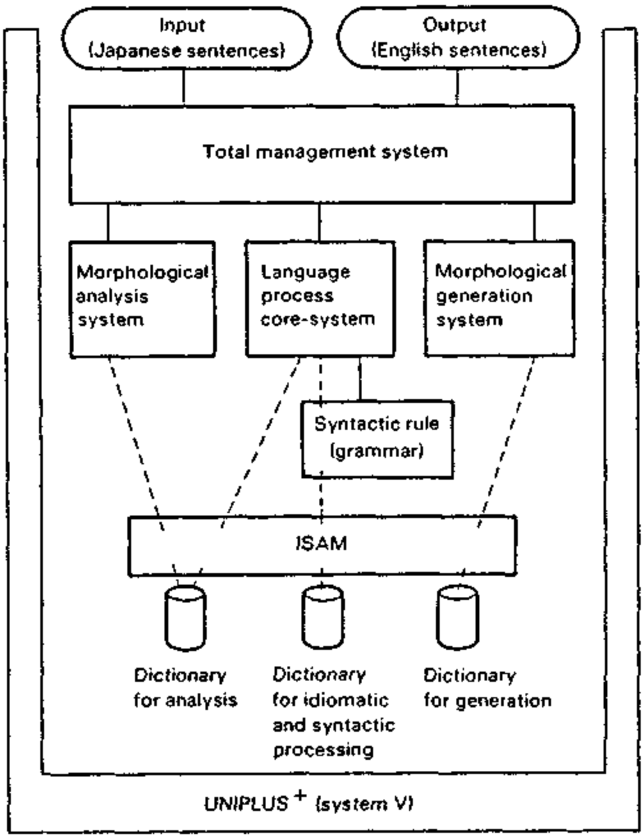

Input (Japanese sentences)
Output (English sentences)
Total management system
Morphological analysis system
Language process core-system
Morphological generation system
Syntactic rule (grammar)
ISAM
Dictionary for analysis
Dictionary for idiomatic and syntactic processing
Dictionary for generation
UNIPLUS+ (system V)

Figure 2 Software Module Configuration

The original Japanese sentences are input using the *i*/1000 Japanese editor (or from a floppy disk file created by the document process system of the *i*/800).

There are two kinds of translation procedures. The first is an "interactive mode" which translate the sentences one by one, and the other is a "batch mode" which translates a certain number of sentences all at once. Since the translation process is basically literal, the output does not always sound natural. The unsatisfactory part of the output can be easily rewritten using an English language editor (post-edit). In cases where substandard output is due to the poor quality of the input (Japanese language), satisfactory results can be gained by rewriting the Japanese sentences (pre-edit) and letting them go through the translation process again.

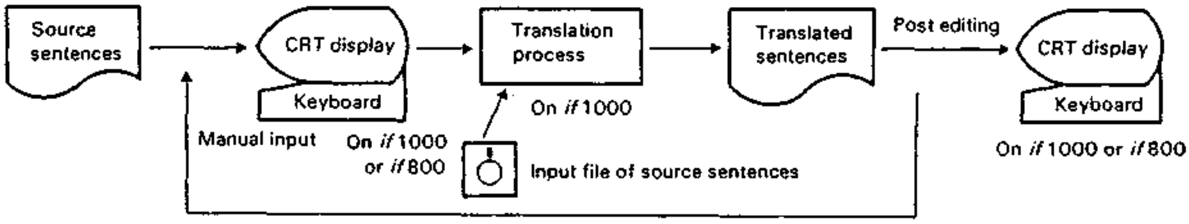


Figure 3 Flow of Translation Procedure

3. Technical Elements of Mahcine Translation and Flow of Translation

Figure 4 shows the technical elements which compose machine translation and the general flow of the translation process.

An input character string is divided into words through morphological analysis. Thus, the part of speech, conjugation type and conjugation form of each word are identified. Since words are not separated by spaces in written Japanese, several possible solutions can exist in sentences which include many Japanese phonetic characters. In general, the longest matching method has been adopted as the guiding principle. In this method, the word which coincides with the longest word in the dictionary is determined as the solution. Since it is possible to generate errors using only this method, it is necessary for accurate analysis to utilize other information for determning the coherence of the various parts of speech, etc.

Syntactic analysis is a procedure which analyzes sentence structure by clarifying word relationships (subject, predicate, modification, qualification, etc.). However, if the analysis is made solely on the basis of information on parts of speech and conjugations obtained through morphological analysis, application of the grammar rules, may still result in several different solutions. This means that there are many cases in which the correct meaning of a sentence cannot be recognized. Some examples are shown below.

"Kare wa dento o keshita" (He turned off the light)
"Kare wa sugata o keshita" (He disappeared.)

These two sentences have the same pattern and the same word "kesu", but the meanings are quite different from each other.

"Kare wa shosai de e o kaite iru." (He is drawing a picture in his study.)
"Kare wa enpitsu de e o kaite iru" (He is drawing a picture with a pencil.)

In the case of Japanese postposition "de", the former indicates the place and the latter the tool using the same words in each sentence.

The procedure used to understand the real meaning of sentences like this is called semantic analysis.

Today, since it is not efficient to carry out syntactic analysis and semantic analysis separately, the use of case grammar which can perform these two analyses at the same time is often adopted.

Context analysis is used to study what words must be supplemented in an elliptical sentence and to understand what demonstratives and pronouns like,"sono" and "sore" refer to by analysis of contextual information.

Generally, however, it is difficult for the machine to understand context information. Also, it is difficult to determine how much of this information is to be accumulated. Therefore context analysis does not work well except in some special cases.

The PIVOT language is an idealized model of a middle experession produced after all analyses have been completed. If the PIVOT language were really to exist, translation between all kinds of languages could be done through it. However, this kind of language has not been fully developed yet. Semantic analysis is the highest practical level which existing machine translation systems can achieve.

In conventional systems, the result of analysis is turned into a middle expression on the source language side, from where it is transferred to the target laguage side. Then the output language is generated through a generation process. This is referred to as the transfer method. On the other hand, in the so-called direct method, words and phrases are translated straight on the basis of the relationship between words and words, structures and structures in both languages. The method employed by the PIVOT language is called the pivot method. At present, most systems use the transfer method.

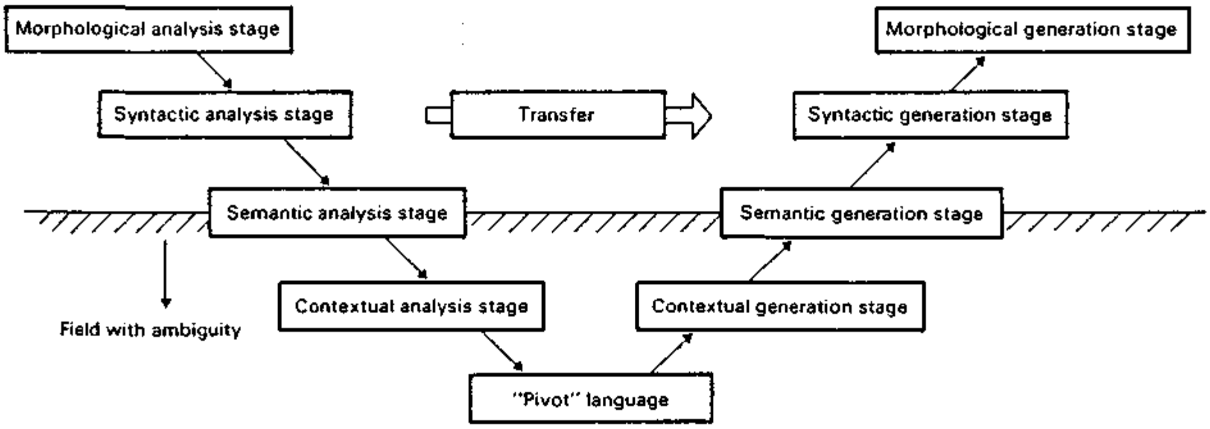

Morphological analysis stage
Syntactic analysis stage
Transfer
Semantic analysis stage
Contextual analysis stage
"Pivot" language
Contextual generation stage
Semantic generation stage
Syntactic generation stage
Morphological generation stage
Field with ambiguity

Figure 4 General Flow of Machine Translation

Since semantic information and context information depend heavily on the cultural background of each country or even district, and on common sense in each technical field, the domain of translation and depth of analysis are in inverse proportion to each other. It can also be said that the process underlying the semantic analysis stage in Figure 4 is in a so-called indefinite territory. Efficiency in machine translation is dependent on putting this indefinite territory into order, extending the translatable domain (sentense pattern, field), and improving its quality (accuracy and smoothness).

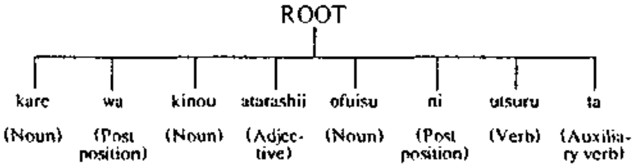
On these points, PENSEE is ranked at an extremely high level.

4. Machine Translation Method Used by PENSEE

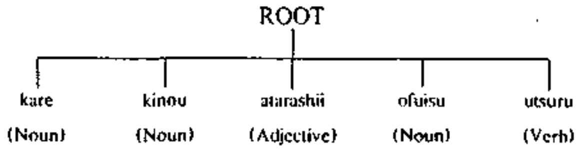
One of the main features of the translation procedure used by PENSEE is its simplified process which operates on only one kind of data. This characteristics has made it possible to reduce the amount of required memory as well as to speed up the process in spite of its deep semantic analysis. This data is in a three structive (called a dependency structure). The semantic process is carried out on the basis of case grammar, in parallel to the syntactic analysis.

The flow of the translation process is shown below using the following sentence 「彼は昨日新しいオフィスに移った」 "Kare wa, kinou, atarashii, ofuisuni, utsuru, ta" (He moved to a new office yesterday).

First of all, the input is separated into words, through morphological analysis.



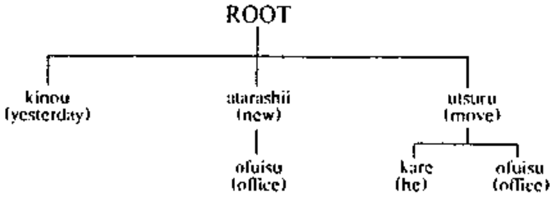
Next, particles which do not correspond to English, the target language, and auxiliary verbs with tense aspect are eliminated from the tree.



"Kare, kinou, atarashii, ofuisu, utsuru" (He, yesterday, new an office, moved to)

The comparison of this sentence to "Iro ga akai ofuisu ni utsutta" (He moved to a red coloured office.) makes it clear that "kare" (he) modifies either "atarashii" (new) or "utsuru" (move). The correct choice cannot be determined without semantic information. In this case "kare" (he) turns out to modify "utsuru" (move) because "kare" is human and nominative.

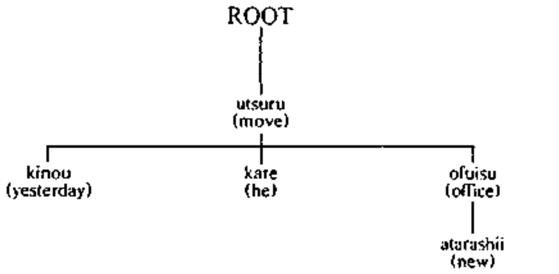
The dependency structure at this moment is one which has a root with declinable words and a free case as shown below.



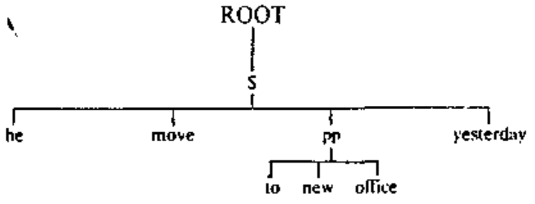
Each declinable word is tied up with the those that have a modifying relationship.

Here, the case structure of "utsuru" (move) governs "kare" (he) and "ofuisu" (office), just as "atarashii" (new) governs "ofuisu" (office). The scope of "kinou" (yesterday) is determined by the semantic label, part of speech of the declinable words, auxiliary verbs and so on. In this case, "ta" which is in the past form, makes it clear that "kinou" (yesterday) belongs to "utsuru" (move).

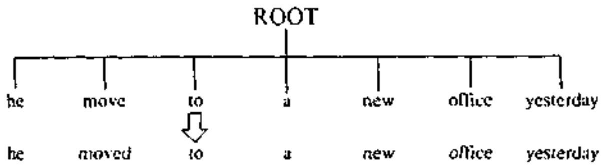
In this way, a dependency structure centered around the main verb "utsuru" (move) is composed.



Now that the deep structure has become clear, this sentence can be transferred into the equivalent tree structure in English. At this stage, the addition of a preposition "to" and the movement of the adverb "yesterday" are also carried out.



Here, S stands for the sentence and pp, the prepositional phrase. Next, the tree structure is transferred into a character string to be output, when the declension of nouns according to number, the declension of pronouns according to case, the conjugation of verbs and adjectives and such things are finally processed. This is called the morphological generation process.



As shown above, instead of using an ordinary transfer method with two middle languages, PENSEE can conduct syntactic and semantic analysis which leads to the constituent structuring of an English tree diagram, using directly the tree structure acquired through Japanese morphological analysis. This method contributes greatly to the reduction of the memory capacity required and to high-speed processing.

5. Dictionary

There are three dictionaries: a system dictionary provided in the system itself; a technical dictionary to which words specific to a given field can be added; a private dictionary to which users can add words as they like. Users are free to compile and register words only in their private dictionary. The system dictionary consists of four dictionaries: a morphological dictionary for Japanese morphological analysis; a morphological dictionary for English morphological generation; a word dictionary of English and Japanese; a syntactic dictionary which has information on case and cooperative relationships for the syntactic and semantic analysis.

Semantic information is provided in each word dictionary. Table 1 shows some examples of semantic information in a private dictionary. Since a private dictionary usually allows its vocabulary only limited usage and is required to have the easiest possible way of registration, it contains only the absolutely necessary semantic information. On the other hand, the word dictionary in the system carries more subdivided semantic information. A syntactic dictionary is basically compiled in the

Table 1 Semantic Categories

Part of Speech	Semantic Categories	Examples
Noun	Soshiki (Organization)	Jiyuto (Liberal Party), Kaisha (Company)
	Ningen (Human)	Bucho (Manager), Kyodai (Brother)
	Shizen (Natural substance)	Taiyo (Sun), Ganseki (Rock)
	Jinko (Artificial substance)	Jidosha (Motor car), Shosetsu (Novel)
	Shinri (Psychology)	Ikari (Anger), Sozo (Imagination)
	Ko-i (Deed)	Hokoku (Report), Hoko (Walking)
	Suryo (Numerical thing)	Ryokin (Fee), Jinko (Population)
	Basho (Space)	Katei (Home), Nanbu (Southern part)
	Jikan (Time)	Gogo (Afternoon), Hangenki (Half life period)
	Chusho (Abstract)	Howa (Saturation), Naiyo (Content)
Verb & adjective	Shunkan dosa (Momentary action with one's will)	Tenka suru (Ignite (vt.)), Sanshutsu suru (Find the answer by calculation)
	Keizoku dosa (Progressive action with one's will)	Hoko suru (Walk), Keisan suru (Calculate)
	Shunkan sayo (Momentary action without one's will)	Hakka suru (Ignite (vi.)), Shototsu suru (Collide)
	Keizoku sayo (Progressive action without one's will)	Hakko suru (Radiate), Seicho suru (Grow)
	Ishiki shinzyo (Mental state)	Gekido suru (Be enraged), Fuan da (Anxious)
	Kankei zyotai (Relation or State)	Itchi suru (Equal), Kyodai da (Huge)

pattern of case frame in case grammer. It also contains other information in order to get higher quality and speed of translation. Some examples are shown below.

1. The surface case pattern
   This is information on what form of adverbial phrase can be taken by each declinable word.
   Examples are shown below.
   noun phrases with postpositional articles
   "kare ga" (he), "kami ni" (on paper), "ji" (letters)
   adverb "tokidoki" (sometimes), "totemo" (very)
   nown phrases on number
   "sankai" (three times), "san bon" (three)
   adverbial form of adjective
   "hayaku" (fast), "utsukushiku" (beautifully)
2. The strength of connection between the verb and the case
   This information shows what sort of nouns are likely to appear in the adverbial phrases of (1). For instance, as for "ataru"; a declinable word, nouns followed by "ga" are often "ningen" (human), "soshiki" (organizations) or groups of that sort, and nouns followed by ("ni") are often ("buttai") (things), ("shokumu") (work, job, function) and so on.
3. Requirements for word selection and a resultant solution
   The informations concerning the requirements and the criteria for word selection is given so that English words, prepositions and sentence pattern can be properly selected according to the Japanese case pattern on the types of nouns in adverbial phrases. Several pairs of requirements and criteria make it possible to translate various Japanese sentence patterns into adequate English ones.
   Furthermore, phrases which can be more conveniently to considered as an entire unit rather than as a single case particle can be processed accordingly since the syntactic dictionary contains the relevant information. For instance, "～ *ni oujite*" is easier to process as one case particle than as a composition of three particles, "ni, oujiru, te". Idioms can also be translated into equivalent English phrases when certain requirements are met. "～ ni ashi o hakobu" is an example of this. The former is called an expression of equivalent phrases, and the latter, an expression of cooperation.

6. Word Selection Based on Semantic Information

Word selection which was impossible before has been achieved by including semantic information in syntactic analysis. This has increased accuracy and the possible domain of translation. Some examples are shown below.

1. Word selection of particles and auxiliary verbs
   "wa"
   "sochi wa ugoka nai" $\xrightarrow[\text{intransitive verb}]{}$ The equipment does not work.
   "sochi wa ugokase nai" $\xrightarrow[\text{causative verb}]{}$ The equipment cannot be used.
   "tame"
   "～ suru tame kojosaseru" $\xrightarrow[\text{causative}]{}$ object (in order to)
   "～ suru tame kojoshita" $\xrightarrow[\text{perfect}]{}$ causation (since)
   "te iru"
   "wakatte iru" $\xrightarrow[\text{momentary action}]{}$ perfect form
   "kaite iru" $\xrightarrow[\text{continuous action}]{}$ progressive form
2. Verb selection
   "kesu"
   "ji o kesu" $\xrightarrow[\text{figure, table, document}]{}$ erase
   "dento o kesu" $\xrightarrow[\text{apparatus, device}]{}$ turn off
   "hi o kesu" $\xrightarrow[\text{things, energy}]{}$ extinguish
   "okiru"
   "jiko ga okiru" $\xrightarrow[\text{phenomenon}]{}$ happen, occur
   "kodomo ga okiru" $\xrightarrow[\text{human}]{}$ wake up
3. Word selection based on cooperation
   "junbi"
   "ryoko no junbi o suru" $\xrightarrow[\text{cooperate}]{}$ prepare for travel
   "junbi ga kanryoshita" → preparation was completed.

Figure 5 shows examples of translation using PENSEE, which prove its function in proper word selection and the automatic addition of omitted nouns.

7. Conclusion

The machine translation system PENSEE in which semantic analysis has been applied using case grammar has been described in this paper.

Previously, only mainframe computers could accomodate this kind of system. Thanks to the concise expression of inner structure, the use of a high-speed algorithm for analysis and a compact dictionary, PENSEE has been implemented as a usable system at the workstation level.

Even mainframe systems could not translate complicated sentences with lots of modifiers or complex or compound sentences, accurately selecting the proper particles and verbs. However, semantic information, case frame and advanced grammatical rules have made all these possible. PENSEE is applicable not only to technical documents but also to general books. At the present time on English-Japanese translation system is under development.

誤ってスイッチを消してしまった場合、最初からやり直す煩わしさはありません。
(Ayamatte suitchi o keshite shimatta ba-ai, saishokara yarinaosu wazurawashisa wa arimasen.)
In case a switch has been turned off by mistake, there is no trouble of doing it over again from beginning.

誤って文書を消してしまった場合、最初からやり直す煩わしさはありません。
(Ayamatte bunsho o keshite shimatta ba-ai, saishokara yarinaosu wazurawashisa wa arimasen.)
In case a document has been erased by mistake, there is no trouble of doing it over again from beginning.

自動発信機能により、電話機に触れることなく文書を送信することができます。
(Jido hassinkino ni yori, denwaki ni fureru kotonaku bunsho o soshinsuru koto ga dekimasu.)
By the automatic transmission function, a document can be transmitted without touching a telephone set.

親展送信機能により、人目に触れることなく文書を送信することができます。
(Shinten soshinkino ni yori, hitome ni fureru kotonaku bunsho o soshinsuru koto ga dekimasu.)
By the confidential transmission function, a document can be transmitted without being conspicuous.

さらに、*if* 1000 UNITOPIA モデル10 Mには強力なグラフィック機能が備えられています。
(Sarani, *if* 1000 UNITOPIA moderu 10M niwa kyoryokuna grafikkukino ga sonaerarete imasu.)
In addition, a powerful graphicate function is set on the *if* 1000 UNITOPIA MODEL 10M.

if 1000 UNITOPIA モデル10Mには、さらに強力なグラフィック機能が備えられています。
(*if* 1000 UNITOPIA moderu 10M niwa, sarani kyoryokuna grafikkukino ga sonaerarete imasu.)
A more powerful graphicate function is set on the *if* 1000 UNITOPIA MODEL 10M.

オフィスの業務処理の技術革新が急速に進むと予想されます。
(Ofuisu no gyomushori no gijutsukakushin ga kyusoku ni susumu to yososare masu.)
It is expected that the technological evolution of the business in the office advances rapidly.

オフィスの業務処理の技術革新が急速に進むと実現されます。
(Ofuisu no gyomushori no gijutsukakushin ga kyusoku ni susumu to jitsugen saremasu.)
If the technological evolution of the business in the office advances rapidly, it is realized.

Figure 5 Examples of Translation

8. References

1. A.L.P.A.C., *Languages and Machines: Computers in Translation and Linguistics*, National Research Council, 1966
2. Information Processing Society of Japan (IPS Japan) Kikaihonyaku (*Machine Translation*), Jouhou Shori, Vol. 26, No. 10, 1985 (in Japanese)
3. Chomsky, N., *Syntactic Structures*, Mouton, 1957
4. Fillmore, C.J., *The Case for Case: Universals in Linguistic Theory* (E. Bach & R.T. Harms eds.) Holt, Rinehart & Winston, pp. 1-88, 1968
5. Shiino, T., Yasuhara, H., Sakamoto, M. and Tanaka, A., *Japanese-English Machine Translation System Implemented In The Personal Computer,* VII-th International Conference on Multiple Criteria Decision Making, 1986
6. Nagasaka, A., et al., *Kikaihonyaku sisutemu no gaiyo (Outline of Machine Translation System, "Rosetta")*, Proc. 30th Annual Convention IPS Japan pp. 1557-1558, 1985 (in Japanese)
7. Okada, K., et al., *Kikaihonyaku sisutemu no nihongo kaiseki bumpo (Grammatical Rules for Japanese Analysis in the Machine Translation System, "Rosetta")*, Proc. 30th Annual Convention IPS Japan pp. 1559-1560, 1985 (in Japanese)