[*Translating and the Computer 24*: proceedings of the…International Conference…21-22 November 2002, London (Aslib, 2002)]

Japanese Language: An Introduction for Language Technology Professionals and Researchers

Carl Kay
Independent Consultant
15-22, Onoharanishi 3-chome
Minoh-shi, Osaka Japan 562-0032
carlkay@mub.biglobe.ne.jp

David Fine
Internationalization Consultant
Philadelphia, PA United States
david.fine2@verizon.net

Background and Synopsis

It is the authors' experience, based on a combined four decades in the translation and localization business, that most people involved in the language business lack knowledge about the basic features of the Japanese language and how it is the same or different from major Western languages. This is quite natural given the many ways that Japan and Japanese are remote presences to most people outside Japan.

Many barriers make Japanese inaccessible to Westerners: a strange and complex written language (and related challenges in computer processing of text); cultural differences; limited opportunities to study the language (and the time such study requires); geographic and time zone gaps, and Japan's shallow interaction with the West compared to exchanges among Western nations.

Still, even in a recession Japan remains one of the major economies of the world. Its information technology market still dwarfs that of the rest of Asia. In software localization Japanese is often a "Tier One" language meaning a target of the first round of localization aimed at the biggest markets not addressable with English product. Japan's government and companies remain major funders of research in machine translation and other language technology. Yet the above-mentioned barriers make it hard for Westerners in the language technology field to handle contact with Japanese as comfortably and effectively as they do with many other languages.

The authors' goal here is to allow language professionals and researchers outside Japan to become more familiar with Japanese and to feel more "at home" when dealing with Japanese in their work.

The structure of this paper is to describe four aspects of Japanese language and culture that are different from Western languages and cultures and have important impact on the language technology business. These differences are:

1. Extremely large Japanese character set, and lack of fit to the language itself
2. Extreme abundance of Japanese homonyms, lack of word spacing, and related input and sorting issues
3. Japanese structural features including S-O-V structure, placement of relative clauses, and omitted or indirect subjects
4. Japan's lack of historical contact with Western nations relative to contact among the Western nations

For each of the above items, the authors will give an explanation of the origins and nature of the situation, the implications for the language technology business, and trends going forward. The emphasis is on a practical overview rather than detailed scholarly discussion. The authors also limit the discussion of Western languages mainly to English while acknowledging that at times we are missing important distinctions between major Western languages. Our emphasis is on the broader contrast of Japanese and the West rather than the finer points that could be made by those with more background.

Feature #1 Extremely large Japanese character set, and lack of fit to the language itself

Origins and nature of the situation:

The Japanese writing system is largely borrowed from Chinese. Around the third or fourth century A.D., when Chinese civilization was quite advanced in terms of written language, government structure and philosophy, Japan was a loosely organized group of clans who had no writing system. The spread of Buddhism into Japan from China and Korea was accompanied by the introduction of the Chinese writing system which consisted of tens of thousands of characters.

Most Chinese characters represent a single word. The visual depiction of some Chinese characters derives from the meaning of the word represented. Other more complex characters are built from smaller elements, which may be characters in themselves though not necessarily. Some of these smaller elements are called radicals and carry a degree of category meaning. For example, many of the words related to botany contain a common element called the grass radical, which gives some framework to the meaning of the character but no clue at all to its pronunciation. Additional elements of each character under a given radical differentiate the many individual characters in the category. In some cases, but not always, these additional elements give some indication of the pronunciation of the character. More abstract words are often represented by a sequence of two characters called a compound.

When the Chinese written language came into Japan, it came packaged with many elements of Chinese Confucian culture and with thousands of new words incorporating the ideas and forms of that culture. Japanese culture even today in the 21st century remains highly shaped by this original importing of Chinese culture and writing system. Of course, many elements were adapted or dropped over time.

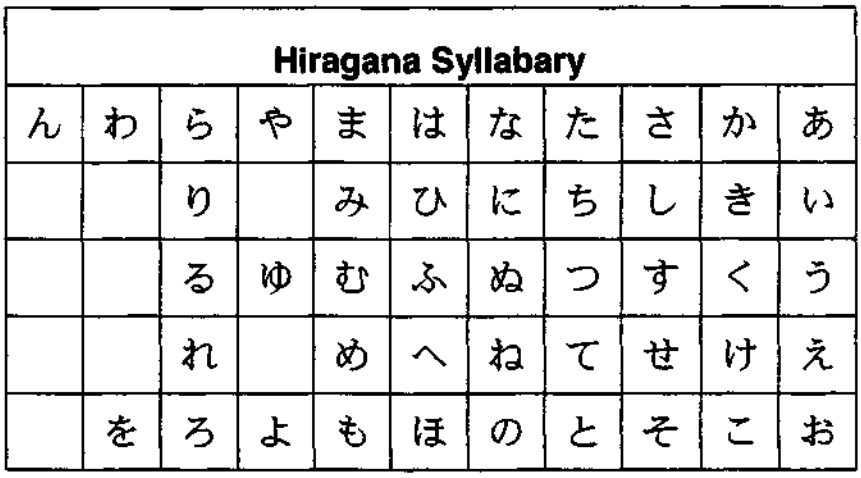
Despite this great contribution China made to Japanese culture, in many ways the Chinese written language is not a good fit to record the Japanese language, which is quite distant linguistically from Chinese. This mismatch has contributed to the extreme complexity of the Japanese writing system in use today.

One such mismatch is in verb structure. Japanese verbs are highly inflected, with many elements added onto the verb stem to indicate tense and active or passive voice. Chinese does not make these distinctions by adding on to the verb word itself. Thus, lacking a writing system, the Japanese seemingly happily borrowed the Chinese character for "eat" as the Japanese character for "eat." (The pronunciation of the two is unrelated, but in both languages the character means "eat.") In Japanese, however, other written markers were needed to add the inflections, such as present or past tense indicators, to the word "eat." Having no writing system in place, Japanese also took these indicators from Chinese as well, this time in the form of Chinese characters that happened to have approximately the same pronunciation as the inflection markers in Japanese (but having unrelated meanings in Chinese, being of interest to Japanese writers only for their pronunciation).

Thus the earliest form of written Japanese consists of borrowed Chinese characters, some of them there because of the meaning borrowed from Chinese and some there because of the sound borrowed from Chinese. Reading ancient Japanese documents was probably challenging in their time and is certainly extremely so now.

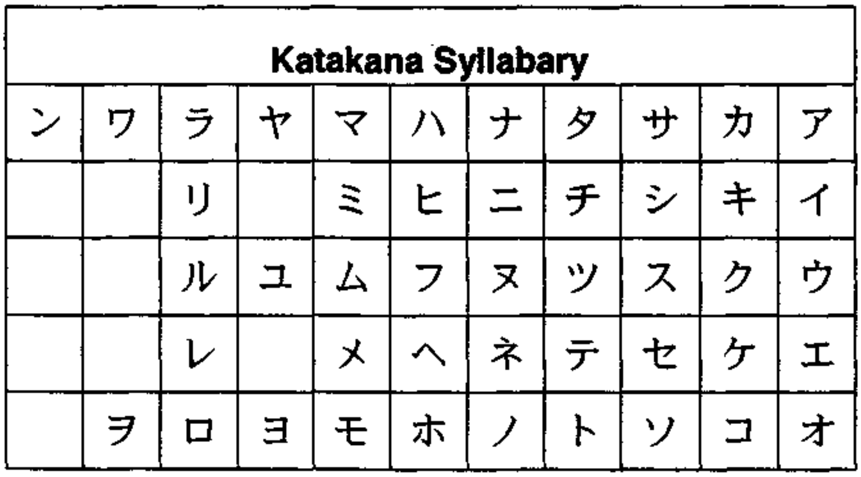
The Chinese characters borrowed for their meaning are called *kanji* ("Chinese characters") in Japanese and remain a major part of the written language. Meanwhile, over time, the Chinese characters borrowed just for their sound were simplified and evolved into a phonetic alphabet of characters called *hiragana*.

All Japanese speech can be recorded with the hiragana character set, but in actual practice a mixture of kanji (for which the pronunciation must be memorized) and hiragana are used. Going back to the above example, the Japanese word for "ate" can be written with the kanji for "eat" pronounced "ta-be" followed by hiragana "ma-shi-ta" indicating past tense, or the word can be written entirely in hiragana: ta-be-ma-shi-ta. There are many rules and conventions about where to use kanji and where to use hiragana, which are called *okurigana* when used to represent the inflections of verbs and adjectives (which are similarly inflected in Japanese). Schools tend to teach children to read and write hiragana first, adding kanji of increasing complexity each year according to a fixed schedule at each grade level.



Hiragana Syllabary										
ん	わ	ら	や	ま	は	な	た	さ	か	あ
		り		み	ひ	に	ち	し	き	い
		る	ゆ	む	ふ	ぬ	つ	す	く	う
		れ		め	へ	ね	て	せ	け	え
	を	ろ	よ	も	ほ	の	と	そ	こ	お

An alternate set of phonetic characters called *katakana* further complicates the situation. Katakana characters duplicate the phonetic coverage of the hiragana character set but are used mainly for words of western derivation, to distinguish them from native Japanese words. For example, the word Ko-n-pyu-ta spells out the word computer in Japanese pronunciation. There is also a Japanese word for computer, keisanki, composed of three kanji characters, but it is rarely used.



Katakana Syllabary										
ン	ワ	ラ	ヤ	マ	ハ	ナ	タ	サ	カ	ア
		リ		ミ	ヒ	ニ	チ	シ	キ	イ
		ル	ユ	ム	フ	ヌ	ツ	ス	ク	ウ
		レ		メ	ヘ	ネ	テ	セ	ケ	エ
	ヲ	ロ	ヨ	モ	ホ	ノ	ト	ソ	コ	オ

Japanese Character Sets for Electronic Processing

The large and complicated Japanese character set makes electronic processing of Japanese text an order of magnitude more complex than that for Western languages. The Japan Standards Association has defined a series of standards which specify codes for Japanese characters. Some of the key Japan Industrial Standard (JIS) character sets are described below.

The JIS X 0201 (Code for Information Interchange) standard defines a 7-bit character set for Roman characters, a 7-bit character set for katakana characters, and an 8-bit character set for both Roman and katakana characters. The JIS X 0201 Roman character set is very similar to the ASCII character set but the backslash has been replaced by the Yen symbol and the tilde has been replaced by the "overline". For anyone who has used Japanese DOS or Japanese Windows this explains why the path separator "\" displays as a Yen symbol.

JIS X 0208 (Code of the Japanese Graphic Character Set for Information Interchange) specifies the set of graphic characters and accompanying codes used in normal written Japanese. The class of graphic characters contained in JIS X 0208 includes special characters, numeric characters, Roman characters, hiragana, katakana, Greek characters, Russian characters, kanji, and line elements. The 6,355 kanji designated in the standard are divided into two levels. The first level contains the most frequently used kanji sorted by representative reading (as will be explained in more detail later, most characters have multiple possible readings), and the second level contains less common characters arranged according to radical and stroke count (based on the number of brush or pen strokes needed to write the character on paper). Each character is uniquely identified by two 7-bit bytes. The first byte defines a *ku*, or group of 94 graphic characters, and the second byte designates a *ten*, or a specific character within the *ku*.

JIS X 0212 (Code of the Supplementary Japanese Graphic Character Set for Information Interchange) is a supplement to JIS X 0208 which defines additional graphic characters. It includes special characters, alphabetic characters, and approximately 5,800 kanji. The JIS X 0212 character set is not widely used in Japanese software.

JIS X 0213:2000 (7-bit and 8-bit double byte coded extended Kanji sets for information interchange) is a more recent extension to JIS X 0208 and serves as a replacement for JIS X 0212. It includes the characters from JIS X 0208 and adds about 4,300 characters. JIS X 0213 defines level 3 and level 4 kanji characters. Unlike JIS X 0212 it fits within the Shift-JIS character encoding (see below). JIS X 0213 has not been widely adopted yet and it remains to be seen how popular it will become.

The Unicode Standard defines a universal character set which contains characters for all the key writing systems of the world. The Unicode Standard is compatible with the international standard ISO/EEC 10646 and corresponding versions of the standards are synchronized. JIS X 0221-1:2001 is a Japanese version of ISO/IEC 10646-1:2000 which is identical in technical content to ISO/DEC 10646-1:2000. Unicode version 3.2 provides codes for more than 95,000 characters and includes all the characters from the JIS X 0208, JIS X 0212, and JIS X 0213 character sets.

In order to reduce the number of required ideographic characters in Unicode, the Chinese, Japanese, and Korean ideographic characters were combined into a single CJK Unified Ideographs region through a process known as Han unification. The arrangement of the ideographic characters is based on their order within four major dictionaries. Japanese symbols, punctuation, hiragana, and katakana are defined in a CJK Phonetics and Symbols code area. Half-width and full-width forms are included in a Compatibility area to maintain compatibility with the pre-existing Japanese character sets.

The Unicode Standard is supported by many operating systems, web browsers, and software applications and is required in many modern information technology standards. A software application that is implemented with Unicode can support character data from all the major languages of the world, and Unicode data is portable across software platforms, languages, and countries. Despite the technical advantages for international software the Unicode Standard has not been widely embraced in Japan.

Japanese character encodings

There are several standard ways to encode Japanese character sets into bit values. Shift-JIS is a popular encoding which is widely used by personal computer software. Shift-JIS operates in an 8-bit environment. The value of a given byte indicates whether it should be interpreted as a single-byte character or as the first byte of a double-byte character. The encoding does not require any escape sequences so software processing text represented in Shift-JIS does not need to remember the state for an arbitrarily large number of characters. Shift-JIS encodes JIS X 0201 and JIS X 0208 characters and has been extended to support the newer JIS X 0213 character set.

Extended Unix Code (EUC) is a multibyte encoding originally developed by AT&T that is commonly used on UNIX based platforms. EUC can handle four different character sets. The primary code set (code set 0) is used for ASCII and the three supplementary code sets (code sets 1, 2, and 3) can be used for other character sets. In Japan the EUC-JP character encoding assigns code sets as described in the following table. Since the advent of JIS X 0213 character set there is also a form of EUC which supports JIS X 0213.

Code Set	**Character Set**
0	ASCII
1	JIS X 0208
2	JIS X 0201
3	JIS X 0212

JIS is a Japanese character encoding based on ISO 2022 which operates in a 7-bit environment. The traditional form of the encoding supports single-byte ASCII and JIS X 0201 Roman characters as well as double-byte JIS X 0208 characters. Escape sequences are used to shift from either single-byte character set to the double-byte character set, or from the double-byte character set to one of the single-byte character sets. A computer program processing text represented in JIS must keep track of the current state. JIS can be used as a medium of exchange in environments which are not 8-bit clean. A form of JIS encoding called ISO-2022-JP was designed to transmit Japanese email messages over the internet. Other forms of JIS are capable of encoding characters from JIS X 0212 and JIS X 0213.

There are three character encoding schemes that are used to represent Unicode text: UTF-8, UTF-16, and UTF-16. The three encodings are all capable of representing the full repertoire of Unicode characters but they encode the data differently in code units of 8,16, or 32 bits. UTF-8 is often used for exchange and storage of Unicode text data. UTF-8 is a multibyte encoding of Unicode where a character is represented by one, two, three, or four bytes. UTF-8 is upward compatible with ASCII and ASCII characters are represented by the identical single-byte value in UTF-8. Japanese characters are typically represented with three bytes in UTF-8. In UTF-16 the most common characters are two bytes and more obscure characters are accessed through pairs of special codes called surrogates. In UTF-32 each character is uniformly four bytes.

It is frequently necessary to transcode or convert between different Japanese character encodings. For example, when moving character data between a Unix system and a personal computer it might be necessary to convert between EUC and Shift-JIS and when sending an email over the internet it is necessary to transcode to ISO-2022-JP. There are standard algorithms to convert among Shift-JIS, EUC, and JIS character encodings.

It is also necessary to be able to convert between Unicode and the key Japanese character encodings. For example, an internationalized software application might process character data internally in Unicode but still need to handle popular Japanese character encodings at system interfaces. Unicode contains all the characters from JIS X 0208, JIS X 0212, and JIS X 0213 and it is possible to do conversions between Unicode and the Japanese character sets. The arrangement of kanji characters is different in Unicode than in the Japanese character sets so a transformation table is used to perform the conversions. A simple and efficient algorithm can be used to convert among the three Unicode encodings UTF-8, UTF-16, and UTF-32.

Processing Japanese text

Historically, software that supported Western European languages used single-byte characters. In order to support the large number of Japanese characters it is necessary to use more than one byte to represent a character. There are two basic programming techniques for supporting the extended range of characters: multibyte characters and wide characters.

A multi-byte character is a character which consists of one or more bytes. A program which processes multibyte character strings must be aware of the boundaries between characters. For example, when traversing a multibyte character string which is encoded in Shift-JIS it is necessary to check whether each byte represents a single-byte character or the lead byte of a double-byte character.

A wide character is a character that is uniformly represented with the same number of bytes and is wide enough to hold each character in the underlying character set. Wide characters permit efficient, uniform processing, but they occupy more memory space. A combination of wide characters and multi-byte characters are sometimes used within the same system. For example, wide characters might be used internally to efficiently process character data in a uniform way while multibyte characters might be used for data storage and exchange.

Implications for the language technology business:

The dramatic gains in recent years in IT processing speed and cost performance, combined with years of research and commercial development, have made computer processing of Japanese much less of a bottleneck than twenty years ago when PCs first appeared. Still, an example of the challenges that remain is that the authors were not able to submit this paper to this international conference electronically if it included Japanese text examples, because the U.K.-based organizer and its printer could not process files containing Japanese characters.

To give another example, a global-scale satellite telephone interface localization project was delayed significantly because the client asked for "kanji localization," while the client's precise intention was to have the interface support Japanese. In fact, all that was needed (and technically possible on the device) was an interface that supported the SO hiragana characters. The Ireland-based localization project manager at the global localization company handling the project did not understand the distinction between kanji and hiragana characters (and didn't realize that she didn't understand). When the localization company's Japan office tried to confirm that what the original client needed was in fact hiragana support, the project manager vociferously argued "no the client said KANJI." She blamed the Japan office for not respecting the wishes of the customer. The Japan office was insulted that a non-Japanese was doubting their expertise about their own language. All of this confusion (and the loss of time that went with it) could have been avoided had the client and the project manager been a little more familiar with Japanese.

Trends going forward:

Awareness of the need to plan ahead for Japanese localization has grown, and with this awareness the demand for software internationalization services has grown. However, as the computing model has moved from standalone applications to shared services and content across networks, the complexity of these issues is only increasing.

Feature #2 Extreme abundance of Japanese homonyms, lack of word spacing, and related input and sorting issues

Origins and nature of the situation:

In the era of typewriting and phototypesetting that preceded computer word processing, each of the approximately 10,000 kanji characters commonly used in Japanese required a separate input which basically meant sliding a print mechanism of some kind over a grid of 10,000 characters that took up a whole desk of space, and picking up each character in order as needed.

The development in the late 1970's of commercial word processors was a major breakthrough in Japan. The key advance was to allow hiragana input (only SO characters) from a keyboard while harnessing the power of the computer to analyze the text and change appropriate hiragana to the correct kanji. Much research and development has gone into this complex problem (described in more detail by author Kay in a previous article). Again, certain specific features of the language add to the challenge.

First is the problem of homonyms. During the aforementioned influx of Chinese culture into Japan, not just writing but also many elements of culture, government organization and other areas were imported from China. For many of the new concepts and structures there was not already a word in Japanese. So the Chinese word for, say, civil service, was imported along with the concept itself. Many thousands of words, most of them two character compounds, came into Japanese this way.

The problem is that in Chinese the pitch of a sound — rising pitch, falling pitch etc.— is a significant element which distinguishes one word from another (two words with same sound but different pitch are completely different, unrelated words in Chinese, each with its own kanji character). This distinction does not exist in Japanese. Thus when Chinese words came into Japanese en masse, words that had the same sound but different pitch all sounded the same to the Japanese ear and became homonyms, words with the same pronunciation. They had the same representation in hiragana but different meaning and, most significant here, were represented by different kanji characters. So Japanese is filled with thousands kanji which share pronunciation/phonetic representation (in hiragana) with from a few to up to dozens of unrelated other kanji. Japanese computers need extremely complex algorithms and dictionaries to correctly convert hiragana input to kanji.

The complexity of the kanji characters has at times made it seem tempting to use just hiragana in Japan. However, with so many homonyms, and no demarcation of word spaces (see below), purely phonetic representation of Japanese is a nightmare. Even among Japanese people who know the characters well, occasionally a speaker will say a word that the listener cannot identify due to the homonym problem, and the speaker will draw the character in the air, using his or her finger like a brush, to clarify the meaning.

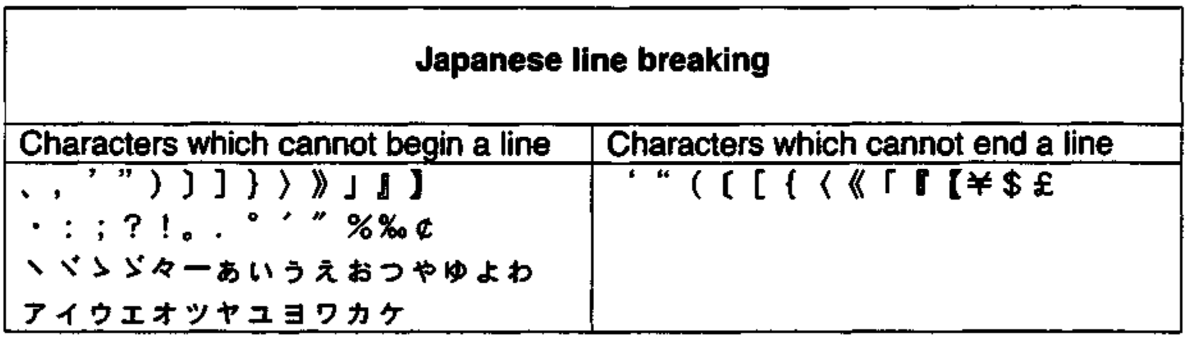
Reading kanji in a text allows Japanese people to grasp the meaning faster than a pure hiragana text and also helps define the boundaries between words which can at times be hard to figure out. Unlike Western languages no space is inserted between words in written Japanese. This is the other major factor, along with the abundance of homonyms, that complicates computer-assisted conversion from hiragana to kanji.

In traditional writing each Japanese character and punctuation mark occupies a square of the same size. Hiragana symbols called *okurigana* are used to inflect the ends of verbs and adjectives. The pattern of kanji and hiragana which appears in Japanese script often provides visual clues that assist the human reader in separating distinct words.

When processing Japanese text by computer one of the fundamental steps is to segment the text into word units. In order to automatically parse text, search text, index keywords, or convert from text to speech it is necessary to first determine the boundaries between words. Computer programs have been developed to segment Japanese text through a combination of dictionary lookup, morphological analysis, and statistical techniques. While kanji input programs are continually improved, incorrect conversions by the software force the operator to manually override, slowing down input and interrupting a writer's train of thought.

The authors had intended to test some Japanese copy in a free Internet Japanese-English translation engine to illustrate the structural differences of the languages covered later in this paper, but the MT system couldn't even parse the word breaks correctly. Thus the result was gibberish in terms of terminology which made meaningless our intention to illustrate structural differences.

Since words are not delimited by spaces the procedure for wrapping lines of text is different from Western European languages. Japanese text basically flows from one line to the next, breaking if necessary in the middle of a word. There are some rules, called *kinsoku shori,* which govern the way lines are wrapped. For example, a line can not begin with a comma or right parenthesis and a line can not end with a left parenthesis. The following table illustrates some characters which can not begin or end a line. JIS X 4051 (Line composition rules for Japanese documents) describes the rules in more detail.



Japanese line breaking

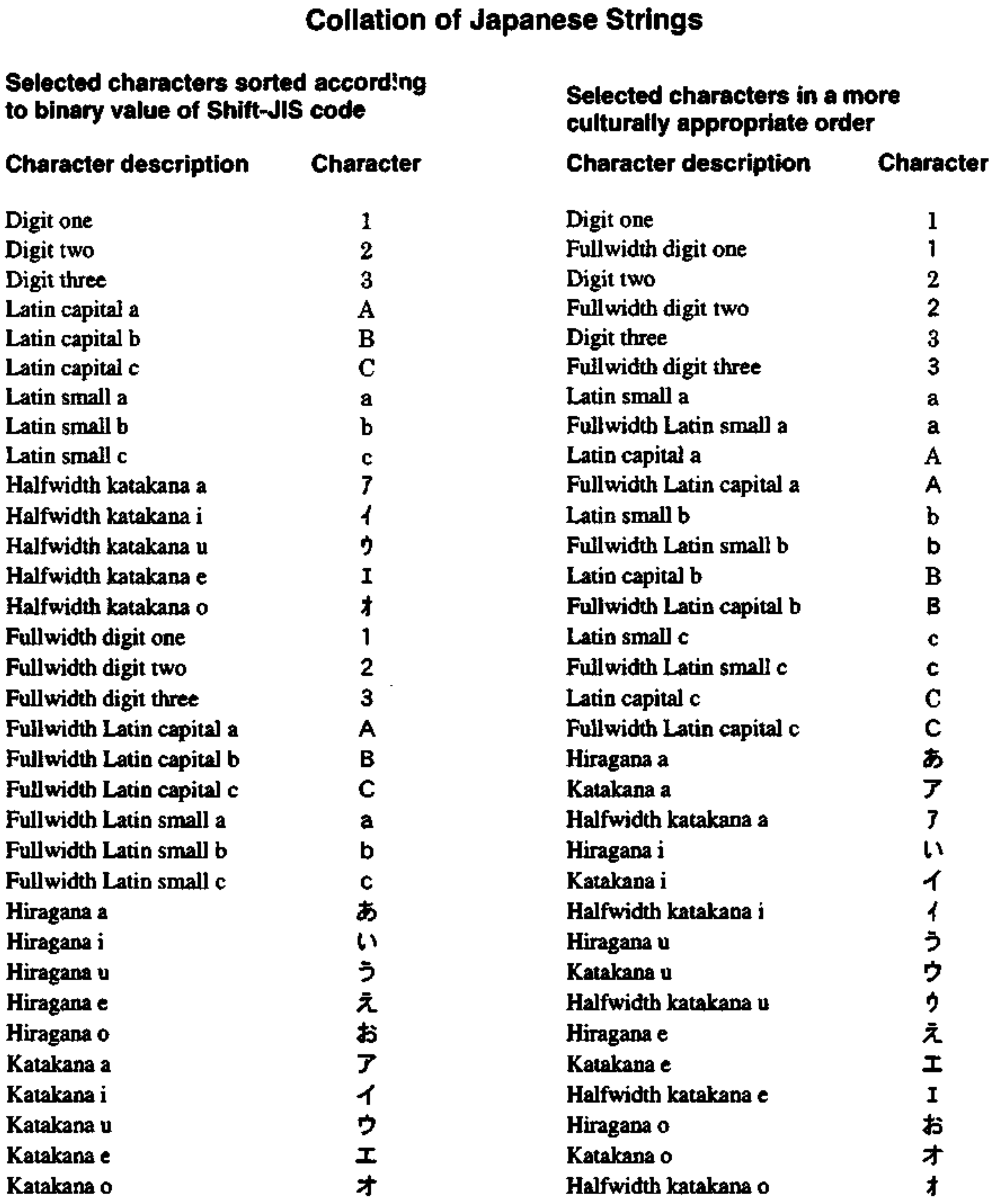
Characters which cannot begin a line	Characters which cannot end a line
、，’”）〕］｝〉》」』】 ・：；？！。．°′″％‰¢ ヽヾゝゞ々ーぁぃぅぇぉっゃゅょゎ ァィゥェォッャュョヮヵヶ	‘“（〔［｛〈《「『【￥＄￡

Among the challenges related to kanji, in particular proper nouns like the names of people and places can have unusual and non-standard readings. Small hiragana or katakana, called *furigana*, are sometimes written above difficult kanji to indicate the reading of the characters. When entering a personal name on a form in Japan it is common to provide the name in both kanji and furigana. Furigana fields sometimes need to be included in software applications to maintain the appropriate pronunciation and to sort character data in a culturally appropriate manner.

There are well-defined rules for sorting hiragana and katakana characters. The symbols are basically sorted in the order that they appear in the Table of Fifty Sounds. Sorting kanji is slightly more difficult because each character possesses many different readings. It is possible to son kanji by representative reading, radical, stroke count, or some combination of these characteristics. A set of Japanese strings which contain a mixture kana and kanji are typically sorted by first associating kana with each string in order to indicate the phonetic value. Then the strings are ordered according to the associated pronunciation. A software application that presents a sorted list of personal names or company names will often need to include a furigana field for each name so that the names can be sorted based on the pronunciation.

Kana and kanji characters have been arranged in the Japanese character sets in a meaningful way. However, a son based on binary value of character set codes rarely produces the expected culturally appropriate order. A culturally appropriate collation algorithm uses multiple levels to properly handle characters from different scripts, upper/lower case characters, diacritics, full-width/half-width characters, hiragana/katakana, and special symbols.

For example, when sorting Shift-JIS text based on the binary value of each character the half-width Roman letters, full-width Roman letters, hiragana, katakana, and half-width katakana characters will all son in separate sections. In a culturally appropriate son the half-width/full-width variations of the same Roman letter would son next to each other and similarly hiragana/katakana symbols with the same pronunciation would son next to each other.



Collation of Japanese Strings

Selected characters sorted according to binary value of Shift-JIS code

Character description	Character
Digit one	1
Digit two	2
Digit three	3
Latin capital a	A
Latin capital b	B
Latin capital c	C
Latin small a	a
Latin small b	b
Latin small c	c
Halfwidth katakana a	ｱ
Halfwidth katakana i	ｲ
Halfwidth katakana u	ｳ
Halfwidth katakana e	ｴ
Halfwidth katakana o	ｵ
Fullwidth digit one	１
Fullwidth digit two	２
Fullwidth digit three	３
Fullwidth Latin capital a	Ａ
Fullwidth Latin capital b	Ｂ
Fullwidth Latin capital c	Ｃ
Fullwidth Latin small a	ａ
Fullwidth Latin small b	ｂ
Fullwidth Latin small c	ｃ
Hiragana a	あ
Hiragana i	い
Hiragana u	う
Hiragana e	え
Hiragana o	お
Katakana a	ア
Katakana i	イ
Katakana u	ウ
Katakana e	エ
Katakana o	オ

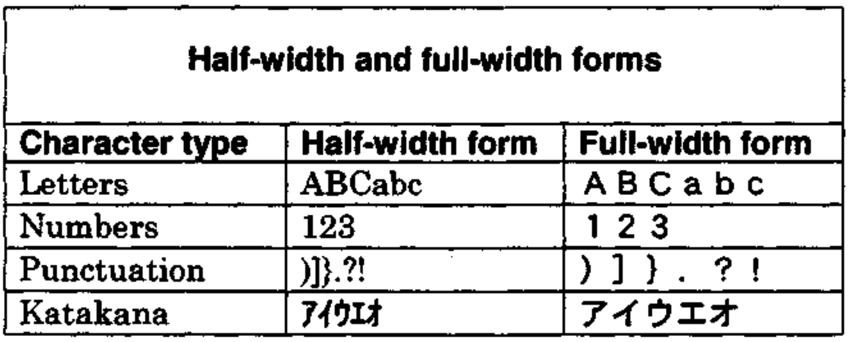
Selected characters in a more culturally appropriate order

Character description	Character
Digit one	1
Fullwidth digit one	１
Digit two	2
Fullwidth digit two	２
Digit three	3
Fullwidth digit three	３
Latin small a	a
Fullwidth Latin small a	ａ
Latin capital a	A
Fullwidth Latin capital a	Ａ
Latin small b	b
Fullwidth Latin small b	ｂ
Latin capital b	B
Fullwidth Latin capital b	Ｂ
Latin small c	c
Fullwidth Latin small c	ｃ
Latin capital c	C
Fullwidth Latin capital c	Ｃ
Hiragana a	あ
Katakana a	ア
Halfwidth katakana a	ｱ
Hiragana i	い
Katakana i	イ
Halfwidth katakana i	ｲ
Hiragana u	う
Katakana u	ウ
Halfwidth katakana u	ｳ
Hiragana e	え
Katakana e	エ
Halfwidth katakana e	ｴ
Hiragana o	お
Katakana o	オ
Halfwidth katakana o	ｵ

There are some national and international standards for collation of text strings. JIS X 4061:1996 [Collation of Japanese character strings] defines a Japanese standard for collation and ISO/TEC 14651 is an international standard for string ordering and comparison. Whenever possible, software developers should take advantage of modern software systems which provide support for culturally appropriate sorts.

Searching for Japanese keywords within plain text is complicated by the lack of word delimiters, redundancy in the writing system, and inconsistent orthography. In order to produce good results a procedure for searching for Japanese text must be flexible enough to handle these issues. First the text that is being searched needs to be segmented so that the search key(s) will be compared against individual word units rather than the end of one word and the start of the next word.

As mentioned previously, the hiragana and katakana syllabaries are redundant in the sense that they represent the same set of fifty sounds. In some cases the search process should distinguish between hiragana and katakana symbols with the same sound but in others it might be appropriate to make no distinction. The Japanese character sets contain both half-width and full-width forms of Roman letters, Arabic numbers, punctuation, and katakana. As with the two forms of kana in some cases it might be desirable to distinguish between half-width and full-width forms of the same character while in other situations it might not.



Half-width and full-width forms

Character type	Half-width form	Full-width form
Letters	ABCabc	ＡＢＣａｂｃ
Numbers	123	１２３
Punctuation	)]}.?!	）］｝．？！
Katakana	ｱｲｳｴｵ	アイウエオ

In practice the same Japanese word is sometimes written in different ways. Foreign loan words can be transliterated into katakana in multiple ways. The portion of a word that is written with kanji and the pan that is written with okurigana can also vary. These inconsistencies pose a challenge when searching for Japanese text.

Implications for the language technology business:

Though Microsoft has come to dominate word processing in Japan as it does in most other countries, other company's word processor products continue to maintain small market shares, based partly on the accuracy or special features of their kana-kanji conversion input engine.

Trends going forward:

The above-mentioned challenge of keyboarding a language with 10,000 characters on a regular typewriter or computer keyboard is made even harder by the limited keyboard of the small cell phones now ubiquitous in Japan. Millions of Japanese users send and receive dozens of messages a day, but typing Japanese with only ten keys (tiny ones at that) is very challenging. Much creative R&D effort is going into input systems to reduce the number of keystrokes needed.

The number of homonyms also adds a challenging element to Japanese voice recognition technology, but the keyboard issues are so central that it is inevitable that voice recognition will play a growing role in Japan in the coming years.

Feature #3 Japanese structural features including S-O-V structure, placement of relative clauses, and omitted or indirect subjects

Origins and nature of the situation:

Compared to most other major world languages, more uncertainty remains about the origins of the Japanese language. There are clear traces of an Altaic heritage through northeast Asia, which may have connections to the ancestors of Hungarian and Finnish. However, there is also strong evidence of influence from Southeast Asia and the South Pacific. The timing and nature of the mixing of these two influences is the subject of a complex scholarly debate which is beyond the author's own expertise and beyond the scope of this paper.

Whatever the origin, the resulting Japanese sentence structure differs greatly from major Western patterns. This makes it impossible to achieve an acceptable level of accuracy in machine translation simply by focusing resources on terms and expressions and then substituting words and phrases in the same location in the sentence. While of course this cannot always be done between English and French or even between Spanish and Italian, word substitutions and very simple sentence element transformations can yield reasonably useful output for MT with Western languages. With Japanese, much more complex algorithms are needed to move sentence elements the greater distance required to cross the language barrier, and lower accuracy results.

The basic Japanese sentence structure is called subject-object-verb or "SOV." In Japanese the verb comes at the very end of the sentence, and the indication whether a sentence is positive or negative comes at the end of the verb. So while in English or French one must commit early to being positive or negative (I do NOT like, Je NE veux pas etc.) in Japanese it is possible to postpone showing one's "yes or no" point of view, while observing the body language of the listener, until arriving at the end of a sentence. It is often said that Japanese people tend to avoid very clear yes-no statements when conflict would result. This feature of the language helps the speaker adjust his or her message to the reaction felt as the statement unfolds.

These very opposite sentence structures of the two languages make simultaneous interpreting from Japanese into English very challenging, because the interpreter either has to produce very awkward English structures— the Japanese interpreted in order of spoken chronology comes out as "As for the proposal which you made on May 15th, whereby you would increase production at your factory etc., etc., we do not accept it"—or alternatively just keep silent through the long sentence (while remembering all of it) and then start speaking in more normal English order only after the final Japanese verb is spoken ("We do not accept..... "). This latter strategy turns simultaneous interpreting more into consecutive interpreting.

Another major feature of Japanese is that modifying clauses precede the noun they are modifying. For example, *kare ga itta mise* (the store he went to) is literally "he went store." In Japanese no preposition appears because the verb ("*itta*") modifies the noun "store" *(mise)* directly. But in English a preposition is required: the store he went to, the store to which he went. The preposition required can be "to" or "for" or a number of others. Sometimes the same English verb can require different prepositions depending on the meaning. This is one factor that makes it very hard to get natural sounding translations in English from machine translation of Japanese.

More subtle kinds of differences exist too, ones that can slow down human translators as well. When translating from English to Japanese, there is no direct way to render abstract nouns as performing concrete actions. A common English sentence like "Advances in technology created new opportunities" sounds very unnatural translated directly into Japanese. Translators must transform it to something like "With advances in technology, opportunities were created" or using the verb implied in the abstract noun, "Because technology has advanced, opportunities have been created."

Treatment of subjects in Japanese is also unique. In many sentences the subject is omitted, to be understood from context. The verb "tabemashita" stands alone as a sentence: "[subject omitted but understood from context] ate." A literal translation of the word on the page - "ate" - is not usually an acceptable English sentence. So the translator, human or machine, must figure out what the subject is and add it to the sentence.

Sometimes the subject is abstracted in a way to emphasize the entire situation rather than the main actor. For example, in the Japanese sentence *"teki ga bo-ru o nageta tokoro o uchikaeshita"* first of all the subject of the sentence itself is omitted. "[Omitted subject- lets use "I" for sake of argument] I hit the ball the opponent threw." What is more interesting here is that the main noun of the phrase that forms the object of the verb "hit" is not actually the word for "ball" but is "tokoro" which means "place, moment, occasion," with the object of "I hit" actually being something more like "the situation in space/time when the opponent threw the ball" rather than just "ball." The emphasis is not just on the ball, but the opponent and the throwing and then suddenly—the moment when I hit the ball.

But we can't quite say in the sports section of an English newspaper "Barry Bonds hit the situation in space time when the opponent threw the ball." The first translation I gave above- "I hit the ball the opponent threw" is a simple translation of the skeleton of the sentence, but a more accurate rendering feeling-wise might be "I hit the ball when the opponent threw." Can you feel the difference in the English between that and "I hit the ball the opponent threw?" It is hard enough for human translators to notice and to capture those differences. From machine translation we should be happy at this stage if we can get as far as the accurate skeleton translation "I hit the ball the opponent threw" without getting tripped up on literally rendering the word "tokoro" as "place" which in fact appears in the version I got back from a leading free web MT site: "The place where the enemy threw the ball was struck back."

You are probably thinking that it is naive of the authors to expect machine translation to translate poetry like this example. The problem is, it is not poetry. It is a very common way to express a quotidian description of a mundane, concrete event. A lot of sentences like this will turn up in material of any genre, from manuals to newspaper articles. This de-emphasis of the linear depiction of the subject as an agent acting on an object, with rather a more detached situational expression, hints at the kinds of linguistic and even world-view differences between Asia and the West that are so exciting but remain so hard to bridge.

Implications for the language technology business:

While there is reasonable demand for Japanese MT products, especially low end web scanners and mid-range products to support human translators, as well as a little demand for outsourced MT services, machine translation makes up only a miniscule portion of the very large Japanese translation market. For all the above reasons, accuracy rates lag those of Western language MT language pairs. This limits the range of applications of MT for Japanese and thus holds down the overall commercial demand level.

The differing order in a sentence of key structural elements between Japanese and other languages can also complicate tasks such as segment alignment in TM tools, localization of links in on-line help and other operations where sub-sentence units must be matched.

Trends going forward:

Investment and progress in MT in Japan continues, but nothing happening now (as known by the authors at least) suggests any substantial breakthroughs in accuracy, cost performance or other metric in the near- to mid-term future. Research into MT between Japanese and Korean, which share many linguistic structural features, is progressing, and accuracy rates approaching those seen in Europe can probably be expected for that language pair.

Feature #4 Japan's lack of historical contact with Western nations relative to contact among the Western nations

Origins and nature of the situation:

Japan's cultural roots are in its local folk culture plus major influences from Asia including Taosim, Confucianism and Buddhism. This distance from the Judeo-Christian tradition of the West was imprinted further during the Edo period from 1660-1860 when Japan was essentially isolated from the rest of the world. During these two hundred years when Europe and North America saw tremendous cross-border cultural, economic, technological, political and military activity, Japan's traditional cultural patterns were deeply reinforced. The result for Japan was a strongly ingrained cultural distance that amplified its geographic distance from the West. The lack of immigration into and emigration out of Japan is further contrast with the history of the West.

Implications for the language technology business:

Without much cross immigration and cultural contact between the West and Japan, some of the basic foundations for the creation of a robust translator community have been absent. When Author Kay joined the American Translators Association in 1982, there were only a few Japanese translator members, and all sessions at the organization's convention used French, Spanish, German or Russian as examples. Over a number of years Author Kay and others built a Japanese Language Division within the ATA. Despite this and other activity by ad hoc groups such as IJET, the Japanese translation community in the West has remained small and scattered compared to the European language translator community, despite the great volume of business and technical translation between Japanese and English. Author Kay now serves as the only non-Japanese Board Member of the Tokyo-based non-profit Japan Translation Federation, an organization which also remains unknown to most non-Japanese despite its role as the industry's leading trade association in Japan.

Given the global footprint of Japanese manufacturers, one would expect that a Japanese translation company with worldwide offices would also exist. In fact, there are none. The largest revenue of a Japanese translation company (or portion of revenue from translation at a larger company) is about $25 million, while there are three or four translation/localization companies in the US and Europe with annual revenues exceeding $100 million. This lack of scale can be partly explained as follows:

- Lack of software developed in Japan for global markets (holds down demand for high volume fast turnaround multi-language projects that force translation companies to scale up and go global)
- Lack of status of service industries in Japan (hard to get financing to grow businesses that lack tangible assets such as factories)
- Lack of qualified Japanese translators at locations around the world, for the reasons mentioned earlier
- Emphasis in Japan on English translation, with very low skill levels in languages such as French, German and Spanish
- Lack of multilingual content management strategy at major Japanese firms

Trends going forward:

A lot of progress has been made in Japanese translation since 1982. However, the recent decline of Japan from economic superpower status to merely an affluent nation is squeezing out both corporate and research activities that did not have at least some economic rationale. The authors hope that language technology research and commercial activity in Japan and between Japan and the West will grow despite the short-term economic challenges.

As Japanese manufacturers move facilities to China, it is likely that a China-based Japanese translation industry will emerge, with very low cost structure though perhaps unstable quality as was seen in Chinese manufacturing during the early stages. This trend could be both a threat to the domestic translation industry in Japan as well as an opportunity for one of the industry leaders to begin to expand in scale and build a global infrastructure.

References and Further Information

American Translators Association Japanese Language Division home page
http://www.ata-divisions.org/JLD/home.htm

Asia-Pacific Association for Machine Translation website http://www.aamt.info/

Hadamitzky, Wolfgang, and Spahn, Mark, *Kanji & Kana*, Charles E. Tuttle Company, Boston and Tokyo, 1990

IJET (Internationa] Japanese/English Translation Conference) This annual conference returns to Europe (Dublin) in May 2003 for first time since 1997 (Sheffield)

Japan Association of Translators (JAT) web site www.jat.org

Japan Translation Federation web site http://www.jtf.jp/ (most of site is in Japanese)

Japanese Standards Association http://www.jsa.or.jp/default_english.asp

Kay, Carl, "Japanese Software Wars," *Language International* 9.3 (1997), John Benjamins, Amsterdam, pp. 10-11,47

Lunde, Ken, *CJKV Information Processing,* O'Reilly & Associates, Sebastopol, CA, 1999

Nakanishi, Akira, *Writing Systems of the World,* Charles E. Tuttle Co, Boston and Tokyo, 1988

Narita, Hajime; Amano, Shinya and Muraki, Kazunori, *Kou Sureba Tsukaeru Kikai Honyaku* ("How to Make Machine Translation Useful"), Babel Press, Tokyo, 1994.

Saraki, Masashi, "Kikai Honyaku no Yakubun Kaizen: Eibun no Meishiku Kouzou—Kouchishuushoku Kouzou no Yakushutsu Houhou" (Improving the sentence output of machine translation: Methods of translating English noun phrase/post-position modifier structure, Workshop held in conjunction with 7th Nationwide Meeting of Gengo Shori Gakkai (The Association For Natural Language Processing), March 30, 2001

Tahara, Toshitsugu, *Einichi Jimu Honyaku no Houhou* (Techniques of English-Japanese Business Translation), Taishuukan Shoten, Tokyo, 2001

The Unicode Standard, Version 3.0, Addison Wesley, 2000

Unicode Home Page: http://www.unicode.org/

Acknowledgements

The authors thank the Asia-Pacific Association for Machine Translation and some of its members for kind assistance with parts of this paper.