From Dictionary to Hyperdictionary

Harvey Abramson, Subhash Bhalla, Kiel Christianson, James M. Goodwin, Janet R. Goodwin, John Sarraille, Randy Sharp, and Jun Yamadera
University of California, Los Angeles
USA

Introduction: What is a Hyperdictionary?

A hyperdictionary is a comprehensive dictionary designed to take advantage of the special characteristics of electronic media: abundant storage space; the ability to create interlocking databases; availability of multiple methods to search for a given item; and multimedia capabilities that allow visual and oral as well as textual representations. Formally the hyperdictionary is defined below:

> 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 the following characteristics of the words: their orthography, pronunciation (voiced by a native speaker), signification, part of speech, use, history, synonyms, 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.

Using a Hyperdictionary to Study Japanese

Although a hyperdictionary may be developed for any single language or set of languages, we have chosen to begin our efforts with a two-way Japanese-English dictionary. Because Japanese is a particularly difficult foreign language for a Westerner to learn, a Japanese<->English hyperdictionary can be an invaluable tool in language teaching as well as in advanced scholarship in various Japanese-studies fields.

While differences in vocabulary, grammar, and culture contribute to the difficulty of learning Japanese, the main problem is the complex writing system with its thousands

of kanji (logographic characters of Chinese origin), each of which must be mastered individually. "Words" are sometimes represented by a single kanji, but often by compounds of two or more kanji. Verbs and adjectives are commonly formed by attaching endings, written in phonetic script, to kanji and sometimes to compounds. The system is further complicated by the fact that almost all kanji can be pronounced in at least two ways: an *on* reading that represents the Chinese pronunciation of the character as it sounded to Japanese ears when the character was introduced in the fifth or sixth century, and a *kun* reading, the native Japanese word semantically equivalent to that character.

Japanese can be written phonetically in either of two syllabaries, called *hiragana* and *katakana*. Unfortunately the large number of homonyms among *on* readings makes it impractical to depend entirely on phonetic representation, even though in theory it would be possible to do so. In the modern language, verb endings and grammatical particles are generally rendered in *hiragana*, while *katakana* is used for emphasis and for words borrowed from foreign languages.

Postwar language reforms attempted to limit the number of kanji in common use, but these limits are ignored in all but the most elementary textual materials. For even the simplest independent reading task one must master about 2000 kanji, plus many more compounds. It takes so long to do this that non-native learners of Japanese, unlike students of most other foreign languages, cannot use the written language to reinforce their acquisition of the spoken language until they are well along in their studies. This is, in part, because dictionary access when dealing with written material is a quite different process than when dealing with the equivalent spoken material.

Most students use at least two types of dictionaries to find the definition of a Japanese word: a bilingual dictionary organized phonetically, and a specialized character dictionary which arranges the characters according to their graphic elements and defines them in the user's language. Unlike bilingual Western dictionaries (such as French<->English), most Japanese dictionaries for non-native speakers require some advanced grammatical knowledge. For example, endings are commonly attached to verbs and adjectives to indicate tense, negation, or increased level of politeness, and the verb or adjective stem itself is often altered in the process. The present tense of a verb is often rendered in the so-called *te* form, rather than in the standard form listed in most dictionaries; the verb *yomu* (read) becomes *yonde imasu*. Verbs and adjectives transformed in this way--through sound changes and added

endings--are usually excluded from both phonetic and kanji dictionaries, even though one constantly encounters these forms in speaking and writing.

Any use of a kanji dictionary, moreover, requires additional advanced knowledge concerning the ways that characters are written. This is because most dictionaries order the characters according to graphical features that are rarely apparent to the untrained eye. There are several different ordering methods. The most common classifies kanji according to 214 basic elements, called radicals. The radicals and the order in which they are listed has been standardized in societies that use the Chinese writing system.

Each kanji can be divided into two parts: its radical, which theoretically carries semantic value, and an element that indicates one of the kanji's *on* readings. The radical serves as an initial search index, and kanji listed under a radical are then arranged in the order of the number of brushstrokes required to write them.

The beginner is thus presented with several problems. She must first determine which element of a kanji is its radical. This is not always easy to do, since many kanji are made up of several elements that can be radicals. For example, the kanji 相 is made up of two elements that can be used as radicals, 木 and 目；only an experienced student knows that in most dictionaries, the kanji is indexed under 目. Moreover, while many radicals can, by themselves, be used as kanji, they are often written very differently when used as components, with their shape varying with location in the character. For example, consider the characters below:

法　水　永

Only an experienced student would know that the three strokes that compose the left-hand part of 法 and serve as its radical are actually a form of the kanji 水, which appears in more recognizable form--also as the radical--in the kanji 永. This means that, in addition to mastering kanji themselves, the student must also learn a number of radicals in their transformed shapes.

Once the correct radical is found, the student must then determine how many brushstrokes are required to write the rest of the kanji. This requires some experience in calligraphy, since the correct number of strokes may differ from the number a naïve student might estimate by eye. If one has never practiced calligraphy, one might not know that the element 口, for example, is written with three strokes rather

than four (or even one!). In other words, using a typical kanji dictionary requires knowledge that the beginning student of Japanese simply does not have.

Even the advanced student finds lookup to be a slow process which may require several separate dictionaries to determine the meaning or pronunciation of a single kanji compound: a *kanwa jiten* (which defines Chinese characters in Japanese), a *kokugo jiten* (a Japanese-language dictionary organized phonetically), and various specialized dictionaries. Morohashi's comprehensive *Dai Kanwa Jiten*, which defines nearly 49,000 characters, lists relatively few compounds and thus must be supplemented by *kokugo jiten* such as the *Kojien* or the *Nihon Kokugo Daijiten*. Conversely, compounds that use non-standard readings for otherwise familiar kanji are very hard to find in even the best *kokugo jiten*. As a result, the advanced student is often stuck at an awkward stage, able to decode a text but not really to read it. Without the crutch of classroom vocabulary lists, moreover, the student must use dictionaries not only for the ordinary uses of characters, but also for unique and idiosyncratic uses in personal and place names. To cite perhaps the worst possible case, the student of pre-modern Buddhist literature may have to consult the following paper dictionaries to read a single text: a multi-volume *kanwa jiten*; a large single-volume or multi-volume *kokugo jiten*; a specialized Buddhist-term dictionary; a dictionary of historical usage; and Japanese<->English character and phonetic dictionaries.

We propose the hyperdictionary defined above to ameliorate these problems. The most important features of the hyperdictionary we are developing are: (1) multiple methods for kanji lookup, so that the user can choose the one most suitable to his needs and level of knowledge; (2) a search engine and interlocking databases that permit multidirectional searches, so that any piece of information about a particular kanji or compound can be used to retrieve all available information about it; (3) extensibility, so that additional search paths and databases can be added at any time (4) inclusion of transformation rules that may be applied if a input verb form is not found in the database; and (5) a simple and familiar user interface.

Developing a Hyperdictionary

Recent technology has enabled the development of a hyperdictionary, in the form of a set of interlocking databases combining many linear dictionaries into a single multi-linear system. Modern storage media such as CD-ROMs or DVDs 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. Furthermore, the Internet provides virtually unlimited space and permits continuous updating and extension of the databases. Using the Internet, it is possible to begin with one type of dictionary, say a kanji->English character dictionary, and expand it by incorporating other databases as they are developed or acquired.

Our experimental web site described below shows how kanji lookup in a hyperdictionary might work. Initially, we have implemented a lookup system for single kanji.

The user interface is a simple browser based query form (written in HTML) enabling one to find the pronunciations and definitions of kanji using several different search keys. (See Figure 1.)

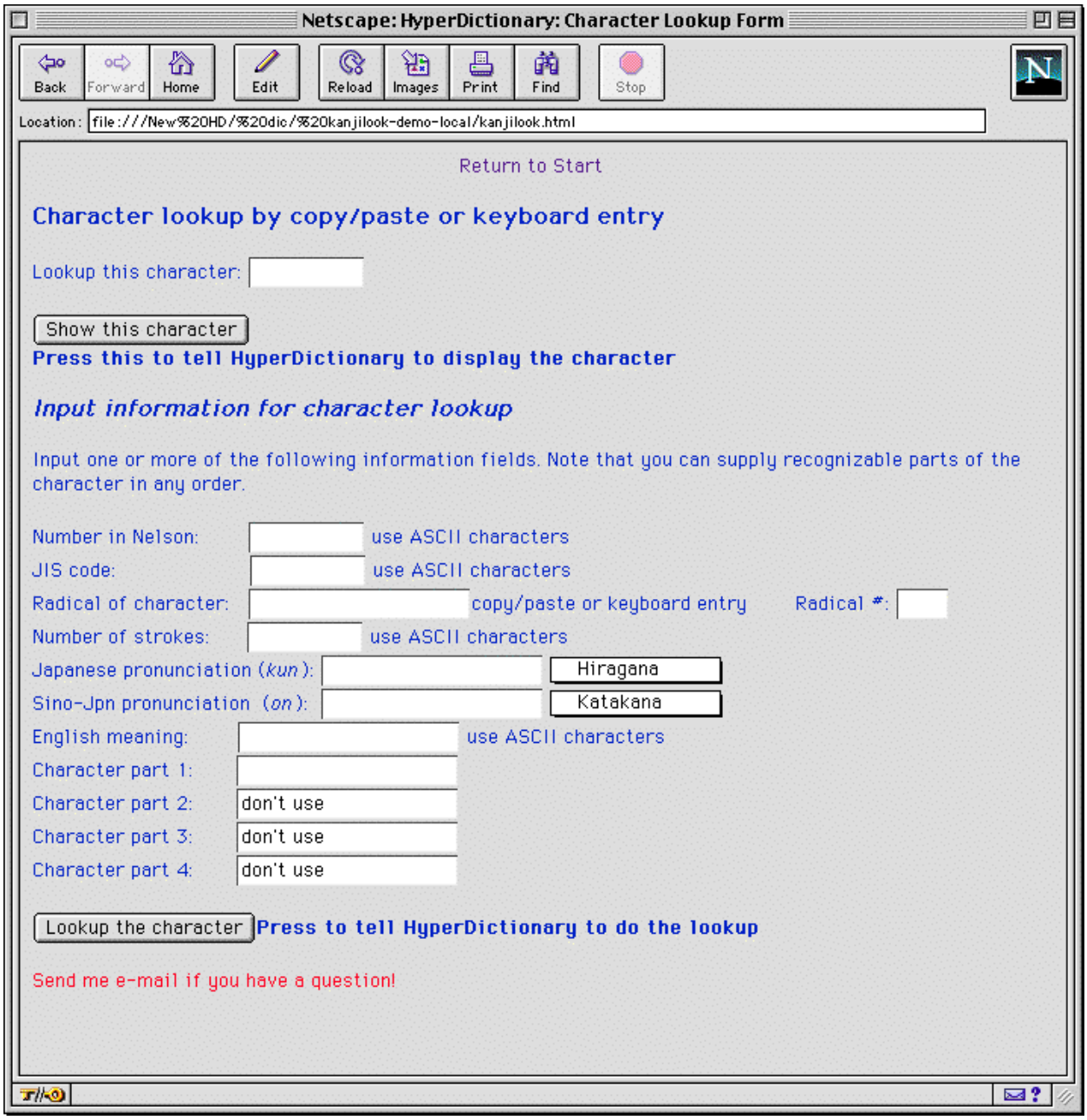


Figure 1: User Query Form

To enter a kanji directly, one may copy the kanji from another electronic source, and paste it into the query form to discover its pronunciations and definitions, or input it directly if the user system supports kanji input. Figure 2a shows the direct input of the character 語.

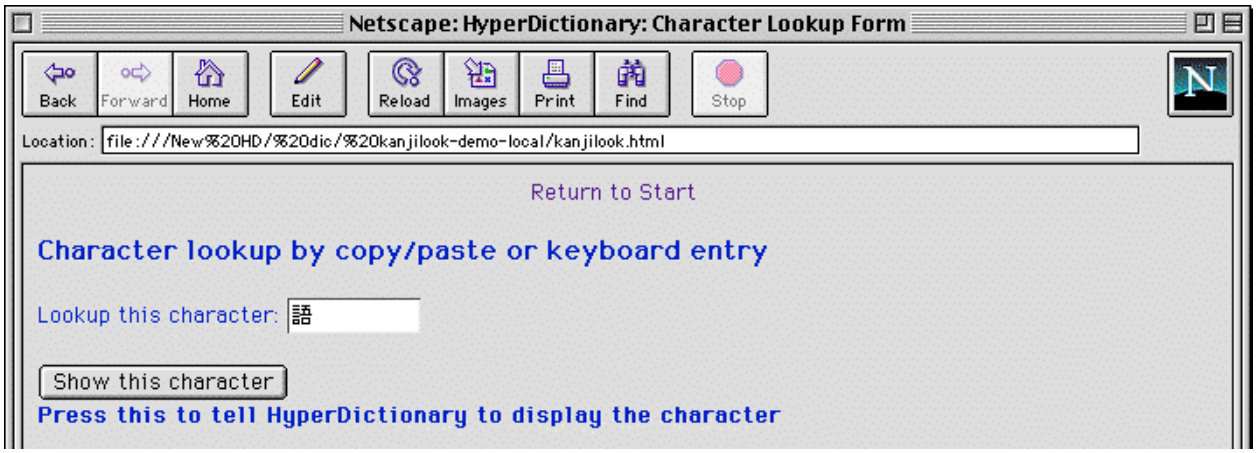


Figure 2a: User Query Form, kanji input section

If the kanji is not available electronically, there are several other search keys which may be used alone or in combination with each other. One may enter the number of strokes in the desired kanji, or, using *hiragana*, *katakana*, or romanized characters (ASCII), one may begin with an *on* or *kun* reading. Another search key is the kanji's radical. It may be found via a separate radical query form on our site, copied, and pasted into the appropriate slot in the kanji query form. Since radicals are usually also stand-alone kanji, the radical may be found from its *on* or *kun* pronunciation. We intend to allow entry by the radical's number as well. As we experiment with the query form and receive suggestions from users, additional search keys can be implemented.

The typical database entry is a "vector", listing all the features related to the entry. Hidden to the user is a search engine based on the computer language Prolog. The engine allows search based on partial information, the fill in of missing information elements, and multidirectional search, or unidirectional information delivery. When some of the elements ("search keys") are specified as a query, the search engine can retrieve one or more entries (up to the total in the database) that include the specified features. The entire entry or any desired part of it can then be displayed using the browser. While all this may seem quite complicated, it takes place entirely behind the scenes. The user sees only his own query and the delivered results. Figure 2b shows the result for the direct input (just shown in Figure 2a) of the character 語.

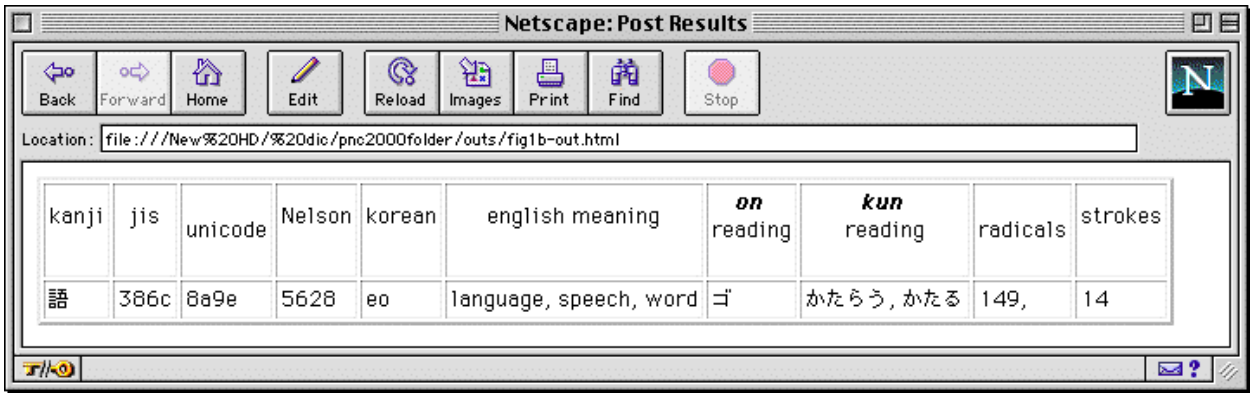


Figure 2b: Results of a Sample Query

If the kanji is simply copied from another electronic source, such as a scanned document or e-mail, the search engine will produce a unique result including definitions and pronunciations, along with keys to other databases. However, if other entry methods are used, the search engine may produce a list of candidate kanji. *On* and *kun* readings, for example, sometimes produce long lists, which can be shortened by providing additional information, as demonstrated below.

The experimental query form and table of results still require some advanced knowledge on the part of the user. To accommodate the beginner, we intend eventually to allow the user to sketch any part of the kanji on a drawing tablet, or to input one or more of a given set of simple graphical features from menus. For example, if the user specifies that the kanji is made up of an enclosure in the form of a box, with an enclosed element made up of three horizontal lines and one vertical line, the kanji *koku* (国) should appear. A Prolog database describing the composition of kanji from their elements and configuration is available to us, but is not yet incorporated into the user interface. We also plan to augment the table of results so that the *on* and *kun* readings are given in romanized characters as well as in kana. (We hope that the Internet character set will be expanded to support the use of macrons over vowels, a critical feature in rendering romanized Japanese.)

Hypderdictionary Web Site: Examples

The figures below are samples of queries and their results. To conserve space, from now on we show only the input section of the query form page, since the rest is unused and repeated without change each time.

Figure 3a shows the data entry form for a query in which the user searches for a kanji based on its Sino-Japanese *on* reading, here entered on the form with roman letters, one of the three available choices for keyboard entry.

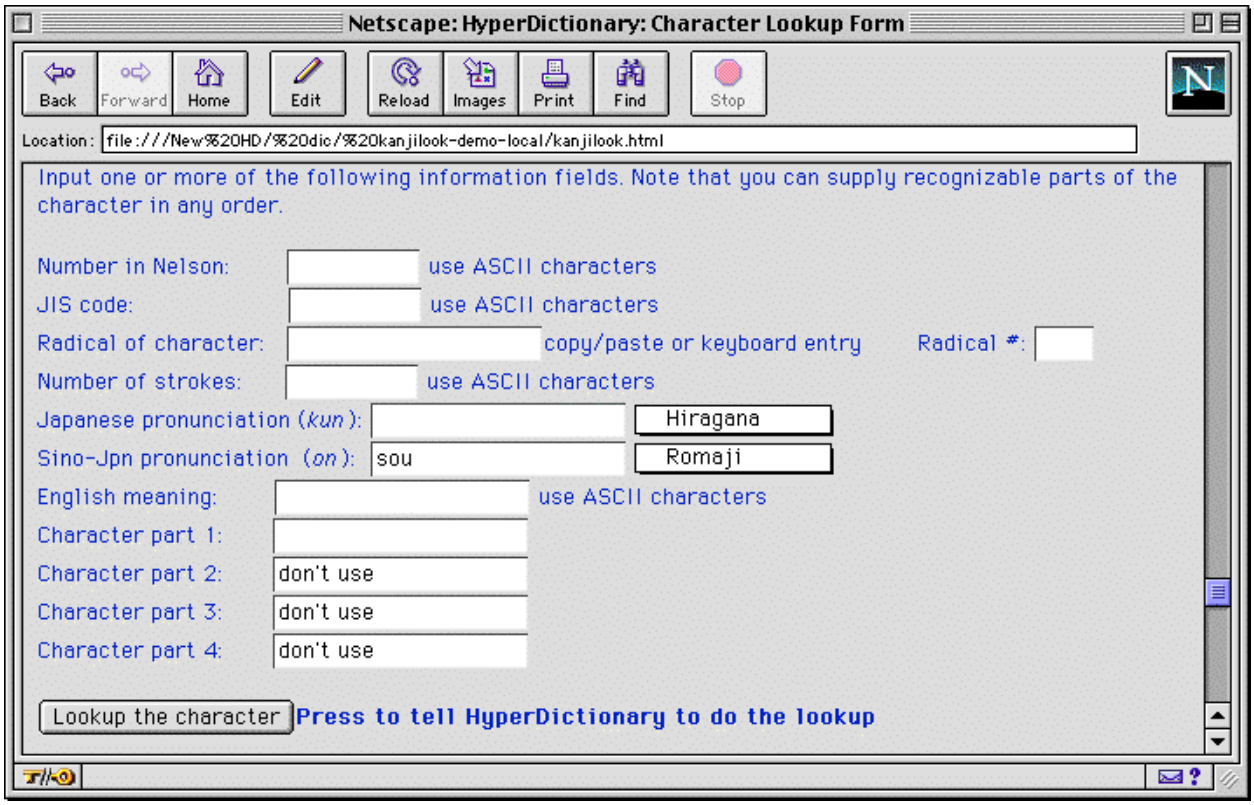
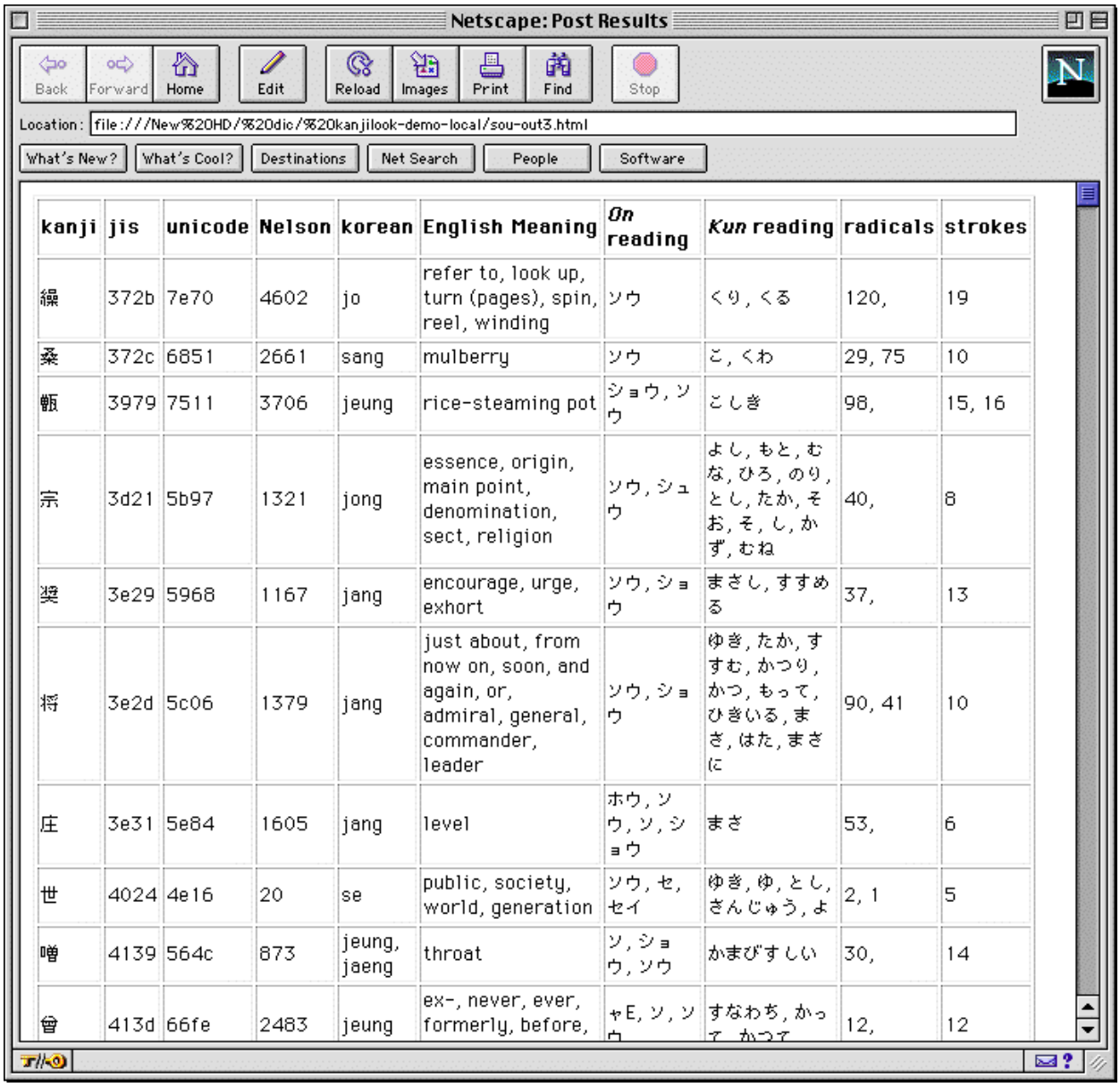


Figure 3a: The on reading sô is entered.

This is not an effective way to find a character, since many have the same *on* reading. As a consequence the search is lengthy, and many candidate kanji are returned -- all those in the database which have this reading. The result is shown (in part) as figure 3b.



kanji	jis	unicode	Nelson	korean	English Meaning	*On* reading	*Kun* reading	radicals	strokes
繰	372b	7e70	4602	jo	refer to, look up, turn (pages), spin, reel, winding	ソウ	くり, くる	120,	19
桑	372c	6851	2661	sang	mulberry	ソウ	こ, くわ	29, 75	10
甑	3979	7511	3706	jeung	rice-steaming pot	ショウ, ソウ	こしき	98,	15, 16
宗	3d21	5b97	1321	jong	essence, origin, main point, denomination, sect, religion	ソウ, シュウ	よし, もと, むな, ひろ, のり, とし, たか, そお, そ, し, かず, むね	40,	8
奨	3e29	5968	1167	jang	encourage, urge, exhort	ソウ, ショウ	まさし, すすめる	37,	13
将	3e2d	5c06	1379	jang	just about, from now on, soon, and again, or, admiral, general, commander, leader	ソウ, ショウ	ゆき, たか, すすむ, かつり, かつ, もって, ひきいる, まさ, はた, まさに	90, 41	10
庄	3e31	5e84	1605	jang	level	ホウ, ソウ, ソ, ショウ	まさ	53,	6
世	4024	4e16	20	se	public, society, world, generation	ソウ, セ, セイ	ゆき, ゆ, とし, ざんじゅう, よ	2, 1	5
噌	4139	564c	873	jeung, jaeng	throat	ソ, ショウ, ソウ	かまびすしい	30,	14
曾	413d	66fe	2483	jeung	ex-, never, ever, formerly, before,	ャE, ソ, ソウ	すなわち, かって, かつて	12,	12

Figure 3b: *sô* produces a very long list of candidate kanji (only the start of the output is shown to conserve space).

In the next example, figure 4a, the query above is refined by also specifying the number of strokes required to write the character. This results in a much smaller and more manageable list of candidate kanji. The result is shown in figure 4b.

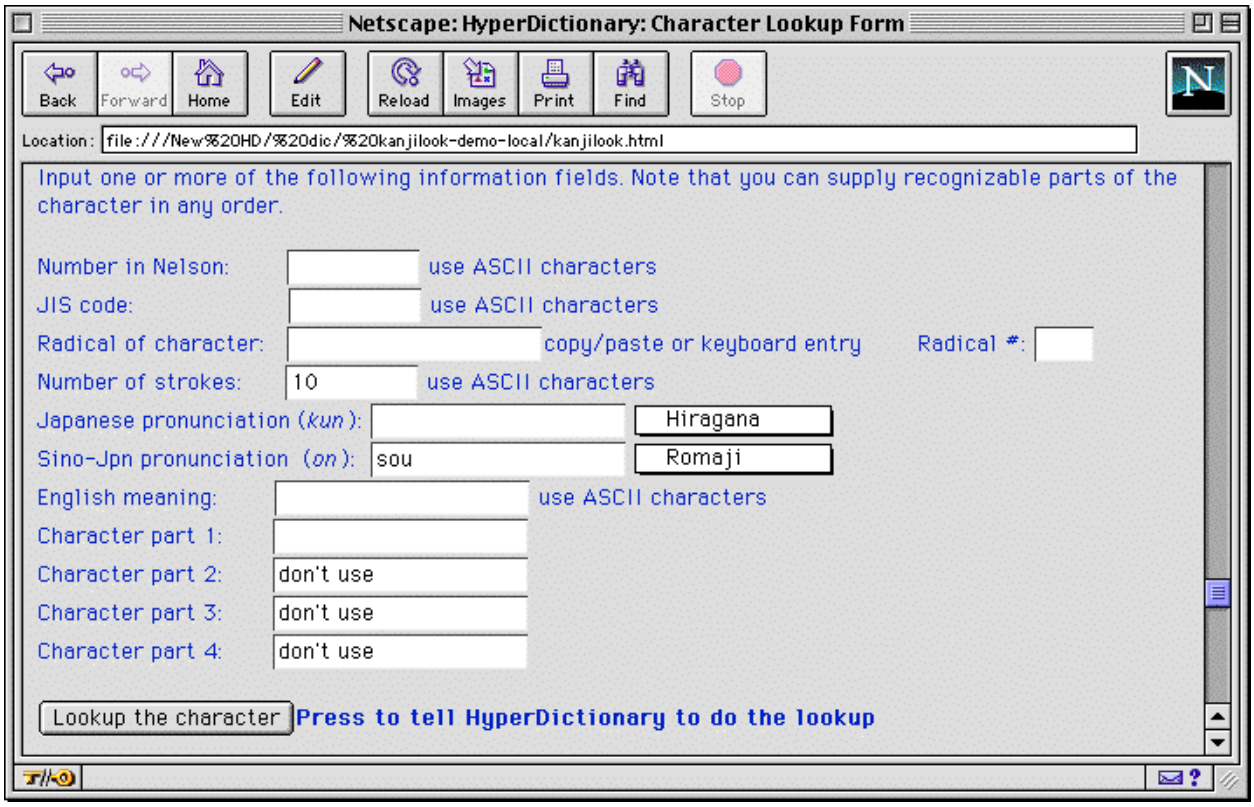
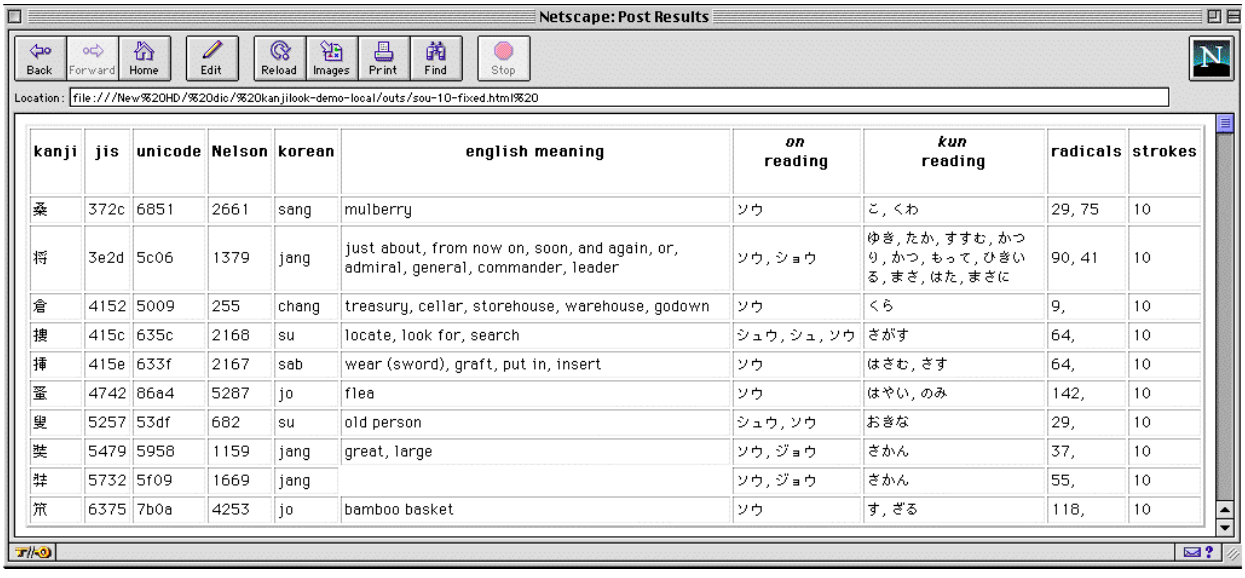


Figure 4a: The number of strokes in the kanji (e.g. 10) plus the *on* reading *sô* are entered.



kanji	jis	unicode	Nelson	korean	english meaning	*on* reading	*kun* reading	radicals	strokes
桑	372c	6851	2661	sang	mulberry	ソウ	こ, くわ	29, 75	10
将	3e2d	5c06	1379	jang	just about, from now on, soon, and again, or, admiral, general, commander, leader	ソウ, ショウ	ゆき, たか, すすむ, かつり, かつ, もって, ひきいる, まさ, はた, まさに	90, 41	10
倉	4152	5009	255	chang	treasury, cellar, storehouse, warehouse, godown	ソウ	くら	9,	10
捜	415c	635c	2168	su	locate, look for, search	シュウ, シュ, ソウ	さがす	64,	10
挿	415e	633f	2167	sab	wear (sword), graft, put in, insert	ソウ	はさむ, さす	64,	10
蚤	4742	86a4	5287	jo	flea	ソウ	はやい, のみ	142,	10
叟	5257	53df	682	su	old person	シュウ, ソウ	おきな	29,	10
奘	5479	5958	1159	jang	great, large	ソウ, ジョウ	さかん	37,	10
壯	5732	5f09	1669	jang		ソウ, ジョウ	さかん	55,	10
笊	6375	7b0a	4253	jo	bamboo basket	ソウ	す, ざる	118,	10

Figure 4b: Using two search keys to produce a much shorter list.

The next example shows the entry of a word using its *kun* reading, again entered in romanized text.

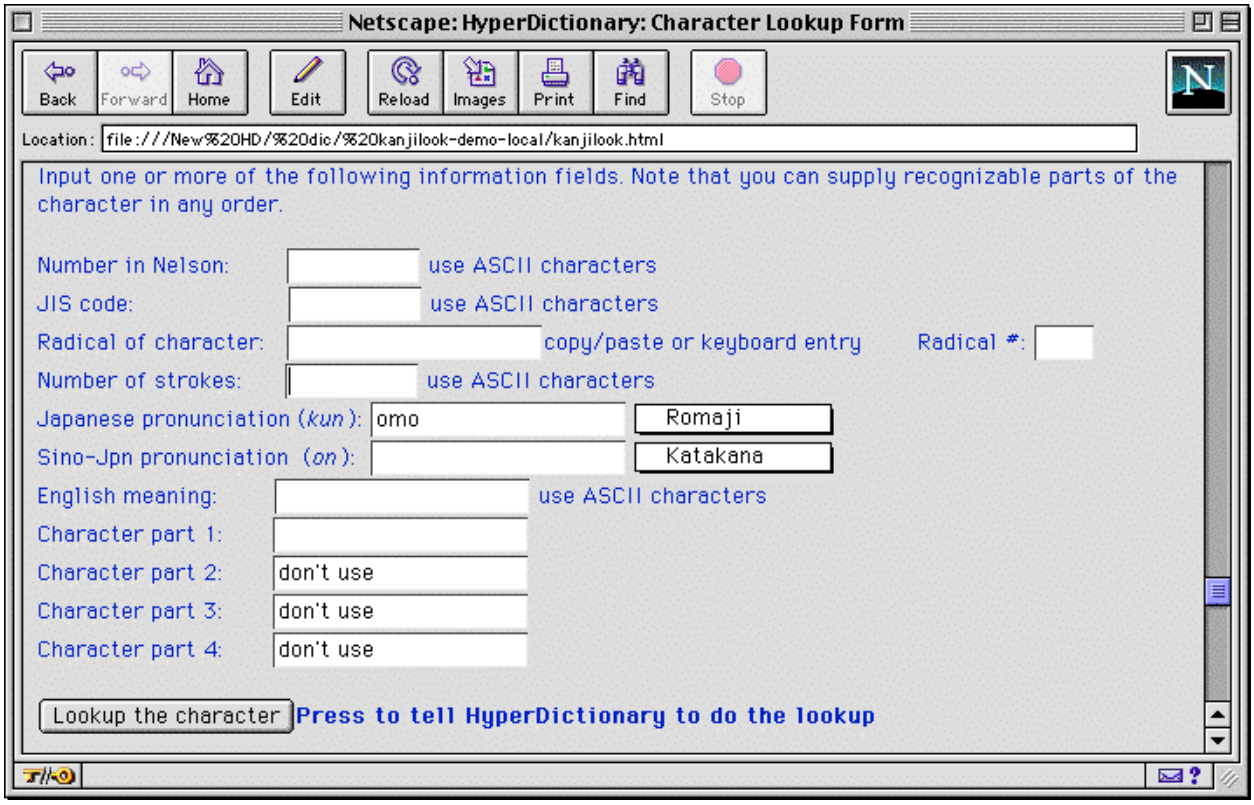
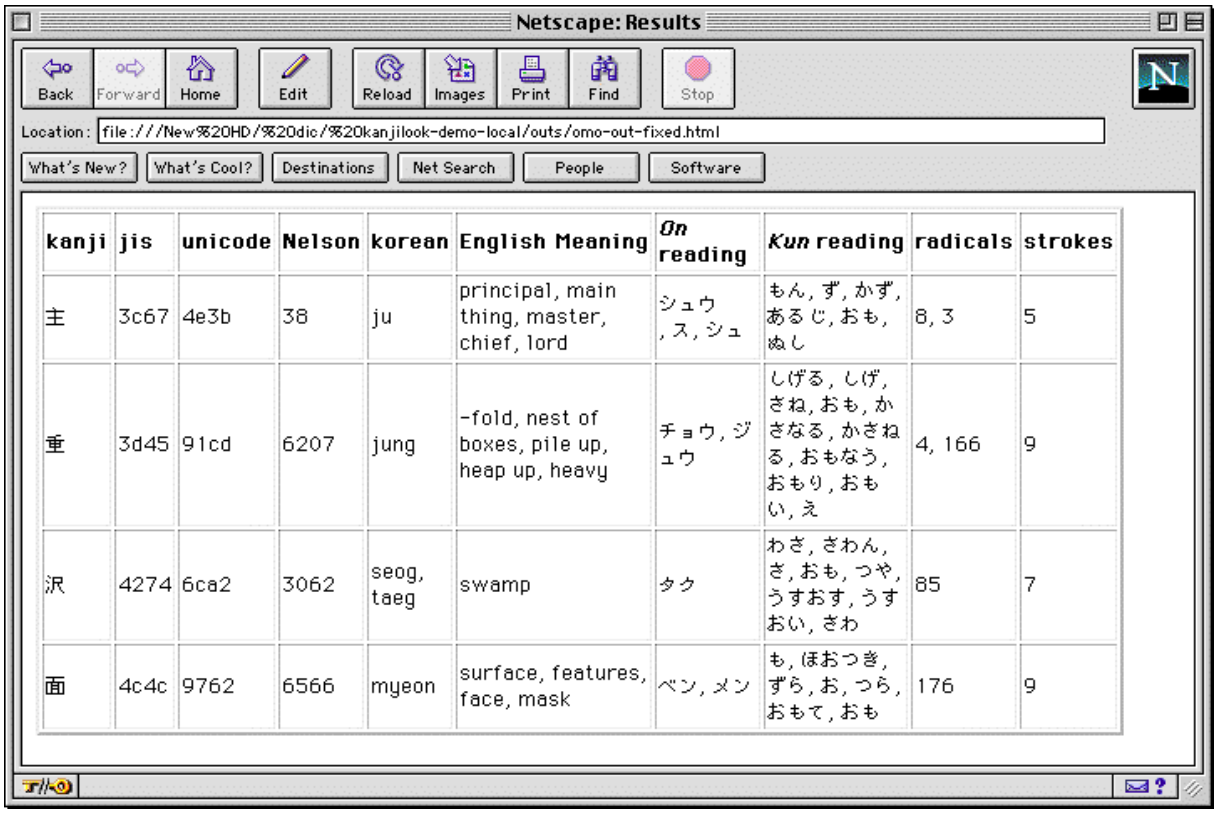


Figure 5a: The *kun* reading *omo* is entered.



kanji	jis	unicode	Nelson	korean	English Meaning	*On* reading	*Kun* reading	radicals	strokes
主	3c67	4e3b	38	ju	principal, main thing, master, chief, lord	シュウ, ス, シュ	もん, ず, かず, あるじ, おも, ぬし	8, 3	5
重	3d45	91cd	6207	jung	-fold, nest of boxes, pile up, heap up, heavy	チョウ, ジュウ	しげる, しげ, さね, おも, かさなる, かさねる, おもなう, おもり, おもい, え	4, 166	9
沢	4274	6ca2	3062	seog, taeg	swamp	タク	わさ, さわん, さ, おも, つや, うすおす, うすおい, さわ	85	7
面	4c4c	9762	6566	myeon	surface, features, face, mask	ベン, メン	も, ほおつき, ずら, お, つら, おもて, おも	176	9

Figure 5b: The *kun* reading *omo* luckily produces a short list.

Additional search keys may be entered to refine the search. We show the addition of the *on* reading, *taku*, (entered, for example, in *katakana*) to narrow the search. In this case a single definition is found, as shown next.

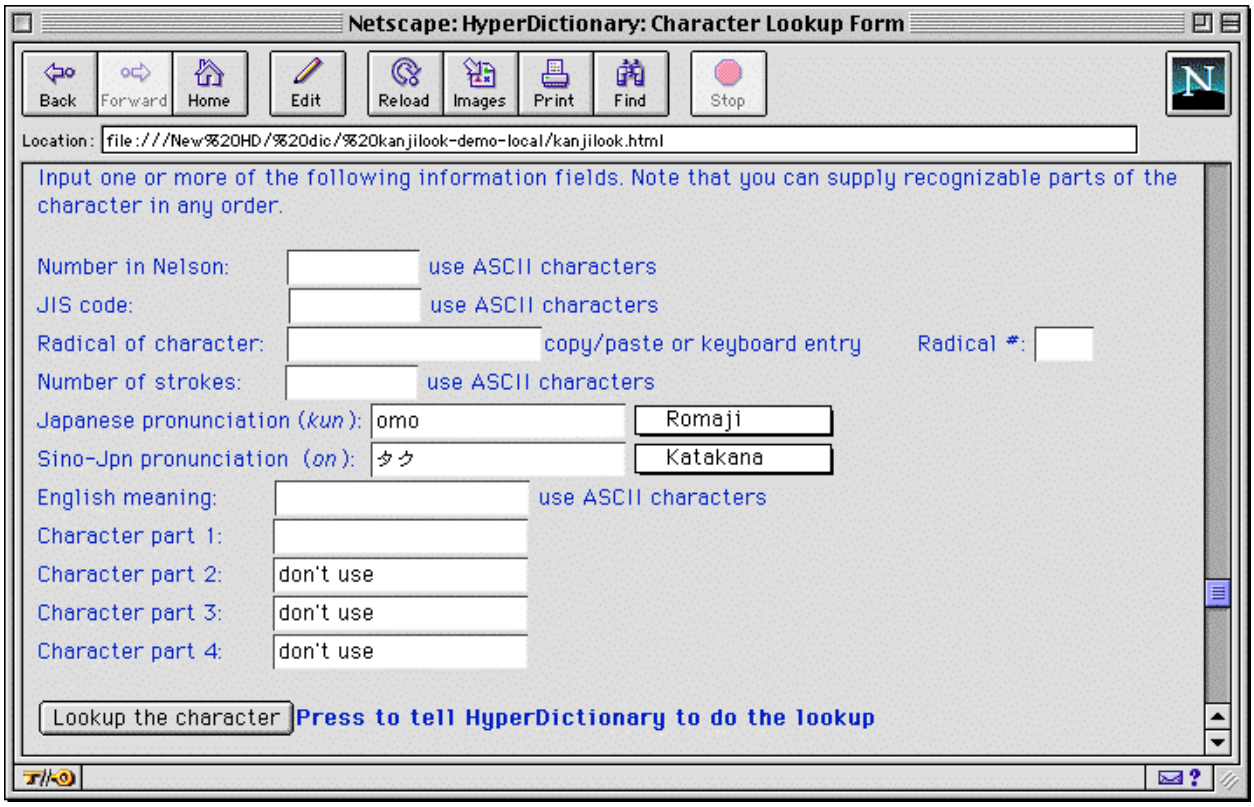
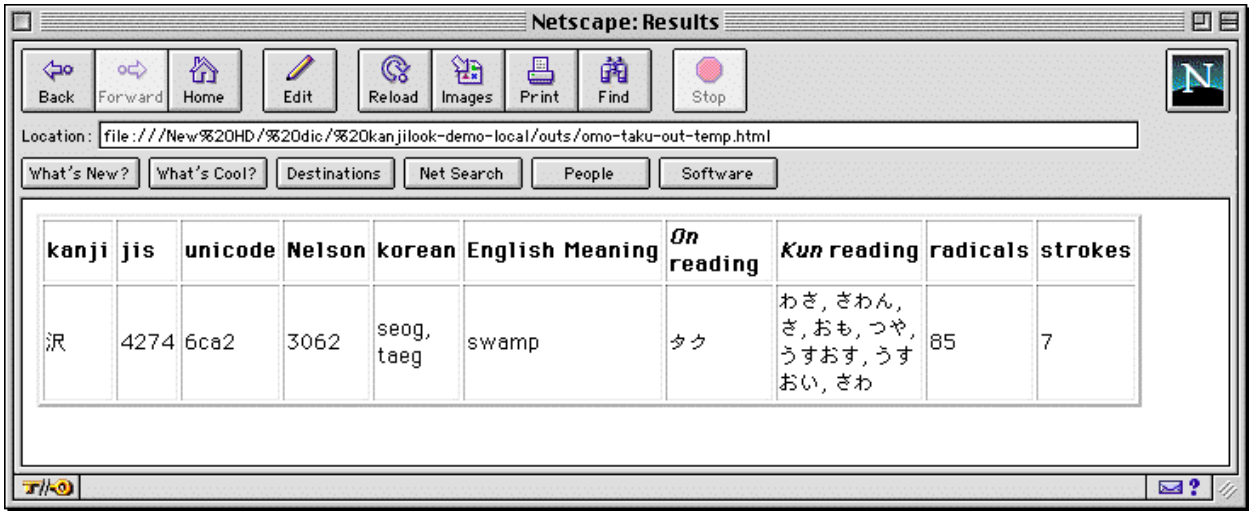


Figure 6a: Both the *kun* and the *on* reading (*taku*) can be entered to narrow the search.



kanji	jis	unicode	Nelson	korean	English Meaning	*On* reading	*Kun* reading	radicals	strokes
沢	4274	6ca2	3062	seog, taeg	swamp	タク	わさ, さわん, さ, おも, つや, うすおす, うすおい, さわ	85	7

Figure 6b: Using both *kun* and *on* readings produces only one result.

Alternatively, we may use the English meaning to narrow the search, as shown in the next example, Figures 7a and 7b.

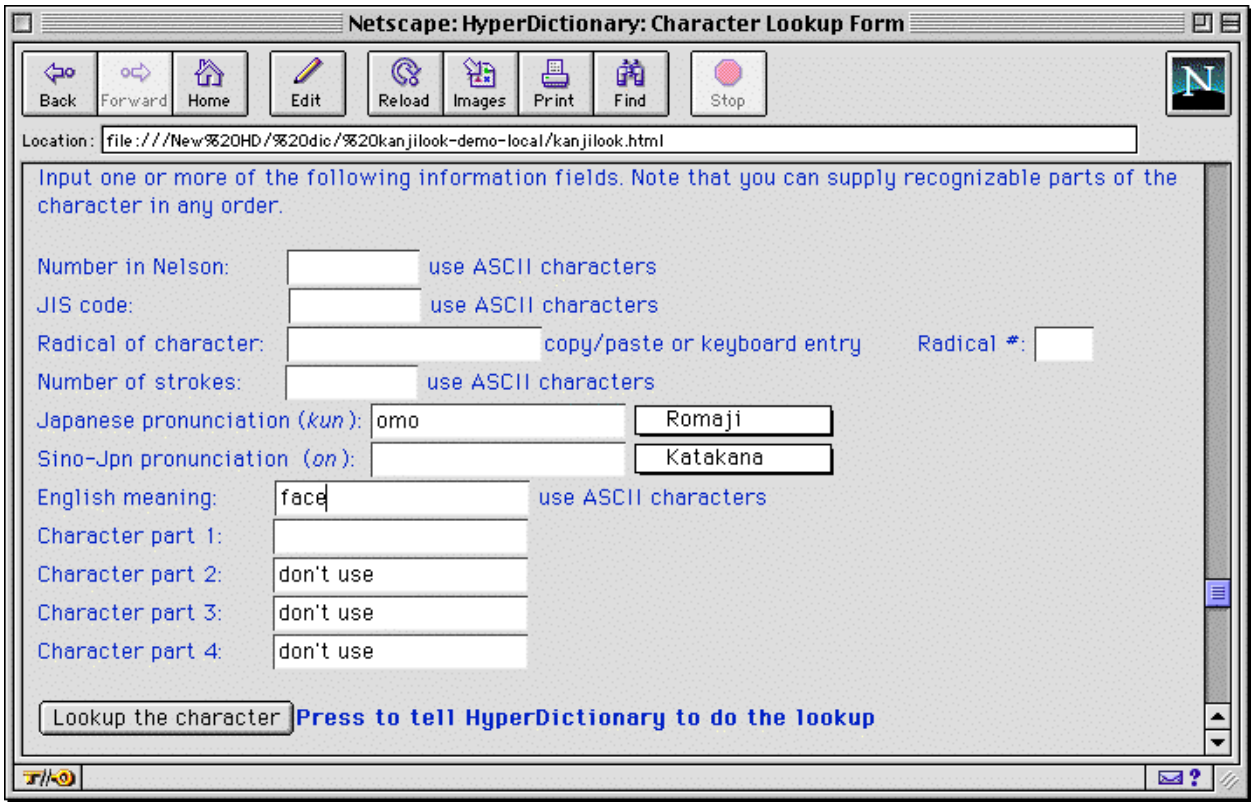


Figure 7a: Entering the *kun* reading and the English definition "face".

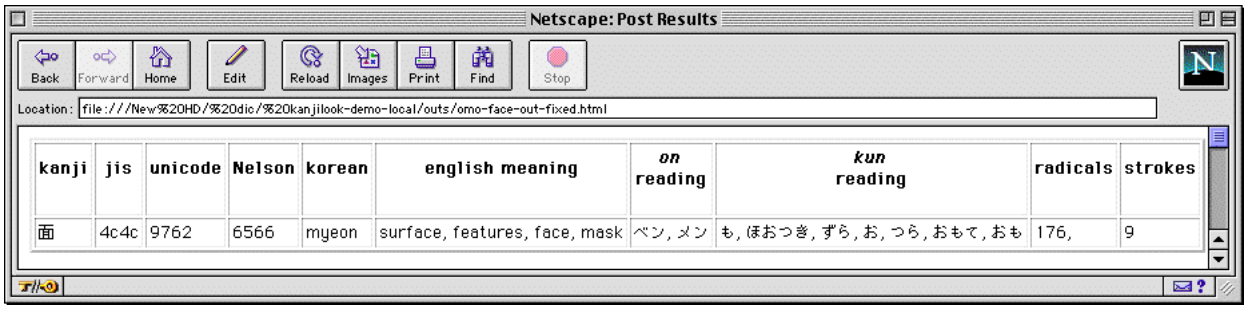


Figure 7b: Entering the *kun* reading and English definition also produces one result.

Still under development are browser interfaces to the lookup systems for the 214 radicals, and for kanji compounds. Compounds will be accessible using any of the member kanji as search keys, not just the initial member of the compound. (The capability to find compounds and their meanings by specifying a limited number of their members in any order can be used as another tool for language study.) Of course, it is possible to use any results from a query to compose a new query by copying and pasting. This is particularly helpful for finding compounds, and forms a simple data entry tool for queries.

If one wants to describe all the kanji in a compound, it will not be necessary to use the same method for each. While it is unlikely that anyone would ever do so, it should be possible to make the following query:

1. first kanji: radical # 72, 0 remaining strokes
2. second kanji: pronounced "moto"
3. third kanji: English meaning "person"

and obtain the result 日本人 (*Nihonjin*; Japanese person).

Adaptations and Extensions

While we are currently working on kanji search methods and using a kanji<->English database, and have available a database of graphical descriptions of kanji, we hope to incorporate other types of databases as well. If implemented fully, the hyperdictionary would link *kanwa jiten*, *kokugo jiten*, and specialized dictionaries to the current database. The ideal hyperdictionary would allow the user to access detailed information about a kanji or kanji compound, such as its etymology or examples of its usage from various historical periods. The hyperdictionary would then become a reference tool for scholars, as well as a learning tool for beginners.

Through the addition of other databases, our current Japanese<->English hyperdictionary can be expanded to include other languages as well. We have already made a modest stab at expansion by providing the romanized Korean pronunciation of kanji in our table of search results. Obvious expansions include providing *hangul* text and Korean definitions, as well as including Chinese definitions, readings in pinyin and Wade-Giles systems, and forms of the characters used in Taiwan and mainland China when they differ from forms used currently in Japan.

A hyperdictionary can be enhanced by various multimedia features. For example, Quicktime movies may show native speakers pronouncing and using words in context. Pictures, sound, and movies can also be used to illustrate features for which words give an inadequate understanding. A hyperdictionary could include virtual-reality walkthroughs of architectural models, three-dimensional views of statuary or ceramics, samples of bird songs, the sounds of musical instruments, or animated diagrams of the movement of sub-atomic particles in a cyclotron.

The features mentioned above can be developed using current technology. The barriers to developing a hyperdictionary are not technical but political and economic

in nature. For example, we must either negotiate permission to use existing databases or assume the labor-intensive task of developing our own. The open-ended nature of the Internet, on the other hand, permits us to begin with a small subset of the hyperdictionary and add databases and multimedia features as we produce or acquire them. By so doing, we hope to develop a tool that will be invaluable for both students and scholars.

Acknowledgements

The CGI scripts connecting the user interface with the Prolog database were written by Masashi Adachi. The database that we have constructed uses elements from Jim Breen's on-line database, and from the graphical analysis of characters by Martin Dürst.

SELECTED BIBLIOGRAPHY

Abramson, H., "A Logic Programming View of Relational Morphology," *Proceedings of COLING-92*, the Fourteenth International Conference on Computational Linguistics, Nantes, France, July, 1992, pp.850-854.

Abramson, H., S. Bhalla, K. Christianson, J. Goodwin, J. Goodwin, J. Sarraille, "Towards CD-ROM based Japanese <-> English Dictionaries: Justification and Implementation Issues," *Natural Language Processing Pacific Rim Symposium*, 95, Seoul, Korea, Dec. 1995.

Abramson, H., S. Bhalla, K. Christianson, J. Goodwin, J. Goodwin, J. Sarraille, L. Schmitt, "The Logic of Kanji Lookup in a Japanese <-> English Hyperdictionary," Association of Literary and Linguistic Computing, and Association of Computers and Humanities, 1996 Joint International Conference ALLC/ACH '96, University of Bergen, Norway, June 1996.

Abramson, H., S. Bhalla, K. Christianson, J. Goodwin, J. Goodwin, J. Sarraille, L. Schmitt, "Multimedia, Multilingual Hyperdictionaries: A Japanese <-> English Example," Association of Literary and Linguistic Computing, and Association of Computers and Humanities, 1996 Joint International Conference ALLC/ACH '96, University of Bergen, Norway, June 1996.

Abramson, H., S. Bhalla, K. Christianson, J. Goodwin, J. Goodwin, J. Sarraille, 1997, "On-line Japanese Character and Word Recognition Using Databases of Identifiable Features of Characters and Compounds," 5th International Conference on Japanese Information in Science, Technology & Commerce, Washington, D.C., 30 July-1 Aug 1997.

Abramson, H., S. Bhalla, K. Christianson, J. Goodwin, J. Goodwin, J. Sarraille, L. Schmitt, "Recognition of Japanese Characters by Non-Native Learners Through a Support Database System," *Proceedings, Second International Conference on Cognitive Technology*, Aizu-Wakamatsu, Japan, August 1997.

Abramson, H., S. Bhalla, K. Christianson, J. Goodwin, J. Goodwin, R. Sharp, J. Sarraille, J. Yamadera, "A Web-based High-level Multi-purpose Tool for Japanese Information Processing: Possibilities and Problems," 6th International Conference on Japanese Information in Science, Technology & Commerce, Stockholm, Sweden, September 1999.

Dürst, Martin. "Coordinate-independent Font Description Using Kanji as an Example." Electronic Publishing 6:3, pp. 133-143.