[*Translating and the Computer 8*. Proceedings of a conference… 13-14 November 1986, ed. Catriona Picken (London: Aslib, 1987)]

Japanese machine translation in Japan and the rest of the world

Peter J. Whitelock

Department of Artificial Intelligence
University of Edinburgh

In this paper I want to discuss the current state of machine translation (MT) between Japanese and other languages, usually English. Not surprisingly, most effort in this field takes place in Japan itself, and the first section of the paper describes the results already available on the market in Japan. In the second section, various relevant research programmes currently underway in Japan are mentioned. Finally, work on Japanese MT in the rest of the world is discussed.

MT IN THE JAPANESE MARKET PLACE

The General Research Laboratories of the Japan Efficiency Society predicted at the start of 1986 that the market for MT systems in Japan that year would be in the range 35,000 to 40,000 million Yen, rising to 250,000 million Yen (about £1,000 million) by 1990.

With a translation industry worth, according to the Japanese Electronic Industry Development Association, 1 trillion Yen (c. £4,000m) in 1986, it is hardly surprising that such a demand for automated assistance exists. What is remarkable, especially compared with MT in the rest of the world, is the number of companies able to meet that demand.

To understand the diversity of the Japanese MT industry, we must look in the first instance not at translation *per se,* but at the distinctive characteristics of the Japanese language. It is the orthography of Japanese that has stimulated the development of computer tools for the monolingual Japanese writer, and this in turn has provided the linguistic and ergonomic foundations for MT.

Japanese is written in characters taken from four different systems. One of these is the Roman set, which is of minor importance. There are two native syllabaries—katakana and hiragana—both of which have characters for the basic syllables (vowel, or unvoiced consonant plus vowel) of the language, and thus number about 50 characters each. With diacritics, and in combination, several times this number of syllables are representable. These characters are in near one-to-one correspondence with the sounds of the language.

Finally, and most distinctively, there is a set of Chinese ideographic characters or kanji. This numbers over 10,000, of which 3,000 or so are in common use. These characters are in a many-to-many relationship with the sounds of the language.

The katakana syllabary is used for the transliteration of words borrowed from Western languages, such as Portuguese, English and German. It is hiragana and kanji that constitute the bulk of Japanese texts. While any word can be written in hiragana, and many common words are, contentful words (open-class or lexical words), such as nouns, verbs and adjectives, can be, and usually are, written using kanji. Hiragana is the character set used for closed-class or grammatical items, such as inflections, postpositions and auxiliary verbs. Grammatical items follow the lexical ones which they modify, and are obligatory in almost all situations. Thus the basic structure of a Japanese text is a series of units—called bunsetsu—each comprising a lexical item (one or more kanji) followed by a series of hiragana characters representing grammatical information of various sorts.

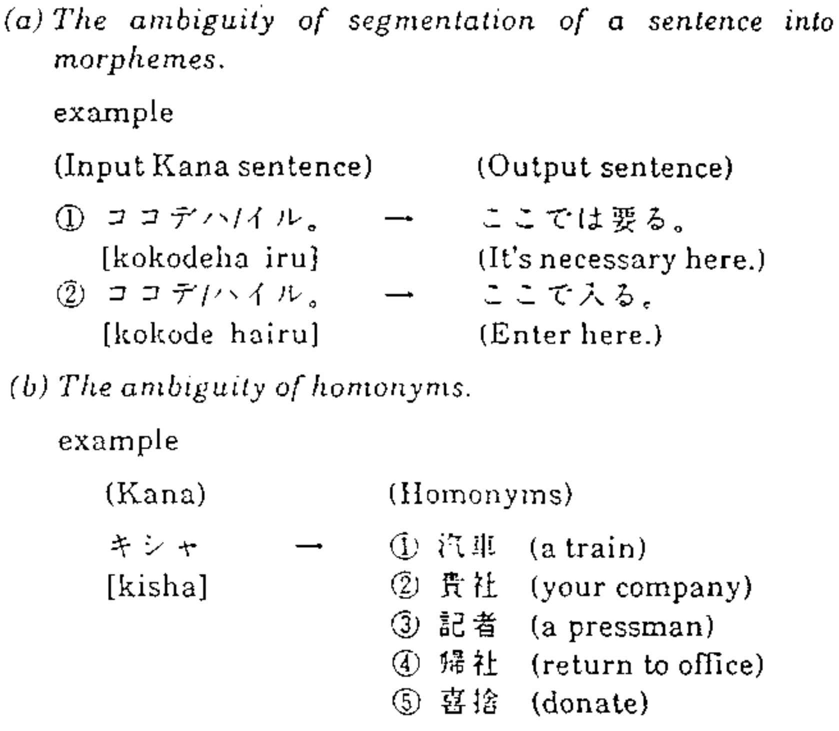
To a native Japanese reader, therefore, the use of the different character sets provides essential information. The change from one character set to another segments the text, giving it a higher level structure, whilst the use of ideographic characters means that many words which are homophones are not homographs (like ‘pain’ and ‘pane’ in English). These information-carrying aspects of the orthographic system more than compensate for the second salient feature of Japanese orthography, the fact that word boundaries are not indicated and punctuation is used only sparingly. Children's books, written entirely in hiragana with spaces between bunsetsu, typically take much longer for an adult reader to comprehend.

The basic problem of Japanese text processing is how to enable a writer to enter any of a very large set of very complicated characters. The original Japanese typewriter provides one key per character, selected by the typist using a mechanical cursor. Without a knowledge of the layout and structure of the character set, such a method is impossibly slow. Even for a skilled user, speeds equivalent to those of a Western typist are never achievable.

Recently, however, word processors made by almost every typewriter company have appeared on the market, some for under £100. Most of these work on the principle of kana-to-kanji conversion. A typist enters characters from one of the non-ideographic sets, corresponding to the way the text would be spoken. Then, with varying degrees of assistance from the typist, these are converted to the appropriate sequences of kanji and kana. There are two aspects to this conversion process. First, the bunsetsu structure of the text must be recovered. Secondly, the word part of each bunsetsu must be converted to kanji. Ambiguities will arise at both of these stages. This is illustrated in Figure 1, taken from Abe *et al.* (1986), a lucid discussion of the use of grammatical, lexical and semantic information to improve the automation of the kana-kanji conversion process.

Within the current state of the art, some form of human assistance is essential for 100 per cent, accuracy, particularly with regard to choosing the correct kanji from a set of possibilities. Nevertheless, the more linguistic knowledge the system can deploy, the more this choice can be constrained, and the typing process facilitated.

The basic information that the machine requires is the set of kanji, each associated with all of its readings. There will usually be two—a native Japanese 'kun' reading, and a Chinese 'on' reading, or more. In the simplest case, the user can indicate precisely which sequences of kana are to become a kanji, and choose from the set of kanji that may have that reading. Unfortunately, there may be several tens of kanji having the same reading,



(a) The ambiguity of segmentation of a sentence into morphemes.

example

(Input Kana sentence) (Output sentence)

① ココデハ/イル。 → ここでは要る。
[kokodeha iru] (It's necessary here.)
② ココデ/ハイル。 → ここで入る。
[kokode hairu] (Enter here.)

(b) The ambiguity of homonyms.

example

(Kana)		(Homonyms)	
キシャ	→	① 汽車	(a train)
[kisha]		② 貴社	(your company)
		③ 記者	(a pressman)
		④ 帰社	(return to office)
		⑤ 喜捨	(donate)

Figure 1. Ambiguity types in kana-kanji conversion (Source: Abe *et al.*, 1986)

particularly if it is an ‘on’ reading. Characters in compound words usually have their 'on' readings, so including a dictionary that shows the possible combinations of kanji into words allows many of these possibilities to be ruled out. If in addition the syntactic categories of kanji and kanji sequences are specified, and the machine knows which affixes form bunsetsu with words of each category, further possibilities are eliminated.

Thus the development of ergonomically satisfactory word-processor technology has forced Japanese manufacturers to confront linguistic issues that have provided an essential basis for MT systems design. From a monolingual dictionary of several tens of thousands of words and their syntactic categories, it is but a short step to adding translation equivalents. And the sort of grammatical information needed to determine valid bunsetsu is an essential part of any Japanese analysis or generation for MT. In generation, it is needed to produce correct Japanese output from the grammatical features computed from an English text. In analysis, it provides a structural basis for fitting case frames to a text, a technique which many MT systems use to resolve ambiguities in the structure of Japanese texts.

So the construction of MT systems working both out of, and into, Japanese has benefited from the technology of kana-to-kanji conversion.

The systems currently available on the Japanese market are summarised in Table 1, taken from Taguchi (1986).

The first seven systems were already on the market at the time this article was written, whilst the other four are given with their projected date of release. Many others are due for release in the near future.

There is little variation in the size of the basic dictionary of general words that the different systems provide, as could be expected from the requirements of kana-to-kanji conversion described above.

The major differences are in the size of machine the systems are intended to run on. The first six systems are all installed on large mainframe computers, have quoted translation speeds in the region of tens of thousands of words per hour (compare human translation at 300wph), and have large term banks as standard. All except the Toshiba system rent at around 500,000 Yen per month, and the Toshiba sells at a flat rate of 6 million Yen. The two Atlas systems of Fujitsu are based on totally different designs, so the Nippon Denki system is currently the only package which integrates systems for translating in both directions. The next two systems are those of the market leader Bravice, who own Weidner, the well-known Western MT company. Their pc-based MicroPack appears unique in the market and has already sold 700 units. The other systems, scheduled for release shortly after the article appeared, are intended for minis, as is the Bravice MediumPack.

One other system that should be mentioned is the Japanese to English version of Systran. This is currently undergoing testing by the Commission for the European Communities.

		Dir.	WPH	Dict.	Terms	On Sale	Cost	Sold
Company	System	(1)	(2)	(3)	(4)		(5)	(6)
Fujitsu	Atlas 1	E-J	60k	50k	250k	Sep 84	350k	80
Fujitsu	Atlas 2	J-E	60k	50k	250k	Jun 85	550k	40
Nippon Denki	Pivot 1	E-J	100k	50k	100k	Sep 86	500k	
Nippon Denki	Pivot 1	J-E	100k	50k	100k	Apr 86	490k	
Hitachi	Hicats	J-E	20k	50k	250k	May 86	550k	
Toshiba	Tauras	E-J	50k	30k	50k	Dec 85	6mfr	
Bravice	Medium	J-E	5k	60k	8k	Jun 84	2mfr +	30
Bravice	Micro	J-E	1k	60k	—	Aug 85	6mfr	700
Sharp	Duet	E-J	5k	50k	20k	Aug 86	lmfr	
Mitsubishi	Thalia	J-E	20k	60k	—	Nov 86	—	
Oki	Rosetta	J-E	4k	50k	20k	Jul 86	—	
Resource Sharing	Star	E-J	4k	15k	16k	Apr 86	—	

Key:
(1) Direction of translation — J-E: Japanese to English
E-J: English to Japanese
(2) Translating speed: words per hour
(3) Basic dictionary size: number of entries
(4) Specialised technical vocabulary: number of terms
(5) Cost k: thousands of Yen rental per month
mfr: millions of Yen fixed rate
(6) Number sold.

Source:
Jun Taguchi, 'MT systems go into business', *NIKKEI Computer*, 4 April, 1986, translated by Mary Gillender

Table 1. MT systems between Japanese and English (on market/ scheduled)

The quality of output of all these systems is roughly equivalent to that of systems commercially available in the west. Thus they should definitely be viewed as pre-translators, or translator's aids, designed to increase throughput. Many manufacturers are pursuing the development of more sophisticated designs in an attempt to improve translation quality. Also, there are several current projects designed to form the basis of long-term improvements in natural language processing techniques and capabilities, involving collaboration between industry, government and academia. Some of these are mentioned in the next section.

RESEARCH PROJECTS IN MT AND RELATED AREAS

The principal MT research project in Japan is the Japanese National Project currently being carried out at four laboratories. At Kyoto University, under the directorship of Professor Nagao, overall leader of the National Project, a sophisticated high-level language for writing all stages of an MT system has been developed. This language, GRADE, has a similar design philosophy to the ROBRA language used in the GETA project at Grenoble (see Boitet, p. 114, above, and Nagao *et al.*, 1985). Other centres involved are the Electrotechnical Laboratories, who are responsible for morphological analysis and synthesis, text input and output, and dictionaries of verbs and adjectives; the Japan Information Centre for Science and Technology (JICST), responsible for the noun dictionaries and technical terms; and the Research Information Processing System of the Agency of Engineering Technology, responsible for overall integration of software and linguistic knowledge bases. The project, which aims to supplement the activities of JICST by automating the translation of scientific abstracts, started in 1982, and a prototype was demonstrated in March 1986. Development of the Japanese-English version will continue until March 1989, and work on an English-Japanese version is scheduled for three years from April 1987.

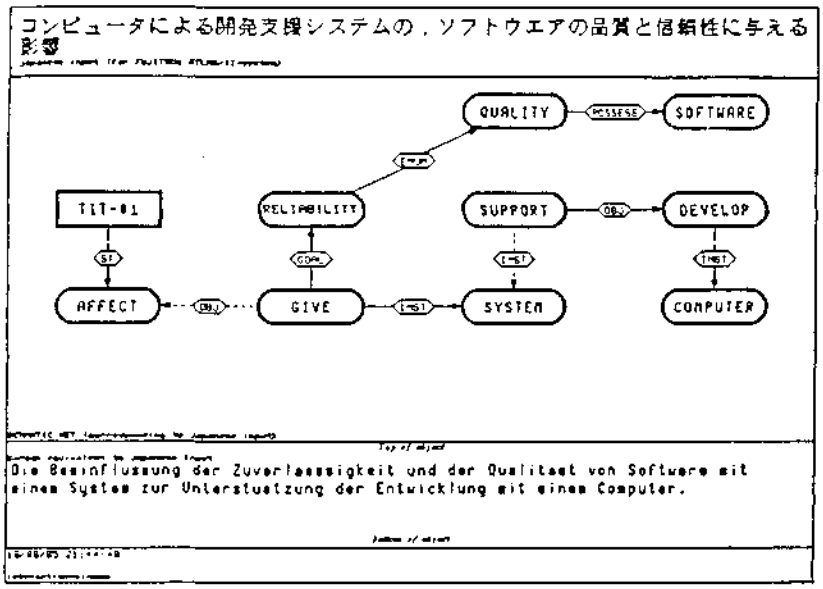
A second important research project recognises the primary importance of the dictionary in MT and other applications of natural language processing. Despite the fact that many companies have already independently developed dictionaries for their own MT systems, the Government and eight companies (Fujitsu, NEC, Hitachi, Toshiba, Matsushita, Mitsubishi, Oki and Sharp) are providing 50 per cent. funding each towards a recently started nine year project to build a much larger and more comprehensive dictionary. It is intended to construct, amongst others, a 200,000 word dictionary of basic language items (English, Japanese, and bilingual), and a dictionary of concepts with 400,000 entries. It appears that the quantity of information in each entry will be much greater than has previously been incorporated into machine-readable dictionaries. The effort to ‘conceptualise’ such a large vocabulary will no doubt have beneficial effects on both MT quality and the practice of lexicography.

The most ambitious research project in natural language is one intended to develop a system for the automatic translation of telephone calls between Japanese and English. A company called ATR (Advanced Telecommunications Research) has been set up with 70 per cent. funding from the Government and 30 per cent. from private companies including NTT, KDD, Hitachi and others. The Interpreting Telephony project will be one of a number to be carried out by ATR, and will cost in the region of 100 billion Yen over a period of 15 years. Even with this sort of timescale and investment, the complexity of this problem is so vast that its literal success is far from being a foregone conclusion. Nevertheless, it provides a framework within which all manner of important research questions in speech and language processing can be addressed. Hopefully, it will have the same stimulating effect on natural language research worldwide as the original Fifth Generation proposals had on research in other aspects of computer technology.

JAPANESE MT IN THE REST OF THE WORLD

Research into Japanese MT in the rest of the world constitutes a fraction of the effort devoted to the matter in Japan, but each of the three projects described in this section addresses questions of relevance to the West or only answerable in the West.

Though certain Japanese companies have started to look seriously at translation between Japanese and Eastern languages such as Malay, Chinese and Korean, there is obviously little linguistic expertise in Japan in Western languages other than English. The SEMSYN project at Stuttgart (Roesner, 1986) is concerned with the development of a German generation module for the Fujitsu Atlas II Japanese-English system. SEMSYN (for semantic synthesis) generates German output from semantic net representations of the titles of papers in Japanese, as shown in Figure 2. Two main knowledge bases are used in this process. One concerns how the concepts that label the nodes of the net may be realised as German words. The other defines the realisation of various patterns that can occur in the net as German syntactic structures. For instance, if a node that represents some action has no agent arc, then the action may appear as the passive form of a verb. Of course, these two types of information interact, since the syntactic category of the word that lexicalises a concept will constrain the realisation



コンピュータによる開発支援システムの，ソフトウエアの品質と信頼性に与える影響

Die Beeinflussung der Zuverlaessigkeit und der Qualitaet von Software mit einem System zur Unterstuetzung der Entwicklung mit einem Computer.

Figure 2. Example translation of the SEMSYN system (Source: Roesner, 1986)

of that concept's environment. When the content and form of a German sentence has been determined, it is passed to another module which performs morphological generation and may also make decisions about the precise order of constituents. The system also embodies a number of heuristics that reconstruct information such as definiteness and number that are not usually explicit in Japanese and thus do not appear in the semantic representation.

Other problems that arise in translating from Japanese to Indo-European languages, such as the widespread omission of personal pronouns, are obviated by the choice of titles as the system's domain.

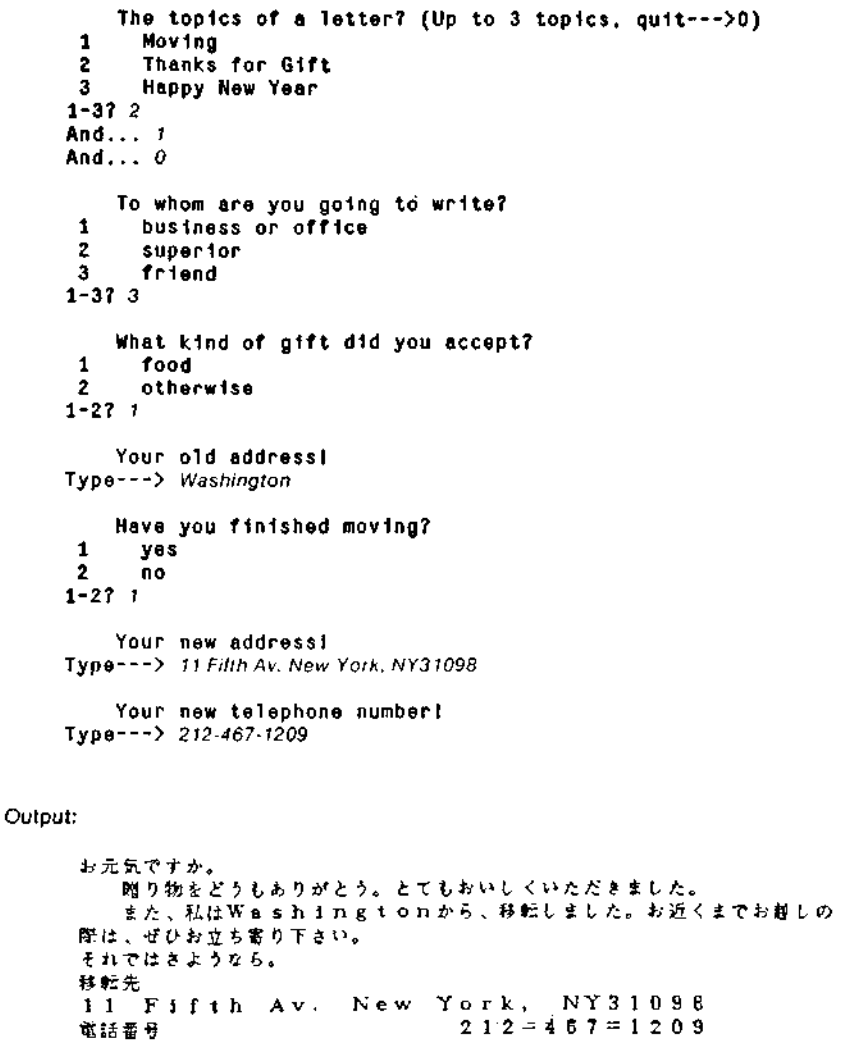
Another interesting line of work is that of Masaru Tomita at Carnegie-Mellon University. Tomita is concerned with the development of what he terms personal MT systems. While most MT systems are designed as translator's aids, a personal MT system is intended for use by a monolingual to translate small documents that would not be worth sending for professional translation. In his thesis Tomita (1985) discusses various aspects of the programming technology required to realise this, such as a fast parsing algorithm that feeds an interactive disambiguation module.

As a small part of his work, Tomita and his colleague Saito have written a program to illustrate an extreme of this particular human-computer interface scenario. The user wishes to compose a letter in Japanese about one of a small number of topics, but knows no Japanese. The machine, on the other hand, knows how to write in Japanese a perfect letter on each of these, given a certain input from the user.

Let us assume that the user wishes to inform a Japanese colleague of a change of address. The machine will ask her a series of questions whose answers fill slots in the stereotypic letter, such as the old address, the new address, the date of the move etc. Without further input from the user, a perfect Japanese letter can be produced. This dialogue and the results are given in Figure 3, from Tomita and Saito (1984).

Of course, what this program is doing is not translation. Nevertheless, it illustrates an important principle that has almost unlimited potential. The machine and the user know different but complementary things, and together they can produce something that neither would be able to alone. As far as the west is concerned, the great lack of expertise in Japanese makes this approach to translation a very attractive one. It forms the basis of the British MT project ‘Read and write Japanese without knowing it’, being carried out at UMIST and Sheffield with funding from the Alvey Directorate and International Computers Ltd.

At Sheffield, a team under George Jelinek are developing a Japanese to English system for an English monolingual. Jelinek's approach is based on a course which has taken place at Sheffield for a number of years, intended to allow librarians, scientists and others to translate texts in their own special fields from Japanese, even though at the start of the course they may



```
    The topics of a letter? (Up to 3 topics, quit--->0)
 1    Moving
 2    Thanks for Gift
 3    Happy New Year
1-3? 2
And... 1
And... 0

    To whom are you going to write?
 1    business or office
 2    superior
 3    friend
1-3? 3

    What kind of gift did you accept?
 1    food
 2    otherwise
1-2? 1

    Your old address!
Type---> Washington

    Have you finished moving?
 1    yes
 2    no
1-2? 1

    Your new address!
Type---> 11 Fifth Av. New York, NY31098

    Your new telephone number!
Type---> 212-467-1209


Output:

    お元気ですか。
        贈り物をどうもありがとう。とてもおいしくいただきました。
        また、私はＷａｓｈｉｎｇｔｏｎから、移転しました。お近くまでお越しの
    際は、ぜひお立ち寄り下さい。
    それではさようなら。
    移転先
    １１　Ｆｉｆｔｈ　Ａｖ．　Ｎｅｗ　Ｙｏｒｋ，　ＮＹ３１０９８
    電話番号　　　　　　　　　　２１２－４６７－１２０９
```

Figure 3. Example session with foreign letter composer (FLC) (Source: Tomita and Saito, 1984)

know no Japanese. At the heart of the course is a ‘Grammar dictionary’, an explicit algorithm for translation of Japanese grammar to quasi-English. In conjunction with a conventional (lexical) dictionary, the user follows the book’s algorithm and produces quasi-English. The user must apply liberal doses of her expert subject knowledge to resolve ambiguities that arise during the process, and of English native speaker knowledge to restructure this quasi-English into fluent, stylistically acceptable English. This particular scenario for the partitioning of knowledge across an MT system is a very convincing modern reinterpretation of Warren Weaver's 'code-cracking' view of translation.

The major problem with the automation of this algorithm occurs at the point of text entry. How does the user who cannot pronounce a kanji enter it? One approach would be optical character recognition, but this is currently not possible, or at least not economically feasible, for arbitrary character sets. Another possible approach could be based on a machine such as that marketed by Matsushita, which will recognise kanji hand drawn on a special pad, making a guess about which is intended and asking the user for confirmation. However, unless a user knows the precise order to draw the strokes making up the kanji, the recognition rate is extremely poor.

To solve this problem, the method used by Nelson's character dictionary (Nelson, 1962) has been elaborated for computer use by Brian Chandler of ICL and UMIST. This system, known as Jelinek—Japanese English Linguistic Interface New Ergonomic Keyboard, requires the user to possess only one technical skill—the ability to count the number of strokes in a character. Even this only needs to be approximate.

The basic organisation of Nelson, like that of Chinese dictionaries, is by radical.* The radical of a character is that part of the character under which it is primarily indexed, so an algorithm for determination of the radical is needed. There is a fixed order in which parts of the character should be compared against the table of 214 radicals. This order is:

(a) the entire character
(b) the left half
(c) the right half
(d) the top half
(e) the bottom half
(f...i) one of the corners
(j) something else

Each radical may determine up to hundreds of kanji. When it has been determined, the number of strokes in the rest of the character are counted. The radical plus number of strokes reduces the search space to a few tens or less, which can then be searched linearly. On computer, any part of a character can play the radical's role of reducing the search space by an order of magnitude. Also the computer can detect the point at which the number of possibilities have been reduced to a point at which they can all be displayed. An example session is shown in Figure 4. Although this may seem a long process, it familiarises the user with the characters, and can gradually be phased into kana-to-kanji conversion as the user learns the

*Japanese dictionaries arc indexed by pronunciation.

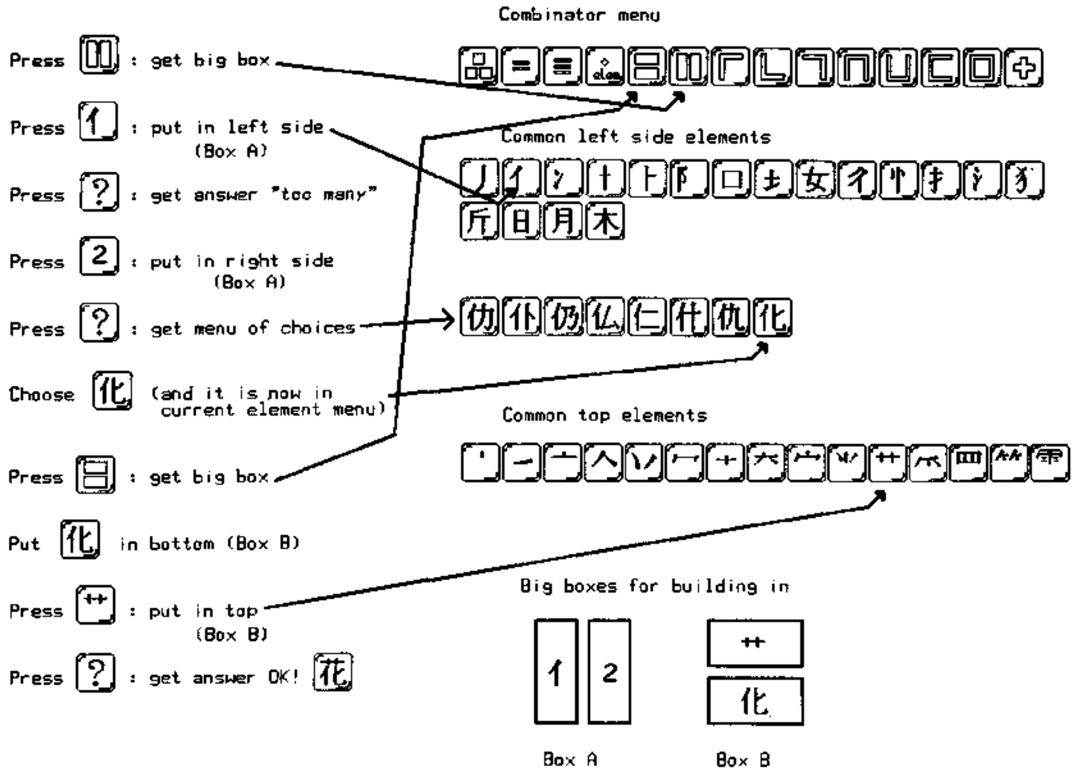

Combinator menu
Press : get big box
Press 亻 : put in left side (Box A)
Common left side elements
Press ? : get answer "too many"
Press 2 : put in right side (Box A)
Press ? : get menu of choices
Choose 化 (and it is now in current element menu)
Common top elements
Press : get big box
Put 化 in bottom (Box B)
Press ++ : put in top (Box B)
Big boxes for building in
Press ? : get answer OK! 花
Box A
Box B

Figure 4. Example session with Japanese – English linguistic interface

readings of the characters she has so painstakingly entered. Any given text can be entered using a combination of the two systems according to the user’s current state of knowledge about kanji and their readings.

The other component of the project, from English into Japanese, is based at UMIST. The intention, like that of Tomita’s personal MT, is to allow the monolingual English speaker to interact with a machine to produce high quality translation. Although in general terms, English into Japanese is an easier direction of translation than the reverse (since a Japanese text does not express many aspects that will need to be inferred for translation into English), this direction is much more difficult from the point of view of use by an English monolingual, since a human will not get ‘the last look’. Thus while a Japanese-English system for an Englishman need have little in the way of an English grammar, an English-Japanese system must possess a generative grammar of Japanese that will produce grammatical text, and which must run without human intervention. However, the human and machine can collaborate on the production of the text representation that constitutes the input to this grammar.

We can recognise two points at which the human can profitably intervene. The first is to resolve the ambiguities of an English text as described by the machine’s grammar of English. These may be genuine ambiguities, or they may be a result of insufficient knowledge on the machine’s part concerning the semantics of words. The forms that interaction could take are described in Whitelock *et al.* (1986). Note that if the ambiguity is genuine, the writer may benefit from having it pointed out. If it is spurious, the user’s responses during interaction can be used to refine the machine’s grammar.

The second point of interaction is during the determination of lexical equivalences. Unlike the previous type of interaction, which was totally source-language specific, this interaction is based on bilingual knowledge. Nevertheless, it is rather easier to organise, since it can be indexed to specific English lexical items, and merely requires the existence of English paraphrases for each of the possible Japanese equivalents.

CONCLUSION

That the Western world should have its own effort in Japanese MT is important from two perspectives. First, there currently exists a massive information flow deficit, in Japan’s favour, between Japan and the West. The design of systems that can be used by English speakers will make important inroads into this deficit. Secondly, the origins of this deficit lie principally with the lack of translators between these languages. Machine aids can assist not only with the production of translations, but with a different type of product—the human translator. The work at Sheffield has demonstrated how an explicit description of a language that they do not know can be used to teach people to translate from that language, and to acquire knowledge of that language. In the future, we can look forward to all manner of machine aids that can assist in this training process.

REFERENCES

Abe, Masahiro, Ooshima, Yoshimitsu, Yuura, Katsuhiko and Takeichi, Nobuyoki, A Kana-Kanji translation system for non-segmented input sentences based on syntactic and semantic analysis in *Proceedings of the 11th international conference on computational linguistics* (COLING), Bonn, 1986.

Nagao, Makoto, Tsujii, Jun-ichi, and Nakamura, Jun-ichi, The Japanese government project for machine translation. *Computational Linguistics,* 11 (2-3) April-September, 1985.

Nelson, Andrew Nathaniel, *The modem reader's Japanese-English character dictionary* (2nd ed.). Rutland, Vermont: Charles E. Tuttle, 1962.

Roesner, Dietmar, When Mariko talks to Siegfried: Experiences from a Japanese/ German machine translation project in *Proceedings of the 11th international conference on computational linguistics* (COLING), Bonn, 1986.

Taguchi, Jun, MT systems go into business, *NIKKEI Computer,* 4 April, 1986, translated by Mary Gillender.

Tomita, Masaru, An efficient context-free parsing algorithm for natural languages and its applications, PhD Thesis, Dept. of Computer Science, Carnegie-Mellon University, Pittsburgh, May 1985.

Tomita, Masaru and Saito, Hiroaki, Automatic composition of stereotypic business letters in foreign languages, Dept. of Computer Science, Carnegie-Mellon University, 1984.

Whitelock, Peter, McGee Wood, Mary, Chandler, Brian, Holden, Natsuko, and Horsfall, Heather, Strategies for interactive machine translation: the experience and implications of the UMIST Japanese Project in *Proceedings of the 11th international conference on computational linguistics* (COLING), Bonn, 1986.

AUTHOR

Peter J. Whitelock, Department of Artificial Intelligence, University of Edinburgh, UK.