Modeling Learners' Cognitive Processes for Improved Dictionary Accessibility

SLAVEN BILAC[†], TIMOTHY BALDWIN[∗] and HOZUMI TANAKA[†]

[†] Tokyo Institute of Technology 2-12-1 Ookayama, Meguro Tokyo, JAPAN
`{sbilac,tanaka}@cl.cs.titech.ac.jp`

[∗] Center for the Study of Language and Information (CSLI)
210 Panama Street
Stanford, CA 94305-4115, USA
`tbaldwin@csli.stanford.edu`

Abstract

Learners of Japanese face great difficulty when trying to lookup words containing kanji in a dictionary, due to the requirement of knowing the correct reading of the target word. We propose a system that imitates the cognitive process learners go through in generating readings for novel kanji strings, and provide direct access to the dictionary entries based on the generated readings. In doing so we remove the correct-reading requirement. The system described here is implemented in a web-based environment and freely available for general use. In this paper we provide an analysis of query and error data collected by our servers.

1. Introduction

Learning a foreign language is a time-consuming and painstaking process, and made all the more daunting by the existence of unknown words (Groot 2000). Without a fast, low-cost way of looking up unknown words in a dictionary, the learning process is impeded (Humble 2001). This is particularly true in non-alphabetic languages such as Japanese, as there is no easy way of looking up the component characters of new words. This research attempts to alleviate the dictionary look-up bottleneck by way of a comprehensive dictionary interface which allows Japanese learners to look up Japanese words in an efficient, robust manner.

The Japanese writing system consists of the three orthographies of hiragana, katakana and kanji, which appear intermingled in modern-day texts. The hiragana and katakana syllabaries, collectively referred to as kana, are relatively small (46 characters each), and most characters take a unique and mutually exclusive reading which can easily be memorized. Kana thus do not present a major difficulty for the learner. Kanji characters (ideograms), on the other hand, present a much bigger obstacle. The high number of these characters (1,945 prescribed by the government for daily use, and up to 3,000 appearing in newspapers and formal publications (NLI 1986)) in itself presents a challenge, but the matter is further complicated by the fact that each character can and often does take on several different and frequently unrelated readings (broadly divided into *on* readings of Chinese origin and *kun* readings of Japanese origin: Nakayama (2002)). These readings often undergo morpho-phonological changes, such as gemination and sequential voicing, in the process of word formation (Tsujimura 1996). The kanji 発, for example, has several readings including *hatsu*[1] and *ta(tsu)*, whereas 表 has readings including *omote*, *hyou* and *arawa(reru)*. Learners presented with the string 発表 *happyou* "announcement"[2] for the first time will, therefore, have a possibly large number of potential readings (conditioned on the number of component character readings they know) to choose from.

With Japanese paper dictionaries, look-up typically occurs in two forms: (a) directly based on the reading of the entire word, and/or (b) indirectly in a kanji dictionary via component kanji characters and an index of words involving those kanji. Clearly in the first case, the correct reading of the word must be known in order to look it up. Quite often, this is an unreasonable assumption. In the second case, the complicated radical and stroke count systems make the kanji look-up process cumbersome and time consuming. For example, to lookup 遷移 *seNi* "transition" without knowing the correct reading the user needs to lookup each character individually (i.e. lookup 遷 via its radical

---

[1] In this paper, we loosely follow the Hepburn system of romanization, with the exception that we romanize long vowels as separate characters giving rise to *hyou* instead of *hyoo* or *hyō* for ひょう. The other notable divergence—taken from (Backhouse 1994)—is the use of the upper-case *N* for syllable-final nasals (corresponding to the kana ん) and lower-case *n* for syllable-initial nasals (as found in な, for example).

[2] Here, *hatsu* undergoes gemination and *hyou* sequential voicing to produce *happyou*.

辷 or stroke count of 15 and 移 via its radical 禾 or stroke count of 11).[3]

With electronic dictionaries – both commercial and publicly available – the options are expanded somewhat. In addition to reading- and kanji-based look-up, for electronic texts, simply copying and pasting the desired string into the dictionary look-up window gives us direct access to the word.[4] Several reading-aid systems (e.g. Reading Tutor[5] (Kitamura & Kawamura 2000; Kawamura 2000) and Asunaro[6] (Nishina et al. 2000; Nishina et al. 2002)) provide greater assistance by segmenting longer texts and outputting individual translations for each segment (word). While these dictionaries and reading aides are a welcome addition to the learner's repertoire, they provide little help to the user when the text is not available in electronic form. To deal with texts available only in hard copy the user still needs to input the word into the dictionary interface. It is often possible to use kana-kanji conversion to manually input component kanji, assuming that at least one reading or lexical instantiation of those kanji is known by the user. Essentially, this amounts to individually inputting the readings of words the desired kanji appear in, and searching through the candidates returned by the kana-kanji conversion system. Again, this is complicated and time inefficient, hence the need for a more user-friendly dictionary look-up method remains. Finally, many electronic dictionaries support the use of regular expressions (REGEXPs) in searches, enabling lookup of words when partial input is possible (Breen 2000). However, such queries often result in a large number of responses, making it hard to locate the desired entry even when it is included in the system output.

In order to allow the user to maximize the use of available knowledge of kanji characters and their readings and remove the requirement that the user possesses the correct reading knowledge of the word he is trying to lookup, we have implemented the FOKS (Forgiving Online Kanji Search) system. The system is a web-based facility that allows the user to enter the estimated reading of a novel word. Based on the input reading the system calculates the dictionary entries that could be perceived as taking that reading and displays the candidates for the user to choose from. Since the (minimum) input is a single field (i.e. the estimated reading), the system can readily be used by small hand-held devices such as web-enabled cellular phones. Furthermore, the system can easily be integrated with a voice recognition/input system to completely remove the need to type the input.[7]

Once the candidate entries are displayed, the user can easily select the target word from the list of candidates to obtain the translation of the word. For example, the user can search for the string 頭上 *zujou* "overhead" by in-

---

[3]The kanji radical and stroke count system leaves a lot of room for error. For example, 遷 also contains the radicals 西 and 己, whereas 移 also contains the radical 夕, potentially leading to confusion as to which radical to look up the character under.

[4]Although even here, life is complicated by Japanese being a non-segmenting language, putting the onus on the user to correctly identify word boundaries.

[5]http://language.tiu.ac.jp/

[6]http://hinoki.ryu.titech.ac.jp/

[7]The limiting factor here is the accuracy of the speech recognition system.

putting the reading *toujou* or *atamajou*, derived from more common readings of the characters 頭 and 上, *tou/atama* and *jou*, respectively. We have previously demonstrated that this system is effective in guiding the user to the target word even when queried with an incorrect reading (Bilac et al. 2003).

In this paper we provide an analysis of query data collected by our dictionary server in an attempt to evaluate the adequacy of the error model used to predict (that is generate and score) erroneous readings and evaluate the effectiveness of the system in leading the user to the dictionary entry based on the incorrect reading.

The remainder of this paper is structured as follows. Section 2. describes common error types and their causes. Section 3. describes the current version of the FOKS system and the error types that it is able to handle. Finally, Section 4. provides an analysis and evaluation of the system performance.

2. Common reading errors

Previously Bilac et al. (2003) proposed a classification of common learning errors according to several basic types. While this classification was adequate for constructing the system, we felt that a more fine-grained classification was necessary to describe the errors actually appearing in the query data. Accordingly we classify the error types as follows:

1. *Inadequate choice of kanji reading.* Even though the learner is aware of the different readings a kanji character can take, he is unable to decide on the proper reading in the given context. For example, 大 can be read as either *tai*, *dai* or *oo(kii)*, so the string 大会 *taikai* "convention, congress" could well be misread as *ookai* or *daikai*.

2. *Confusion as to the vowel length.* For example, 主催 *shusai* "organization, sponsorship" can be mistakenly read as *shuusai*. This error type is common in speakers of languages which have no vowel length distinction.

3. *Inadequate palatalization.* Palatalized sounds are replaced with their non-palatalized counterparts (or the opposite.) For example, *ryuu* in 亜流 *aryuu* "epigone" is erroneously replaced by *ruu* giving rise to the incorrect reading of *aruu*.

4. *Incorrect voicing.* The learner substitutes an unvoiced consonant with its voiced counterpart or the other way around. For example, 団 *dan* "group" and 塊 *kai* "lump" form a compound 団塊 *dankai* "baby boom" rather than *dangai*.

5. *Incorrect application of gemination rules compound governing reading formation.* For example, 脱 *datsu* "remove,leave" and 出 *shutsu* "leave" form the compound 脱出 *dasshutsu* "escape", but readings such as *datsushutsu* or *dashutsu* could equally arise from the component character readings.

6. *Insufficient knowledge of other phonological changes governing word formation.* The learner is aware of the correct reading of each character in the compound but cannot derive a correct reading to compound specific phonological change. For example knowing 春 *haru* "spring" and 雨 *ame* "rain" does not entail knowing 春雨 *harusame* "spring rain".

7. *Errors due to graphic similarity of characters.* This is a common error type in learners from non-kanji language backgrounds. For example, 墓 *bo, haka* "grave" and 基 *ki, moto* "base" are graphically very similar, leading to the possible substitution of readings among words containing them (e.g. 墓地 *bochi* "graveyard" and 基地 *kichi* "base").

8. *Errors due to graphic and phonetic similarity of characters.* When kanji have similar-sounding readings, the probability of error increases. For example, besides sharing the pronunciation of *in*, the character 闇 *yami,in* "darkness" incorporates the character 音 *oto,on,in* "sound" leading to the possibility of "borrowing" other readings of 音, namely *on* or *oto*.

9. *Errors due to character co-occurrence.* The learner applies a reading from a kanji character occurring in the same context. E.g., a user unfamiliar with the character 激 *geki, hage(shii)* "violent, fierce" might apply a known reading of the character 厳 *geN, kibi(shii)* "strict,severe" due to the common kana suffix (i.e. 激しい *hageshii* "violent" vs. 厳しい *kibishii* "strict").

10. *Insufficient knowledge of proper nouns (personal and place names).* In Japanese, kanji characters in proper nouns often take on archaic readings which are hard to predict.

11. *Insufficient knowledge of idiomatic expressions.* Also common are the cases of *ateji* where kanji characters are assigned to a concept based on their reading without regard for their meaning (e.g. 珈琲 *koohii* "coffee") or the characters are assigned based on meaning while the original (foreign) pronunciation is maintained (e.g. 煙草 *tabako* "tobacco" is formed from 煙 *kemuri* "smoke" and 草 *kusa* "grass"). These strings need to be learned case by case and therefore cause significant problems even for native speakers.

12. *Errors due to character-level semantic similarity.* Characters like 右 *migi* "right" and 左 *hidari* "left" have a similar meaning and as such are easily confused by learners, resulting in an erroneous reading. Semantic confusion sometimes occurs at the word level, too, such as between 火事 *kaji* "fire" and 火災 *kasai* "(disastrous) fire".

13. *Inadequate kana content.* Kanji characters often occur with kana endings (*okurigana*) in free text. Based on the available knowledge, learners can choose to ignore the kana content when postulating the reading[8] For example, 滑稽 *kokkei* "comical" can be read as *suberukei* due to knowledge of 滑る *suberu* "(to) slide,slip".

---

[8]As mentioned in Section 1., kana characters are not considered as viable reading error sources so we assume that the user knowingly ignores the kana content.
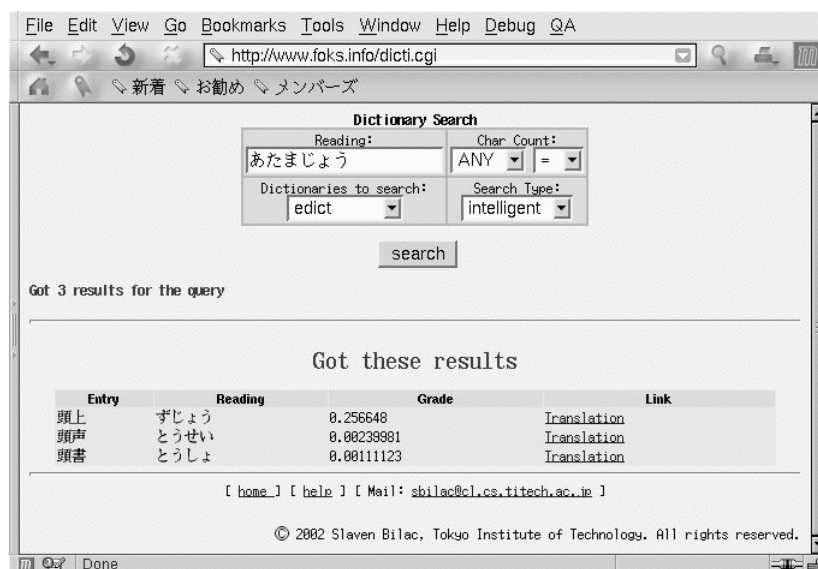
Figure 1.: Results displayed by FOKS system as a response to the query *atamajou*

14. *Other*. Any leftover cases which do not fit into the above classes are lumped into this class.[9] An example of an actual system query which falls into this class is *chiNke* which the system was (inexplicably) able to predict as an incorrect reading for the desired 塵芥 *chiriakuta* "garbage".

As can be seen from this classification, a larger number of causes can affect the derivation of an incorrect reading for a target entry. Quite commonly, several causes combine simultaneously, making the classification difficult. In Section 4., we will look into the observed distribution of these error classes. But first, we will introduce the FOKS system in greater detail.

3.   System description

The FOKS system was implemented at the Tokyo Institute of Technology as a means of improving dictionary accessibility for learners of the Japanese language. It is based on the notion that learners acquire Japanese character readings gradually, starting with the most common characters and readings and then moving on to less frequent ones. Due to such ordering of the learning process they might be unable to construct the prescriptively correct reading for a novel string, even though familiar with some (or all) of the characters contained in the string.

---

[9]When we generate novel readings for a dictionary entry, we apply several substitutions/changes (based on the data extracted from the dictionary and training corpus) one after the other, potentially making the derivation process and the base form unintelligible to our judge.

The majority of electronic dictionaries require accurate reading knowledge in order to look up a desired dictionary entry, since they employ direct matching sometimes augmented with REGEXPs (see above) as the lookup criteria. Unlike most other dictionary interfaces, the FOKS system does not assume correct reading knowledge of the target string, but instead tries to estimate what string the user is looking for based on the input reading. The system judges the plausibility (in the form of a likelihood score) of each reading-dictionary pair based on the probability of each kanji character taking a particular reading and the overall reading undergoing further morpho-phonological changes. The corpus frequency is then combined with the calculated probability to produce the overall plausibility score of the reading given the desired dictionary entry.

The system is built under the assumption that the cognitive process a user goes through in deriving a reading for a novel string is the following:

1. For each kanji in the word postulate a plausible (possibly erroneous) reading.

2. Form an overall reading for the word by combining the individual readings of all components. If the overall reading is not satisfactory (and other readings for individual kanji are available) re-postulate individual kanji readings to improve the overall reading.

3. When necessary, apply any phonological/morphological changes to the overall reading to get the final reading postulate.

Depending on the proficiency of the learner, the number of choices available at each step varies.

The system imitates this cognitive process by first calculating the probabilities of each reading given a kanji character and the probabilities of various morpho-phonological changes affecting the overall reading based on dictionary data and a training corpus (Bilac et al. 2002). Then, for each dictionary entry we apply the extracted readings and their calculated probabilities to generate novel readings which we score with a likelihood measure based on the above probabilities and corpus frequencies. The score is assigned under the assumption of segment independence, thus failing to take into account the interaction of various readings and phonological changes. Although this deviates slightly from the observations of the researchers (Itô & Mester 1995; Itô & Mester 2002; Frellesvig 1995) it simplifies the calculations significantly and should still be adequate for modelling learners of the language.

Currently, FOKS handles all the error types given in Section 2. except for types 7 and 12, although the handling of types 8 and 14 is limited.

3.1. Implementation Details

The base dictionary for the FOKS system is the publicly-available EDICT Japanese-English electronic dictionary.[10] We extracted all entries containing at least one kanji character and created a set of novel (potentially erroneous) readings, which we scored for plausibility as described above. Corpus frequencies calculated over the complete set of 200,000+ sentences in the EDR Japanese corpus (EDR 1995) were used to obtain the final plausibility measure. Once the complete set of readings is generated, it is stored in a relational database and queried through CGI scripts. Since the readings and scores are pre-calculated, there is no time overhead in response to a user query.

Figure 1. depicts the system output for the query *atamajou*.[11] We can see that the target string 頭上 is the highest ranking candidate, meaning that the user can easily access its translation even though the initial query was based on an incorrect reading. When the user clicks on the desired entry the original input and target entry pair is stored in system logs (as full queries, see below). Notice that conventional electronic dictionaries employing a direct match search criterion would not return any candidates since *atamajou* is not a valid word in Japanese.

In addition to reading-based search, the system provides the means to limit the search space by various additional constraints (number of characters, prefix, etc.). Also, the users can search in "simple" mode whereby the candidates are selected based on direct match with only the correct readings, effectively reducing the search ability to that of a conventional dictionary system.

The system is available for public use and easily accessible through any Japanese language-enabled web browser. Currently, only a Japanese-English dictionary is included but it would be a trivial task to add links to translations in alternative languages provided dictionaries are available. Furthermore, since the whole generation and scoring process is fully automatic it would be straightforward to apply the same process to a different Japanese dictionary.[12]

4. Evaluation

The FOKS website was initially made accessible to the public in November 2001. Since that time we have collected logs for all queries to our system (94,180 queries to date). Based on the sequence of user input, queries can be divided into two groups: **full queries** with input reading and target dictionary entry pairs recorded[13] and **partial queries** where it is not possible to

---

[10] `http://www.csse.monash.edu.au/~jwb/edict.html`

[11] This is a screen-shot of the system as it is visible at `http://www.foks.info/`.

[12] For example, we have employed the same methodology to allow searching ENAMDICT, a proper name dictionary in EDICT distribution, on the same principles.

[13] Here, the user entered a reading and subsequently selected the target entry from the list of candidates displayed.

Target Entry	Incorrect Input	Frequency
市井 *shisei* "town"	*ichii*	120
乱高下 *raNkouge* "fluctuations"	*raNkouka*	99
山車 *dashi* "festival car"	*yamasha*	78
幕間 *makuai* "intermission"	*makukaN*	58
山車 *dashi* "festival car"	*yamaguruma*	51
殺陣 *tate* "sword battle"	*satsujiN*	37
山車 *dashi* "festival car"	*yamakuruma*	35
心太 *tokoroteN* "gelidium jelly"	*shiNbuto*	27
お土産 *omiyage* "souvenir"	*odosaN*	16
⋮	⋮	⋮
玄人 *kurooto* "expert"	*geNjiN*	1
百方 *hyappou* "in every way"	*hyakue*	1
石筍 *sekijuN* "stalagmite"	*kokuiN*	1
踏襲 *toushuu* "follow"	*fumushuu*	1
返戻 *heNrei* "giving back"	*heNmodosi*	1
工作 *kousaku* "construction"	*kosaku*	1
地点 *chiteN* "spot"	*jiteN*	1
一息 *hitoiki* "one breath"	*ichii*	1
候間 *sourouraida* "as ..."	*koukaN*	1

Table 1.: Example full queries

determine the target entry.[14] The latter group is ignored for the purposes of evaluation since we only aim to analyze how the system models user errors at the reading-level. We thus use only the full queries in evaluation.

The complete set of full queries contains 5,820 input/target entry tokens. Of these, we analyzed 4,675 token pairs comprising 2,658 distinct types. In 2,076 cases (1,158 distinct types), or 44.4% of the data, the input reading is not the correct dictionary reading of the target entry. The high percentage of queries with an erroneous reading clearly shows that the ability to handle reading errors helps the user get to the target entry in a large number of cases. Table 1. gives some examples of stored pairs with their occurrence frequency.

The remaining 2,599 pairs (1,594 distinct types) were cases where the input reading was the correct dictionary reading of the target entry. In such cases our system returned 12.9 candidates on average and the target entry mean rank was 3.2. This shows that the number of candidate entries is low enough not to obscure the correct ones in the list and that the additional search ability does not unduly penalize competent users.

As a second step, we looked at each unique input/target pair where there the input reading did not correspond to the correct reading of the target entry (1,158 pairs in all), and classified each entry as corresponding to one or more of the error types listed above. The error analysis was conducted by a teacher

---

[14]Here, the user either queried the system with a string containing kanji characters or a regular expression to obtain the translation directly, the user was only interested in the reading and did not click through to the translation, or the target entry was not available in the candidate listing.

Error Type(s)	Frequency	Example Target String	Example Input
1	708 (61.1%)	家裁 *kasai* "family court"	*yasai*
1,5	50 (4.3%)	祝勝 *shukushou* "victory celebration"	*shukkatsu*
2	44 (3.8%)	阻止 *soshi* "obstruct"	*soushi*
5	44 (3.8%)	大会 *taikai* "convention, congress"	*dakkai*
1,11	38 (3.3%)	水無月 *minazuki* "June"	*mizunaitsuki*
14	37 (3.2%)	塵芥 *chiriakuta* "garbage"	*chiNke*
1,4	37 (3.2%)	着信 *chakushiN* "arrival"	*kijiN*
1,2	35 (3.0%)	笑い事 *waraigoto* "laughing matter"	*shoiji*
4	30 (2.6%)	端子 *taNshi* "terminal"	*taNji*

Table 2.: Most common error types (multiple error types reflect the existence of multiple errors in the input query)

Error Type	Frequency	Percentage
1	926	79.9
5	101	8.7
2	87	7.5
4	84	7.3
11	48	4.1
14	37	3.2
10	16	1.4
3	15	1.3
6	14	1.2
13	13	1.1
7	9	0.8
8	3	0.3
12	3	0.3
9	3	0.3

Table 3.: Individual error types by frequency

of Japanese as a foreign language with over 30 years of teaching experience. Table 2. gives the most common error types with representative examples. We can see that (uniquely) type 1 errors are the most common, accounting for 61.1% of the data, and an additional 18.8% when instances classified according to multiple error types are factored in. From this we can conclude that choosing the correct reading for a character based on the available context is the hardest problem the learner has to deal with. Such error distribution is adequately reflected in the reading set we generate as the majority of the probability mass is assigned to type 1 errors.

From Table 2. we can see that instances of multiple error classification are relatively frequent (19.5% of all types), underlining the effectiveness of our model at modeling the effects of compound errors. To ascertain the relative impact of the individual error types, we calculated the proportion of queries

for which a given error type was evident, as given in Table 3.[15] From this, we can once again see that error type 1 (reading choice) is the most prevalent source of reading confusion, but also that phonology (error types 4, 5 and 6) leads to errors in 17.2% of the cases. Note the appearance of error types 7 and 12 in Table 3. despite there being no explicit handling of them in the reading generation process. This is due to other error types conspiring to produce readings which coincide with the effects of graphic and semantic similarity, respectively.

For 3.2% of input readings, our judge could not determine the source of reading error (and hence assigned error type 14). There are two possible explanations for this. One is that the error model used in our system allows for the application of various sources of confusion in a layered fashion, ultimately masking the individual error types. The other possible explanation is that foreign language learners sometimes obtain and store reading knowledge based on its context and self-derived rules which native speakers (even teachers of the language) cannot readily identify. However, this hypothesis would need further exploration on larger data sets.

The most significant shortcoming of our evaluation is that we only have full queries in cases where our system offered the target entry as a candidate. The failure of the system to return the target entry can be ascribed to: (a) the target entry not being in the dictionary, (b) the system not handling the error type present in the query, or (c) the provided input being too ill-motivated to make the connection with the target entry. With the current system setup, we have no way of differentiating these three effects, and moreover, we are unable to determine which of the partial queries were successful (but the translation was not accessed) and which were not.

5. Conclusion and future work

The FOKS system is a Japanese dictionary interface aimed at removing the presupposition of complete and correct reading knowledge in the word look-up process. By allowing access to dictionary entries based on (predictably) incorrect readings, FOKS encourages the user to maximally use available knowledge. In this paper we have explained the basic model employed in guiding the user to the dictionary entry based on the input reading, and analyzed the error distribution in actual input data obtained from system query logs. The system logs contain initial input and target entry pairs which were classified according to the error type (if any) appearing in the initial reading input. Using this data we demonstrated that the cognitive process users go through when postulating a reading for a novel string containing kanji is adequately captured by our probabilistic model.

Finally, we identified two major types of error common in learners of Japanese but currently not handled adequately: errors due to graphic or se-

---

[15]Percentages do not add to 100% since a single input can involve more than one error type.

mantic similarity of kanji. In the future, we would like to expand our model to incorporate handling of these factors.

Acknowledgements

References

Backhouse, A. E.: 1994, *The Japanese Language: An Introduction*, Oxford University Press.

Bilac, S., T. Baldwin & H. Tanaka: 2002, 'Bringing the dictionary to the user: the FOKS system', in *Proc. of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan.

Bilac, S., T. Baldwin & H. Tanaka: 2003, 'Improving dictionary acessibility by maximizing use of available knowledge', *Traitement automatique des langues*, **44:2**.

Breen, J.: 2000, 'A WWW Japanese Dictionary', *Japanese Studies*, **20**: 313–317.

EDR: 1995, *EDR Electronic Dictionary Technical Guide*, Japan Electronic Dictionary Research Institute, Ltd., (In Japanese).

Frellesvig, B.: 1995, *A Case Study In Diachronic Phonology, The Japanese Onbin Sound Changes*, Aarhus University Press.

Groot, P. J. M.: 2000, 'Computer assisted second language vocabulary acquisition', *Language Learning & Technology*, **4**(1): 60–81.

Humble, Ph.: 2001, *Dictionaries and Language Learners*, Haag + Herchen.

Itô, J. & R. A. Mester: 1995, 'Japanese phonology', in J.A. Goldsmith, ed., *The Handbook of Phonological Theory*, Blackwell, chap. 29, pp. 817–838.

Itô, J. & R. A. Mester: 2002, 'The phonological lexicon', in Natsuko Tsujimura, ed., *The Handbook of Japanese Linguistics*, Blackwell, chap. 3, pp. 62–100.

Kawamura, Y.: 2000, 'The role of the dictionary tools in a Japanese language reading tutorial system', Ljubljana University International Seminar, (In Japanese).

Kitamura, T. & Y. Kawamura: 2000, 'Improving the dictionary display in a reading support system', in *International Symposium of Japanese Language Education*, Seoul, Korea, (In Japanese).

Nakayama, Mineharu: 2002, 'Sentence processing', in Natsuko Tsujimura, ed., *The Handbook of Japanese Linguistics*, Blackwell, chap. 14, pp. 398–424.

Nishina, K., M. Okumura, S. Sugimoto, Y. Yagi, T. Abekawa, N. Totsugi & F. Ryang: 2000, 'Development research on multilingual Japanese reading aid for foreign students with scientific background', *Research Report of Telecommunications Advancement Foundation*, **15**: 151–159, (In Japanese).

Nishina, K., M. Okumura, Y. Yagi, N. Totsugi, F. Ryang, S. Sugimoto & T. Abekawa: 2002, 'Development of Japanese reading aid with a multilingual interface and syntax tree analysis', in *Proc. of the Eight Annual Meeting of The Association for Natural Language Processing (NLP2002)*, Keihana, Japan, pp. 228–231, (In Japanese).

NLI: 1986, *Character and Writing system Education*, vol. 14 of *Japanese Language Education Reference*, National Language Institute, (in Japanese).

Tsujimura, N.: 1996, *An Introduction to Japanese Linguistics*, Cambridge, Massachusetts: Blackwell, 1st edn.