An On-Line Japanese Handwriting Recognition System integrated into an E-Learning Environment for Kanji

Steven B. Poggel
steven.poggel@gmail.com

Wednesday 17th February, 2010

Contents

Chapter 1

Japanese Script

The Japanese writing system has a long history. It goes back to around 800 A.D. The Japanese script is in fact a writing system, as Japanese is denoted in a combination of three different scripts: *Hiragana*, *Katakana* and *Kanji*. Kanji is a conceptual script, where each character bears the meaning of one or more semantic concepts and represents morphemes. Hiragana and Katakana are both syllabic scripts, and the individual characters do not bear reference to concepts or even words, but merely to phonological units, usually two phonemes.

In this chapter, the development of the script will be reviewed in section 1.1. In section 1.2 the current Japanese writing system will be exemplified, with a focus on the Kanji in section 1.2.2. Hiragana and Katakana will be reviewed in section 1.2.1, which centres around the Kana scripts. Machine processing of the different Japanese scripts and the difficulties that go along will be demonstrated in section 1.2.4. The difficulties of learning to use the Japanese script will be illustrated in section 1.3.

1.1 A Short History of the Japanese Script

The historical development of the Japanese script is tightly connected to the history of the Kanji characters. Kanji, in Japanese 漢字 (Jap. pron. カンジ / Kanji; Eng. lit. *Han characters*) refers to the 'characters of the Han', meaning the Han Dynasty (206 B.C.-220 A.D.; simplified Chinese: 汉朝; traditional Chinese: 漢朝) (Foljanty 1984). In Mandarin the same characters are referred to as *Hànzì* (simplified Chinese: 汉字; trad. Chinese: 漢字). Note, that the first character 漢 (Chin. 'han', Jap. 'kan', Eng. 'Han') of both the words *Han dynasty* and *Kanji* is identical in Japanese and traditional Chinese, even though it has a different reading in the Chinese and Japanese language. In traditional Chinese the character with the same meaning (汉) has a different shape. This apparent oddity will be explained in greater detail in section 1.1.1.

1.1.1 Historical Development

The Kanji script as developed and coined by the Han is in principle still valid today. It is used alone or in combination with phonetic spelling in China, Japan, Taiwan, Hong Kong. In Vietnam it was used before it was replaced with the Vietnamese alphabet (Viet.: 'quốc ngữ', Eng. lit. 'national language', Eng. 'national script'), a script based on the Latin alphabet. In South Korea the Han characters were in use until they were replaced with Hangul (Kor. with Han characters 韓國語; Eng. 'Korean') (Foljanty 1984).

The Kanji characters were brought to Japan by Koreans living in Japan around 300-400 A.D. Since the Kanji were used by the Koreans to write Hangul they also used it to write Japanese. There was no other Japanese script before that time. Reports about an original Japanese script called *Jindai Moji* (Jap. 神代文字; Eng. 'scripts of the age of the gods') could not be proved. They are now assumed to be a political and speculative invention by Japanese Nationalists in the early 19th. century (Foljanty 1984). According to (Lange 1922) the *Kogo Shūi* (Jap. 古語拾遺; a historical record of the Inbe clan), which was written around 800 B.C. denies the presence of a Japanese native script before the introduction of the Han characters. However, the questions seems irrelevant in the sense, that no longer text or document has been found, written in that script.

In the Christian year 712 an ancestral act of writing was performed at Japanese emperor Temmu's court. Hieda no Are, a member of the guild of the *Kataribe* or reciters, basically a Japanese Griot, dictates the *Kojiki* (Jap. 古事記; Eng. 'Record of Ancient Matters') to Ō no Yasumaro. Ō no Yasumaro wrote the Kojiki, which is not the first written document found in Japan, however it is Japan's oldest attempt to write down spoken Japanese (Grassmuck 1997; Chamberlain 1982).

At the time the Han characters were first used to write Japanese, they were already a developed script. The script was more than 1,000 years old since the characters stabilised to their modern form within the Han period. The first Chinese characters were found on oracle bones from the Shang Dynasty (Chinese 商朝), which ruled

over China some 500 to 600 years within the time period between 1600 B.C. and 1046 B.C. (Grassmuck 1997; Guo et al. 2000).

According to the Kojiki, a scholar called Wani (Jap. 王仁) from Korea brought two foundational Chinese books to Japan, the *Lunyu* (Simplified Chin. 论语; trad. Chinese: 論語; Eng. 'Analects'), also known as *The Analects of Confucius* and the *Qianziwen* (Chin./Jap. 千字文; Jap. pron. センジブン/senjibun; Eng. 'The Thousand Character Classic'), which is a Chinese poem used as a primer for teaching Chinese characters to children. It contains exactly one thousand unique characters (Grassmuck 1997). However, (**?**)Lange1922) demurs strongly against this view and assumes an evolving adoption of the Kanji in Japan. This view seems more plausible, since it has been proved that Japan had contact to Korea even in the time before common era and had contact to China at least from the first century A.D. The Chinese language comprehends more than 40,000 Hànzì characters lexicographically. Only around 25% of those including about 250 *Kokuji* (Jap. 国字; Eng. 'national characters') are in Japanese dictionaries. Only around 2,000-3,000 of those are part of the common characters (Foljanty 1984).

The Japanese Ministry of Education issued a list of 1,850 standard Kanji in 1946 under the name of *Tōyō kanjihyō* (Jap. 当用漢字表; Eng. 'list of Kanji for general use'). The list of Tōyō Kanji was slightly revised and extended in 1981 and comprised 1,945 Kanji as the Jōyō Kanji (Jap. 常用漢字; Eng. 'often used Kanji') (Foljanty 1984). As of 2010 a revised list of 2,131 characters is in official use (Noguchi 2009).

In China there had been a spelling reform in the 1950s, affecting many of the general use characters, resulting in simplified Chinese. In Japan, the Ministry of Education issued it's own reform when the Tōyō Kanji list was introduced. However, the Japanese reform affected a smaller set of characters of only a few hundred and resulted in Shinjitai (Jap. shinjitai: 新字体; Jap. kyūjitai: 新字體; Jap. pron. シンジタイ/shinjitai; Eng. 'new character form'), which replaced the Kyūjitai (Jap. shinjitai: 旧字体; Jap. kyūjitai: 舊字體; Jap. pron. キュウジタイ/kyūjitai; Eng. lit. 'old character forms'). This explains how some characters are still identical in traditional Chinese and Japanese, because they were not affected by any spelling reform, like the aforementioned 漢 (Jap. pron. カン/kan; Chin. pron. 'hàn'), while other characters are different, like the simplified Chinese 'hàn': 汉. Henceforth, and throughout this document, all Japanese characters are in the new character form shinjitai.

1.2 The Modern Japanese Writing System

The Japanese writing system has a complex structure. The three scripts *Hiragana* (section 1.2.1.1), *Katakana* (section 1.2.1.2) and *Kanji* (section 1.2.2), are combined to one writing system. Each script has its task within the system:

- Kanji are used to write lexical morphemes, i.e. content-bearing morphemes.
- Katakana are used to transcribe foreign words, borrowings and nonstandard areas.
- Hiragana are used to write grammatical morphemes and anything else that is not written in one of the other two scripts, e.g. the spoken syllables of a word that should be written with a Kanji character unknown to the writer of that word.

The actual writing system is mainly based on Kanji and Hiragana, catenated to Kanji-Kana blended writing.

(1) a. マリア： 山田さん、 「火の鳥」 という アニメ を もう 見ました
Maria: Yamada-San, [Firebird] say anime OBJ-PARTICLE already seen
か。
QUESTION-PARTICLE
'Maria: Ms Yamada, say, have you seen the Firebird cartoon yet?'
Taken from (Katsuki-Pestemer 2006)

b. マリア：山田さん、「火の鳥」というアニメをもう見ましたか。

In example (1a), the blending of the different scripts can be seen:
Both the foreign name マリア (Eng. 'Maria') and the borrowing アニメ (*anime*, Jap. short for Eng. 'animation') are written in Katakana. Kanji are used for:

- The Japanese name 山田 (*Yamada*).
- The nouns 火 (Eng. 'fire') and 鳥 (Eng. 'bird').
- The verb stem 見 (Eng. 'see').

The rest is written in Hiragana:

- The politeness ending さん (*san*; Eng. equiv. 'Mr/Ms/Mrs') for addressing a person with their name.
- The genitive particle の (*no*) between 火 (Eng. 'fire') and 鳥 (Eng. 'bird'), to yield 火の鳥 (Eng. 'Fire-bird').
- The interjection とうい (Eng. 'say').
- The object particle を (*wo*).
- The adverb もう (Eng. 'already').
- The past tense conjugation and politeness ending of the verb ました (*mashita*).
- The question particle か (*ka*).

Three different scripts are used next to each other in one sentence, indistinguishable for the untrained eye. Example (1b) shows the sentence as it is printed in (Katsuki-Pestemer 2006). Without prior knowledge of the different Japanese scripts it is hard to even distinguish the individual word tokens, as blanks are usually absent in Japanese writing. Other than actually knowing the words, which is not the usual case for a beginner of learning Japanese, often the change of script is the only way to recognise a new token. However, Kanji and Hiragana are often used within the same word, too.

Despite those complexities, other features of the Japanese writing system are simpler than in latin-based alphabetic scrips. For example, there is no capitals or lowercase letters. Each character has a reserved space of roughly the same size. In the following sections, the different scripts will be presented in greater detail. Their composition and use will be discussed.

1.2.1 Kana かな

If the Japanese had abolished the Chinese characters after formation of the syllabic scripts and used only those, studying the Japanese script would be a less complex task. (Lange 1922) reports about attempts to remove the Kanji from Japanese and use the Kana or even Latin script, so called ロマジ (*Romaji*, a Latin or 'Roman' transcription of Japanese; Eng. lit. 'Roman characters'). However, non of those attempts succeeded and both Kana scripts serve as auxiliary scripts to the predominant Kanji characters.

The Chinese characters have been used in two ways in Japanese. Firstly, in order to express the morphological content of a character, but also in order to use the sound of the character as a syllable. The characters that have been used as syllables were transformed to two separate short-hand notations. One way used cursive writing of the sound Kanji, reducing the character graphically, such that its original shape became virtually in-cognisable. This development resulted in the Hiragana script. An example of that kind of reduction is shown in figure 1.1. The other method of reduction uses only one of the character's Graphemes in order to represent the whole

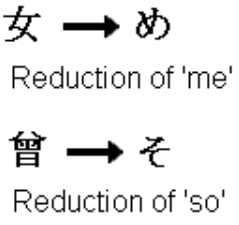

Figure 1.1: Reduction from Kanji to Hiragana

character. That reduction process (see figure 1.2) lead to the development of the Katakana script.

Using cursive or reduced characters became popular in the 9th century already. Both Hiragana ('smoothened Kana') and Katakana ('fragmented Kana') can easily be distinguished from the more complex Kanji. They represent a different linguistic content than the Kanji, namely rather a syllable than a morpheme. The fact that two parallel scripts came into existence can be explained by their use of different social groups. Hiragana are a product of the literately active court ladies, while Katakana were developed in the Buddhist seminaries. The current system knows 46 Hiragana and Katakana for identical syllables (Foljanty 1984). See appendix A.1 for a complete list of characters.

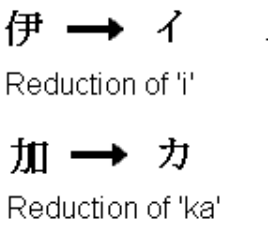

Figure 1.2: Reduction from Kanji to Katakana

1.2.1.1 Hiragana ひらがな

Hiragana is one of the three syllabic scripts. The third one, *Hentaigana* can be neglected, as it is not in active use any more. In principle each character represents a syllable, which can contain either a vowel (like 'u', う), a consonant and a vowel (like 'ta', た) or the nasal consonant ん ('n'). Hiragana are used for any words for which no Kanji exist, like から (*kara*, Eng. 'from') or grammatical particles like the object particle を (*wo*, works like a case marker). Additionally, Hiragana are used, when the Kanji is not known to the writer or reader (Foljanty 1984).

For instance, a writer may even passively know a Kanji, but not be able to actively produce it. Say the Kanji in question was 新 (Jap. pron. atara/あたら and shin/シン; Eng. 'new'). Example (2) shows first the spelling with the Kanji character in (2a), where the Hiragana parts are underlined. Then the alternative spelling without the Kanji character in (2b) (with blanks in between, in order to visualise the individual characters and their readings). The underlined last part is identical, just the Kanji character has been replaced with an alternative Hiragana spelling.

(2) a. 新しい
atarashii
'new'

b. あ た ら し い
a ta ra shi i
'new'

Hiragana can be modified with

- **Dakuten** (濁点, Jap. pron. ダクテン/dakuten; Eng. 'turbid'; Jap. coll. *ten-ten*, Eng. 'dot dot') for syllables with a voiced consonant phoneme. The dakuten glyph (゛) resembles a quotation mark and is directly attached to a character (Foljanty 1984).
- **Handakuten** (半濁点, Jap. pron. ハンダクテン/handakuten; Eng. 'half-turbid'; Jap. coll. *maru*, Eng. 'circle') for syllables with a /p/ morpheme. The glyph for a 'maru' is a little circle (゜) that is directly attached to a character (Foljanty 1984).

(3) a. た → だ, き → ぎ, ふ → ぶ, へ → べ, そ → ぞ
ta → da, ki → gi, fu → bu, he → be, so → zo

b. は → ば, ひ → び, ふ → ぶ, へ → べ, ほ → ぼ
ha → pa, hi → pi, fu → pu, he → pe, ho → po

Example (3) shows the dakuten in use. In (3a) the turbidity, i.e. the transformation from a fortis to a lenis consonant is exemplified. The dakuten can only be applied to characters that begin with the sounds /k/, /s/, /t/ and /h/. For the consonants 'k', 's' and 't' this is natural in the way that the place of articulation remains the same for the consonants 'g', 'z' and 'd' and their corresponding sounds /g/, /z/, /d/, compared to /k/, /s/ and /t/. Assuming the place of articulation, a changed intensity and use or no use of the vocal cords as the natural connection between the characters with and without a *ten-ten*, there is a difference for the sound /b/: Since there is no Hiragana character to match a /p/ sound, the characters with an /h/ sound serve as a basis to form the modified characters with a /b/ sound. Quite logically, since the Handakuten only yields a /p/ sound, the same set of characters, namely, the 'h'-row in the Hiragana character set, is modified in order to obtain the 'p'-row. In (3b) the transformations from the /h/ and /f/ sounds to /p/ are shown in a complete list[1].

[1]The /f/ sound in Japanese is 'lighter' than in English. Concretely, the place of articulation for /f/ in English is labio-dental, whereas in Japanese it is virtually bi-labial. Therefore, the phonation stream creates less friction and the allophone of the /f/ phoneme that is typical for the Japanese language sounds similar to an /h/ sound.

1.2.1.2 Katakana カタカナ

Just like the Hiragana, the Katakana form a syllabic script. They are mainly used to transcribe foreign words like names and borrowings, like マリア (*maria*) and アニメ (*anime*) from example (1). Another use of the Katakana is called *Furigana*, little notations next to Kanji in order to indicate their reading. They were developed around the period of the Heian, from *Manyogana*, Kanji characters that were used to denote pronunciation. For example the character 加, which was used to represent the pronunciation *ka*, was shortened to カ, by leaving out the second part 口 (Hadamitzky 1995). Compare also figure 1.2. For a full table of Katakana characters see appendix A.1.2. The dakuten can be applied to Katakana as well and have the same meaning.

1.2.2 Composition of the Kanji 漢字

1.2.2.1 Typology

In order to study the Kanji and their composition, it is useful to know how they were first indented and built. Integrated as the integral part into the Japanese writing system, despite the reform and the choice different subsets of what is considered the standard character set, the characters are still mainly composed the way as intended by the scholars of the Han period.

From the religious writings on the oracle bones mentioned in section 1.1.1 a secular script emerged. In parallel, the process of graphical abstraction advanced and finished around 100-200 A.D., leaving aside the modern reforms of the 20th century. The invention of the paint-brush around 100 B.C. improved and simplified writing, also the writing surfaces in their order of appearance, bone, stone, metal, wood and then paper, brought further simplification and spreading of writing. Paper and paint-brush offered the possibility to write without hindrance and technical coincidences, therefore it was possible to standardise the characters and improve them from artistic and aesthetic viewpoints.

Figure 1.3: Kanji pictograms

Radical 1	Radical 2	Result	Meaning
日 ('sun')	月 ('moon')	明 ('bright')	both the sun and the moon are 'bright'
人 ('man')	木 ('tree')	休 ('rest')	a man is 'resting' beside a tree
田 ('rice field')	力 ('power')	男 ('man, male')	'a man' is powerful on the rice field
女 ('woman')	子 ('child')	好 ('love')	a woman 'loves' a child
木 ('tree')	木 ('tree')	林 ('wood, grove')	two trees make a 'wood'
木 ('tree')	林 ('wood')	森 ('forest')	three trees form a 'forest'
門 ('gate')	日 ('sun')	間 ('between')	the sun can be seen 'between' the doors of the gate

Table 1.1: Kanji ideograms

The Kanji can be classified according to their building principle:

1. **Pictograms** are graphically simplified images of real artefacts. The examples in figure 1.3 after Kano et al. (1990) show the graphical reduction process. Pictograms are only a small minority among the Kanji, their number ranges around 120. Another 100 pictograms appear as a part of more complex characters (Foljanty 1984).

Class radical	Sound radical /ki/	meaning
土 ('earth')	奇	埼 'spit, promontory, cape'
山 ('mountain')	奇	崎 'promontory, cape, spit'
石 ('stone')	奇	碕 'cape, promontory, spit'
玉 ('jade')	奇	琦 'gem, precious stone'
糸 ('thread')	奇	綺 'figured cloth, beautiful'
馬 ('horse')	奇	騎 'riding on horses'
宀 ('roof')	奇	寄 'to gather'
金 ('metal')	奇	錡 'cauldron, chisel' (Chinese only)

Table 1.2: Kanji phonograms

2. **Ideograms** are combinations of two or more pictographical characters. They often bear a more abstract meaning than a simple pictogram. The abstract meaning of the complex character is meant to be associated with the content of the individual parts. The number of ideograms is fairly small, too. Abstract terms like 'top' (Jap. 上, pron. うえ/ue), 'bottom' (Jap. 下, pron. した/shita), 'left' (Jap. 左, pron. ひだり/hidari), 'right' (Jap. 右, pron. みぎ/migi) and numbers like 'one' (Jap. 一, pron. いち/ichi), 'two' (Jap. 二, pron. に/ni), 'three' (Jap. 三, pron. さん/san), 'four' (Jap. 四, pron. し/shi), 'five' (Jap. 五, pron. ご/go) and so forth can be regarded as parts of the ideograms (Foljanty 1984). See table 1.1 (after (Kano et al. 1990)) for some examples on Kanji pictograms.

3. **Phonograms** are combinations of two Kanji characters. One of those refers to a concept class (for *class characters* or *radicals*, see section 1.2.2.2), while the other character exclusively bears a phonetic value. The content of the second part of a phonogram is not relevant and can be ignored.

   In table 1.2 the character 奇 (Jap. pron. き/ki, Eng. 'strange') is used for the purpose of pronunciation only (/ki/), while the radical defines an object class. Object classes can be categories like 'human and human actions', 'metal', 'horse', 'roof / under a roof' etc. The semantic identity within the *Morphogram* is assembled with two reference figures. The pronunciation part is identical for all characters, it serves as a selection criterion within a semantic class selected by the class radical (Foljanty 1984).

As a character type, phonograms are predominant among the Hànzì. Therefore, the phonogram concept, including the radical concept, was transferred to all Chinese characters. Pictograms that are class radicals themselves, are interpreted as characters with an empty sound radical. As the phonograms are historically the last development step of the Han characters, they constitute a different quality in the Chinese script. Phonograms mark the transition between a non-linguistic pictographic script that does not represent linguistic units, but rather images of objects, to a linguistic script. In principle, there is no difference to an alphabetical or syllabic script. Except, morphemes are represented instead of phonemes or syllables. However, one character often denotes more than one morpheme (Foljanty 1984).

In Japanese, basically the same relation between the Kanji characters and morphemes can be observed. However, the correspondence between the morphemes and the syllables (and thus the characters and syllables) is often missing. Since Chinese has a monosyllabic morpheme structure, a one-to-one correspondence between morpheme, character and syllable can be observed. Congruence of character, morpheme and syllable in Japanese can only be found for Chinese borrowings, but not all of them. The original Japanese vocabulary has multisyllabic morphemes, therefore some impreciseness arises in the graphical reproduction of the morpheme structure (Foljanty 1984).

1.2.2.2 Radicals

In section 1.2.2.1 the function of the class characters or Radicals has been mentioned. The Radicals usually derive from pictograms. The Chinese systematics with 214 Radicals is used in Japan, as well. Among the 1945 Jōyō Kanji, only around 50% of the 214 Radicals is in autonomous use. 22 of them denominate empty classes, 46 denominate classes of only one Kanji. The Radicals are grouped according to their graphical position inside a Kanji. Figure 1.4 shows the groupings of the Radicals after (Foljanty 1984). The groupings of the Radicals account for a preference of certain Radicals to appear in a certain category. The selection of the class Radical is primarily based on semantic criteria. For example, the character 杉 (Jap. pron. スギ/sugi; Eng. 'cedar') both graphemes have the ability to serve as a class Radical. Therefore there is a choice of using either the Radical 木 (Jap. pron. キ/ki; Eng. 'tree') in the *hen* position, or the Radical 彡 (Jap. pron. カミカザリ/kamikazari; Eng. 'hair ornament') in the *tsukuri* position (see Figure 1.4). This is very pictographic, since the leaves of a cedar are long and thin, (see figure 1.5), therefore the character is most likely an ideogram. Now,

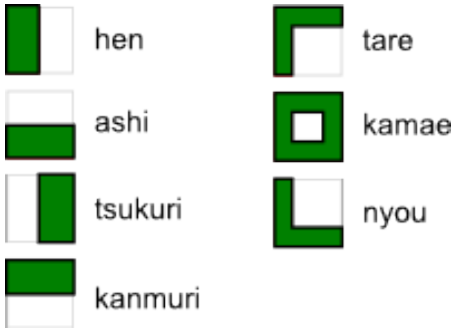

Figure 1.4: Radical positions in Kanji characters

Figure 1.5: Cedar leaves

the categorisation is purely a semantic choice. The cedar has certainly more qualities of a tree than qualities of hair, therefore, the class radical is 木 (tree) and not 彡 (hair). (Hadamitzky 1995) files it under the 木 (tree) class Radical.

1.2.2.3 Readings

Each character has at least one reading. The Kanji script, being a morphemic script does not have a correspondence from character to spoken sound in speech, like alphabetic scripts. Certainly, alphabetic scripts are often far from a one-to-one correspondence between sound and letter, however, the Kanji have their very own ways. There are two main readings of the Kanji:

- *On'yomi* (音読み, Jap. pron. オンヨミ/onyomi; Eng. 'sound reading'), which roughly based on its original Chinese reading.
- *Kun'yomi* (訓読み, Jap. pron. クンヨミ/kunyomi; Eng. 'concept reading, native reading'), which represents the original Japanese word that happens to be written with a Chinese character. The Kun'yomi are further classified into different types of readings.

Because of this, the Kanji that have been developed in Japan only have a Kun-reading. A bit more than half of the official Jōyō-Kanji (999 of 1945) have two readings. The Japanese Kun-reading was part of the adaption process to the Han-characters, where the character was taken for its meaning and underlaid with the Japanese word. The Sino-Japanese reading originated from Chinese borrowings. The complex Chinese pronunciation, however, was adapted to the Japanese phoneme inventory.

Take for example the Kanji 山 (Eng. 'mountain'). The original Japanese word for *mountain* is *yama* (やま), therefore the character 山 has this reading attached as a Kun-reading. The Chinese borrowing is *san* (サン), with the meaning *mountain*. The reading that is used depends on the context (Foljanty 1984). 山 is read Sino-Japanese *san* or in a phonetic assimilation *zan* in:

- 山水, Jap. pron. サンスイ/**san**sui, Eng. 'landscape'
- 火山, Jap. pron. カザン/ka**zan**, Eng. 'volcano'

- 下山, Jap. pron. ゲザン/ge**zan**, Eng. 'descent'
- 富士山, Jap. pron. フジサン/fuji**san**, Eng. 'Mount Fuji'[2]

山 is read in the Japanese Kun-reading in:

- 小山, Jap. pron. コヤマ/ko**yama**, Eng. 'hill'
- 山山, Jap. pron. ヤマヤマ/**yamayama**, Eng. 'mountains'

(Hadamitzky 1995)

1.2.3 Structure of the Japanese Writing System

Having demonstrated the Hiragana in 1.2.1.1, the Katakana in 1.2.1.2 and the Kanji in section 1.2.2, it is now possible to report about the structure of the writing system as such. It now becomes apparent, why a sentence like the one presented in example (1) in section 1.2 is written and spelt the way it is.

(4) 帰りたくない
kaeritakunai
go-home-want-not

'I do not want to go home'

An important part of the blended writing in Japanese are the *Okurigana* (送り仮名, Jap. pron. オクリガナ/okurigana; Eng. 'accompanying characters'). Okurigana are Kana that form a word together with a Kanji character (Lange 1922). For instance, the word in example (4) has only one Kanji character: 帰 (Jap. pron. カエ/kae; Eng. 'go-home'), while the other characters are all Kana characters, that are part of the same word and modify the verb.

- り (*ri*): Flexion of the verb used for *tai*
- たく (*taku*): derivation of *tai* = want
- ない (*nai*): Negation ending

Knowing the two syllabic scripts by heart helps distinguish the Kanji from them. Therefore it is possible to use the three scripts next to each other, even without blanks. The Okurigana further help distinguish different versions of the same Kanji, as they sometimes even partially describe the sounds of a verb stem.

(5) a. 変ーえる
ka-eru
change

'to change something' (transitive)

b. 変ーわる
ka-waru
change

'to change oneself' (reflexive)

Compare for example (5a) and (5b). The Japanese verb ending is る (Hiragana: *ru*). A part of the verb stem are expressed with Hiragana, too. In (5a) it is the え (Hiragana: *e*) while in (5b) it is the わ (Hiragana: *wa*). The Kanji are further used to write nouns, the Hiragana are used for particles and other grammatical functions, while the Katakana are used for foreign words.

[2] 'Mount Fuji' is often called 'Fujiyama' in several languages. Thus, the Kun-reading *yama* of 山 has emerged as a borrowing into other languages, even though in Japan the On-reading is used in the particular case of 'Fuji san'.

1.2.4 Machine Writing of Japanese

Machine processing of the Japanese scripts has been an issue, ever since humans started to automate their writing. Lange (1922) reports in the foreword to the first edition (1896) of his work about difficulties during the printing process. Firstly, there was only one printing office in Germany that was able to print these kind of things at all. Also, there were technical issues, like incompatible letter size and missing characters among the letters available. The manufacturing of the appropriate letters delayed the publication.

> Bei der Drucklegung des Werkes in der Reichsdruckerei, der einzigen Druckerei in Deutschland, welche dergleichen zu drucken im stande ist, stellten sich leider einige Übelstände heraus. Einerseits sind die Kana-Zeichen, welche dieselbe besitzt, zu gross, um als erklärende Zeichen bequem neben die chinesischen Zeichen gesetzt werden zu können, andererseits fehlten unter den 10000 chinesischen Zeichen, die aus Shanghai gekommen, eine Anzahl in Japan ganz gebräuchlicher Zeichen. Diese mussten entweder aus den vorhandenen Teilzeichen, die eine andere Form als die übrigen Typen haben, zusammengesetzt werden oder sie mussten erst geschnitten werden; haupsächlich durch den letzteren Umstand hat sich die Fertigstellung des Werkes verzögert.
>
> When the work was printed in the Reichsdruckerei, the only print shop in Germany capable of printing this kind of work, some mischiefs became obvious. On one hand, the Kana characters that the Reichsdruckerei had were to large to be put next to the Chinese characters as explanatory characters, on the other hand some characters that are very common in Japan were missing among the 10000 Chinese characters that came from Shanghai. The missing characters had to be composed from the partial characters that had a different shape then the other types, others had to be manufactured. Mainly because of the last-mentioned circumstances the completion of the act has been decelerated.

This quote from Lange shows the difficulties an author was facing when trying to print a book. Difficulties in machine writing Japanese also occurred at later stages. The technical and practical problems remain to this day, because of the great number of characters. (Foljanty 1984) reports that in Japan the number of machines needed to print Japanese characters was small, because even business correspondence was mainly done by hand up to the 1970is. Computers worked with Katakana only. There were devices like mechanical-type Japanese typewriters, with around 2000 keys for the different Kanji. A Kanji synthesis system worked with the internal structure of the Kanji. The 'multi-radical lookup' in Breen's (2004) system *WWWJDIC* has a similar concept. A user chooses a number of Radicals from a list and the system computes all Kanji that contain these Radicals. In order to handle word processing the Industry standard are Input Method Editors that come with the operating system. Input Method Editors (IME) use Romaji, the Latin transcription of Japanese and a language model in order to generate a Kanji list, appropriate to the spelling of a word. That is necessary since one Kanji can have several readings and Japanese is rich in homophones (Yuan et al. 2005; Hadamitzky 1995). Certainly the most modern input method for Japanese characters is handwriting recognition. The difficulties of other input methods show that there is great need for a better and faster method of input. See section **??** for a description of research efforts in order to provide technology for using handwriting as an input method for Japanese.

1.3 Difficulties of the Japanese Script for Learners

There are at least four ways to learn the Japanese Script. With *learning the Japanese script* we mean *learning the Kanji*. For the Hiragana and Katakana are only small sets of syllabic characters, the effort to study those is only slightly larger than learning the Latin alphabet. The number of different learning methods and the diversity of advice that can be found on this topic on the Internet and in teaching books suggests that learning the Kanji is regarded a difficult issue. It is apparent that the difficulties lie in the complexity of the Kanji, but also on the structure of the writing system as such. Stahlmann (2004) reports the subjective impression of many learners that Chinese is among the most challenging languages and even seems daunting. Japanese is no different, as most of the difficulties of the Chinese language are characteristics of Japanese as well. The pronunciation of Japanese is comparatively simple for a European learner, as the phonetic inventory of the language is similar to that of English (Tsujimura 2007). On the other hand, the Japanese writing system with its three different scripts and several readings of a character has other interweavements as demonstrated in the previous sections. Despite the difficulties, the interest in the Japanese language and culture is unbowed:
There are many books available centred around cultural interchange between the western and Japanese culture. For example, (Haschke and Thomas 2008) present German words of Japanese heritage, i.e. a miniature etymological lexicon. The sheer existence of that type of book on the free market suggests that Japanese language and culture radiate fascination into the western world. Therefore, there is a great need for learning methods.

The four groups of learning methods are:

1. **Writing Repetition** - Each Kanji has to be written several times until its meaning and the readings are full known to the learner. Most Kanji books generally follow this concept.

2. **Flash Cards** - Instead of writing the Kanji, the flash card method teaches the passive knowledge. Flash cards for Kanji are readily available and can be home made easily.

3. **Mnemonic methods** - Methods that use mnemonics in order to help a learner memorise the Kanji.

   (a) **Textual mnemonics** The *Heisig method* assigns a unique mnemonic to each Kanji in order make it more memorable. These mnemonics are little stories oriented toward the shape of a Kanji and do not necessarily have a connection to the actual meaning of a Kanji. They are just mnemonics. The Heisig method is a two-part method. In the first part the learner studies all the Kanji and their translations, in the second part the Japanese readings are introduced (Heisig and Rauther 2007).

   (b) **Visual mnemonics** *Kanji Pict-O-Graphix* is another mnemonic method. In this method the Kanji are depicted with images that represent the meaning of the Kanji. The stories are therefore more visualised than in the Heisig method (Rowley 1992).

1.3.1 Learner's Problems

1. **Similar Kanji.** In order to be able to distinguish characters like 営 in 営業 and 管 in 管理 a high level of concentration is necessary. Especially, when Japanese is mostly written on the computer and not by hand, the muscular memory does not help any more. Therefore the more Kanji are known, the more difficult it becomes to distinguish similar Kanji.

2. **Compounds.** Often, new vocabulary is studied, rather than new Kanji. The vocabulary are often composed of Kanji characters that are already known. Studying time is needed in order to study the vocabulary and compounds. Therefore it is difficult to study new Kanji at the same time.

3. **Unusual readings.** Since Kanji have several readings, it is useful for a learner to study the most frequent ones. For example, the Kanji 雪 (Eng. 'snow') has two standard readings (see section 1.2.2.3 for more on readings):
   The Chinese reading (*on-reading*) is セツ/setsu, while the Japanese reading (*kun-reading*) is ゆき/yuki. These two readings occur in very usual Japanese words, having a relation to *snow*. *Setsu* occurs in 雪害 (Jap. reading セツガイ/*setsu*gai; Eng. 'damage through snow') and 新雪 (Jap. reading シンセツ/shin*setsu*; Eng. 'fresh snow'). *Yuki* occurs in 初雪 (Jap. reading ハツユキ/hatsu*yuki*; Eng. 'first snow of the season'), 大雪 (Jap. reading オウユキ/ou*yuki*; Eng. 'heavy snow') and 雪合戦 (Jap. reading ユキガッセン/*yuki*gassen; Eng. 'snowball fight'). Similarly, the Kanji 崩 (Eng. 'break down, destroy') has two standard readings:
   The Chinese reading (*on-reading*) is ホウ/hou, while the Japanese reading (*kun-reading*) is くず/kuzu. The Kanji 崩 in the reading *kuzu* does not appear as an individual word, in the following examples it is used as a verb stem only. These two readings occur in Japanese words related to *break down*. *Hou* occurs in 崩御 (Jap. reading ホウギョ/*hou*gyo; Eng. 'death of the emperor'). *Kuzu* occurs in 山崩れ (Jap. reading ヤマクズレ/yama*kuzu*re; Eng. 'landslide') and 切り崩す (Jap. reading キリクズス/kiri*kuzu*su; Eng. 'erode'). The compound of the two, forms a logical unit 雪崩 (Eng. 'avalanche', which could be described semantically as a *'snow landslide'). The reading however, is quite unexpected. Instead of a combination of 'yuki' and 'kuzu(re)' to yield **yukikuzure*, the correct reading is *nadare* (Jap. ナダレ). Unexpected readings like that can be a frustrating and exhausting experience for a learner. The reading *nadare* in this example is a *Jukujikun*. Jukujikun (Jap. 熟字訓; pron. ジュクジクン/jukujikun; Eng. 'compound kun readings') are specialised Kun-readings that only occur in fixed compounds, comparable to *irregular* words in English (Wydell 1998).

4. **Alternative Kanji.** Some Kanji have the same meaning and reading, yet look differently. For instance the number *one* can be written 壱 or simply 一.

5. **Homophones.** Japanese has several homophones. An extreme example is きょう (*kyou*): Around 80 Kanji have at least an alternative reading *kyou*. Also, for example, there are three Kanji that can be read アツイ/atsui,namely, 熱い (Eng. 'hot' - for objects or feelings), 暑い (Eng. 'hot' - for the weather) and 厚い (Eng. 'thick' - for clothes; 'kind, warm' - as a characteristic of a person). This can be confusing for a learner.

6. **Infrequent Jōyō Kanji.** There are Kanji that are in the official list of daily use Kanji, but in fact they are actually not frequently used. These can be forgotten easily by a learner, since those Kanji do not appear often in Japanese texts.

7. **Non-Jōyō Kanji.** The fact that a character is not in the list of frequently used characters does not mean that the character is negligible. It may well be necessary to know, at least passively.

Appendix A

Japanese Language

A.1 Kana かな

A.1.1 Hiragana ひらがな

Hiragana is a syllabic script. For a description of use and structure of Hiragana see section 1.2.1.1. A full table of the Hiragana script can be found in table (A.1). Table (A.1) is adapted from (Hadamitzky 1995).

	a	i	u	e	o	n
∅	あ	い	う	え	お	ん
k	か	き	く	け	こ	
s	さ	し	す	せ	そ	
t	た	ち	つ	て	と	
n	な	に	ぬ	ね	の	
h	は	ひ	ふ	へ	ほ	
m	ま	み	む	ね	も	
y	や		ゆ		よ	
r	ら	り	る	れ	ろ	
w	わ				を	

Table A.1: The full set of Hiragana characters

A.1.2 Katakana カタカナ

Katakana is the second syllabic script in use in the Japanese language. For a more detailed description of what it is used for, see section 1.2.1.2. A full table of the Katakana script is depicted in table (A.2). Table (A.2) is adapted from (Hadamitzky 1995).

	a	i	u	e	o	n
∅	ア	イ	ウ	エ	オ	ン
k	カ	キ	ク	ケ	コ	
s	サ	シ	ス	セ	ソ	
t	タ	チ	ツ	テ	ト	
n	ナ	ニ	ヌ	ネ	ノ	
h	ハ	ヒ	フ	ヘ	ホ	
y	ヤ		ユ		ヨ	
r	ラ	リ	ル	レ	ロ	
w	ワ				ヲ	

Table A.2: The full set of Katakana characters

List of Figures

Listings

List of Tables

References

Breen, J. (2004, August). Multiple Indexing in an Electronic Kanji Dictionary. In M. Zock and P. S. Dizier (Eds.), *post COLING Workshop on Enhancing and Using Electronic Dictionaries*, Geneva, Switzerland. COLING: International Committee on Computational Linguistics.

Chamberlain, B. H. (1982). *The Kojiki: Records of Ancient Matters* (2nd ed.). Boston, USA: Tuttle Publishing.

Foljanty, D. (1984). Die japanische Schrift (in german). In Institut für deutsche Sprache Mannheim, T. Kaneko, and G. Stickel (Eds.), *Japanische Schrift, Lautstrukturen, Wortbildung*, Volume 1 of *Deutsch und Japanisch im Kontrast*, Chapter 2, pp. 29--63. Heidelberg: Julius Groos Verlag.

Grassmuck, V. (1997). Die japanische Schrift und ihre Digitalisierung (in german). In W. Nöth and K. Wenz (Eds.), *Reden über Medien*, Volume II of *Reihe Intervalle. Schriften des WZ*. Kassel: Hochschulverlag.

Guo, Z., K. Liua, X. Lua, H. Maa, K. Lia, S. Yuanb, and X. Wub (2000, October). The Use of AMS Radiocarbon Dating for Xia–Shang–Zhou Chronology. In *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, Volume 172, pp. 724--731. 8th International Conference on Accelerator Mass Spectrometry: Elsevier B.V.

Hadamitzky, W. (1995). *Kanji und Kana*, Volume 1 of *Kanji und Kana*. Berlin: Langenscheidt KG.

Haschke, B. and G. Thomas (2008). *Kleines Lexikon Deutscher Wörter Japanischer Herkunft - von Aikido bis Zen (in German)* (1st ed.). Munich: Beck.

Heisig, J. W. and R. Rauther (2007). *Die Kanji lernen und behalten (in German)* (2 ed.), Volume 1 of *Die Kanji lernen und behalten*. Göttingen: Klostermann.

Kano, C., Y. Shimizu, H. Takenaka, and E. Ishii (1990, January). *Basic Kanji Book*, Volume 1. Bonjinsha.

Katsuki-Pestemer, N. (2006). *Grundstudium Japanisch 2 (in German)* (2nd ed.), Volume 2 of *Grundstudium Japanisch*. Troisdorf: Bildungsverlag EINS.

Lange, R. (1922). *Einführung in die Japanische Schrift (in German)* (2nd ed.). Number 15 in Lehrbücher des Seminars für orientalische Sprachen. Berlin: Walter de Gruyter.

Noguchi, M. S. (2009, October 21). Get set for next year's overhaul of official kanji. *The Japan Times Online*. Online. Retrieved 2010-01-14 from `http://search.japantimes.co.jp/cgi-bin/ek20091021mn.html`.

Rowley, M. (1992, August). *Kanji Pict-O-Graphix: Over 1,000 Japanese Kanji and Kana Mnemonics*. Berkeley CA, USA: Stone Bridge Press.

Stahlmann, R. (2004). Didaktische, inhaltliche und funktionelle Optimierung einer selbst entwickelten Chinesischlernsoftware (in German). Master's thesis, Offenburg University of Applied Sciences, Offenburg, Germany. Manuscript committee: Prof. Dr. Roland Riempp (supervisor), Prof. Dr. Thomas Breyer-Mayländer.

Tsujimura, N. (2007). *An Introduction to Japanese Linguistics* (2nd ed.). Blackwell Textbooks in Linguistics. Oxford: Blackwell Publishing.

Wydell, T. N. (1998). What Matters in Kanji Word Naming: Consistency, Regularity, or On/Kun-Reading Difference? In K. Tamaoka and C. K. Leong (Eds.), *Cognitive Processing of the Chinese and the Japanese Languages*, Volume 14 of *Neuropsychology and Cognition*, pp. 359--373. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Yuan, W., J. Gao, and H. Suzuki (2005, September). An Empirical Study on Language Model Adaptation Using a Metric of Domain Similarity. In R. Dale, K.-F. Wong, J. Su, and O. Y. Kwong (Eds.), *Natural Language Processing – IJCNLP 2005*, Lecture Notes in Artificial Intelligence, pp. 957--968. Heidelberg: Springer.

Document created on Wednesday 17th February, 2010 at 16:22