Philipp Strazny
Tim Nishimura

Any text document consists of content and layout. The document translation process aims at recreating a document in the target language that is equivalent to the source document in both content and layout. Thus, the document translation process has two main subprocesses: content translation and layout adjustment. Content translation must be performed by native speakers of the target language, and since this is apparent to most people, it generally is.

The situation is different in the case of layout adjustments. Modern translation tools are so good at extracting the translatable text portions from source documents while protecting the non-translatable formatting elements that layout adjustments may not even be needed. This is typically the case for the translation of web formats, such as HTML or XML. Since web layout is rather fluid, with a large part of the actual presentation controlled by the web browser, it is generally sufficient to simply replace the source with the target text. If the goal is to produce translated print documents, however, the translated text often has to be forced into a predetermined, fixed layout. Due to time constraints, cost considerations, or other logistic factors, desktop publishers often find themselves confronted with the task of touching up a document of which they are unable to read a single word.

Although one may deplore this situation as a violation of best practice, it is nevertheless common enough to warrant treatment as an integral part of the translation process. As such, it requires support material to help non-readers in their layout adjustment task.

In this issue, we will look at Japanese. The first concern a desktop publisher may have is text directionality. As many people will know, Japanese books are traditionally read from right to left, in a top-to-bottom column format. However, scientific and technical publications including user manuals for hardware and software are always written left to right in the same format as English documents, and the Web appears to be spreading this format still further (see Figure 1 for a sample from a Japanese government web site). Thus, when English-language technical documentation is translated into Japanese, the source text should simply be replaced with Japanese, and the document layout shouldstay as is.

Figure 1: a page from www.e-gov.go.jp – left-to-right, top-down layout

When space is tight in print documentation, it is often necessary to adjust line breaks manually. For a non-reader, Japanese text appears daunting at first glance, since words are often not separated by spaces. However, Japanese writing has a number of surface characteristics that can be a useful guidance.

First, Japanese uses punctuation marks to delimit sentences (period: 。), subclauses (comma: 、) and insertions (parentheses). Thus, just like in English, it is always safe to insert a line break after a period, or a comma, a closing parenthesis, or before an opening parenthesis. When foreign words are transcribed into Japanese script, spaces are indicated either with the ・ character or a one-byte space. Inserting a line break immediately after this dot character or the space is acceptable.

The Japanese writing system uses three different sets of characters, each one for a specific purpose. Chinese characters, called *Kanji*, are used to convey concepts or word meanings – they are *logographic* symbols. Thus, *Kanji* carry the main meaning of Japanese texts. *Kanji* are fairly easy to recognize, since most of these symbols look fairly intricate, for example: 控訴審判決. Since Japanese uses many hundred *Kanji*, a complete listing is impractical.

Hiragana are symbols of Japanese origin which form a syllabary. This means that, like English letters, they stand for a speech sound rather than a word meaning. However, while English letters generally represent a single sound, *Hiragana* represent a whole syllable. This is the complete set of *Hiragana*:
あいうえおかきくけこさしすせそたちつてとなにぬねのはひふへほまみむめもやゆよらりるれろわゐゑをんがぎぐげござじずぜぞだぢづでどばびぶべぼぱぴぷぺぽゃゅょ
Hiragana are used to represent grammatical information, i.e. they roughly correspond to English prepositions, conjunctions, and similar function words. *Hiragana* are generally attached at the end of a

word, i.e. *Hiragana* typically form a unit with preceding *Kanji*.

Katakana look like an “edgier” form of *Hiragana*:
コサシスセソタチツテトナニヌネノハヒフヘホマミムメモヤユヨラリルレロワャュョィェォンガギグゲゴザジズゼゾダヂヅデドバビブベボパピプペポヴ

Katakana are used for transcribing foreign words and names.

In some cases, as in product names or entity names, Japanese also uses Western script. Besides, Arabic numbers are commonly used in Japanese just like in English.

Since these fairly easily distinguishable symbol sets are used for such different purposes, it is possible to make some useful generalizations for basic layout adjustments:

- Try not to separate adjacent *Kanji* symbols, i.e. adjacent *Kanji* should stay together as much as possible. However, when a long series of Kanji (three or more) extends beyond the line limit, you may separate them.
- Less than three adjacent *Hiragana* and adjacent *Katakana* should always stay together. Any series of three or more *Hiragana* or *Katakana* characters can generally be separated. The only exception are character combinations that form a single syllable. These character combinations, however, are also easily identifiable, as the second and third character is smaller in size than the first one, for example, チェッ, ちぇっ, タッ, or たっ. Furthermore, the long vowel sign “ー” that is attached to *Hiragana* or *Katakana* should never be separated from the preceding characters, as this sign constitutes a part of the same syllable.
- never separate *Kanji/Katakana/Western script* from immediately following *Hiragana*
- never separate *Arabic* numerals from immediately following *Kanji*
- you can separate *Arabic* numerals from immediately following *Hiragana/Katakana/Western script*
- you can separate *Hiragana* from immediately following *Kanji/Katakana/Western script/Arabic numerals*
- you can separate *Katakana* from immediately following *Kanji/Western script/Arabic numerals*
- you can separate *Kanji* from immediately following *Katakana/Western script/Arabic numerals*
- you can separate *Western script* from immediately following *Kanji/Katakana/Arabic numerals*

Here is some sample text. All Kanji are shown in blue, Katakana in light blue, Hiragana in green:

神奈川県の米軍厚木基地内の工事を巡り、談合により入札価格が不当につり上げられたとして、米政府が日本の建設会社など１３社に総額約６億８０００万円の損害賠償を求めた訴訟の控訴審判決が５日、東京高裁であった。

テニス・ＡＩＧオープン第４日（５日・有明テニスの森公園＝ＡＩＧグループ特別協賛、読売新聞社後援）－－男子シングルス３回戦は、世界ランク１位のロジャー・フェデラー（スイス）が、昨年優勝のウェスリー・ムーディ（南アフリカ）を圧倒し、８強入り。

Here is the same text again, with # inserted to mark possible line breaks:

神奈川県の#米軍厚木基地内の#工事を#巡り、#談合により#入札価格が#不当につり#上げら#れた#として、#米政府が#日本の#建設会社など#１３社に#総額約#６億#８０００万円の#損害#賠償を#求めた#訴訟の#控訴審判決が#５日、#東京高裁で#あった。#

テニス・#ＡＩＧ#オープン#第#４日#（５日・#有明#テニスの#森公園＝ＡＩＧ#グループ#特別#協賛、#読売新聞社後援）#―男子#シング#ルス#３回戦は、#世界#ランク#１位の#ロジャー・#フェ#デラー#（スイス）#が#、昨年優勝の#ウェス#リー・#ムーディ#（南#アフリカ）#を#圧倒し、#８強入り。#