[MT Summit III, Washington, DC, USA, July 1991]

# EJ/JE Machine Translation System ASTRANSAC - Extensions toward Personalization

Hideki Hirakawa, Hiroyasu Nogami, Shin-ya Amano
R&D Center
Toshiba Corporation
1 Komukaitoshiba-cho, Saiwai-ku, Kawasaki, 210 Japan
E-mail: hirakawa@isl.rdc.toshiba.co.jp

**Abstract**

The demand for personal use of a translation system seems to be increasing in accordance with the improvement in MT quality. A recent portable and powerful engineering workstation, such as AS1000 (SPARC LT), enables us to develop a personal-use oriented MT system This paper describes the outline of ASTRANSAC (an English-Japanese/Japanese-English bi-directional MT system) and the extensions related to the personalization of ASTRANSAC, which have been newly made since the MT Summit II.

## 1 Introduction

ASTRANSAC is an English-Japanese/Japanese-English machine translation system which works on a TOSHIBA engineering workstation AS1000 (SPARC LT). AS3000 and AS4000 series. Its design philosophy is, as shown in our paper for the MT Summit II [Amano *et* al., 1989], the improvement in overall translation efficiency, from input of the source text to output of the target text. The automatization of a translation process in a translation office, the so-called translation factory, where co-operative work among human translators, operators and MT systems is realized, is one of the best means of achieving the highest efficiency in translating industrial documents (manuals). ASTRANSAC is designed for this particular kind of applications; in fact, that is currently the primary needs for the MT system. However, we envision a future when the MT system is used in a more personal environment; that is, non-professional users utilize MT systems on their own computers. We call this the "personalization" of the MT system. In such a situation, the following features are required of the MT system besides the accuracy of translation, which is the most important:

* Small-size powerful computer

* Full support of translation (OCR,DTP)

* Easy customization (grammars and dictionaries)

« Easy operation

This paper mainly describes the extensions related to the personalization of ASTRANSAC, which have been newly made since the MT Summit II.

## 2 Outline of ASTRANSAC

### 2.1 Translation Method

Translation part of ASTRANSAC is based on the transfer method. All components except for the morphological analyzer and the morphological generator are language-independent; the same software is used for both English-to-Japanese and Japanese-to-English translation. The subsequent sections expound the translation process.

#### 2.1.1 Dictionaries and Morphological Analysis

ASTRANSAC has three kinds of dictionaries:

(1) Common word dictionary (50,000 words)

(2) Technical term dictionary (maximum of 200,000 technical terms)

(3) User-defined word dictionary (maximum of 200,000 words for one field)

The entry words and their target equivalents in the user-defined dictionary have the highest priority in translation. The morphological analyzer constructs word lattice for input sentences and produces a sequence of word structures, which in turn will be the input of the syntactic analyzer. The Japanese morphological analyzer utilizes word connection information to eliminate morphologically implausible word sequences.

#### 2.1.2 Syntactic Analysis

Syntactic analysis and semantic analysis do not function sequentially, but proceed interactively. Their roles can be clearly divided as a module. The syntactic analyzer derives only one syntactic structure for a string of categories in a sentence. Lexical ambiguities are resolved in the normal manner of syntactic parsing, that is, by eliminating category values that do not permit coherent word category combinations. Structural

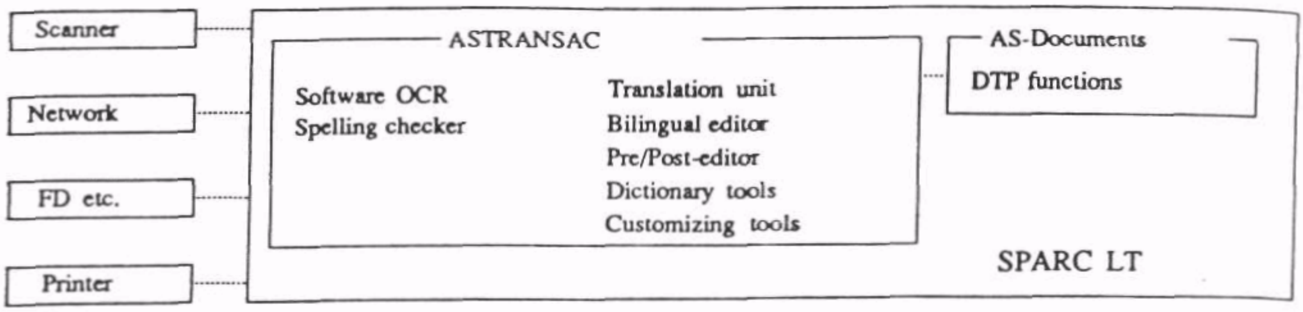


Fig. 1 System Configuration

ambiguities are implicitly represented in the syntactic structure. Semantic analysis will enable constructing a plausible conceptual structure, resolving such implicit ambiguities. The syntactic analyzer employs an ATN-like fashion.

### 2.1.3 Semantic Analysis and Transfer

The semantic analyzer including the transfer processor constructs a semantic interpretation and conceptual structures for the target language. The proposed semantic analysis method is lexical-based. Each lexicon has a semantic interpretation and transfer rules. In the semantic analysis process, structural ambiguities (mentioned in 2.1.2) and semantic ambiguities are disambiguated by preference scoring and optimum solution searching. Semantic ambiguities of this kind are treated in the semantic analyzer. To illustrate, the Japanese particle "は" can mark two or more semantic roles, such as agent, object, location etc. The following two sentences have the same syntactic structure but have very different semantic structures.

(1) 私 (I) は買う。 (I buy 0-PRONOUN)
(2) 本 (book) は買う。 (0-PRONOUN buy a book)

The particle "は" marks the agent role in (1), the object role in (2). In the semantic analysis phase of ASTRANSAC, the noun phrase followed by the particle "は" (i.e. "私 (I)" and "本 (book)") has multiple semantic roles in relation to the predicate "買う (buy)". Preference scoring by semantic interpretation of the noun-verb relation gives a different semantic structures for (1) and (2).

This approach is "fail-safe" regardless of semantic information; semantic information is not used for restricting interpretation of a sentence. Also, it excels in efficiency since syntactic rules are independent of semantic rules.

Transfer rules are written in a tree-to-tree conversion format. As mentioned above, each lexicon has its transfer rules for determining its semantic interpretation, i.e. its target word, target linguistic structure and linguistic features. The output of the lexical transfer is transformed by the structural transfer part. Structural transfer rules have the same format as lexical rules. Basically, Structural transfer realizes a contrastive grammar

### 2.1.4 Syntactic Generation

The main role of syntactic generation is to determine the word order in a target language structure (tree) This is performed by traversing the tree structure according to syntactic generation rules written in an extended context-free grammar form. These rules can access various linguistic features in the tree structure, such as the weight of a substructure.

## 2.2 System Configuration

The entire software is written in C and runs under UNIX on AS1000 (laptop engineering workstation SPARC LT), AS3000 and AS4000 series. The translation speed is 10,000 - 15,000 words/hour both for EJ/JE versions when SPARC LT (13.2 MIPS) is used. The speed depends on the complexity of the source text.

The total system configuration is shown in Fig. 1. Details of software OCR "ASREADER" and DTP software "AS-Documents" are given in Section 3. ASTRANSAC is equipped with three editors (pre-editor, bilingual editor and post-editor), the dictionary management tools and the customizing tools. The translation unit and the bilingual editor run in parallel.

# 3 Extensions toward Personalization

## 3.1 Built-in Software OCR

ASTRANSAC is equipped with a built-in software OCR (ASREADER) for English documents. This tool requires no special hardware for recognizing English characters. By connecting the image scanner to the workstation, users can input English documents into ASTRANSAC. This OCR is rather strongly connected with the translation part so that users may obtain the printed output from an original document by a few operations This is convenient for getting rough translation quickly. The specification of this OCR is as follows:

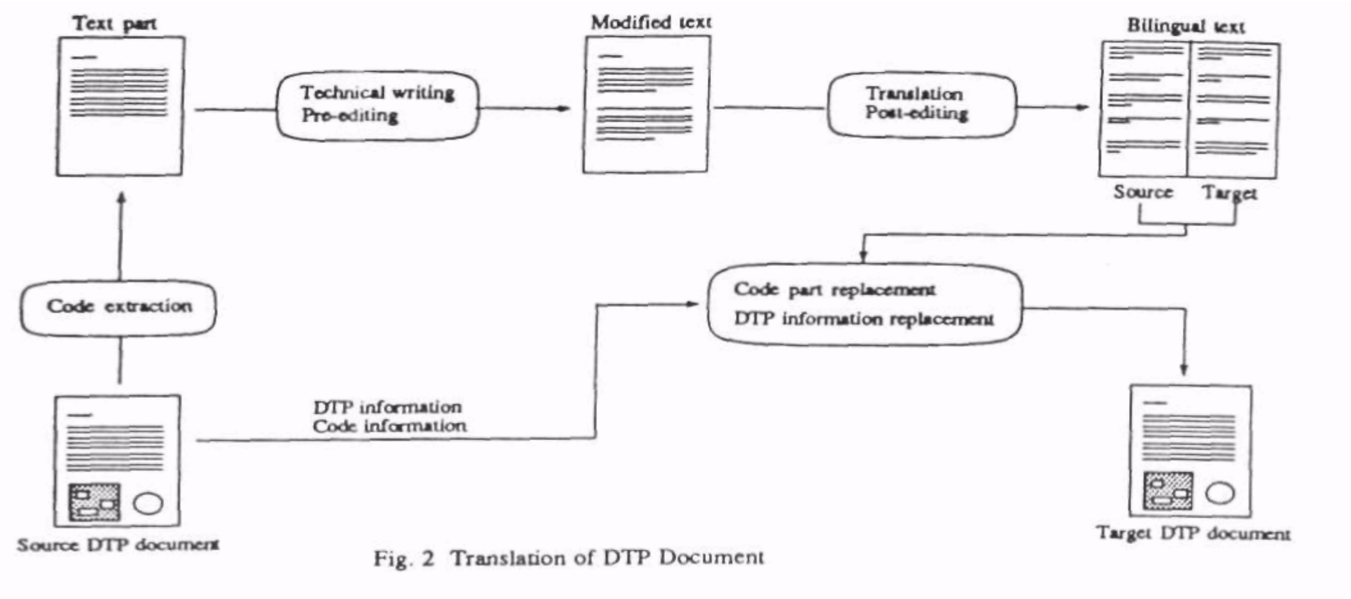


Fig. 2 Translation of DTP Document

Recognition speed: 50-60 characters/sec (SPARC LT)
Recognition rate : 99.7% for printed documents
Fonts : Omnifonts

The features of this tool are: automatic document structure recognition and picture exclusion, word verification and spelling check.

(1) Automatic document structure recognition and picture exclusion
Usually, input texts include complicated forms, such as tables, figures, pictures, headers and footnotes, and also employ a multi-column format. This OCR automatically recognizes a document structure, extracts the text parts, and decides the text flow of a document. Text flow can be interactively edited by the user.

(2) Word verification and spelling check
Words with low recognition certainties and words not in the MT dictionaries (common, technical, user-defined) are searched and displayed on the screen with their scanning images. Then, the user can check and correct OCR errors easily by using this editing function.

### 3.2 DTP Document Processing

DTP software is coming to be used personally in recent years. ASTRANSAC has provided the text level interface to DTP software AS-Documents, However, this interface treats only code information. We have enhanced ASTRANSAC (the Japanese version of FrameMaker) to treat MIF (Maker Interface Format) level interface for maintaining DTP information (font, format etc.) through translation[Itoh *et al.* 1991].

There are several ways to treat DTP information in MT system. In designing this function, we have set up the following specifications:

- Inter-edit proof
- No DTP information manipulation during translation

Inter-edit here means the editions by the operator through translation process such as pre-edit and post-edit. Input of the translation part may differ from the original DTP text because of some modifications in pre-editing or other stages. In Japanese-to-English translation especially, technical writing of the original text is often applied to improve the total efficiency of translation. For example, an unnecessarily long compound sentence consisting of simple sentences connected by meaningless conjunctions or connective verbal forms ("REN-YOU CHUUSHI HOU") is divided into shorter and clear sentences by the pre-editor. Even in this case, format information should be correctly reflected on the output of the MT system. Here, we call this feature "inter-edit proof." DTP interface has to naturally deal with such a kind of inter-edit processing,

Also, users want to be free from the DTP information during translation. Therefore, it should be neglected or hidden from them when they concentrate on translation, or edition, using the bidirectional editor.

Our DTP interface meets these requirements by adopting the following procedure (Fig. 2):

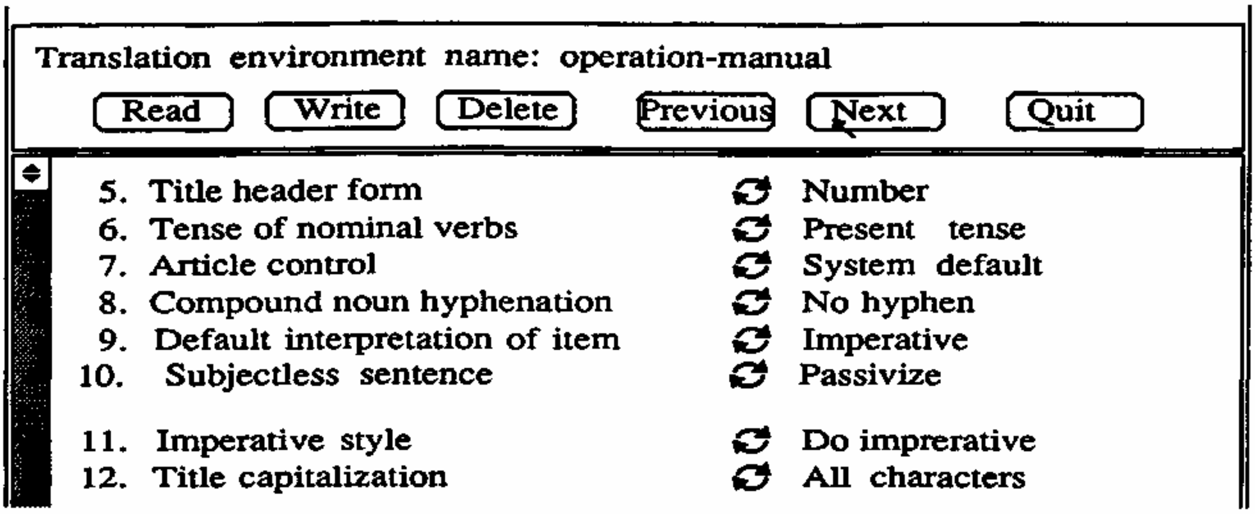
(1) Extract code information from the DTP file

(2) Carry out translation (including pre-editing and post-editing)

(3) Replace the codes of the DTP file with their target counterparts

(4) Replace necessary information in the DTP file

The first step creates a text file by extracting code information (text information) from the original



Translation environment name: operation-manual

5. Title header form — Number
6. Tense of nominal verbs — Present tense
7. Article control — System default
8. Compound noun hyphenation — No hyphen
9. Default interpretation of item — Imperative
10. Subjectless sentence — Passivize
11. Imperative style — Do impreratíve
12. Title capitalization — All characters

Fig. 3 Example of Customizing Parameter Set-up Screen

DTP file. This text file is pre-edited, translated and post-edited the using source text editor and the bilingual editor equipped with ASTRANSAC. The output of the second step is a bilingual data file (collection of source/target sentence pairs). In the third and fourth steps, a target DTP document is created from the original DTP document and the bilingual data file. The processing in the third step is not straightforward due to the modification of the source text in inter-editing. The part of the DTP file, where a target sentence should be replaced, is searched by comparing the source sentences in the DTP file and the source sentences which have been modified using technical writing methods. The translation of the latter will be replaced by the corresponding translation in the DTP file. In this search, completely matched sentences (those to which no inter-edition is applied) are used as a key to identifying the positions of unmatched sentences. The fourth step converts font-type information in the DTP file. For example, English font "TimesRoman12" is converted into Japanese "MINCHOU10".

There remain some future problems to be solved for the treatment of DTP documents. One is the automatic adjustment of size and positioning of a document format. It is necessary to prevent an overflow of characters caused by the difference in total length between the original text and the target text.

## 3.3 Customizing Functions

Customization of the MT system is one of the most crucial factors in its personalization. There are many types of customizing requirements, such as those in linguistic aspects and man-machine interface aspects. This section describes three of the customizing functions of ASTRANSAC.

### 3.3.1 Customization of Translation Part

Customization of translation part concerns linguistic, or translation-related, features of the MT system. For example, analysis grammar, transfer rules and generation grammar are the targets of the customization. In principle, these general knowledges, including the common word dictionary, are not open to users in ASTRANSAC. One reason is that it is almost impossible for end users to improve such knowledge sources; what they want is not the presentation of the knowledge sources, but the improvement in translation quality. Another reason is that we guarantee compatibility of the system against version-ups of ASTRANSAC. If general rules are open to end users, MT system suppliers cannot make a version-up of the MT system without effecting the modifications made by users.

In consideration of the above, ASTRANSAC adopts parametrization of linguistic feature controls for customizing general rules. In this approach, all linguistic rules can be in principle controlled by the user's parameter selection. A set of linguistic parameters are provided by MT system suppliers. The following gives some examples of linguistic parameters:

- English to Japanese

  (P1) Default translation for participial constructions:
  して (chronological sequence) |
  するので | (reason) | しながら (same time)

  (P2) "You" omission:
  yes | no

  (P3) Japanese sentence style:
  polite | normal

- Japanese to English

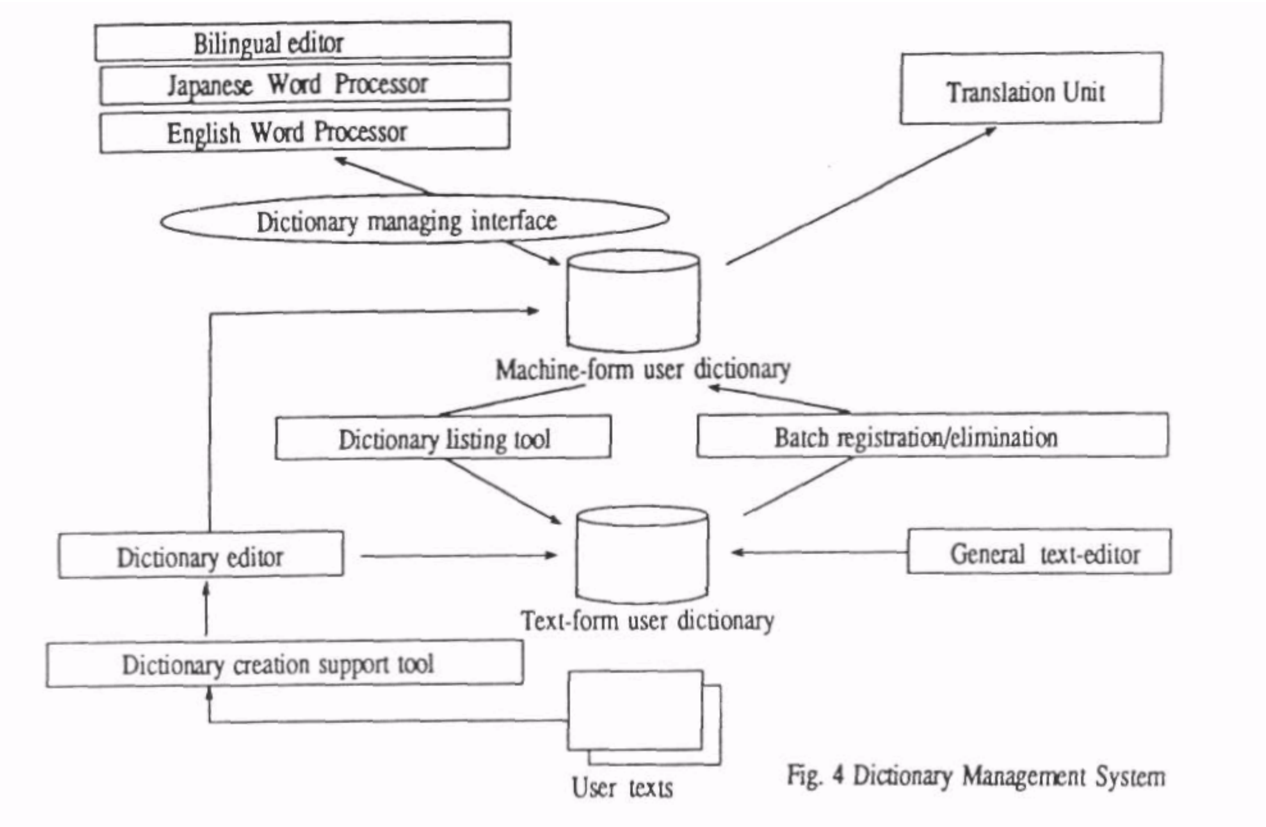


Fig. 4 Dictionary Management System

(P4) Default translation for subjectless sentences:
passive | personal pronouns | “it” |
imperative | user-defined string

(P5) Default articles;
no articles | determiners | “the” |
system default

(P6) Default tense for verbal nouns:
present | past

(P1) specifies the default translation for participial constructions which are rather difficult to automatically analyze without world knowledge and domain-specific knowledge. (P2) specifies whether the word “you” is omitted in the Japanese sentence generation. In translation of manual documents, “you” usually is omitted. (P3) specifies “normal/polite” style in the Japanese morphological generation. Polite style is suitable for letters. (P4) treats subjectless sentences, which are very common phenomena in Japanese, Although Japanese has imperative expressions, declarative sentences are often used as their substitutes. Using the parameter (P4), users can select the following expressions,

“キーを押します。”
(= 0-PRONOUN push the key)

=> The key is pushed. (passive)
=> I push the key. (personal pronouns)
=> It pushes the key. (“it”)
=> Push the key. (imperative)
=> # pushes the key. (user defined string — “#”)

(P5) specifies the treatment of articles, which are usually undecidable from Japanese source sentences. In translating operation manuals, default determiners should be "the" because they describe the operation method for a certain real tool.

Most of linguistic parameters are designed to determine the default value of linguistic choices. Therefore, if the MT system can accurately predict the appropriate choice, the default preference is ignored. For example, even if the user specifies that an English sentence should be translated in the passive voice, this rule is not applied when not appropriate as below:

“The remote files cannot be removed.”

⇒?? リモートファイルは削除されることができない。
(Passive form)

⇒ リモートファイルは削除することができない。
(Active form)

Currently, these linguistic parameters are mostly for the customization of transfer rules and generation rules. We plan to develop more for analysis grammar. Part of sub-grammar concepts is realized through the parametrization of analysis grammar rules.

Fig. 3 shows a screen display of customizing parameter set-up. The user can easily choose parameters by mouse operation. A set of parameter settings are named and stored in the system’s catalog. Therefore, end-users usually select only the cataloged names such as “operation manual” and “trouble report”.

### 3.3.2 Customization by User Dictionary Development

Development of user dictionaries is the basic method for improving the quality of the MT system. ASTRANSAC provides several functions for helping users develop their own user dictionaries. There are two formats for the user dictionary. One is a machine accessible form, and the other is a text form which can be edited by a general text editor (vi etc.) or a Japanese word processor equipped in ASTRANSAC. The followings are the dictionary management tools:

(1) Batch-mode dictionary registration/elimination tool
This tool registers/eliminates words in a text-form user dictionary from a machine-form user dictionary in batch mode operation,

(2) Dictionary listing tool
This tool creates a text-form user dictionary from a machine-form user dictionary in batch mode operation.

(3) Dictionary editor
This tool provides an interactive dictionary maintenance environment.

(4) Dictionary creation support tool
This tool scans user texts and creates dictionary entry candidates with their target words if possible. User can edit these candidate using the dictionary editor and can create a user dictionary before translation of the texts.

The configuration of the dictionary management system is shown in Fig. 4.

### 3.3.3 Customization using User Texts

In principle, analogy-based MT approach utilizes the information in the (user) texts through the whole translation process [Sato and Nagao, 1990] [Sadler, 1989). We consider this concept is very important also for personalization of MT systems since this approach inherently produces the customized translation output. In this sense, user texts can be sources for improving the quality of a MT system. This section shows two examples of utilizing user text information introduced to the experimental version of ASTRANSAC.

One of the functions utilizing the user texts is related to the proper target word selection. In the traditional target word learning method, the operator selects the appropriate target word from the candidates provided by the MT system editor. After this operation, the selected word has the highest priority. This method has the following drawbacks:

- The operator has to carry out operations word by word appeared in the source text.
- These kinds of operations are likely to be neglected by users who are not professional translators, or who require only rough translations.

From the viewpoint of translation factory, these disadvantages may not be so serious since the system manager prepares the high quality user-defined dictionaries However, in view of personalization, these pose grave problems.

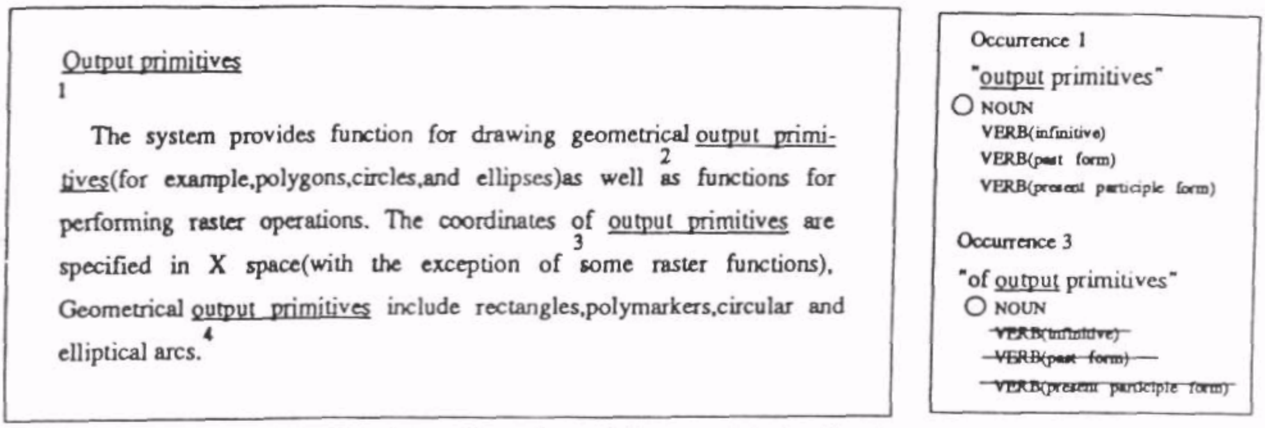
The experimental version of ASTRANSAC provides the automatic target word selection function based on the user's texts. The user prepares a source text and a text written in the target language which is in the same field with the source text. Henceforth, we shall call the latter the reference text. Word selections in translation are controlled by the vocabulary in this reference text. For example, the word "terminal" has several target word equivalents as:

terminal: ターミナル, 端末,
端末装置, (=> computer equipment)
端子, (=> electrical parts)
終着駅 (=> station)

If the given target language text contains the word " 端末 ", but not "ターミナル", "端末装置", "端子" and 終着駅 ", then the system chooses "端末"as the first candidate for the source word "terminal". The first three have the same meaning, i.e. computer terminal. Selecting the correct target word from these three equivalents is important in industrial translation where terms should be properly selected according to the user's convention. On the other hand, selection from words with different meanings, for example " 端末 " (computer equipment)," 端子 " (electrical parts), and "終着駅" (station), is crucial for both industrial translation and rough translation.

In our experiment, approximately 60% of nominal terms in a source text has more than one occurrence of their target words in a reference text in case this reference text is extracted from the same manual as the source text. It has been shown that around 85-90% of these words are unambiguously selectable by the above strategy [Nogami *et al.*, 1991], The vocabulary control information can be stored and utilized in the succeeding translation.

The source text is another source for improving the quality of a translation system. Most of MT systems currently carry out translation sentence by sentence; the scope of analysis, transfer, and generation is restricted to intrasentence. Recently, utilization of the information outside the sentence, i.e. the information in the whole document, has been proposed [Inagaki *et al.,* 1990][Tanaka *et al.,* 1990]. Thus, ambiguities in a sentence, which cannot be disambiguated using the MT system's knowledge and the information obtained from the sentence itself, may be disambiguated. Based on this concept, we have developed the experimental version of ASTRANSAC.



Output primitives
1

The system provides function for drawing geometrical output primitives(for example,polygons,circles,and ellipses)as well as functions for performing raster operations. The coordinates of output primitives are specified in X space(with the exception of some raster functions). Geometrical output primitives include rectangles,polymarkers,circular and elliptical arcs.

Occurrence 1
"output primitives"
○ NOUN
VERB(infinitive)
VERB(past form)
VERB(present participle form)

Occurrence 3
"of output primitives"
○ NOUN
~~VERB(infinitive)~~
~~VERB(past form)~~
~~VERB(present participle form)~~

Fig.5 Disambiguation of Categorial Ambiguity

The following are some examples of the ambiguities which are the target of disambiguation by text information:

- Categorial ambiguity
- Word boundary ambiguity
- Syntactic ambiguity

In English, a word usually has multiple categories, In parsing English sentences, resolving this categorial ambiguity is a serious, hard problem, because the incorrect choice of the category for one word greatly decreases the accuracy of overall translation. For example, in "output primitives", "output" has four categorial (including morphological interpretation) ambiguities: noun, verb (infinitive, past form, present participle form),

The second ambiguity, peculiar to agglutinative languages like Japanese, appears when a morphologically and syntactically plausible word boundary exists. For example, the kanji sequence "今日本人" can be analyzed either as "今 (now)／日本人 (Japanese)" or "今日 (today) ／本人 (he or she)". Disambiguating this kind of ambiguity often requires semantic, pragmatic and domain-specific knowledge sources.

The third ambiguity includes PR-attachment problems and conjunction-scope problems. For example, the following sentence has PP-attachment ambiguity on the phrase "on rectangles".

"The same header file defines some interesting macros on rectangles."

=> The same header file defines [some interesting macros [on rectangles]],

=> The same header file defines [some interesting macros][on rectangles].

The following sentence has conjunction-scope ambiguity.

"Painting panels and individual items"

=> [Painting panels] and [individual items]

=> Painting [panels and individual items]

These ambiguities can be resolved highly accurately if proper expressions are found in other parts of the text. Fig. 5 shows examples of the disambiguation of categorial ambiguity- Here, the first expression "Output primitives" is, as described above, categorially ambiguous. However, the second and third occurrences of "output primitives" strongly suggest that the part-of-speech of "output" is the noun modifying "primitives".

Fig. 6 shows the translation flow for utilizing source text information. We call this the two-path method since the original text is scanned twice. In the first path, the translation system only extracts information from sentences and stores them in a database. In this processing, only highly accurate information (positive and negative facts) are collected.

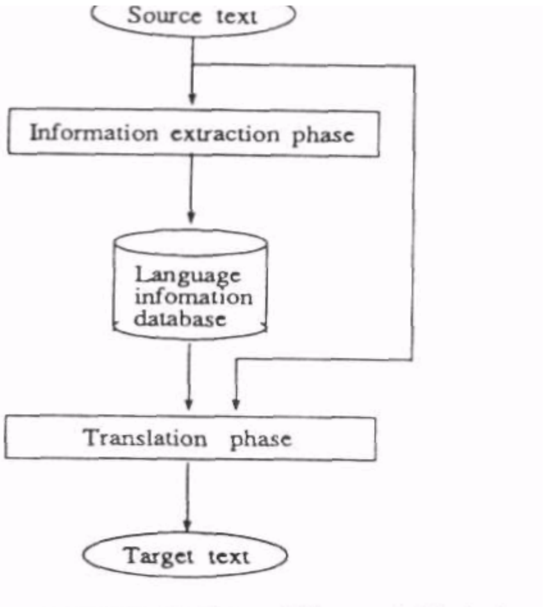


Fig. 6 Translation Flow of Two-path Method

For example, “output primitives” is analyzed as a compound noun phrase because the expression “of output primitives are” in the text. In the second path, actual translation is carried out utilizing the information in the database. In this example, “noun” is first selected for the part of speech of “output” in the categorial disambiguation phase.

This kind of processing is merely heuristic, but surely improves the accuracy and coherency of translation without user operations.

## 4 Conclusion

A brief explanation of the personalization of AS-TRANSAC has been presented. The growth of hardware resources will encourage the personalization of the MT system, where customizing techniques, including linguistic and HI features, are indispensable.

## References

[Amano *et al*., 1989] Amano, S., Hirakawa, H. and H. Nogami. The TAURAS Design Philosophy. *Proceedings of the MT SUMMIT II*, 1989.

[Inagaki *et al.,* 1990] Inagaki, H., Miyahara. S., Nakagawa, T. and F. Obashi. Disambiguation by Document-Oriented Preference Sets. *Proceedings of COLING-90,* 1990.

[Itoh *et al.,* 1991] Itoh, E., Takeda, K., Hirakawa, H. and S. Amano. Machine Translation System which Saves the Format of DTP-Documents (in Japanese). *Information Processing Society of Japan*, 3-37, 1991.

(Nogami *et al.,* 1991] Nogami, H., Kumano, A., Tanaka, K. and S. Amano. Learning of Translation Words Using Target-language Documents (in Japanese). *Information Processing Society of Japan,* 3-29, 1991.

[Sadler, 1989] V. Sadler. *Working with Analogical Semantics.* Dordrecht: Foris Publications, 1989.

[Sato and Nagao, 1990] Sato, S. and M. Nagao. Memory-based Translation. *Proceedings of COLING-90,* 1990.

[Tanaka *et al.,* 1990] Tanaka, K., Nogami, H., Hirakawa, H. and S. Amano, Machine Translation System Using Information Retrieved from the Whole Document (in Japanese). *Information Processing Society of Japan* pp. 405-406, 1990