Faculty of Computing and Information Technology

Department of Robotics and Digital Technology

Technical Report 93/13

A Japanese Electronic Dictionary Project (Part 1: The Dictionary Files)

J. W. Breen

November 30, 1993

Enquiries:-

Technical Report Coordinator
Robotics and Digital Technology
Monash University
Clayton VIC 3168
Australia

tr.coord@rdt.monash.edu.au +61 3 565 3402

Contents

Abstract

Electronic multi-lingual dictionaries have seen considerable development in the last decade. The standardization of coding systems for the orthography of many Asian languages in the same period, combined with the increased availability of low-cost micro-electronic storage and display systems has opened up considerable demand and potential for dictionary systems in these languages. This report describes an on-going project to develop and maintain a comprehensive electronic Japanese-English dictionary capable of use within a variety of search-and-display, electronic-text reading support, and machine translation environments. The project files are being developed in the public domain. The dictionary files have, at the time of writing, attained the status of being the major freely available electronic repository of Japanese-English dictionary material in the world.

Keywords

dictionary, lexicography, Japan, JIS, WNN, SKK, kanji, hiragana, katakana, Unicode

Chapter 1

Introduction

1.1 Project Overview

This paper describes an on-going project to develop and maintain a comprehensive electronic Japanese-English dictionary capable of use within a variety of search-and-display, electronic-text reading support, and machine translation environments. The project files are being developed in the public domain. The project, which was initiated and coordinated by the author, has been assisted by the volunteer labour of many people who have selected, translated, formatted and edited material for the dictionary.

The dictionary files developed in the project consist of:

1. EDICT, the main dictionary file of Japanese words with their pronunciations and English translations;

2. KANJIDIC, a file of explanatory information and indices for the 6,353 kanji characters in the JIS Level 1 and 2 character sets described in the JIS X 0208-1990 standard.

The EDICT and KANJIDIC files have, at the time of writing, attained the status of being the major freely available electronic repository of Japanese-English dictionary material in the world, and are being used by thousands of researchers, teachers and students via a large number of software packages which have been developed to use or incorporate them. The EDICT file now has approximately 70,000 entries, of which some 30,000 are Japanese proper names.

An adjunct to the development of the dictionary files is the development of software which accesses and displays the material in a variety of contexts. An additional paper is in preparation which describes a number of software systems developed by the author and others to carry out this task.

1.2 Japanese Orthography and Text Processing

The Japanese language has a unique system of orthography which poses particular challenges to both computer text processing and the lexicographer. This paper assumes on the part of the reader an elementary knowledge of Japanese orthography, an overview of which is included in Appendix A. For a fuller description of the Japanese writing systems, the reader is directed to an introductory text such as Neustupný [1], or O'Neill[2]. An overview is also in Unger [3] and a more historical treatment in Seeley [4].

The ability to store and process the many thousands of characters used in Japanese orthography is central to an electronic dictionary system. A basic knowledge of the mechanisms used to store and process Japanese characters is assumed in this paper. An overview of the coding systems is included in Appendix B, and the reader is referred to Lunde [5] for a thorough treatment of the subject.

Chapter 2

Japanese - English Dictionaries

Japanese dictionaries take two general forms. The first is a word-oriented dictionary of the form found in most languages. Such dictionaries are a lexically ordered set of entries, each of which begins with a head-word and contains synonyms, explanations, etc. In Japanese, such dictionaries usually have their headwords in *kana*, and are ordered according to the standard *gojuon* [1] order. Both *katakana* and *hiragana* headwords are used, and are treated as equal for the purpose of ordering. There are literally hundreds of such dictionaries in print and use in Japan. The major and most authoritative is the *kojien* [6] (広辞苑), which is the Oxford English Dictionary of Japanese dictionaries.

The second form of dictionary, the character dictionary, is unique to languages such as Chinese and Japanese, which use non-phonetic symbols. Japanese character dictionaries, known as *kanwajiten* (漢和辞典: Chinese-Japanese dictionary), typically have the characters grouped by radical, that is, an identifiable element of the character, usually occurring at the left or top of the character. Within each radical group, characters are ordered by stroke count; either the total number of strokes, or the number of strokes excluding the index radical. The radicals used for this purpose are usually based on the 214 classical radicals of Chinese orthography. Use of such a dictionary requires some skill and practice in both identifying the index radical and counting the strokes that comprise the character. Often a separate index is available which enables each character to be found by either its *on* or *kun* reading. The information in each entry typically includes the full set of *on* and *kun* readings, the meaning(s) associated with the character and a selection of the major compounds which begin with the character. Larger dictionaries also include the major variants of the character and some etymological information.

Again, there are many *kanwajiten* in use in Japan. The major reference dictionary is the 13-volume Morohashi *daikanwajiten* (大漢和辞典)[7].

Dictionaries which combine Japanese with another language such as English obviously need to have sections or parts which cover both Japanese-English and English-Japanese. They fall into two groups: those prepared for Japanese speakers and those prepared for English speakers. There are many dictionaries in the first group, and relatively few in the second. The reason for the existence of the two groups of dictionaries lies with the nature of written Japanese. Without at least familiarity with the *kana*, and in many cases a large number of the *joyo* kanji, dictionaries prepared for the Japanese domestic market are of

[1] 五十音: fifty sounds. The standard 5 by 10 table of *kana* which is usually used for syllable ordering.

little use to non-speakers of Japanese. The following examples from two such dictionaries illustrate this point:

> **universe** n. **1** 宇宙 (cosmos). **2** 万有, 天地万物, 森羅(しんら)万象. **3** 世界 (world); (転じて) 満天下(の人々), 全人類 (all mankind), etc. etc. [8]

and

> しんり ［真理］a trial. その事件は目下〜中だ the case is *under (on) trial.* [9]

Dictionaries prepared for those with limited Japanese reading skills usually rely on the use of *romaji*, particularly in the headwords. As such dictionaries tend to be aimed at the market of less skilled users, both the quantity and quality of the entries are limited. The following examples are from two such dictionaries.

> **universe** (n.) uchū [宇宙]; zen-sekai [全世界]. [10]

and

> **uchū** 宇(う)宙(ちゅう) N. universe, cosmos [11]

As students progress in Japanese studies, the dictionaries prepared for Japanese speakers become more accessible. This applies particularly to Japanese-English dictionaries. English-Japanese dictionaries remain a problem, and it is often claimed that there is no English-Japanese dictionary available which adequately serves the requirements of moderately advanced students. One recent dictionary attempts to bridge this gap. It is essentially a version of a domestic English-Japanese dictionary which has had the English pronunciations removed, and the readings of all Japanese words which are in kanji added (in the form of *furigana*, i.e. small *hiragana* above each character). While this is a considerable advance, it still lacks sufficient contextual information in English to enable a non-speaker of Japanese to select between the various translations. The following example is from that dictionary:

> **universal** a. 宇宙[うちゅう][2]の, 万有[ばんゆう]の, 全世界[ぜんせかい]の, 普遍的[ふへんてき]な, 万能[ばんのう]の. [12]

For several decades there have been quite adequate character dictionaries available to English speakers. The Nelson [13], which sets the benchmark in this type of dictionary, has details of over 5,000 characters, and 70,000 compounds. In recent years several other dictionaries have been published, including the Spahn & Hadamitzky[14] which broke new ground by also listing compounds which have the character in reference in other than the first position.

[2]The hiragana denoted [..] appears as furigana.

Chapter 3

The Main Dictionary File

3.1 The EDICT File

The initial dictionary used in the project was the EDICT file supplied as part of the MOKE (Mark's Own Kanji Editor) PC text editor package, developed by Mark Edwards. The original EDICT, despite its acronym (English DICTionary), was a basic Japanese-English dictionary text file with the following format:

kanji-headword [*yomikata* [1]] /English-1/English-2/../

or

kana-headword/English-1/.....

This format is similar to that used in MOKE's *henkan* [2] files, and was based on the file format used in the SKK [3] FEP [4].

As originally released with MOKE, EDICT had under 2,000 entries, compiled by Edwards. Like MOKE's reference files, it was encoded using EUC. MOKE provides a facility the search the file sequentially for a text string occurring in the English fields, and also to match on a kanji headword encountered in a text file. It also can append extra records to the EDICT file.

As compatibility with MOKE, which at the initial stages of the project was the only readily available MS-DOS Japanese text editor, appeared to be desirable, the EDICT file format was retained. Edwards gave his permission for the original EDICT entries to be included in the project file. Hindsight indicates that the EDICT format has been less than ideal, but as there is now a considerable number of software packages developed around that format, it is unlikely to change.

[1] 読み方: way of reading. The term commonly used for a phonetic transliteration of a word.

[2] *henkan*, or conversion files play an important part in Japanese text editors, as they provide the mappings needed to translate kana sequences into the appropriate kanji.

[3] Simple Kana to Kanji, an FEP package developed by Masahiko Sato at Tohoku University for Nemacs.

[4] FrontEnd Program; a term commonly used in Japan for software which takes keystrokes (either romaji or kana) and converts the text into the appropriate kanji.

3.2 File Structure

A number of file design goals were developed for the project:

- the file structure would be accessible by users with appropriate text editors. This meant, in effect, that the master files would be held in a text form. No indexing information would be held in the file itself, and it would be the task of software using the file to reformat the file into, say, an indexed database structure, or to provide an index system external to the file. As part of this, it was necessary that there be ready identification of tokens consisting of kanji and/or kana strings or English words for the purposes of indexing and searching.

- the file would lend itself, as far as possible, to bidirectional use, i.e. in support of both Japanese-English and English-Japanese searches. The one-to-many nature of dictionary entries presented a problem which could not be totally resolved without the introduction of a file structure which included a complex indexing arrangement. This would, however, violate the first goal.

  After some experimentation with various indexing, search and display techniques, it was concluded that this goal could be met adequately with a file of the EDICT structure, provided the English translation fields for each Japanese entry contained enough contextual information to enable a user, following a search on an English keyword which resulted in a display of all the entries which contained that word, to select the appropriate Japanese entry.

3.3 Lexicographic Details

In the development of the contents of the file, the following lexicographic principles have been followed:

- no use is made of the romaji system, except when it is appropriate to include it in an English field, such as with proper nouns, and with objects such as *sukiyaki* and *sushi*, which are well-known by their Japanese names. This decision was made in part on principle: the Japanese language is rarely written seriously using romaji, and partly to encourage the learning and use the kana at an early stage in language acquisition.

- as is the normal practice with Japanese dictionaries, all Japanese verbs and adjectives are listed in their dictionary form [5]. Users of such dictionaries are expected to be able to identify the dictionary form of a word. With electronic dictionary files,this presents a problem only when an attempt is made to match entries against text which contains inflected forms of such words. English verbs have been included in their infinitive form, thus removing the need to mark that they are verbs. In a few cases it has been necessary to distinguish between transitive and intransitive (Japanese) verbs with the (vt) and (vi) marker.

[5] Japanese has no infinitive form for verbs. The usual practice is to record verbs in dictionaries in their familiar or "plain" non-past form.

- adjectival nouns or quasi-adjectives (形容動詞: *keiyodoshi*), also sometimes referred to as na-adjectives because their nonpast prenominal form ends with a な (na), are recorded without their な ending, but will have the modifier (an) [6] included as a grammatical indicator.

- words, such as compounds and loanwords, which are commonly combined with the verb する (*suru*) are given the grammatical marker (vs). Such entries do not have the する in either the headword or the *yomikata*, and the English fields are defined in the form of nouns or participles.

  Example: 料理 [りょうり] /cooking (vs)/

- nouns which may take the genitive case particle の (*no*) are marked with the modifier (a-no).

- differing Japanese spellings of a headword arising from the inconsistent use of *okurigana* [7] are handled by including separate entries for each common spelling. (This is an inelegant but expedient way of handling what the Japanese refer to as the "okurigana problem", as it enables text-reading software to make appropriate matches with entries. As it adds to the size of the file, and could be seen to be encouraging poor writing practices, it is hoped that at some time such software will incorporate methods for the deterministic identification of *okurigana* patterns, thus enabling the entries in the file to be confined to the standard spellings.)

- it was decided to include proper nouns in the file; both the family and given names commonly used in Japan, and place names. Such entries do not commonly occur in dictionaries; however it was considered worthwhile given that the dictionary file is intended to assist, *inter alia*, with text reading.

- in keeping with the nature of the project, the spelling of the English fields in the entries has been left in the style of the various contributors. Thus, the file contains a mixture of British and North American spellings. As far as possible, translations have been edited to ensure that they are broadly comprehensible to speakers of all varieties of English, and that major national variants are identified.

- a number of abbreviated indicators are used in parentheses in the English fields to provide grammatical and other information about the entries. Among these are:

vt	transitive verb
vi	intransitive verb
id	idiomatic expression
col	colloquialism
vul	vulgar expression or word
pn	person name (family or given)

[6] an: adjectival noun

[7] *okurigana* is the term for writing part of a verb form in kana. Practices vary in okurigana usage, for example the word *ikebana* (flower arrangement) is often written both 生け花 and 生花. The former is preferred by purists, as it indicates the reading of the 生 kanji, but the latter is also often seen.

pl	place name
giv	given name
fam	familiar language
pol	polite (丁寧語: *teineigo*) language
hum	humble (謙譲語: *kenjogo*) language
hon	honorific or respectful (尊敬語: *sonkeigo*) language
pref	prefix
suf	suffix
uk	word usually written using kana alone
uK	word usually written using kanji alone

Here are examples of entries in the file:

さよなら /good-bye/

しくしく泣く [しくしくなく] /to sob/to weep/

アーキテクチャ /architecture/

蛇口 [じゃぐち] /faucet/tap/

移籍 [いせき] /changing household registry/

萎縮 [いしゅく] /withering/atrophy/contraction/

衣川 [ころもがわ] /Koromogawa (pn,pl)/

遺詠 [いえい] /posthumous song or poem/

医者 [いしゃ] /doctor (medical)/

育つ [そだつ] /to raise (child)/to be brought up/to grow (up)/

3.4 File Compilation

As mentioned above, the original EDICT file had fewer than 2,000 entries. A further 2,000 entries were keyed by the author from a first-year student dictionary [15] published by Swinburne Institute of Technology (with the kind permission of the authors), and several hundred more entries were added by the author from lists compiled during Japanese reading and study.

Since the release of this initial version of EDICT in mid-1991, there has been steady flow of entries from persons interested in adding to the coverage of the EDICT file. Several individuals had compiled extensive personal vocabulary lists, which they contributed. The Sony company in Japan permitted the inclusion of a significant in-house file. Many entries were derived from a Japanese-German dictionary file in the EDICT format compiled by Helmut Goldenstein from the *Langenscheidt* edited by Hadamitzky[16]. As Goldenstein had permission to make this file freely available, the entries which were not already in the EDICT file were translated into English.

A major task in developing such a dictionary file is the identification and entry of the tens of thousands of kanji compounds which make up the written language, along with their *yomikata*. Fortunately this task has been greatly assisted by the prior work of the two main public domain FEP systems in Japan: the above-mentioned SKK system, and the even larger WNN[8] FEP system developed jointly by the Kyoto University Research Institute for Mathematical Sciences, OMRON Corporation and ASTEC Corporation. As part of the WNN project, a public project was undertaken in which academics, researchers and students throughout Japan submitted kanji/kana entries to go into the large *henkan* file, now known as the WNN Pubdic. MOKE's main *henkan* file was compiled from the SKK and WNN files, and the existence of these files in the public domain has been a major factor in the development of EDICT.

Many thousands of entries in EDICT have been derived directly fom the SKK and WNN sources. These entries are:

1. loanwords (外来語: *gairaigo*). As one feature of these systems is the ability to generate a loanword from its English original, there are many English/loanword pairs which only required reformatting to be included in EDICT.

2. proper nouns. The WNN Pubdic has separate files of *jinmei* (人名: person name) and *chimei* (地名: place name) words, which were suitable for immediate conversion into the EDICT format. In the case of proper nouns, the English translation field is used for the romanized transliteration of the words, e.g.

   広島 [ひろしま] /Hiroshima (pl)/

   Note that a slightly modified version of the common Hepburn system of romanization has been used throughout the EDICT file. In Japanese the lengthened form of the vowel "o" (お) is almost always written おう (ou), whereas in romaji it is usually written "oo" or "ō", and on occasions "oh". In order to reflect the Japanese orthography more accurately, and to be compatible with the usage of common FEP systems, the "ou" convention has been followed throughout. Thus 東北 (とうほく) will appear as Touhoku, not Toohoku or Tōhoku. The only exceptions to this are selected entries for Tokyo, Osaka and Kyoto, which are well-known in that form, and may not be recognized in the more orthographically correct versions of Toukyou, Oosaka and Kyouto.

[8] An acronym formed from Japanese sentence "Watashi no Namae wa Nakano desu" (my name is Nakano), as a design goal of the WNN project was to be able to enter such a sentence, and have the correct kanji and kana [私の名前は中野です] selected automatically.

Chapter 4

The Character Dictionary File

The second file necessary for a complete dictionary system is the character file, containing information about each of the kanji commonly used in Japanese. Such a file fulfils a number of roles in a dictionary system:

1. to record against each character the key items which can be used to identify the character, e.g. radical, stroke count or reading;
2. to record other information of interest in scholarship or translation, such as the English meaning(s) associated with the character, references to major dictionaries, coding schemes, etc.
3. to provide a starting point for dictionary software, once a character has been identified, to search the main dictionary file and display compounds containing that character.

4.1 Character File Structure

In keeping with the goals of the main dictionary file, the character file was compiled as a simple text file (KANJIDIC) with one line for each of the 6,353 kanji characters in the JIS X 0208-1990 standard.

Each line consists of a number of space-separated fields. The fields in the lines of information are:

1. the kanji itself.
2. the 4-byte ASCII representation of the hexadecimal coding of the two-byte JIS encoding of the character. This field is to facilitate editing and manipulation of the file.
3. information fields, beginning with a unique identifying letter or pair of letters. At the time of writing the field types in use are:
   - U[hexnum] - The Unicode encoding of the kanji, one per line. (See Appendix C for information on the Unicode system.)

- N[num] - the index number in the Nelson dictionary. There will be one per line unless the character is not in Nelson, or is considered to be a non-standard version, in which case there will be a {see Nnnn} cross-reference appended to the meanings fields.
- B[num] - the radical (部首: *bushu*) number, one per line. As far as possible, this is the radical number used in Nelson. In a number of cases Nelson classifies the character differently from the classical or historical usage. In these cases, the classical or historical radical numbers follow as a separate C[num] entry or entries.
- C[num] - the historical or classical radical number (where this differs from the B[num] entry.)
- S[num] - the stroke count, at least one entry per line. If there is more than one field, the first is considered the accepted count, while subsequent ones are common variants.
- G[num] - The *joyo* grade level (i.e. the elementary school grades in which the characters are taught in Japan.) There is at most one per line, with the following coding:

  G1 through G6 indicate *joyo* grades 1-6.
  G8 indicates the remaining (seondary school) general-use characters.
  G9 indicates *jinmeiyo* (人名用: for use in names) characters. These kanji lie outside the *joyo*, but are permitted in names.
- H[num] - the index number in Halpern [17]. There is at most one per line.
- F[num] - the frequency-of-use ranking, as reported in [17]. At present, the 2,135 most-used characters have a ranking. Those characters that lack this field are not ranked.
- P[code] - the SKIP pattern code, similar to that used in Halpern[17]. The code is of the form P[num]-[num]-[num]. See Halpern for a description of his SKIP pattern code [1].
- Q[num.num] - the four-corner encoding of the kanji. This is a Chinese classification system which categorizes the stroke combinations in the corners of the kanji. Some older Japanese character dictionaries, such as Morohashi, include a four-corner index.
- MN[num] - the index number in the Morohashi *daikanwajiten*. In the few cases where JIS kanji are not in Morohashi, the index indicates the last number with the same Bushu and stroke count.
- MP[num.num] - the volume and page in Morohashi on which the entry is to be found for the kanji.
- E[num] - the index number in Henshall[18]. Only the *joyo* kanji are covered by this book.

[1] At the time of writing, this field is not included in the distributed version of the file, as Halpern has claimed patent and copyright protection for the system.

- Y[xxxxx] - the PinYin (Mandarin Chinese) reading of the character, including the tone. There may be several PinYin fields for each kanji, however the *kokuji* (国字: characters developed in Japan) do not have these fields.

4. the readings used for the characters. As is a common practice in character dictionaries, the *on* readings are in *katakana* and the *kun* readings in *hiragana*. Readings used as prefixes and suffixes are indicated by a trailing and leading "-" respectively, and a "." is used to separate a reading from its *okurigana*. Note that the readings are not in any particular order, and are based on common usage rather than any officially approved lists.

5. the English translations and/or notes. Each field is encapsulated by a pair of braces {..}.

Here are examples of the entries for a number of common kanji in the KANJIDIC file:

雨 312b U96e8 N5042 B173 S8 G1 H3561 F595 P4-8-1 Q1022.7 MP12.0001 MN42210 E3 Yyu3 Yyu4 ウ あめ あま- -さめ {rain}

屋 3230 U5c4b N1392 B44 S9 G3 H3098 F318 P3-3-6 Q7721.4 MP4.0149 MN7684 E236 Ywu1 オク や {roof} {house} {shop} {dealer} {seller}

映 3147 U6620 N2118 B72 B73 S9 G6 H892 F362 P1-4-5 Q6503.0 MP5.0806 MN13838 E81 Yying4 Yang3 エイ うつ.る うつ.す は.える -ば.え {reflect} {reflection} {projection}

4.2 Character File Compilation

The compilation of the KANJIDIC file began in late 1991 when a contributor to the main file obtained and provided the basic (*bushu*, stroke count and reading) information for the JIS Level 1 kanji. This was extended shortly afterwards to include the JIS Level 2 kanji by the provision of a file of the *bushu* and stroke counts from a public source in Japan. The Sony dictionary file provided many of the Nelson indices, readings and initial English translations.

In the following eighteen months there was a large amount of painstaking editorial work carried out by several contributors to verify and expand the information fields. During this time the Nelson indices and the Grade codes were completed and verified; the Halpern and Henshall indices, the usage fequencies, and the SKIP codes compiled; and the readings and translations established and verified for all but a small number of the characters.

In early 1993, the author became aware of a similar file which had been compiled by Christian Wittern at Kyoto University with some material provided by Urs App from Hanazono University. Wittern made this file available to the project, and it is from this source that the Morohashi, four-corner and PinYin fields are derived. A further verification of the Nelson indices was also made with this source.

Chapter 5

Copyright Issues

Dictionary copyright is a complex issue, because clearly the first lexicographer who published "犬 (*inu*) means dog" could not claim a copyright violation by all subsequent Japanese dictionaries. What makes each dictionary unique (and copyrightable) is the particular selection of words, the phrasing of the meanings, the presentation of the contents (a very important point in the case of these files), and the means of publication. Advice to the author from people with experience in copyright matters is that EDICT and KANJIDIC are in no different a position than any other new dictionary. While there is a popular view that material in the public domain keyed in by volunteers is somehow exempt from copyright law, the more considered opinion is that the same restrictions apply as for any other form of publication.

Copying material from published dictionaries has been discouraged during the development of these files, but with a voluntary effort being carried out simultaneously in many countries, it has been difficult to police this matter. It has also been the established practice for lexicographers to use previous works as reference material in the development of their own dictionaries. Nelson, in the foreword to his character dictionary, acknowledges his indebtedness to many other dictionaries and reference works, the list of which covers more than a page. While there is a certain amount of "honour among thieves" in commercial lexicography, dictionary compilers frequently include bogus entries in order to trap pirates of their material. Several such entries were noticed in Spahn & Hadamitzky[14], where obscure kanji were given English meanings from Lewis Carroll's "Jabberwocky" (brillig, slithy, borogove, etc.)

Two interesting points of copyright arise in relation to the character dictionary file:

- the file contains indices identifying the characters in a number of major dictionaries (Nelson, Halpern, etc.) It is important to consider whether those indices are protected by copyright, and whether their inclusion in the file can be viewed as a breach of the copyright of those publications.

  The position taken with KANJIDIC is that the indices are not intrinsic components of the information content of those publications. They are pointers to the information in the publications, and their presence in the KANJIDIC file merely facilitates access to the relevant information, in the same manner as a page number. The practice of including indices to other dictionaries is well-established. Nelson includes the indices of the Fuzanbo dictionary, and the author's copy of the pre-war

Rose-Innes [19] character dictionary includes indices for the contemporary edition of the Ueda *daikanwajiten*[20].

- a more difficult point arose with the SKIP codes mentioned above. Halpern states in his dictionary that patent protection has been sought for the encoding method, and specifically advises that the codes must not be used without permission to order or index characters in other dictionaries. While a number of users of the files have expressed doubt to the author about the validity of such a patent, as the encoding system is quite similar to others used previously by other authors, it is clear that the presence of the SKIP codes in the file could, with appropriate software, be used to identify the kanji, and thus would violate a valid copyright. For this reason the distribution of the SKIP codes was discontinued until the issue could be resolved.

The copyright of the project dictionary files presumably belongs to the myriad of contributors. However with the placing of the project in the public domain from the beginning, it is clear that the protection sought is minimal. The only restrictions placed on the distribution of the files is that no charge must be made other than a nominal charge for the medium, and that the files must not be incorporated in commercial products. With regard to this latter point, there have been discussions with several prospective developers of commercial packages about access to the files. The agreed position has been that products may access the files, but they may not include the dictionary files themselves for more than a nominal charge, and they must provide the facility for users to obtain and incorporate updated versions of the files.

Chapter 6

Conclusion

The project described in this paper represents a major development in computer lexicography, and a considerable achievement, considering that its compilation has been carried out largely by volunteers who are employed in fields other than linguistics and language instruction.

Software packages drawing upon the files have been compared favourably with expensive commercial products. A technical translator of Japanese working in the USA remarked in correspondence with the author recently that the files are "the" information source increasingly being used in his field.

Chapter 7

Acknowledgements

The total number of people who have contributed to the project is approaching a hundred. Their work is most gratefully acknowledged, and their names are recorded in the documentation files of the project.

The author is particularly indebted to the following people, who have made significant contributions:

Urs App, Stephen Chung, Lee Collins, John Crossley, Hitoshi Doi, Mark Edwards, Mike Erickson, Curtis Eubanks, Jeffrey Friedl, Magnus Halldorsson, Ken Lunde, Theresa Martin, Kevin Moore, Yasuaki Nakano, Clifford Olling, Alfredo Pinochet, Harold Rowe, Iain Sinclair, Rik Smoody, Kurt Stueber, Hidekazu Tozaki, Scott Trent, Richard Walters, Christian Wittern.

Appendix A

The Japanese Writing System

The following is a very brief outline of the essential elements of Japanese orthography.

The language is written in three sets of graphical symbols:

- *kanji*. These are the characters invented in China, and introduced to Japan during the first millenium AD. Instead of representing a vowel or consonant, a kanji character represents a word or a morpheme. Of the tens of thousands of kanji, the Japanese Education Ministry has established a set of 1,945 *joyo* (常用: general-use) kanji for use in domains such as schools, textbooks and newspapers. This set forms the basis of modern Japanese literacy, although several thousand additional kanji are in use in special subject areas. Each kanji has, in general, two sets of pronunciations or readings. The *kun* reading is usually derived from an archaic Japanese source, and applies when characters are used singly in nouns, verb roots, etc., or are used in combinations to form proper nouns. Such words are known as *wago* (和語: native Japanese words). The *on* readings usually derive from Chinese sources, and are believed to be the result of the adoption of many Chinese words and concepts into Japanese during the assimilation of the writing system centuries ago. *On*-readings are used when two or more characters are used together to form a new word or compound. These words are known as *kango* (漢語: Chinese word). As an example of the different types of readings, the character 東 has the meaning of east, and is used alone to write the word meaning east (*higashi*). Thus the *kun* reading is *higashi*. When used in a compound with the character 京, which means capital, a new word with the meaning eastern-capital is created, with the pronunciation *tokyo*. Thus the *on* reading is *to*.

- *hiragana* and *katakana*. These are two syllabaries, collectively known as *kana*, which were developed in Japan during the second half of the first millenium AD. Both were developed from kanji characters which were traditionally used phonetically. Hiragana is more cursive in shape, and katakana more angular.

  Example: ひらがな (hiragana), カタカナ (katakana)

  In their modern forms, each syllabary has 46 basic symbols: 5 representing the five vowels (a i u e o), 40 consonant+vowel syllables (ka ki .. sa shi ..) and a single consonant (n). Other symbols are created through the use of diacritic marks to indicate voiced or labial consonants. Several small kana are also used as vowel and consonant modifiers, bringing the total number of distinct syllables in modern use to over one hundred.

Although either syllabary can be used effectively to write any Japanese text, in modern Japanese the three sets of symbols are used together in distinctly different roles:

1. *kanji* are used to write the base forms of most nouns, verbs, adjectives, adverbs, etc.

2. *hiragana* is used to write most of the grammatical elements, such as the inflecting portions of verbs and adjectives, connectives and particles, as well as demonstratives, many personal pronouns, and a number of other words such as auxiliary verbs.

3. *katakana* is used primarily for the many thousands of loanwords (外来語: *gairaigo*) which have entered Japanese from other languages in recent centuries. Most of these have come from English, although a number have also come from Chinese, French, Portuguese, Dutch and German. It is also used for emphasis and for many onomatopoeic words.

The following simple sentence demonstrates a combination of all three:

雪子はデパートで着物を買いました。 (Meaning: Yukiko bought a kimono at the department store.)

The kanji components of this sentence are:

- 雪子 (*yukiko*) - proper name (the characters mean *snow child*);
- 着物 (*kimono*) (the characters mean *wearing object*);
- 買 (*ka*) - the root of the the verb *kau* (to buy).

The hiragana components are:

- は (*wa*) - particle marking the sentence topic;
- で (*de*) - the particle indicating location of an action;
- を (*wo*) - the particle indicating the preceding phrase is the object of the verb.
- いました (*imashita*) - the "past tense, polite" inflection appropriate for the verb *kau*.

The katakana component is:

- デパート (*depaato*) - the abbreviated loanword derived from the English department store.

Japanese can also be written using the Latin alphabet. The representation of Japanese syllables in this alphabet is called *romaji* (ローマ字), of which there are several forms. In this paper the Hepburn system is used. Romaji is used in Japan where it may assist foreigners, such as on railway station signs, and in much advertising and shop signage where it has a certain chic appeal. Apart from some elementary language texts for foreign students, it has virtually no role in written Japanese language.

Appendix B

Japanese Text Processing

[Part of the information in this appendix was drawn from the document "Electronic Handling of Japanese Text", by Ken Lunde, which is in a freely available public domain file: "JAPAN.INF". Lunde has greatly expanded this document in producing his recent text[5].]

The advent of computers, and the relatively limited codesets which prevailed in the early decades of computerization, represented a considerable challenge to would-be Japanese computer users. Until the 1970s most Japanese computer systems were restricted to using romaji or one of the kana, usually katakana, for input and output. Some commentators even expressed the view that the incapability of computers to handle the large symbol sets used in Chinese and Japanese would hasten the demise of those writing systems.

The invention of output devices, such as laser printers, capable of handling a large variety of symbols with sufficient precision, combined with the reduction in the manufacturing costs of programmable graphics display devices and micro-electronics in general enabled, by the early 1980s, a progressive incorporation of traditional Japanese orthography into Japanese computer systems.

One of the essential elements in the computerization of any symbol set is the standardization of the coding system used to represent the symbols. After considerable experimentation and development by Japanese computer companies, a Japanese Industrial Standard (JIS C 6226-1978) Code of the Japanese Graphic Character Set for Information Interchange was released in 1978, with further revisions in 1983 and 1990 to track the introduction and update of the jōyō kanji set and its extensions. The current standard is JIS X 208-1990. In this standard, a two-byte code is used for each symbol. The code ranges were selected to enable them to be incorporated as graphical extensions to the common compatible coding systems used world-wide (ASCII, CCITT No. 5 Alphabet, AS1776, etc.). Also the codes were selected to lie within the printable characters of the ASCII set, and thus they could, with the insertion of suitable shift-in/shift-out sequences, be mixed in text and electronic mail files with characters from the other codes. The total number of symbols capable of encoding by this system is 8,836 (94 by 94). However this is ample to provide for all the characters in common Japanese use.

The code standard covers 6,877 characters: 6,353 kanji in 2 levels (level 1: 2,965 kanji arranged by on-reading; level 2: 3,388 kanji arranged by radical), 86 katakana, 83 hiragana, 10 numerals, 52 English characters, 147 symbols, 66 Russian characters, 48 Greek characters, and 32 line elements. The separation of the kanji into 2 levels was to accommodate the systems which did not want the expense of storing the graphical representation of many rarely-used characters. The adoption of the on-reading order for the level 1 kanji was to assist the users of the very early word-processor systems, where character entry was by ordinal index. The kana with diacritic marks are included as separate characters, as are a selection of small kana.

A supplemental character set (JIS X 0212-1992) has also been established, which contains a further 5,801 rarely-used kanji.

The raw JIS codes cannot be used in text documents as they coincide with normal ASCII codes. Three code-modification systems are commonly employed to distinguish between the encoded Japanese and normal 7-bit codes:

1. JIS-code. This is a 7-bit encapsulation method recommended in the JIS standard document. In it, a 3-byte kanji-in shift sequence is prepended to a segment of Japanese text, and a 2-byte kanji-out sequence appended. The raw JIS codes are used between the two shift sequences. This method is commonly used for document files which are being transmitted over communications networks, as only 7 bits of each byte are used, but is less often used within computer systems as the shift sequences complicate the text-handling.

2. EUC (Extended Unix Code, also known as AT&T JIS). This is an 8-bit code in which 128 is added to the two raw JIS codes, thus setting the most significant bit of each of the bytes to 1. EUC is used in Unix systems, and in some MS-DOS PC systems.

3. Shift-JIS, also known as MS-KANJI. This is also an 8-bit code, and also has the most significant bit of the first byte set to one, but distributes the information content of the raw JIS codes unevenly over the two bytes. This is done to maintain compatibility with the earlier half-width encoding system for kana. Shift-JIS is used in Macintosh systems, the NEC 9800 series of PCs, and DOS/V (the Japanese version of MS-DOS PC operating system.)

Appendix C

The Unicode Coding System

(The following information about Unicode was provided by Lee Collins at Taligent.)

The Unicode sequences are the final official mapping to JIS of the CJK-JRG's (Chinese, Japanese, Korean - Joint Research Group) "Unified Repertoire and Ordering Version 2.0" which is the unified Han character set of ISO 10646 and Unicode. All of the Unicode companies (Apple, IBM, Microsoft, NeXT, Taligent, etc) are now using this mapping. There has been some confusion because of difference in nomenclature. Unicode people call it UniHan, the Chinese sometimes call it HCS (Han Character Set) and ISO calls it "Ideographic CJK Character Unified Repertoire and Ordering". ISO cannot use the term "Han character" because Japan was very sensitive to this (even though it is a direct translation of "Kanji") and it cannot be called a character set because only ISO WG2 is empowered with the authority to encode characters. Problems of naming aside, they are all the same thing.

The CJK-JRG was formed under the aegis of ISO in 1990 to investigate and propose a unified Han character set for inclusion in ISO 10646. It brought together various experts on Han characters from China, Hong Kong, Japan, Korea, Taiwan and the United States selected by the national bodies participating in ISO WG2. Including the initial work in the US on Unicode and in China on GB 13000, which were merged and became the basis for the URO, the task spanned about 4 years The work was completed in April 1992. It contains 21,000 Han characters from all of the major standards used in East Asia, including JIS X 0208-1990 and JIS X 0212-1990.

The Unicode consortium provides a cross-reference file for all of the source sets.

For further details about the URO/UniHan, refer to the "The Unicode Standard Version 1.0 Vol II" published by Addison Wesley, ISBN 0-201-60845-6. For a slightly different presentation of the characters, a copy of 10646 or of the "Ideographic CJK Character Unified Repertoire and Ordering Version 2.0" might be available through the local representative on ISO WG2.

Bibliography

[1] J. V. Neustupný, *Introduction to Japanese Writing.* Melbourne: Japanese Studies Centre, 1984.

[2] P. G. O'Neill and S. Yanada, *An Introduction to Written Japanese.* London: English Universities Press, 1963.

[3] J. M. Unger, *The Fifth Generation Fallacy.* New York: Oxford University Press, 1987.

[4] C. Seeley, "The japanese script since 1900," *Visible Language*, pp. 267–302, July 1984.

[5] K. R. Lunde, *Understanding Japanese Information Processing.* Sebastapol, CA: O'Reilly & Associates, 1993.

[6] I. Shinmura, *Kojien.* Tokyo: Iwanami Shoten, 1983.

[7] T. Morohashi, *Daikanwajiten.* Tokyo: Taishukan Shoten, 1967.

[8] T. Iwasaki and J. Kawamura, *New English-Japanese Dictionary.* Tokyo: Kenkyusha, 1960.

[9] M. Shimizu and S. Narita, *Japanese-English Dictionary.* Tokyo: Kodansha, 1976.

[10] M. Takahashi, *Romanized English-Japanese Dictionary.* Tokyo: Taiseido, 1976.

[11] F. Tamamura, *Practical Japanese-English Dictionary.* Tokyo: Association for Overseas Technical Scholarship, 1970.

[12] *Kenkyusha's Furigana English-Japanese Dictionary.* Tokyo: Kenkyusha, 1990.

[13] A. N. Nelson, *The Modern Reader's Japanese-English Character Dictionary.* Tokyo: Tuttle, 1974.

[14] M. Spahn and W. Hadamitzky, *Japanese Character Dictionary.* Tokyo: Nichigai Associates, 1989.

[15] A. Skoutarides and G. Peters, *Nihongo Dictionary for Japanese I.* Melbourne: Swinburne Institute of Technology, 1988.

[16] W. Hadamitzky, *Lehrbuch und Lexikon der japanischen Schrift.* Koeln: Langenscheidt, 1991.

[17] J. Halpern, *New Japanese-English Character Dictionary*. Tokyo: Kenkyusha, 1990.

[18] K. G. Henshall, *A Guide to Remembering Japanese Characters*. Tokyo: Tuttle, 1988.

[19] A. Rose-Innes, *Beginners' Dictionary of Chinese-Japanese Characters*. Yokohama: Yoshikawa Shoten, 1937.

[20] Z. Ueda, *Kanwajiten*. Tokyo: Kodansha, 1966.