Multimedia, Multilingual Hyperdictionaries: A Japanese↔English Example

Harvey Abramson[1], Subhash Bhalla, Kiel Christianson, James Goodwin, Janet Goodwin[2], John Sarraille[3] & Lothar Schmitt[2]

[1] *Unisys Corp., G140-9, 2476 Swedesford Road, Paoli, PA 19301, USA*
[2] *University of Aizu, Aizu Wakamatsu 96580, Japan*
[3] *California State University, Stanislaus, Turlock, California 95381, USA*

KEYWORDS: computational lexicography, hyperdictionary, language study

E-MAIL: harvey@tr.unisys.com
{bhalla, kiel, james-g, janet-g, lothar}@u-aizu.ac.jp
john@ishi.csustan.edu
FAX NUMBER: +81-242-37-2735 [2]
PHONE NUMBER: +81-242-37-2713 [2]

Extended Abstract

1 Introduction

Japanese is often characterized as the most difficult foreign language for a Westerner to learn. While differences in vocabulary, grammar, and culture contribute to this difficulty, the main problem is the complex writing system which, utilizing many thousands of characters, requires such an expenditure of time that for a very long period, the student of Japanese simply *cannot read*. This learner illiteracy in the target language represents a nearly insurmountable barrier to acquisition (c.f. Krashen, 1987, 1989, 1994). In addition to the usual bilingual dictionaries, the student must use dictionaries not only for the ordinary uses of characters, but also for unique and idiosyncratic uses of characters in personal and place names. Modern storage media such as CD-ROMs and magneto-optical disks permit large amounts of data to be stored so that the information contained in a set of bilingual and character dictionaries can be compactly represented. The information thus represented can be treated as a linguistic database which can be deductively accessed, combining many linear dictionaries into a single multi-linear dictionary. Furthermore, traditional methods of character lookup can be generalised easily. The applicability of the notion of a deductive dictionary for human users is, of course, not confined to any single language or language pair.

2 Japanese Orthography

The Japanese writing system, deftly characterised by James D. McCawley as 'without doubt the screwiest writing system in the world' (Constantine, 92), uses thousands of kanji (characters of Chinese origin), hiragana and katakana – two sets of kana syllabaries of 46 characters each (highly simplified Chinese characters which are used phonetically) and sometimes even the Roman alphabet. Although more characters are used in writing Chinese, the Japanese writing system is more complicated than the Chinese writing system because there are usually several pronunciations associated with each kanji. For example: the same kanji – individually or in combination with others – can be given different pronunciations, each meaning the same thing, but used in different social contexts in decreasing order of formality:

明日　みょうにち
myounichi – tomorrow

明日　あした
asu – tomorrow

明日　あす
ashita – tomorrow

A student attempting to read Japanese thus requires not only the bilingual dictionaries familiar to any language student, but also a character dictionary which gives the various pronunciations of the characters and their definitions in the student's language, as well as pronunciations and definitions of compounds, words which are written with two or more characters.

One major initial problem faced by a student of Japanese is how to search for a kanji in a dictionary. Some dictionaries are arranged phonetically or have phonetically-organized indexes, but this is useful only if one knows how to pronounce the kanji in the first place. Otherwise, a given kanji is simply a graphic that must be reduced to primitives in order to facilitate a search.

In some newer dictionaries, kanji are arranged by shape, but this is an arbitrary classification scheme that is not apparent to all users. The most common search method is to break the kanji into two parts: one called a radical that carries (at least in theory) semantic value, and another that indicates pronunciation – not necessarily in Japanese, however, but rather in the original Chinese. There are 214 radicals, organised in order of the number of brush strokes required to write them; these provide the initial search index. Kanji with a given radical are

then organized further according to stroke count. Thus a student who wishes to use a character dictionary must first memorise the 214 radicals (or be condemned to a linear search through a chart). S/he must also learn some basic rules of writing kanji, since the number of strokes sometimes differs from the number a naive user might calculate by eye. Even then, however, difficulties remain, since most complex kanji contain more than one standard radical, and the student must determine which of these radicals is the true index of the kanji. Moreover, since most "words" in Japanese combine two or more kanji, the user must then search through another list, this time of combinations.

Current Japanese↔English dictionaries, whether in book or electronic form, do not serve English-language users very well. Because so much learning "overhead" is required, they cannot be used until the learner is at a fairly advanced stage. As indicated above, they require too much advance knowledge for the beginning user, who often finds that all kanji of more than four strokes look alike. The more sophisticated user still finds look-up a slow process which may require several separate dictionaries to determine the meaning of a single kanji combination; the result is that she or he is often stuck at an awkward stage, able to decode a text but not really to read it. Since the better English↔Japanese dictionaries are produced with the Japanese user in mind, he is in a slightly better position, but still receives inadequate help on matters of usage. Thus there is a need for a dictionary that gives all levels of users rapid access to complete and accurate information. This can be done only through a system that utilizes multiple-access paths and interlocking databases such as we propose below.

3 What is a hyperdictionary?

The problems discussed above can be ameliorated by introducing the concept of a "hyperdictionary" as defined below:

> **Hyperdictionary**
> A relational and deductive database containing the words of a language or languages, together with an open-ended set of access and display methods so as to present at least their orthography, pronunciation, signification, part of speech, and use, their history, synonyms, homonyms, antonyms, derivation, relationships to one another, and any other aspect of the words which may be necessary for reference, teaching or study purposes. Additional information about the language or languages, including grammar, morphology, semantics, pragmatics, machine tractable representations, etc., as well as relevant information concerning geography, names, literature, society, culture, history and so on, is not excluded from the database.

This hyperdictionary should be an encyclopedic database of the words of a language or languages coupled with an open-ended set of deductive access methods and multi-media display methods. Current hardware and software technology, with the exception of very high quality portable screens and reliable miniature optical character readers, are adequate for the purpose, so the real problems of implementing hyperdictionaries lie in the design of the database and the access and display methods. The design process must be carried out from the fundamental point of view that the resulting hyperdictionary will be implemented not as a book, that is, as a list of words, but with media which permit efficient random, non-linear associative access and a variety of powerful input and display techniques.

It is absolutely essential to design the hyperdictionary in such a way that will allow the user to access information in any number of ways, according to the user's cognitive and learning styles and the demands of the task at hand. In doing this, we need to consider the findings of previous studies focusing on the orientation of users within a hypermedia environment, useful features to include in a hyperdictionary, and various methods of dictionary organization (cf. Tripp and Roby, 1992, 1994; Tinkham, 1993), as well as the failure of traditional dictionaries in preventing communication breakdown due to lexical deficit (cf. Christianson, 1995).

Finally, numerous studies have been conducted into dictionary use and misuse by language learners. For instance, Christianson (1995) found that over 40% of words looked up in various dictionaries by Japanese EFL learners were used incorrectly in their writing. A hyperdictionary has the potential to drastically reduce such numbers, as it would allow for fast, extensive access to cross-references and grammatical information.

4 Implementation Issues

The need to find kanji in a variety of ways requires providing multiple input methods. The traditional lookup methods involving pronunciation, stroke count, and radical must be provided as well as lookup by any of the JIS coding systems. These search methods, though, were invented as a means of indexing into linear lists of characters. It is preferable to free the user from any such kind of procedure, for example by facilitating direct input of characters on a screen. This requires either a video camera or optical scanner for input of characters which are already written or printed on

some other medium (e.g., in a book or on a sign) or a penlike device with which characters can be hand written on the screen. Alternatively, a menu might provide graphical possibilities for synthesising the character. Fundamental component shapes can be hyperlinked to a graphical display of more common primitives. Together with spatial information, and with mechanisms for couplings with other similarly generated components, partial queries to a shape database may automatically be generated. Prolog-style matching and search may be used to access a small set of candidate kanji to be presented for user selection, with direct links to the dictionary database. In this way character lookup can be accomplished not only by being able to recognize the radical, but rather, given a suitable character analysis recorded in the database, by any reasonably identifiable constituent of the character. For example, the character 訓 should be discoverable by searching either on 言 or on 川, or indeed even on 口 or on any of the horizontal or vertical lines. (See (Abramson, 95) and (Abramson et al, 95) for further discussion of this.) In order to find kanji in this manner, it is necessary to have an analysis of all characters into their identifiable components (eg, 訓 = 言 + 川) (see (Dürst, 93)) and also of all drawn characters into components which are parsable in the sense of two dimensional grammars for the specification of graphic languages. The capability to search for kanji containing similar components, which hardly exists with present approaches, can be an advantage in vocabulary building and language study.

Various access paths, multiple search methods to be supported, and the contents of the dictionary database all depend on user access needs for locating a reference in the hyperdictionary.

In order to support multiple accessing requirements, a database needs to be created to provide multithreaded access to stored data resources. Based on analyses of individual access needs, an optimum search plan is implemented. In addition, the essential database components that may be necessary are identified. On completion of the analysis of individual access plans, the final content of the database includes:

- words, meanings and encoded pronunciations;
- access information such as JIS codes;
- shape encoding information to facilitate search based on similarity of kanji shapes;
- radical based encoding to create a family of kanji having a common radical;
- special encodings as may be necessary to support the identified access paths;
- multiple indexes to support search paths.

The hyperdictionary must also support complex searching options that are not provided by conventional dictionaries. For example,

- a complex search stating ‘identified radicals’ and ‘a range of number of strokes’ for a character.
- all compounds containing a given kanji.
- thesaurus-like search to provide a collection of words with similar senses and meanings.

For supporting the complex searching options, query language support needs to be provided that can allow use of ’AND’, ’OR’, and ’NOT’ options over the given access methods. More complex language manipulation capability may also be needed to simplify the user interface for unskilled users. For example, the notion of dictionary lookup can be extended by providing a general morphological analyser as a front end to the deductive database.

We illustrate these ideas by showing the core of an exemplary database, using Prolog as a specification and prototyping language. The database consists of definitions of radicals, kanji, and compounds. (In the actual database, there may be more fields than those shown below.)

```
radical(    '黽',
            strokes_in_radical( 13 ), number_of_radical( 205 ),
            comment( [ japanese( 'aogaeru' ), english( 'tree', 'frog' ) ] ),
            nickname( [ english( 'frog' ) ] )).

kanji(      character( '鼈' ), category( 'rare' ),
            radical( '黽' ), number_of_radical( 205 ),
            strokes_in_radical( 13 ), additional_strokes( 12 ),
            meanings( [ chinese( 'betsu' ),
                  japanese( 'suppon' ),
                  english( 'snapping', 'turtle' ),
                  english( 'mud', 'turtle' ),
                  english( 'fresh', hyphen, 'water', 'soft', hyphen, 'shell',
                        'turtle' ) ] )).

compound( '鼈甲色',
      meanings( [ japanese( 'bekkou-iro' ),
                  english( 'amber', 'color' ) ] )).
```

Starting from this core, further components of the database may be derived:

- English, Japanese and Chinese wordlists with pointers to the radicals, kanji, and compounds in which they occur.
- From each compound, definitions of which characters are in the compound. For example, from '鼈甲色' the following are derived:

```
in_compound('鼈','鼈甲色').
in_compound('甲','鼈甲色').
in_compound('色','鼈甲色').
```

- From the Chinese and Japanese pronunciations of radicals, kanji, and compounds, katakana and hiragana equivalents are generated. (The hyphen in the roman is present only as

an aid to students who are learning pronunciation.)

chinese ('betsu') 'ベツ'

japanese ('bekkou-iro') 'べっこういろ'.

The specification of in_compound facilitates searches for compounds on the basis of any character in the compound. This, combined with the kanji lookup by any identifiable constituent (see [Abramson et al, 1996] in this conference), makes it possible to look up compounds provided that some part of each (or any) character in the compound is recognized. By "some part" we mean either radical or other constituent structure, stroke count, pronunciation (indicated in romaji or kana) or meaning. Furthermore, the hyperdictionary may be extended in may ways. Audio for the pronunciation of Japanese and/or English might be added directly to the database, or alternatively, speech generators could generate sound from the kana or English spelling already in the database. Space does not permit a listing of all the possible probes into or extensions to the database, but we hope this brief description gives an indication of what is possible.

5 Wider implications

The difficulties faced by non-native readers of Japanese have created barriers between the Japanese and other peoples that are hard to surmount. These barriers contribute to a feeling of isolation on the part of the Japanese and to a myth of uniqueness that deceives both Japanese and foreigners alike into believing that Japanese culture is impenetrable. The psychological distance between Japan and most other nations can be attributed in part to this communication problem, and has continuing implications for issues of trade, diplomacy, and scholarly exchange. A multimedia dictionary can help to bridge this communication gap.

References

Abramson, H. (1995). The Web of Kanji, Deductive Dictionaries and Logic Programming, Proceedings Natural Language Understanding and Logic Programming Conference, Lisbon, May 26–29 l995, pp. 1–22.

Abramson, H., Bhalla, S., Christianson, K., Goodwin, J., Goodwin, J., Sarraille, J., Schmitt, L. (l995) *The Logic of Kanji Lookup in a Japanese↔English Hyperdictionary*. Proceedings Joint International Conference ALLC-ACH '96, Bergen, Norway, June 25–29, 1996.

Christianson, K. (1995). When FL writing errors occur despite dictionary use. Paper presented at the 21st Annual Conference of the Japanese Association of Language Teachers (JALT), Nagoya, Japan, November 4, l995.

Constantine, P. (1992). *Japanese Street Slang.* New York: Tengu Books. Foreword by James D. McCawley.

Dürst, M. (1993). Coordinate-independent Font Description using Kanji as an Example. *Electronic Publishing,* Vol. 6(3), pp. 133–143.

Krashan, S. D. (1987). *Principles and practice in second language acquisition*. New York: Prentice-Hall.

—. (1989). Language teaching technology: A low-tech view. In: Altatis, J. E. (Ed.). *Georgetown University Round Table on Languages and Linguistics 1989, pp.* 393–407. Washington, D.C.: Georgetown University Press.

—. (1994). Beyond the input hypothesis. Plenary session presented at the 28th Annual TESOL Conference, Baltimore, MD, March 11, 1994.

Tinkham, T. (1993). The effect of semantic clustering on the learning of second language vocabulary. *System*, 21(3), pp. 371–80.

Tripp, S. D. & W. Roby. (1992). The effects of congruent orienting strategies on learning from hypertext bilingual lexicon. *Journal of Hypermedia and Multimedia Studies*, 2(2), pp. 6–12.

—. (1994). The effects of various information resources on learning from a hypertext bilingual lexicon. *Journal of Research on Computing in Education*, 27(1), pp. 92–103.