University of Ljubljana
FACULTY OF ARTS

Acta Linguistica Asiatica

Volume 2, Number 3, December 2012

Lexicography of Japanese as a Second/Foreign Language
Part 2

ACTA LINGUISTICA ASIATICA
Volume 2, Number 3, December 2012

Published by: Znanstvena založba Filozofske fakultete Univerze v Ljubljani
(Ljubljana University Press, Faculty of Arts)

Issued by: Department of Asian and African Studies

For the publisher: Andrej Černe, the Dean of the Faculty of Arts

Journal's web page:
http://revije.ff.uni-lj.si/ala/
The journal is published in the scope of Open Journal Systems

ISSN: 2232-3317

Abstracting and Indexing Services:
COBISS, Directory of Open Access Journals, Open J-Gate and Google Scholar.

Publication is free of charge.

Address:
University of Ljubljana, Faculty of Arts
Department of Asian and African Studies
Aškerčeva 2, SI-1000 Ljubljana, Slovenia

E-mail: mateja.petrovcic@ff.uni-lj.si

TABLE OF CONTENTS

FOREWORD

Having received a lively response to our call for papers on the lexicography of Japanese as a second language, the editorial board decided to dedicate two issues of this year's ALA to this theme, and I am happy to introduce the second round of papers, after the first thematic issue published in October this year.

This issue is again divided into two parts. The first two papers offer analyses of two aspects of existing dictionaries from the point of view of Japanese language learners, while the following four papers present particular lexicographic projects for learners of Japanese as a foreign language.

The first paper, by **Kanako Maebo**, entitled *A survey of register labelling in Japanese dictionaries - Towards the labelling of words in dictionaries for learners of Japanese*, analyses register labelling in existing dictionaries of Japanese, both in those expressly intended for learners of Japanese as a second language and those intended for native speakers, pointing out how register information provided by such dictionaries is not sufficient for L2 language production. After stressing the usefulness of usage examples for learners trying to write in Japanese, she offers an example of a corpus-based register analysis and proposes a typology of labels to be assigned to dictionary entries, calling for the development of corpora of different genres to be used for lexical analysis.

In the second paper, *An analysis of the efficiency of existing kanji indexes and development of a coding-based index*, **Galina N. Vorobeva** and **Victor M. Vorobev** tackle one of the most time-consuming tasks learners of Japanese are confronted with: looking up unknown Chinese characters. After a comprehensive description of existing indexes, including less known indexing systems developed by Japanese, Chinese, Russian and German researchers, they compare the efficiency of these systems using the concept of selectivity, and propose their own coding-based system. Although searching for unknown characters is becoming increasingly easy with the use of optical character recognition included in portable electronic dictionaries, tablets and smart-phones, not all learners have yet access to such devices. Efficient indexes for accessing information on Chinese characters are therefore still a valuable tool to support language learners in this most tedious task, while the ability to decompose a character into component parts remains an important basis for character memorisation.

The second part of this issue presents four projects aimed at supporting particular lexical needs of learners of Japanese as a second language.

In the first paper, *Development of a learners' dictionary of polysemous Japanese words and some proposals for learners' lexicography*, **Shingo Imai** presents a new lexicographic approach to the description of polysemous words. As Imai rightfully stresses, the most basic and common words learned by beginning language learners are actually often very polysemous; being deceivingly simple at first glance, they are often introduced with simple glosses or basic prototypical examples at the first stages of learning, and later treated as known words in intermediate or advanced textbooks, even if used for less common senses which are still unknown to the learners, causing much

confusion. In the dictionary series presented here, polysemous headwords are thoroughly and systematically described within their semantic networks, where the connections between core and derived meanings are schematically visualised and exemplified.

The following two papers present two of the first and most popular web-based systems for Japanese language learning support, both of which have been developing for more than a decade, supporting Japanese language learners all over the world.

Reading Tutor, a reading support system for Japanese language learners, presented by **Yoshiko Kawamura**, is a widely known and used system based at Tokyo International University, which offers automatic glossing of Japanese text with Japanese definitions and examples, and translations into 28 languages. After introducing the system, its development, functionalities and its tools for signalling the level of difficulty of single words, characters, or whole Japanese texts, the author describes its possible uses in language instruction and autonomous learning, and one concrete example of its application to the development of learning material for a specific segment of learners, foreign candidates to the Japanese national examination for certified care workers, mostly Filipino and Indonesian nurses working in Japan. The author concludes with suggestions for fostering autonomous vocabulary learning.

The other Japanese language learning support system with an equally long and successful tradition, developed at Tokyo Institute of Technology, is presented by its initiator, **Kikuko Nishina**, and one of its younger developers, **Bor Hodošček**, in *Japanese Learning Support Systems: Hinoki Project Report*. The article presents the many components of this successful system, including Asunaro, a reading support system aimed especially at science and engineering students and speakers of underrepresented Asian languages, Natsume, a writing assistance system using large-scale corpora to support collocation search, Natane, a learner corpus, and Nutmeg, an automatic error correction system for learners' writing.

The last project report, by **Tomaž Erjavec** and myself, introduces resources and tools being developed at the University of Ljubljana and at Jožef Stefan Institute: *JaSlo: Integration of a Japanese-Slovene Bilingual Dictionary with a Corpus Search System.* The dictionary, corpora and search tools are being developed primarily for Slovene speaking learners of Japanese, but part of the tools, particularly the corpus of sentences from the web-harvested texts, divided into five difficulty levels, can be used by any learner or teacher of Japanese.

I hope you will enjoy reading these articles as much as I did, and wish you a peaceful New Year.

Kristina Hmeljak Sangawa

RESEARCH ARTICLES

A SURVEY OF REGISTER LABELLING IN JAPANESE DICTIONARIES: TOWARDS A BETTER LABELLING FOR LEARNERS OF JAPANESE

Kanako MAEBO[*]
Hitostubashi University
xiangcai2@gmail.com

Abstract

Writing by learners of Japanese as a foreign language often contains words that do not fit the style of their context. One possible reason for this is the lack of information on word connotation and usage labels in existing dictionaries. The present study examines the current state of connotational information and register labels in Japanese learner's dictionaries, Japanese language dictionaries and dictionaries of synonyms, and proposes a possible technique for analysing the words' descriptions. The study reveals that Japanese learner's dictionaries and dictionaries for Japanese native speakers use different register labels and assign them from a different perspective. In the case of synonyms, presenting them in their context of use appears to be more useful than only listing them. Finally, the *Balanced Corpus of Contemporary Written Japanese* (BCCWJ - *Gendai Nihongo Kakikotoba Kinkoo Koopasu*) is used to show how corpora can be a very important linguistic resource for the analysis of lexical register.

Keywords

register; style; Japanese learner's dictionary; Japanese language dictionary; corpus

Izvleček

Izdelki piscev, ki se učijo japonščino kot tuj jezik, pogosto vsebujejo besede, katerih slog se ne ujema s slogom sobesedila. Eden od možnih razlogov za to je pomanjkanje oznak o rabi in informacij o konotacijah besed v obstoječih slovarjih. Članek analizira informacije o konotacijah in oznake o rabi v učnih slovarjih japonščine kot tujega jezika, enojezičnih slovarjih japonščine in slovarjih japonskih sinonimov. Ugotavlja, da slovarji za učence japonščine kot tujega jezika in slovarji za govorce japonščine kot maternega jezika uporabljajo različne oznake in označujejo besede z drugačnih zornih kotov. Opisi sinonimov, ki vključujejo tudi sobesedila, v katerih se pojavljajo, se izkažejo za bolj koristne od golih seznamov. Analiza s pomočjo uravnoteženega korpusa sodobne pisane japonščine (BCCWJ) pokaže, kako so lahko korpusi zelo pomemben jezikovni vir pri analizi leksikalne zvrstnosti.

Ključne besede

zvrstnost; slog; učni slovar japonščine; slovar japonščine; korpus

[*] Translated by Jasmina Krofič and Kristina Hmeljak Sangawa

1. Introduction

The writing of learners of Japanese as a foreign language (hereafter referred to as learners) often exhibits problems at various levels, including grammar, vocabulary or expression, composition and argumentation. The choice between colloquial or typically written language expressions which fit the style of writing is often a problem even for advanced level learners.

The cause of this lies in the specific characteristics of the Japanese language. Japanese is considered to be a language which strictly distinguishes between written and spoken language. According to Nakamichi (1989) "in Japanese there is an extremely high demand to maintain a fixed style within a text. The stylistic value of a word sometimes exhibits as much regulatory force as its semantic value. Furthermore, it may even happen that stylistic suitability is given priority over grammatical correctness or appropriate meaning."[1]

The major problem leaners face here is that despite these rigorous restrictions on the style of texts and words, information on the stylistic characteristics of words readily available for learners is limited. While producing texts, learners make use of dictionaries to investigate word meaning and usage. However, judging from learners' writings, it is often clear that they did not get enough information about the stylistic peculiarities of the words they use. Their knowledge of a word's stylistic characteristics is usually acquired through lists of differences between written and spoken language which can be found in textbooks on writing for foreign language learners, or from their teachers' instructions during composition classes.

This is hardly a desirable state of affairs for learners and teachers. The compilation of a learner's dictionary enabling the learners to acquire sufficient information regarding a word's stylistic characteristic is therefore highly desirable.

2. The stylistic characteristics and register (*isō* 位相) of words

A word's stylistic characteristics are "those characteristics held by each particular word which form the style of an entire text"[2] (Miyajima, 1972). Miyajima (1977) divides words into three classes, with the pivot class of stylistically neutral "everyday words" (*nichijōgo* 日常語), alongside which exist "written language words" (*bunshōgo* 文章語) and "slang words" (*zokugo* 俗語). Each of these classes is further divided into subclasses.

---

1 日本語においては、文章を通して一定の文体が維持されることは非常に強い要求であって、語の文体的特徴は、時として語義的特徴と同等の規制力を発揮することがある。さらには、文法的・意味的な正しさよりも文体的なふさわしさが優先されることさえある。

2 文章全体としての文体をなりたたせるような、個々の単語の持っている特徴のこと

The stylistic characteristics of words have also been described as one aspect of *linguistic phase* (*isō* 位相 in Japanese) or register. According to Yonekawa (2002) a linguistic phase (*isō*) refers to the speaker or writer's social attributes (such as age, occupation, gender etc.), the mode of expression and communication used, the setting, the speaker or writer's psychological and language awareness etc[3]. Considering this definition, we may say that a word's stylistic characteristics are those aspects of a word which are focused on and which become evident in different settings and modes of expression.

When considering the use of words which fit the style of their context, it is necessary to take into consideration who the speaker or writer is and in what environment the words are being used. In the following sections, the word *register* shall be used as an equivalent of the Japanese term *isō* ("*phase*") to refer to this aspect of word use.

3. Aim and research method

The present paper aims at describing the present state of register information in existing dictionaries, and discuss methods that can be used to acquire such information. It reports on a survey of register information in learners' dictionaries, and on a second survey of register information in dictionaries targeted at Japanese native speakers.

In order to describe register information in existing dictionaries, the following dictionary elements are considered and analysed: editorial policy, explanatory notes and directions for use in the dictionary front- or back-matter, lists of symbols and abbreviations used in the dictionary, and mentions of register within dictionary entries. Four pairs of adverbs were selected from lists of expressions highlighting differences between "spoken expressions" and "written expressions" found in textbooks on writing reports and research papers targeted at foreign learners of Japanese. The eight adverbs are: *tabun* [たぶん "perhaps"] and *osoraku* [おそらく "probably"], *zenzen* [全然 "(not) at all"] and *mattaku* [まったく "entirely"], *zenbu* [全部 "all, wholly"] and *subete* [すべて "completely"], *ichiban* [一番 "most"] and *mottomo* [最も "most"]. These adverbs are presented in the textbooks as synonyms, but to be used in different types of texts. They were therefore chosen as the object of this analysis, because it is reasonable to expect that they are words necessitating some information on register in their lexicographical description.

3 表現主体の社会的属性(年齢、職業、性別等)、表現主体が使用する表現様式、伝達様式、場面、表現主体の心理・言語意識など

4. Results of the surveys

4.1 Survey of learners' dictionaries

For English, there are learners' dictionaries based on large scale corpora, which present the senses of each headword in order of frequency, and include common collocations, style labels and plenty of other information that is deemed useful to learners of English as a second language. In contrast, there are only very few learners' dictionaries of Japanese produced in Japan which can be of use to learners above the intermediate or advanced level. Dictionaries with at least 10,000 headwords include *The Kenkyusha's English-Japanese Japanese-English Learners' Pocket Japanese-English Learners' Dictionary – Kenkyūsha einichi nichiei poketto jiten* [研究社 英日・日英ポケット辞典] published by Kenkyusha in 1996 and reprinted in 2009, *Kodansha's Furigana Japanese Dictionary - Furigana eiwa-waei jiten* [ふりがな和英・英和辞典] published by Kodansha in 1999 and reprinted in 2008, the *Informative Japanese Dictionary - Nihongo o manabu hito no jiten* [日本語を学ぶ人の辞典 "Dictionary for people learning Japanese"] published by Shinchosha in 1995, and *Tuttle Concise Japanese Dictionary - Tuttle konsaisu eiwa ―waei jiten* [タトル・コンサイス英和・和英辞典] published by Tuttle in 2008. Considering that the receptive vocabulary of adult native speakers of Japanese is considered to be in the range of 40,000 to 50,000 words (Nihongo-kyōiku Gakkai, 1987, p. 295), the size of these dictionaries is clearly unsatisfactory for advanced learners.

Let us now consider to what extent information on word register can be obtained from these dictionaries. Of the dictionaries mentioned above, the following three have labels indicating register, and were further analysed in detail.

1) Kenkyusha's English-Japanese Japanese-English Learners' Pocket Dictionary (hereafter referred to as Kenkyusha's)
2) *Kodansha's Furigana Japanese Dictionary* (hereafter referred to as *Furigana*)
3) *Informative Japanese Dictionary* (hereafter referred to as *Informative*)

Kenkyusha's and *Furigana* are Japanese-English bilingual dictionaries, and they include both a Japanese-English and an English-Japanese dictionary bundled together, making it possible to look up words starting from either language. *Informative* is a bilingualised dictionary with Japanese headwords, definitions and examples in Japanese, and English and Chinese translations of most headwords.

The register labels used in these dictionaries are summarised in Tables 1 and 2.

Table 1: Register labels used in three learners' dictionaries
(dots ○ indicate that the label is used, minus-signs indicate that the label is not used)

	Kenkyusha's	***Furigana***	***Informative***
Formal	○	○	—
Informal	○	—	—
semi-formal	—	○	—
colloquial	○	○	—
Honorific	○	○	—
Humble	○	○	—
Polite	○	—	—
Crude	—	○	—
Rude	○	—	—
Brusque	○	—	—
Literary	○	—	—
hanashikotoba [話しことば - "spoken language"]	—	—	○
kakikotoba [書きことば - "written language"]	—	—	○

Table 2: Definitions of register labels in the three learners' dictionaries

Label	**Explanation**
Formal	[*Kenkyusha's*] a word used in formal official situations. [*Furigana*] Words marked as formal are characteristics of formal situations and are not likely to be used in casual conversation. This category includes predicate words that are more polite than semi-formal.
Informal	[*Kenkyusha's*] a word used in relaxed and friendly situations
semi-formal	[*Furigana*] This label refers to predicate words in what the Japanese refer to as ですます体 [*desu-masu-tai*, "*desu-masu style*"]. The semi-formal style, in contrast to the informal style, expresses politeness toward the person(s) the speaker is addressing.
Colloquial	[*Kenkyusha's*] an informal word used in conversation. [*Furigana*] Words marked as colloquial are characteristic of casual conversation and not likely to be used in formal situations.
Honorific	[*Kenkyusha's*] a word indicating respect for others [*Furigana*] This label is used for two types of words. One type is predicate words which express respect for someone by honoring the subject of a sentence. The other type is nouns that express respect for their referents.

Label	Explanation
Humble	[*Kenkyusha's*] a word indicating humility. [*Furigana*] This label is used for two types of words. One type is predicate words which express respect for someone else by humbling the subject of a sentence. The other type is nouns that express respect for someone else by humbling their referents.
Polite	[*Kenkyusha's*] a polite word
Rude	[*Kenkyusha's*] a potentially impolite or offensive word.
Crude	[*Furigana*] Words marked as crude are felt to be inappropriate in polite conversation.
Brusque	[*Kenkyusha's*] a potentially rough or abrupt word.
Literary	[*Kenkyusha's*] a word used in the written language.
hanashikotoba [話しことば "spoken language"]	[*Informative*] おもに話しことばとして使う。[Omo ni hanashi kotoba to shite tsukau - "Mainly used as spoken language"]
Kakikotoba [書きことば "written language"]	[*Informative*] 論文、レポートなどに使う。日常会話などは使わないもの。[Ronbun, repōto nado ni tsukau. Nichijō kaiwa nado wa tsukawanai mono - "Used in essays, reports etc. Not used in daily conversation"]

As can be seen in Table 2, *Kenkyusha's* and *Furigana* both use the labels *formal*, *informal/semi-formal*, *honorific*, *humble*, *colloquial* and *rude*/*crude*; *Kenkyusha's* also uses three additional labels, *polite*, *brusque* and *literary*, while *Informative* only uses two labels, *spoken* and *written*. These are labels for words which require some consideration regarding the setting in which they are used and the relationship between speaker, listener and referent. Considering the fact that honorific language and expressions indicating the speaker's respect for the listener or referent are very developed and frequent in Japanese, labels such as "honorific" and "humble" are indispensable in a dictionary. However, in these labelling schemes surprisingly little attention is given to words used in writing. Despite this information being as much important, only two of the three dictionaries surveyed above use any label for words used in writing, *literary* in *Kenkyusha's*, and *kakikotoba* ("*written language*") in *Informative*. The labels *formal* and *informal/semi-formal*, judging from their explanation, seem to be used for words used in spoken rather than written interactions.

Label explanations are most exhaustive in *Furigana*. Such explanations include the information needed by learners about the situations in which the words are actually used. However, as no examples are given, it is difficult to understand to what type of words these labels are applied. Examples would be helpful not only from the standpoint of learners, but also for teachers, to grasp the range of vocabulary indicated by each label.

In the following paragraphs, dictionary entries are analysed to observe whether labels are used to differentiate between synonyms used in different settings. Dictionary entries for the four pairs of adverbs listed in section 3 were found not to contain the information necessary to know which word should be used in what context, despite the

fact they clearly belong to different registers. Semantic descriptions were also unhelpful from this point of view, as can be seen in the following example in Table 3.

Table 3: Dictionary entries for ichiban and mottomo in Furigana
(translitteration in square brackets added by present author)

一番	[*ichiban*] the most (indicating a superlative) [→最も[*mottomo*]] 日本で一番高い山は富士山です。 [*Nihon de ichiban takai yama wa Fujisan desu.*] The highest mountain in Japan is Mt. Fuji.
最も	[*mottomo*] the most (indicating a superlative) [→一番[*ichiban*]] 水泳は最もおもしろいスポーツだと思う。 [*Suiei wa mottomo omoshiroi supōtsu da to omou.*] I think swimming is the most interesting sport. エベレストは世界で最も高い山です。 [*Eberesuto wa sekai de mottomo takai yama desu.*] Everest is the highest mountain in the world.

Table 3 shows parts of the entries for the words *ichiban* and *mottomo* in *Furigana*. Each entry is accompanied by a cross reference to their respective synonym. However, as the meaning definition is exactly the same and examples are exceedingly similar, these entries give no clue to the difference in register between *mottomo* and *ichiban*. Cross references to synonyms should be accompanied by some information on the differences between them, not only with regard to meaning, but also register. Usage examples should also be chosen or edited to indicate that words are used in different circumstances.

This survey of Japanese learner's dictionaries thus revealed that register labels are present, but that the information learners actually need, regarding the concrete situations of use, is unsatisfactory.

4.2 Survey of dictionaries for native speakers of Japanese

As there are currently hardly any dictionaries for learners of the Japanese language above the intermediate level, many foreign language learners in Japan use dictionaries designed for Japanese native speakers. Register labelling was therefore also surveyed in Japanese monolingual dictionaries and thesauruses.

4.2.1 Japanese dictionaries

Register labels found in 20 general dictionaries targeted at native speakers of Japanese, including concise monolingual dictionaries and thesauruses, were classified into the following 11 categories of labels: “archaism” [古語 *kogo*], “elegant, refined word” [雅語 *gago*], “written language” [文章語 *bunshōgo*], “spoken language” [口語 *kōgo*], “colloquial/slang” [俗語 *zokugo*], “jargon” [隠語 *ingo*], “vulgarism” [卑語 *higo*], “dialect” [方言 *hougen*], “male language” [男性語 *danseigo*], “female language” [女性語 *joseigo*] and “baby talk” [幼児語 *yōjigo*], as reported in Maebo (2009). The results of this survey seem to indicate at first sight that many more labels are used in dictionaries for native speakers than in learners’ dictionaries, but not all of these labels are used in all dictionaries; each dictionary surveyed contained on average 4 to 7 labels.[4] However, factors such as the lack of clear definitions of the registers mentioned, and the use of different labels for the same words in different dictionaries, make it difficult for learners to find the information they need regarding register in dictionaries targeted at native speakers. These dictionaries probably do not include such information because it may be considered to be something known by any adult native speaker of Japanese.

Let us then consider dictionaries targeted at younger native speakers of Japanese. Three dictionaries targeted at junior-high school students were surveyed:

1) *Benesse shinshū kokugo jiten* [ベネッセ新修国語辞典] (“*Benesse’s Japanese Dictionary New Edition*”, 2nd Ed., 2012),
2) *Gakken Gendai hyōjun kokugo jiten dai 2 han* [学研現代標準国語辞典第 2 版] (“*Gakken Contemporary Standard Japanese Dictionary 2nd ed.*”, 2011) and
3) Sanseido’s *Reikai shin kokugo jiten dai 8 han* [例解新国語辞典第 8 版] (“*New Japanese dictionary with examples, 8th ed.*”, 2012).

Register information contained in these dictionaries is discussed below.

4 Some of the terms in the above list of labels were not used as labels, but information on register was included in entry definitions. Such information can be equally of help to learners.

1) *Benesse shinshū kokugo jiten* [ベネッセ新修国語辞典]

Table 4: Register labels and explanations in *Benesse shinshū kokugo jiten*

位相の注意（使う場面と相手に注意することば） 使う場面と相手を考え、注意して使ったほうがよい語に次のマークを付けました。  友人同士でおしゃべりをするときになどには使えるが、目上の人や文章に使わない、くだけたことば　例）そいつ  友人同士でも使わないほうがよい、あまりよくないことば　例）ぐる  現在はあまり使われない、古い感じのすることば。文章につかわれることがある。 例）今宵
Translation: Register notes (for words that should only be used paying attention to the situation and to the person addressed) We added the following marks to words that should be used with care, considering the situation and the person to whom one is talking.  Informal words that can be used when talking between friends and the like, but not towards higher ranking persons nor in writing, e.g. *soitsu*.  Not very good words that should better not to be used even among friends, e.g. *guru*.  Old-fashioned words that are not used much at present. They are sometimes used in writing.

2) *Gakken Gendai hyōjun kokugo jiten dai 2 han* [学研現代標準国語辞典第 2 版]

Table 5: Register labels and explanations in *Gakken gendai hyōjun kokugo jiten dai 2 han*

用法…ことばの使い方の注意点。たとえば次のような注意点があります。 ・「文学的」…主に文学作品、詩歌で使われることば ・「文章語的」…主に書くときに使われることば ・「俗な言い方」…仲間うちでよく使われる俗語や隠語など ・「くだけた言い方」…主に話しことばとして使う

Translation:

Usage… Notes about the usage of words. For example, the following notes are used.

- "literary" … words used mainly in literary works and poems.
- "written" …words mainly used when writing.
- "slang" … slang and jargon frequently used among friends
- "informal language" …mainly used in spoken language

Other labels are not listed in the legend above, but can be found in the entries themselves, such as "rather old expression" [やや古い言い方 *yaya furui iikata*], "old expression" [古い言い方 *furui iikata*], "formal expression" [改まった言い方 *aratamatta iikata*], "rough, rude expression" [ぞんざいな言い方 *zonzai na iikata*].

3) Sanseido's *Reikai shin kokugo jiten dai 8 han* [例解新国語辞典第 8 版]

In Sanseido's *Reikai shinkokugo jiten* (Reikai's new Japanese dictionary, 8th ed.), there are no special notes on labels. However, headword explanations include phrases such as "formal expression for ..." [ーのあらたまった言いかた ... *no aratamatta iikata*], "formal word for ..." [ーの形式ばったことば ...*no keishikibatta kotoba*] , "old-fashioned word for ..." [ーの古めかしいことば ... *no furumekashii kotoba*] . There are also notes labelled "phrase" which includes supplementary explanations about the circumstances of usage.

As could be seen in the above survey of label explanations in the three dictionaries for junior-high school students, register labels found in these dictionaries were similar to and not more numerous than those found in dictionaries for adult speakers of Japanese. In other words, labels in dictionaries targeted at younger users also use a typology corresponding to Miyajima's (1977) strata of lexical style characteristics and are not assigned from the point of view of the concrete situation and medium of use. However, in *Shinshuu kokugo jiten*, register information does consider the circumstances of use and the person addressed, which is a point of view to be considered when labelling words according to register.

Let us now see what type of information is offered regarding the four pairs of adverbs listed in section 3. The entries for these headwords in the three dictionaries for Japanese junior-high school students contained hardly any information regarding the different use of these synonyms. There were only explanations regarding the fact that the adverb *zenzen* [全然 "(not) at all"], which is only used with predicates in the negative form in standard Japanese, is also used with affirmative predicates with the meaning "very" in casual colloquial language.

Since junior-high school students only rarely write short research papers or reports, information to distinguish between words typically used in academic writing and other everyday words could be considered unnecessary. However, dictionaries

targeted at adult users, which are also generally used by high-school students, do not carry such information either, supposedly because it is superfluous for adult native speakers of Japanese who already know how to distinguish between written and colloquial language. In other words, information about which words to use when writing cannot be found in Japanese dictionaries aimed at junior-high school students nor in general dictionaries for adult speakers of Japanese.

4.2.2 Thesauruses

Thesauruses are helpful for understanding the proper use of synonyms. Four Japanese thesauruses were surveyed for register information, and the following labels were found.

① Kadokawa ruigo shin jiten [角川類語新辞典] ("Kadokawa new dictionary of synonyms"), Kadokawa [角川] 1981/2002.

Register labels included: "literary" [文語 *bungo*], "elegant, refined" [雅語 *gago*], "written language" [文章語 *bunshōgo*], "everyday word" [日常語 *nihijōgo*], "spoken language" [口語 *kōgo*], "slang" [俗語 *zokugo*], "jargon" [隠語 *ingo*], "dialect" [方言 *hōgen*], "old-fashioned expression" [古風な表現 *kofūna hyōgen*], "male language" [男性語 *danseigo*], "female language" [女性語 *joseigo*], and "baby talk" [幼児語 *yōjigo*]. No notes within definitions or descriptions.

② *Tsukaikata no wakaru ruigo reikai jiten shinsouhan* [使い方のわかる類語例解辞典新装版] ("*Dictionaries of synonyms with examples to understand their use, new edition*"), Shōgakukan [小学館] 2003.

No register labels. Each dictionary entry describing a group of synonyms has a column labelled "Meaning and use of each word" with explanations on differences in meaning and usage.

③ *Ruigojiten* [類語辞典] ("*Dictionary of synonyms*"), Kodansha [講談社] 2008

Register labels included: "written language" [文章語 *bunshōgo*], "elegant, refined" [雅語 *gago*], "slang" [俗語 *zokugo*], "vulgarism" [卑語 *higo*], "baby talk" [幼児語・児童語 *yōjigo - jidōgo*], "female language" [女性語 *joseigo*], "male language" [男性語 *danseigo*].

Additional information is provided within definitions and in notes to single entries.

④ *Ruigigo tsukaiwake jiten* [類義語使い分け辞典] ("*Dictionary of synonym use*"), Kenkyūsha [研究社] 2007

No register labels. Entries contain explanations regarding registers.

As could be seen above, some thesauruses contain register labels, while others contain register information within single entries. Usage labels are used more often than in monolingual dictionaries, but their types are largely the same as in the monolingual dictionaries described in the previous section. Information on register is present not only in the form of labels, but also within additional explanations for single entries, and more detailed information can be found here than in monolingual dictionaries.

The same four pairs of adverbs listed in section 3. were analysed in these thesauruses, and the following information was found. Each piece of information is preceded by the number of the dictionary containing it.

Table 6: Register information for four pairs of adverbs in four different thesauruses

tabun	① — ② — ③ — ④ *Tabun* is a softer expression than *osoraku*.
osoraku	① — ② — ③ Has a rather formal nuance. ④ *Osoraku* is used in a formal polite style.
zenzen	① colloquial ② Comparing *zenzen* and *mattaku*, *zenzen* is a more informal expression. ③ — ④ —
mattaku	① — ② Comparing *zenzen* and *mattaku*, *zenzen* is a more informal expression. ③ — ④ —
zenbu	① Colloquial ② — ③ Used in a more spoken context than *subete* and in more concrete cases. ④ Z*enbu* means all parts and can be thought of as the spoken language variety of *subete*. However one part of it overlaps with the usage of *minna*, which is misleading.
subete	① colloquial ② — ③ — ④ Z*enbu* means all parts and can be thought of as the spoken language variety of *subete*. However one part of it overlaps with the usage of *minna*, which is misleading.

ichiban	① colloquial ② *Mottomo* is the more common word ③ An expression often used in everyday conversation, with the meaning of *mottomo*. ④ *Ichiban* is an expression used in spoken language, while *mottomo* is used in written language.
mottomo	① colloquialism ② *Mottomo* is the more common word. ③ An expression often used in everyday conversation and carries the meaning of *mottomo*. ④ *Ichiban* is an expression used in spoken language, while *mottomo* is used in written language.

When looking at the description of each word we see that register information is added to almost all words. Furthermore, such information is often given not only as a label, but in the form of an explanation. The thesauruses also differ from monolingual dictionaries in structure: synonyms referring to the same meaning are gathered in one entry, and there are therefore many comparisons between synonyms. The strength of this kind of description is the possibility of gaining information about the register of the searched word and its synonyms at the same time. This kind of description method would be welcome also in learner's dictionaries.

5. Information gained from corpora

In the preceding sections, register information was surveyed in learner's dictionaries, Japanese dictionaries and thesauruses, but it was found to be insufficient from the point of view of the situation in which a word is used. The following paragraphs discuss a method to describe such information in detail.

In recent years, corpora have been compiled for Japanese, and in 2011 the National Institute for Japanese Language and Linguistics constructed the Balanced Corpus of Contemporary Written Japanese (BCCWJ - *Gendai nihongo kakikotoba kinkō kōpasu* 現代日本語書き言葉均衡コーパス, Maekawa, 2008, 2011). The online corpus search application *chūnagon* has also been developed, making it possible to easily find collocation and other information by combining different search criteria. However, it is not possible to download the entire text with *chūnagon*[5], making it difficult to analyse register information in detail. Although there is also a CD which contains the whole text, the absence of searching tools makes it difficult to analyse the text without a certain knowledge of natural language processing or programming.

[5] https://chunagon.ninjal.ac.jp/

In 2012 however, the online searching system NINJAL-LWP for BCCWJ [6] (hereafter abbreviated to NLB) was developed (Pardeshi, 2012). The main feature of this system is corpus search tool for lexical profiling which offers a comprehensive picture of the collocational and grammatical patterns of lexical words (nouns, verbs, adjectives and adverbs). The system also offers information regarding the frequency of each word in each of the subcorpora of BCCWJ (books, Diet minutes, Yahoo! Q&A (*Chiebukuro*), Yahoo! Blogs) in terms of tokens per million words. Within the books subcorpus, the system can also show separately the number of tokens per million words only within dialogues or only within prose. This information is useful for grasping the situations and contexts in which each word tends to be used.

The four pairs of adverbs were analysed using NLB to determine what information can actually be acquired about them using this corpus and tool.

Table 7: Frequency of occurrence of the four pairs of adverbs analysed (no. of tokens per million words)

	Books		**Diet minutes**	**Yahoo! Q&A**	**Yahoo! Blogs**
	Prose	**Dialogues**			
tabun	24.8	84.93	33.23	121.07	60.02
osoraku	84.52	89.41	112.39	68.97	44.6
zenzen	23.26	142.27	70.09	130.62	107.95
mattaku	231.99	289.19	217.12	266.31	200.48
zenbu	59.9	199.96	191.54	106.82	96.28
subete	376.14	299.05	164.75	224.52	229.65
ichiban	57.32	129.01	19.74	46.67	30.84
mottomo	218.31	84.39	77.54	42.73	80.02

Table 7 shows the tendency of occurrence for each word.

In the book corpus, the adverbs *tabun*, *zenzen*, *zenbu* and *ichiban* tend to occur frequently in the dialogues, while the remaining prose contains more occurrences of *subete* and *mottomo*. *Osoraku* and *mattaku* were found more or less equally often in both types of texts. In the diet minutes, which record questions and answers in a very formal setting, the following tendency can be seen: *tabun* < *osoraku*, *zenzen* < *mattaku*, *zenbu* >*subete* and *ichiban* < *mottomo*. The texts of Yahoo!Q&A and Yahoo!Blogs are written by ordinary people, without any editing or monitoring. Yahoo!Q&A (*Yahoo!Chiebukuro*) consists of questions on any topic written by anonymous users and answers to these questions by other anonymous users; it can therefore be expected that these texts are written with some consideration and respect for the expected readers. The Yahoo!Blog subcorpus consists of texts in which writers

[6] http://ninjal-lwp-bccwj.ninjal.ac.jp/

often express their own opinions. Both the Yahoo!Q&A and the Yahoo!Blogs subcorpus contain texts of very different styles, and it is therefore difficult to draw any conclusion from the numbers in Table 7. It is however apparent that *tabun*, *mata* and *subete* tend to occur frequently.

On the basis of the above results, the following observations can be made about the tendencies of use of the four pairs of adverbs.

- *Tabun* occurs in dialogue-like texts. It also tends to occur in subjective texts.
- *Osoraku* occurs in explanatory texts. A tendency of usage in informal spoken settings can be observed.
- *Zenzen* occurs in dialogue-like texts.
- *Mattaku* occurs in dialogue-like and explanatory texts. However, it also tends to be used in formal spoken language and when the writer is very much aware of the readers of a text.
- *Zenbu* occurs in dialogue-like contexts.
- *Subete* tends to occur in explanatory texts and when the writer is very much aware of the readers of a text.
- *Ichiban* occurs in dialogue-like contexts.
- *Mottomo* occurs in explanatory contexts. It also tends to be used in formal spoken settings.

The texts gathered in these four corpuses are not necessarily homogeneous. It is therefore only possible to observe certain tendencies of each word's occurrence. However, by surveying these tendencies and frequencies, and further analysing register differences as they are reflected in patterns and tendencies of word usage in BCCWJ, it should be possible to describe the register of single words.

6. Conclusion

The present paper presented an overview of register labelling in learner's and monolingual dictionaries and thesauruses.

Learners' dictionaries contain relatively plentiful register information from the viewpoint of interpersonal relationships. However, the following kinds of information are still insufficiently provided: (1) information on how different synonyms are used differently in writing according to different registers and (2) register labels such as "archaic" an/or "literary" for words that are slightly outside everyday use but within target vocabulary for middle or advanced learners of Japanese.

On the other hand, there are more types of register labels in Japanese monolingual dictionaries for native speakers than in learners' dictionaries, but register labels tends to be assigned mostly to words with a clearly limited usage, such as "archaic", "elegant (poetic)", "slang", and "(technical) jargon", while there is hardly any information on register regarding everyday words. Also, register labels provided are often based on Miyajima's three-tier distinction of "everyday words" in the middle, with "written words" and "colloquial words" on both ends, without much regard to interpersonal aspects and situation of use, which would be more informative for learners of Japanese.

Thesauruses contain more register information than monolingual Japanese dictionaries, which is especially useful when comparing synonyms, to know which synonym is to be used in which context. Such discussions of the differences between synonyms of similar meaning would be welcome also in dictionaries for learners of Japanese as a second language.

Considering register information as has been found in the surveys described above, the following five groups of register labels can be considered necessary in dictionaries for learners of Japanese as a foreign language:

1) labels related to interpersonal relationships, for words which need to be used with careful consideration of the listener or reader, such as expressions indicating the speaker's respect for the listener or referent, or words which can be unpleasant to the listener;

2) labels for words used in spoken conversation, related to the setting in which the word is used, such as formal or informal occasions;

3) labels for words used mainly in writing, related to different settings, such as letters or other texts requiring formal formats, or texts such as reports which need to be written objectively;

4) labels for archaic expressions, rare words, literary and other expressions with limited use in contemporary language;

5) labels related to terminology from specialised fields.

By further dividing these five groups of labels and adding such detailed information to dictionary entries, it would be clear which words are used in what circumstances. With the help of such detailed register information, it would also be possible to clearly describe the use of synonyms, when including them in an entry. In addition, examples showing typical settings in which each word is used would also be of great benefit to learners.

In order to assign such detailed labels to dictionary entries, it is necessary to know in what contexts each word is used. This can be gleaned from corpora. In this paper, NLB was used to observe tendencies of use of some adverbs in different subcorpora. For a comprehensive and detailed lexicographic description, other specific and

homogeneous corpora are needed, such as spoken corpora or corpora of academic writing.

7. Further research

With the creation of large-scale corpora for Japanese, the resources needed for the analysis of lexical meaning, use, and collocations are gradually taking shape. However, in order to analyse lexical register, homogeneous specialised corpora need to be constructed, and this, in turn, requires the analysis and categorisation of textual characteristics. Such an analysis of different types of texts, aimed at further refining the labelling types mentioned above, will be the subject of further research.

References

Maebo, K. [前坊香菜子] (2009). Go no buntaiteki tokuchō ni kansuru jōhō ni tsuite no ichi kōsatsu – Kokugo jiten to ruigo jiten no chōsa kara [語の文体的特徴に関する情報についての一考察—国語辞典と類語辞典の調査から—]. *Hitotsubashi nihongo kyōiku kenkyū hōkoku* [一橋日本語教育研究報告]. 3: 50-60.

Maekawa, K. (2008). Balanced Corpus of Contemporary Written Japanese. In Huang, C.-R., Mikami, Y., Hasida, K., & Tokunaga, T. (Eds.) *Proceedings of The 6th Workshop on Asian Language Resources (ALR 6)*, 101-102. Retrieved July 2012 from http://www.aclweb.org/anthology-new/I/I08/I08-7.pdf#page=109.

Maekawa, K. (2011). Development of Japanese Corpora at the National Institute for Japanese Language and Linguistics: With Emphasis on Five Sources of Difficulty in Japanese Corpus Development. *Lexicography:Theoretical and Practical Perspectives (ASIALEX2011 Proceedings)*,17-26.

Miyajima, T.[宮島達夫] (1977). Tango no buntaiteki tokuchō [単語の文体的特徴]. In: *Muramatsu Akira kyōju kanreki kinen Kokugogaku to kokugoshi* [村松明教授還暦記念国語学と国語史]. Tōkyō: Meiji shoin [明治書院]. 871-903.

Nakamichi, M. [中道真木男] (1989). Go no buntaiteki tokuchō [語の文体的特徴]. In: *Keesu sutadei Nihongo no goi* [ケーススタディ日本語の語彙]. Tōkyō: Ōfūsha [桜楓社].

Nihongo-kyōiku Gakkai [日本語教育学会編] (1987). Nihongo-kyōiku jiten [日本語教育事典]. Kihon-go, Kiso-go [基本語、基礎語]. 295. Taishukan shoten [大修館書店]

Pardeshi, P. (2012). Compilation of Japanese Basic Verb Usage Handbook for JFL Learners: A Project Report. *Acta Linguistica Asiatica*, *2*(2), 37-63.

Yonekawa, A. [米川明彦] (2002). Dai 6 shō Isō to isōgo [第 6 章　位相と位相語]. In *Asakura Nihongo kōza 4 – Goi – imi* [朝倉日本語講座4　語彙・意味]. Tōkyō: Asakura shoten [朝倉書店].

Dictionaries

Benesse shinshū kokugo jiten [ベネッセ新修国語辞典] (2nd Ed.). (2012). Tokyo: Benesse [ベネッセ].

Gakken gendai hyōjun kokugo jiten [学研現代標準国語辞典] (2nd Ed.). (2011). Tokyo: Gakken Kyōiku Shuppan [学研教育出版].

Informative Japanese Dictionary - Nihongo o manabu hito no jiten [日本語を学ぶ人の辞典]. (1995). Tokyo: Shinchosha [新潮社].

Kadokawa ruigo shin jiten [角川類語新辞典]. (2002). Tokyo: Kadokawa [角川].

Kenkyusha's English-Japanese Japanese-English Learners' Pocket Japanese-English Learners' Dictionary – Kenkyūsha einichi nichiei poketto jiten [研究社 英日・日英ポケット辞典] . (1992). Tokyo: Kenkyusha [研究社].

Kodansha's Furigana Japanese Dictionary - Furigana eiwa-waei jiten [ふりがな和英・英和辞典]. (1999). Tokyo: Kodansha [講談社].

Tuttle Concise Japanese Dictionary - Tuttle konsaisu eiwa-waei jiten [タトル・コンサイス英和・和英辞典] (2008). Rutland, Vermont, & Tokyo: Tuttle.

Reikai shin kokugo jiten [例解新国語辞典] (8th Ed.). (2012). Tokyo: Sanseido [三省堂].

Ruigojiten [類語辞典]. (2008). Tokyo: Kōdansha [講談社].

Tsukaikata no wakaru ruigo reikai jiten shinsouhan [使い方のわかる類語例解辞典新装版]. (2003). Tokyo: Shōgakukan [小学館].

AN ANALYSIS OF THE EFFICIENCY OF EXISTING KANJI INDEXES AND DEVELOPMENT OF A CODING-BASED INDEX

Galina N. VOROBEVA[*]
Kyrgyz National University
g.vorobyova@yahoo.com

Victor M. VOROBEV
Kyrgyz National University
v.vorobyov@yahoo.com

Abstract

Considering the problems faced by learners of Japanese from non-kanji background, the present paper discusses the characteristics of 15 existing kanji dictionary indexes. In order to compare the relative efficiency of these indexes, the concept of selectivity is defined, and the selectivity coefficient of the kanji indexes is computed and compared. Furthermore, new indexes developed by the present authors and based on an alphabet code, a symbol code, a semantic code, and a radical-and-stroke-number code are presented and their use and efficiency are explained.

Keywords

kanji index; search; efficiency index; selectivity; kanji coding

Izvleček

Pričujoči članek obravnava lastnosti 15 obstoječih kazal v slovarjih kitajskih pismenk. Da bi primerjali relativno učinkovitost teh kazal, definiramo koncept izbirnosti ter izračunamo in primerjamo koeficient izbirnosti teh kazal. Nadalje predlagamo in predstavljamo rabo in učinkovitost novih kazal, ki smo jih osnovali na kodiranju z latiničnimi črkami, simbolnem kodiranju, semantičnem kodiranju in kodiranju na osnovi pomenskega ključa in števila potez.

Ključne besede

kazalo kitajskih pismenk; iskanje; indeks učinkovitosti; selektivnost; kodiranje kitajskih pismenk

[*] Translated by Boštjan Bertalanič, Andrej Bekeš and Kristina Hmeljak Sangawa

1. Background of the study

The most commonly known indexes used to look up kanji (Chinese characters) in character dictionaries are the radical index (部首索引 *bushu-sakuin*), the stroke-number index (総画数索引 *sōkakusuu-sakuin*) and the readings index (音訓索引 *onkun-sakuin*). The radical index is used when searching kanji by means of their radical; it consists of a list of all radicals arranged by increasing number of strokes, where each radical is followed by a list of characters belonging to this radical, and each list of characters with the same radical is usually arranged by increasing number of strokes. The stroke number index is used to look up characters when their number of strokes is known; it consists of all characters contained in a dictionary, arranged by increasing total number of strokes. The readings index is used to look up characters by their reading, and consists of a list of all readings of the characters contained in the dictionary, usually arranged in standard *gojūon* kana order.

However, it is also generally known that learners from non-kanji background, i.e. learners who are not familiar with Chinese character writing, find it difficult to use traditional character dictionaries. When searching in a traditional radical index, it is sometimes difficult to determine which part of the character is to be considered its radical, as in the case of 巨 where 工 is the radical. Traditional ordering is complicated also because it does not classify characters consistently according to shape, but rather takes into account meaning, such as in the case of the characters 間 (“between”), 閉 (“close”), 開 (“open”) which are indexed under radical 門 (“gate”), while the character 問 (“question”) is indexed under radical 口 (“mouth”) and 聞 (“listen”) under radical 耳 (“ear”). The user should therefore already know in advance the meaning of a character in order to look it up, which is seldom the case. Indexes ordered by number of strokes are also troublesome to use, since we tend to make mistakes when counting, and there are many characters with the same number of strokes. In order to use reading indexes, on the other hand, the user needs to know the reading of a given character in order to look it up, but most users from a non-kanji background generally look up characters exactly because they do not know how to read them.

2. Previous research

Previous research has aimed at developing more efficient search methods. Both in the cultural sphere using Chinese characters and outside of it, diverse types of search methods have been developed and are being used besides the above mentioned and well known radical index, stroke index or readings index. To mention a few, other methods include the “five step arrangement kanji table” (五段排列漢字表 *godan hairetsu kanji hyō*) developed by the Russian researcher Rosenberg (1916), the Four corner method (四角号碼 *Sì jiǎo hào mǎ*) developed in China (Wang, 1925), the phonetic key index (音符 *onpu*) developed by Shiraishi (1971/1978), the index of katakana shapes, the initial stroke pattern index and the index of meaning symbols

developed by Kanō (1998), the index of stroke order patterns by Wakao and Hattori (1989), the index of character shapes by Sakano, Ikeda, Shinagawa, Tajima and Tokashiki (2009), the key words and primitive meanings index by Heisig (1977/2001), a radical index consistently based on shape by Hadamitzky and Spahn (1981), the system of kanji indexing by patterns (SKIP) developed by Halpern (1988), and the Kanji Fast Finder system by Matthews (2004). The authors of the present paper have also contributed to research on character indexing (Vorobeva, 2009, 2011).

3. Research aims

The goal of this paper is to 1) describe existing character search methods and analyse their efficiency; and 2) develop an efficient character search method based on a coding of character which appropriately expresses character form.

4. Research method

Given the variety of existing character indexes, we considered that an evaluation and comparative analysis of their efficiency was necessary. In order to compare the efficiency of the various character indexes, in this study we decided to use the concept of *selectivity* which expresses the processing efficiency of computer data. With reference to character indexes, we introduced the concept of *selectivity coefficient* and used it to compare and assess the efficiency of existing character indexes by calculating their selectivity coefficient.

We then built some new types of character indexes on the basis of character codes which accurately express character form. We constructed a database with four kinds of character codes for all new *Jōyō kanji*, sorted the data according to the order used in dictionary compiling, and developed four indexes: an alphabet code index, a symbol code index, a semantic code index and a radical and stroke code index.

Finally, we compared and evaluated the efficiency of these four code indexes.

5. Types and characteristics of existing character indexes

In the following paragraphs we introduce 12 existing types of character indexes, omitting the *on-kun* readings index, the radical index and the stroke-number index already described in the introduction.

5.1 Five step arrangement kanji table - 五段排列漢字表

The “graphic system” search method developed by the Russian researcher Vasil’ev (1867) and the “five step arrangement kanji table” developed by Rosenberg (1916) are generally known as the “Russian graphic system” (“Kod_Rozenberga”, 2012).

This is a method of arranging characters by form. Users only need to remember some simple rules and can use this method even when they cannot determine which part is the radical, have problems counting the number of strokes and do not know how to read the character. The method was developed by professor Vasilij Pavlovič Vasil´ev, a Chinese language scholar at Kazan University in Russia, who extracted 19 kinds of graphic elements or strokes (see Figure 1) which compose Chinese characters, and classified each character according to its last stroke, i.e. the stroke which is written last.

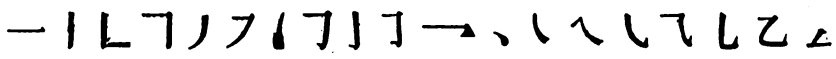


Figure 1: Vasile’v’s 19 graphic elements of Chinese characters

On the basis of these graphic elements, he compiled a new type of character dictionary with a unique character ordering and search method: the first Chinese-Russian dictionary, entitled Графическая система китайских иероглифов. Опыт первого китайско-русского словаря [*Grafičeskaâ sistema kitajskih ieroglifov. Opyt pervogo kitajsko-russkogo slovarâ. - “Graphic System of Chinese Characters. An attempt at the First Chinese-Russian Dictionary.”*] (Vasil’ev, 1867, see figure 2). This was the beginning of a far-reaching reform in dictionary structuring and organisation.

Figure 2: Vasil’ev’s *First Chinese-Russian Dictionary*

Two decades later, professor D. A. Pešurov at Saint Petersburg University published a Chinese-Russian dictionary entitled Китайско-русский словарь. по графической системе (*Kitajsko-russkij slovar´ po grafičeskoj sisteme* - *"Chinese-Russian Dictionary. Based on the Graphic System"*, Pešurov, 1891) on the basis of Vasil'ev's graphic method (see figure 3).


漢俄書法合壁字彙

КИТАЙСКО-РУССКІЙ

СЛОВАРЬ.

(ПО ГРАФИЧЕСКОЙ СИСТЕМѢ.)

Д. А. ПЕЩУРОВА.

ЦѢНА 2 РУБ.

САНКТПЕТЕРБУРГЪ.

1891.

Figure 3: Pešurov's Chinese-Russian Dictionary

The Russian graphic system was subsequently also used in a Japanese character dictionary, compiled by professor Otto Rosenberg of St. Petersburg University, who emphasised the importance of kanji learning stating as follows: "… especially for the person who is going to study the literary arts of Japan and Japanese civilisation of former ages, the knowledge of Chinese characters should become the core subject of learning"[1] (Rosenberg, 1916, p. 7). At the same time, he also discussed the difficulties to be faced when learning Chinese characters and using character dictionaries, and wrote the following comment regarding characters being listed according to their radicals: "I feel considerable inconvenience and difficulty, because Chinese characters do not possess an ordering such as alphabets[2]" (Rosenberg, 1916, p. 1).

On the basis of the graphic system developed by Vasil´ev, Rosenberg constructed a search system that was based on the shape of the characters and was completely different from the traditional indexes based on radicals, stroke number and readings, and used it to compile the 五段配列漢字典 (*Godan hairetsu kanjiten* - *"Five Step*

[1] 特に前代日本の文明，前代日本の文学芸術を知らんとする人に取りては，漢字の知識は，その主要科目たるべきなり。

[2] 非常なる不便と困難とを感じたり。それは主として漢字にアルファベットの如き順序なきによれるなり。

Arrangement Character Dictionary") a dictionary with a novel character arrangement and search system published in Japan (Rosenberg, 1916, see figure 4).

Figure 4: Rosenberg's 五段配列漢字典 (*Godan hairetsu kanjiten*, 1916)

The basic idea of the five-step arrangement search system is that "One look at the shape of the character [...] is enough. It can then immediately be found rapidly and reliably[3]" (Rosenberg, 1916, Explanatory notes 1).

Russian speakers are used to the notation in Cyrillic and Roman letters, and even in the case of Chinese characters they feel the need for a systematisation similar to the alphabet. Rosenberg took that into account when classifying and arranging characters according to the last written stroke. In contrast to Vasil'ev's system with 19 strokes (Vasil'ev, 1867), Rosenberg used 24 strokes and divided them into 5 groups. In order to implement a Chinese character ordering system based on stroke order, Rosenberg extracted 24 types of strokes, and grouped them into 5 categories, according to the direction in which they are written. Finally, he selected one stroke to represent each of the five groups, as shown in table 1 (Rosenberg, 1916, p. 20).

Table 1: Basic strokes in Rosenberg's system

Direction	Basic strokes
↗	丿
↘	乀
↙	丿
↓	丨

[3] 文字(中略)の形を, 一見したるのみにて, 十分なり。而して直ちに迅速確實に検出することを得べし。

→	一

Rosenberg exemplified the 5 types of strokes using the 5 strokes of the character 本, as shown in figure 5.

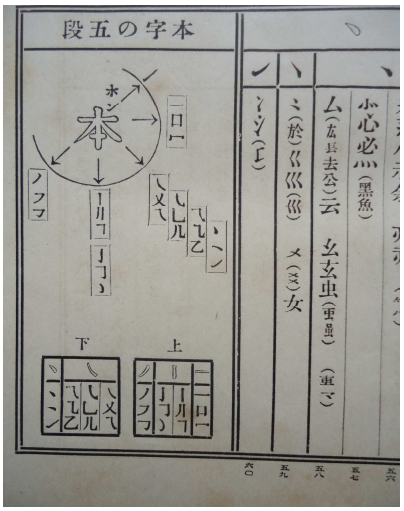


Figure 5: Rosenberg's five types of strokes, exemplified by the character 本

The Chinese character chart (漢字の字母表 *kanji no jibohyō*) in Rosenberg's (1916) dictionary is shown in figure 6, while figure 7 shows one page of the Five step arrangement Chinese character index (五段配列漢字表 *godan hairetsu kanjihyō*), and figure 8 shows one page of the dictionary's main entries.

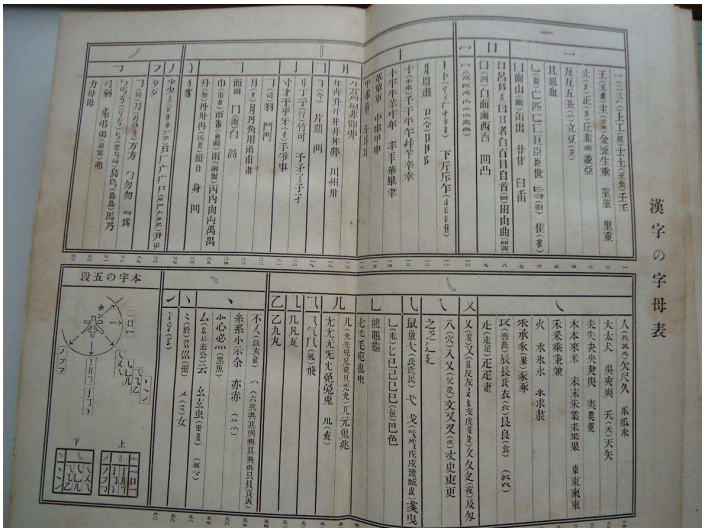


Figure 6: The Chinese character chart (漢字の字母表 *kanji no jibohyō*) in Rosenberg's dictionary

The Chinese character type chart (see figure 6) includes the five basic strokes, beneath them all 24 strokes classified by shape as belonging to one of the basic five strokes, and finally, beneath them, 567 character types (Chinese characters and character patterns) classified according to the 24 strokes and divided into 60 columns. Characters listed in the Five step arrangement Chinese character index (五段配列漢字表 *godan hairetsu kanjihyō*, see figure 7) are arranged according to the order of their strokes in the character chart. Figure 8 shows an example dictionary page.

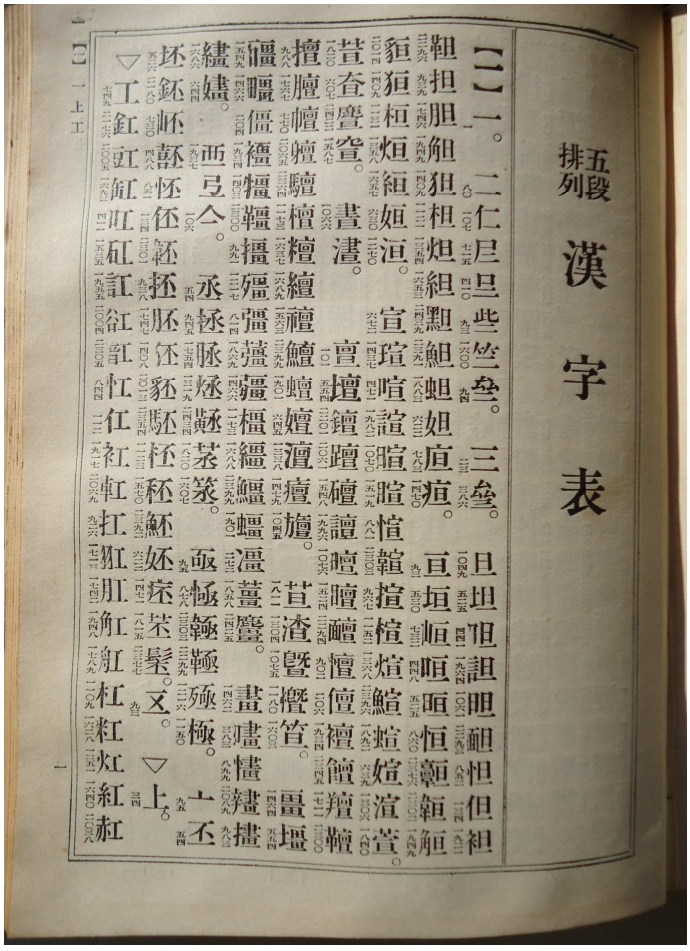


Figure 7: First page of the Five step arrangement Chinese character index (五段配列漢字表 *godan hairetsu kanjihyō*) in Rosenberg's dictionary

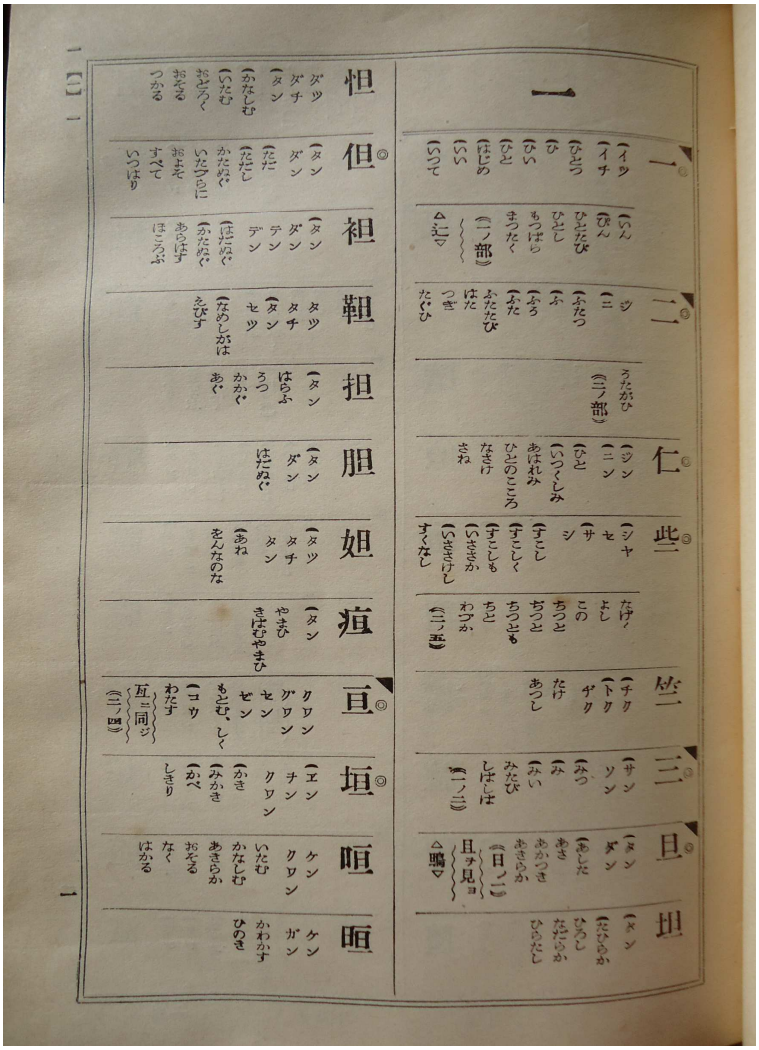


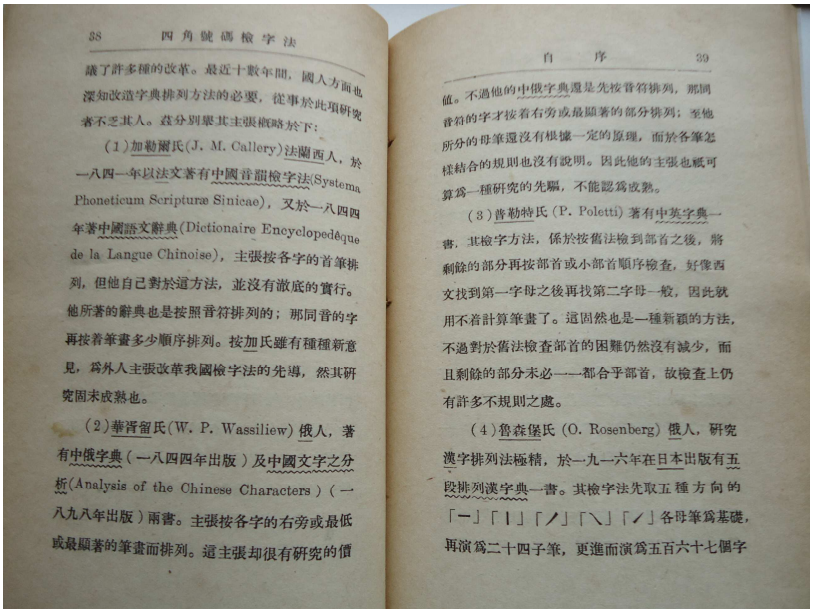
Figure 8: Examples of main entries in Rosenberg's dictionary

In the 19th century Russian researchers focused on the form of characters and, adhering to the principle of coherent classification, proposed a novel character sequencing and search method. In the former Soviet Union, this "Russian graphic system" was the basis for the Большой китайско-русский словарь (*Bol´šoj kitajsko-russkij slovar´* - "*The Great Chinese-Russian Dictionary*", Panasyuk & Suhanov, 1983). Later, the "Russian graphic system" for arranging and searching characters according to their stroke characteristics was also influential in the creation of the "Four corner method" (四角号碼 *Sì jiǎo hào mǎ*, see figures 9 and 10), developed in China during the 1920s. (Wáng, 1934, p. 38).

5.2 The Four corner method

The Four corner method (Chinese 四角號碼 *sì jiǎo hàomǎ*, Japanese 四角号碼 *skikakugōma*) is one of the Chinese character search systems and in the context of Japanese language means the "code based on four corners". It was developed in China by Wáng Yún Wǔ, who published his 號碼檢字法 (*Hàomǎ jiǎnzì fǎ - "Code based character search method"*) in 1925, 四角號碼檢字法 (*Sì jiǎo hàomǎ jiǎnzì fǎ - "Four corner code based character search method"*) in 1926, and finally 四角號碼檢字法· 附檢字表 (*Sì jiǎo hàomǎ jiǎnzì fǎ: fu jiǎnzì fǎ biao - "Four corner code based character search method: with a character look-up table appendix"*) in 1934. Just like the "Russian graphical system", the Four corner method does not depend on the radical, stroke number, stroke order, reading or meaning of a character, but rather allows for character look-up by means of a code which is based on the shape of the strokes in the four corners of a character. The stroke shapes in each of the four corners are assigned numbers from 0 to 9, and in order to distinguish between those Chinese characters which happen to end up with the same quadruplet of assigned digits, an extra "corner", named 附角 (*fujiǎo*) is additionally assigned. Each Chinese character can thus be uniquely coded and ordered by a five digit number. In order to use this Chinese character index, it is not necessary to have any knowledge of traditional search systems based on radicals, stroke number or stroke order.

Figure 9: Front page of 四角號碼檢字法 附檢字表
(Sì jiǎo hàomǎ jiǎnzì fǎ: fu jiǎnzì fǎ biao - "Four corner code based character search method: with a character look-up table appendix") (Wang, 1934)



38 四角號碼檢字法

議了許多種的改革。最近十數年間，國人方面也深知改造字典排列方法的必要，從事於此項研究者不乏其人。茲分別舉其主張概略於下：

（1）加勒爾氏(J. M. Callery)法蘭西人，於一八四一年以法文著有中國音韻檢字法(Systema Phoneticum Scripturæ Sinicae)，又於一八四四年著中國語文辭典(Dictionaire Encyclopedêque de la Langue Chinoise)，主張按各字的首筆排列，但他自己對於這方法，並沒有澈底的實行。他所著的辭典也是按照音符排列的；那同音的字再按着筆畫多少順序排列。按加氏雖有種種新意見，爲外人主張改革我國檢字法的先導，然其研究固未成熟也。

（2）華肯留氏(W. P. Wassiliew)俄人，著有中俄字典（一八四四年出版）及中國文字之分析(Analysis of the Chinese Characters)（一八九八年出版）兩書。主張按各字的右旁或最低或最顯著的筆畫而排列。這主張却很有研究的價

自 序 39

值。不過他的中俄字典還是先按音符排列，那同音符的字才按着右旁或最顯著的部分排列；至他所分的母筆還沒有根據一定的原理，而於各筆怎樣結合的規則也沒有說明。因此他的主張也祇可算爲一種研究的先驅，不能認爲成熟。

（3）普勒特氏(P. Poletti)著有中英字典一書，其檢字方法，係於按舊法檢到部首之後，將剩餘的部分再按部首或小部首順序檢查，好像西文找到第一字母之後再找第二字母一般，因此就用不着計算筆畫了。這固然也是一種新穎的方法，不過對於舊法檢查部首的困難仍然沒有減少，而且剩餘的部分未必一一都合乎部首，故檢查上仍有許多不規則之處。

（4）魯森堡氏(O. Rosenberg)俄人，研究漢字排列法極精，於一九一六年在日本出版有五段排列漢字典一書。其檢字法先取五種方向的「一」「丨」「丿」「㇏」「㇀」各母筆爲基礎，再演爲二十四子筆，更進而演爲五百六十七個字

Figure 10: Introduction to 四角號碼檢字法 附檢字表
Sì jiǎo hàomǎ jiǎnzì fǎ: fu jiǎnzì fǎ biao - “Four corner code based character search method: with a character look-up table appendix” (Wang, 1934, p. 38)

To give an example, the character 法 is assigned the code 34131, by assigning these numbers to the strokes in each corner, in the following sequence:

3 4
法 1
1 3

The Four corner method was adopted in the compilation of 大漢和辞典 (*Dai Kan-Wa Jiten*, Morohashi, 1960), *The Great Chinese Character - Japanese Dictionary* in 13 volumes published in Japan. The coding rules used in this dictionary are described in Morohashi (1984, p. 1038).

5.3 Katakana shape based classification

The Katakana shape based classification system (カタカナ字形分類索引 - *katakana jikei bunrui sakuin*, Kanō, 1998, p. 1007) sorts all the 1945 *jōyō kanji* according to the similarity of their component strokes to Japanese kana syllabary character shapes, with Chinese characters accordingly arranged as kana in the “a, i, u, e, o” order. Chinese characters are listed under the part that shares the same shape with some katakana character, as can be seen in table 2. The position of the katakana-like shape within each character is not questioned.

Table 2: Examples of Chinese characters listed under each katakana character

katakana form type	examples of corresponding Chinese characters
ア	了, 子, 孔, 好, …
イ	仙, 代, 付, 休, …
エ	工, 功, 式, 江, …
オ	才, 材, 財, 閉, …

5.4 Initial stroke pattern index

The Initial stroke pattern index (書き出しパターン索引 *kakidashi pataan sakuin*, Kanō, 1998, p. 1020) shares similarities with the aforementioned Five step arrangement Chinese character table developed by Rosenberg, but while Rosenberg operates with the strokes written at the very end, the Initial stroke pattern index deals with the strokes that are written first. The Initial stroke pattern index defines the six initial stroke patterns given in table 3.

Table 3: Initial stroke patterns in Kanō's Initial stroke pattern index (1998)

1	2	3	4	5	6
一	丨	ノ	丶	フ	レ

Characters with the same initial stroke pattern are ordered by their number of strokes. For example, characters belonging to pattern 1 (一) appear in the following order, according to their number of strokes 一, 二,丁, 三, 天, … .

5.5 Stroke order index

The Stroke order index (筆順索引 *hitsujun sakuin*) was implemented by Wakao and Hattori (1989) in their dictionary for the decipherment of cursive style characters (くずし解読字典 *Kuzusi kaidoku jiten*). Characters in this dictionary are ordered according to their stroke order and stroke direction. Brush movement is represented by arrows pointing to eight directions, with each direction being assigned a number (code) ranging from 0 to 7 (Wakao & Hattori, 1989, p. 469), as shown in table 4.

Table 4: Brush stroke direction patterns in Wakao & Hattori's Stroke order index (1989)

0	1	2	3	4	5	6	7
↑	↗	→	↘	↓	↙	←	↖

Each character is assigned a code, based on the direction of the first four strokes (起筆 *kihitsu* "first stroke", 第二筆 *dainihitsu* "second stroke", 第三筆 *daisanhitsu* "third stroke" and 第四筆 *daiyonhitsu* "fourth stroke"). For example, the first strokes of the character 仙 are written in the following directions in the cursive style: ↙ (5), ↗ (1), ↓ (4). Here the second stroke, ↗ (1), not visible in the printed form of this character, is included in cursive stroke counting because the brush is brought back to its starting point after the first stroke (Wakao & Hattori, 1989, p. 466). This coding system could easily be applied to standard character forms in general character dictionaries. In this case, the character 仙 would be coded according to its standard stroke order, ↙, ↓, ↓, ↓, resulting in the four digit code 5-4-4-4.

5.6 Key Words and Primitive Meanings Index

The Key Words and Primitive Meanings Index (Heisig, 2001, p. 506) assigns a unique meaning or interpretation to each character or character component part. English words representing these meanings are ordered alphabetically, making it possible to look up any character according to these English translations of assigned meanings. In order to use this index, the user must first learn the assigned meanings of each component part.

5.7 System of Kanji Indexing by Patterns (SKIP)

Halpern (1988/1990, 1999), developing his System of Kanji Indexing by Patterns (SKIP), assigned to each character a numeric code. In order to do this he first divided characters into four patterns, numbered from 1 to 4:

1 - ◧ Characters that can be divided into left and right parts;
2 - ⬒ Characters that can be divided into top and bottom parts;
3 - ◻ Characters that can be divided by an enclosure element;
4 - ■ Characters that cannot be classified under patterns 1, 2 or 3

These pattern numbers are used as the first digit in a code assigned to each characters. Characters of type 1 to 3 are divided into two parts. The number of strokes in each part of the character is then used as the second and third component of the character code. For example, the character 相 consists of a left and a right part and is thus categorised as type 1; the left part, 木, has 4 strokes and the right part, 目, has 5 strokes, resulting in its SKIP code 1-4-5. In the case of characters belonging to type 4, the second component of their code is the total number of strokes, while the third component is a code number, ranging from 1 to 4, assigned according to their shape. For details, readers can refer to the dictionary's detailed front matter. Further examples of SKIP codes are given in table 5 below.

Table 5: Examples of System of Kanji Indexing by Patterns (SKIP) coding

Type	Kanji	Number of strokes	SKIPcode
1 ◧	八	2	1-1-1
	相	9	1-4-5
2 ⬒	二	2	2-1-1
	父	4	2-2-2
3 ◻	山	3	3-2-1
	間	12	3-8-3
4 ■	火	4	4-4-4
	女	3	4-3-4

The dictionary contains a SKIP index constructed by ordering characters according to their SKIP codes in ascending order.

5.8 Fast Finder

Matthews (2004), in a way similar to Halpern (1988/1990) and Halpern (1999), assigns a pattern to each character, but does not create codes. Characters are divided into 8 patterns on the basis of their constituent components: left of the left-right, right of the left-right, top of the top-bottom, bottom of the top-bottom, three types of enclosures and non-divisible characters. Characters belonging to each pattern are listed together on pages beginning with the pattern itself, followed by all characters belonging to it. Lists of characters with complex shapes are further minutely subdivided and ordered according to the complexity of their shape, in ascending order of their number of strokes.

5.9 Index by Radicals

Hadamitzky and Spahn (1981) adopted an index based on traditional radicals, but reduced the number of radicals in order to simplify character search. They selected 79 of the 214 generally used radicals, which are also standardised in Unicode (Unicode, 2012), and used them to construct their Index by Radicals. Those radicals whose shapes were deemed too complex were omitted and characters traditionally assigned to them were assigned to other radicals. Some examples are given in Table 6 below. As a result of the reduced number of radicals, the number of characters listed under each radical has increased.

Table 6: Examples of radical substitution in Hadamitzky and Spahn (1981)

214 radicals in general use		79 radicals proposed by Hadamitzky and Spahn (1981)	
Radical number	**Radical**	**Radical number**	**Radical**
176	面	3s	口
177	革	3k	艹
178	韋	3d	口

5.10 Index by meaning symbols

The Index by meaning symbols (意味記号索引 *imikigō sakui*) lists characters according to 495 meaning symbols used by Kanō (1998), ordered according to their increasing number of strokes. Kanō (1998, p. 6) explains that "… We indicate the part that represents the meaning of a character as its 'meaning symbol'… Some of the 'meaning symbols' have the same shape as radicals, but are, as radicals, called differently (e.g. 水 *mizu* 'water' and 氵 *sanzui* 'three [drops of] water'). [...] It also happens sometimes that a character does not contain a 'meaning symbol'. In such cases we have shown the original character, on which the current character is based." For example, the radical of 斎 (nowadays usually substituted by 斉) is 斉, and its 'meaning symbol' is 示 (the shape of an altar) (cf. Kanō, 1998, p. 952).

5.11 Index by character shapes

The Index by character shapes (字形索引 *jikei sakuin*) is implemented in a text book including 512 Chinese characters (Banno, Ikeda, Shinagawa, Tajima & Tokashiki, 2009) and is similar to a radical index. However, the 215 "character shapes" (*jikei*) used include both radicals and other shapes which are traditionally not considered as radicals. The index lists "character shapes" ordered by their number of strokes, each followed by characters containing it, with their assigned numbers corresponding to the order in which they are introduced within the textbook.

5.12 Index by principal semantic determiners

The Index by principal semantic determiners (意符 *ifu*) developed by Shiraishi (1971/1978) is similar to a radical index, but instead of radicals relies on 243 characters representing principal semantic determiners. These determiners also include radicals, characters and character constituent elements that are not radicals. Principal semantic determiners are ordered by stroke number.

5.13 Final considerations

The above survey revealed a great variety of different types of character indexes. Among those which are based on character constituent elements and strokes, there are also indexes which assign numerical codes to characters. In order to compare and evaluate these indexes, we introduce the notion of "selectivity" (選択性 *sentakusei*). A comparison and evaluation of indexes based on selectivity is presented in the next section.

6. Evaluation and comparison of existing character indexes

6.1 Shared characteristics of existing character indexes

A point shared by all aforementioned character indexes is that only one character element or property is selected as the basis upon which the index is built. Elements or properties employed for this purpose are the number of strokes, radicals, initial stroke pattern etc. Considering the necessity to evaluate and comparatively analyse each of these indexes, we introduced the notion of selectivity of character indexes and used it to compare and evaluate existing indexes.

6.2 Definition of the "coefficient of selectivity"

To compare and evaluate the efficiency of character indexes, we will use a notion used to express computer processing efficiency, i.e., "selectivity" (Vorobeva, 2009, p. 72). The "selectivity" of an index has previously been defined as follows.

> "The ratio of the number of distinct values in the indexed column / columns to the number of records in the table represents the selectivity of an index.
>
> Example with good selectivity: a table having 100,000 records where one of its indexed columns has 88,000 distinct values, then the selectivity of this index is 88,000 / 100,000 = 0.88
>
> Example with bad selectivity: if an index on a table of 100,000 records has only 500 distinct values, then the index's selectivity is 500 / 100000 = 0.005 and in this case a query which uses the limitation of such an index will return 100000 / 500 = 200 records for each distinct value." (Akadia, 2008)

Based on the notion of selectivity, the "Coefficient of Selectivity" (hereafter CS) as a measure of index efficiency is defined as:

$$CS=V/N\times 100\%$$

where V stands for the number of distinct values in the indexed column(s) and N is the total number of records in the table.

Here, in the case of an index based on character shape, N is the total number of characters in the index, and V is the number of different groups to which characters are assigned. A group is for example a group of all characters under the same radical or a group of all characters with the same number of strokes. For example, in the case of an index by stroke numbers, V is the number of groups containing characters with the same number of strokes. In the case of an index based on radicals, V is the number of different radicals. On the other hand, in a phonetic based *on-kun* (loan and native reading) index, N is the total number of all loan (*onyomi*) and native (*kunyomi*) readings associated with all characters covered by the index, while V is the number of all distinct loan (*onyomi*) and native (*kunyomi*) readings.

To give an example of how to compute the coefficient of selectivity, let us use the old version of the *jōyō kanji* list (not the new *shin jōyō kanji*). The 1945 characters in the *jōyō kanji* list can be divided into 23 groups according to their number of strokes, or 201 groups of characters with distinct radicals (only 201 radicals are used in the 1945 jōyō kanji). Thus, for example, for the stroke number index and radical index we can compute the CS as follows:

stroke number index: V=23, N=1945, CS=23/1945×100％=1.2％

radical index: V=201, N=1945, and CS＝201/1945×100％=10.3％

From this we can conclude that the radical index is about 10 times more efficient than the stroke number based index. In the following subsections we are going to compare and evaluate the individual indexes according to this notion of “selectivity”.

6.3 Evaluation of efficiency of existing character indexes

In Vorobeva (2009), ten types of existing character indexes were compared, while in the present paper the object of comparison are 15 indexes. Existing indexes were compared on the basis of the CS (Table 7). In the table, comparison of indexes based on the character shape, character readings and character meaning is given.

It is clear from the above analysis that the CS of character shape based indexes varies between 1.2% and 25.4%, and is thus generally low. The possible reason for this is that such indexes are generally based only on one character property or element. For example, for each character, the radical index relies only on one element - the radical, the stroke number index relies only on one property - the stroke number, the initial stroke index relies only on the initial stroke, etc. Groups thus defined, i.e. groups of characters with the same radical, the same number of strokes or the same type of initial stroke, each contain a large number of different characters. On the other hand, the CS of indexes based on character reading vary in the range of 27.6~40.6％, and are more efficient if compared to character shape based indexes. However, in order to use these indexes, one has to know the readings in advance.

Table 7: Coefficients of selectivity of various character indexes

Index type	CS (%)
Indexes based on character form	
Stroke number index (Henshall, 1988)	1.2
Index by katakana shapes (Kanō, 1998)	2.6
Radical index (Hadamitzky & Spahn, 1981)	4.1
Initial stroke pattern index (Kanō, 1998)	6.1
Five step arrangement kanji table (Rosenberg, 1916)	7.1
Four corner method (Morohashi, 1984)	10.2
Radical index (Henshall, 1988)	10.3
Stroke order index (Wakao & Hattori, 1989)	10.7
Index by principal semantic determiners (Shiraishi, 1978)	12.4
Fast Finder (Matthews, 2004)	14.1
SKIP (Halpern, 1988)	15.4
Index by meaning symbols (Kanō, 1998)	25.4
Indexes based on character reading	
Index by principal phonetic determiners (Shiraishi, 1978)	27.6
On-kun reading index (Henshall, 1988)	40.6
Index based on character meaning	
Key Words Index and Primitive Meanings (Heisig, 2001)	100.0

Indexes based on character meaning, such as the Key Words Index and Primitive Meanings (Heisig, 2001) have a CS reaching 100%, but require the user to know in advance the meaning of all characters.

Based on these findings it can be said that indexes which rely on the coded total shape of a character, such as the two indexes above, are much more efficient then indexes relying on a single element or property. It can therefore be concluded that such an index is needed for efficient character searching. Thus, in the following section, a new efficient index will be proposed and developed. For this purpose, character were structurally decomposed and coded.

7. Structural decomposition and coding of characters

Three coding systems based on the structural decomposition of characters and three character databases built on the basis of these systems are introduced by Vorobeva (2011). In the present paper, we will discuss four systems of character coding. All characters contained in the *jōyō kanji* list, and also those added in the *shin jōyō kanji* list were encoded, and four databases were built on the basis of the following codes:

(1) an alphabet code
(2) a symbol code (containing roman letters and digits),
(3) a semantic code (expressed by words),
(4) a radical and stroke codes.

Structural decomposition of characters and semantic analysis of their constituent elements stimulate a deeper understanding of character meaning. Structural decomposition of characters can be achieved at two levels, i.e., decomposition into strokes (書記素 *shokiso*) and decomposition into constituent elements (the smallest meaningful constituent elements of Chinese characters, 構成要素 *kōseiyōso*).

Decomposition into strokes follows the stroke order, as in the following example:

女 → {く, 丿, 一}

Decomposition into constituent elements also follows the stroke order, but to obtain the smallest meaningful constituent elements, as in the following example:

露 → {雨, 足, 夂, 口}

After having defined the alphabet and number codes of strokes and constituent elements of each character, and the alphabet and number codes for each whole character, the complete *shin jōyō kanji* list was encoded.

7.1 Type of strokes and their encoding

7.1.1 The alphabet code of strokes

According to Zadoenko and Khuan (1993), Chinese characters used in China can be decomposed into strokes belonging to 24 different types. Fazzioli (1987) also uses almost the same strokes. In an analysis of the shapes of all characters in the *shin jōyō kanji* list, we found that the 24 types of strokes proposed by Zadoenko and Khuan (1993) are necessary and sufficient. We therefore coded these 24 shapes into Latin alphabet letters from A to Z based on the similarity of their shapes, so that for each stroke a corresponding letter could be guessed on the basis of its shape (cf. Table 8).

A	B	C	D	E	F
G	H	J	K	L	M
N	O	P	Q	R	S
T	U	V	W	Y	Z

Table 8: The 24 character strokes and their alphabet codes

7.1.2 The alphabet coding of characters and its use

An alphabet code was obtained for each character by following its stroke order and transforming each stroke into its corresponding alphabet code as given in table 8. Thus, the stroke order for each character could also be expressed by an alphabet code, as in the following examples:

三（AAA） 川（PBB） 玉（ABAAQ） 女（KPA） 小（JLQ）

All characters in the *shin jōyō kanji* list were encoded according to this procedure, and the alphabetic order of the code thus obtained was used to arrange the entries in a new character dictionary index. In this way a new index of character entries could be developed. An excerpt of the index is given in section 8, table 12.

7.2 Types of constituent elements and their encoding

7.2.1 Types of constituent elements

There are two types constituent elements which can form characters: elements which are radicals and elements which are not traditionally considered radicals but are patterns which correspond to radicals in some aspects. We name the latter “graphemes”. Constituents combine in complex ways to form individual characters.

7.2.2 Types and coding of radicals

Traditionally, 214 types of graphic patterns have been considered as radicals, as mentioned in the previous sections. Unicode 6.1.0 (Unicode, 2012) includes a table of these 214 radicals, and each radical is assigned a unique number. Analysing both the *jōyō kanji* list and the *shin jōyō kanji* list, we found that of these traditional 214 radicals, 201 types could be found as constituent elements of characters in the *jōyō kanji* list and 202 types as constituent elements of characters in the *shin jōyō kanji* list. In the present study, we assigned three different codes to each radical:

(1) an alphabetic code composed of the alphabetic codes of all strokes which make up the radical;

(2) a symbol code corresponding to the traditional numbering of radicals from 1 to 214;

(3) a semantic code consisting of a word expressing the principal meaning of each radical.

Examples of these codes are given in table 9.

Table 9: Examples of coded radicals

Radical	Alphabet code	Symbol code (number)	Semantic code (Principal meaning)
一	A	1	one
丨	B	2	stick
人	PO	9	man
山	BEB	46	mountain
馬	BABAAGLQQQ	187	horse

7.2.3 Types and coding of graphemes

Vorobeva (2011, p. 11) structurally decomposed each character from the *shin jōyō kanji* list and extracted all graphemes included in them, resulting in 161 different types of graphemes. This extraction was based on an intuitive notion of what would constitute a grapheme. In the present study, we first defined some rules to determine what part of the character can be extracted as a grapheme. We then structurally decomposed each character in the *shin jōyō kanji* list according to these rules, extracted their graphemes and found 220 different types of graphemes. Combining this list with the 202 traditional radicals found in the previous analysis, we found that the 2136 characters in the *shin jōyō kanji* list are composed of 422 different constituent elements.

The rules used to structurally decompose characters and extract graphemes are based on Stalph (1989, p. 69) and are given below.

1. Graphemes are limited to characters or character constituent elements included either in Unicode 6.1.0 (Unicode, 2012) or the "*Mojikyō tankanji*" list (Mojikyō Kenkyūkai, 2002). Unicode 6.1.0 contains 74,617 characters listed as CJK Unified Ideographs, and the "*Mojikyō tankanji*" list contains 110,000 characters. The reason for using these lists is that when characters are decomposed, the decomposition may result in shapes that do not appear in the *shin jōyō kanji* list.

2. When the decomposition of characters or constituent elements into smaller elements reaches a point where a stroke does not constitute an independent character, the decomposition is stopped and such a shape is considered a grapheme. For example: decomposing 及 further would give two elements: 乃 and 𠂉, but since 𠂉 is not an independent element, the decomposition is stopped and 及 is considered itself a grapheme. In shorthand notation: 及 ≠ 乃 + 𠂉

Attention is necessary in the case of 一 and 乙. Here they are considered as characters and not as strokes.

3. When the decomposition of a character or a constituent element reaches an element which by itself is not a radical but is a constant shape always used in

combination with a radical or other graphemes, this shape is considered a grapheme, without further decomposition. For example, the constituent element 辟 appears in characters such as 避, 壁, 癖, and 璧. It contains the element 尸 which by itself is not a radical and which regularly accompanies the radical 辛. This constituent element 辟 is not further decomposed, i.e. 辟 ≠ 尸 + 辛, and 辟 is considered as a grapheme.

4. When decomposing a character or a constituent element, if it is necessary to cut one stroke to obtain two independent shapes and if the total number of resulting strokes is then higher than in the original character or constituent element, then such character or constituent element is considered as a grapheme and is not decomposed further. For example:

出 ≠ 山 + 山, 重 ≠ 千 + 里.

Thus, 出 and 重 are considered as graphemes.

5. When decomposing a character or a constituent element, if it is necessary to insert a stroke into a decomposed element to make it complete and if the total number of resulting strokes of the two parts exceeds the number of strokes in the original character, then such a character or constituent element is considered a grapheme and not decomposed. For example: 雀 ≠ 少 + 隹 and thus 雀 is considered a grapheme.

6. When decomposing a character or a constituent element, if it is necessary to split two crossing strokes to arrive at a grapheme or a radical, such a character or constituent element is considered a grapheme. For example: 必 ≠ 心 + ノ, and thus 必 is a grapheme.

7. A character or constituent element which is deemed impossible to further decompose is considered a grapheme. For example: 奈, 喬, 兼, etc, are considered graphemes.

As the next step, all graphemes were assigned three types of codes according to the same procedure used for radicals. Details about encoding are given in Vorobeva (2011, p. 21). Examples of codes for graphemes which are component parts of characters in the *shin jōyō kanji* list are given in table 10 below.

Table 10: Examples of coding of graphemes included in characters in the *shin jōyō kanji* list

Grapheme	Alphabet code	Symbol code	Semantic code
丁	AJ	2AJ	street
丂	AN	2AN	snare
マ	YQ	2YQ	chop-seal
人	POA	3POA	meeting

7.2.4 Symbol codes, semantic code, radical and stroke code, and their use

Symbol codes and semantic codes of character are described in Vorobeva (2007, p. 22). The symbol code of each character is obtained by listing the symbol codes of each constituent element (cf. 7.2.2), following the stroke order.

The semantic code of a character is defined as the semantic code of its first two constituent elements, following the stroke order. Only the first two elements are used in order to avoid long codes.

Further, codes based on radicals and strokes are obtained by listing the numbers of elements which correspond to traditional radicals and the symbol codes (roman letters) assigned to each of the remaining strokes, following standard stroke order. Letters indicating strokes and numbers indicating radical shapes are divided by slashes, as in the following examples.

決 → {氵, 冂, 大} → 85 / H / 37,

今 → {亼, 一, フ} → 9 / 1 / Y。

Numbers and letters used in the above code are, for radicals, standard radical numbers from Unicode 6.1.0 (Unicode, 2012), and for strokes, the alphabet codes assigned to basic strokes (cf. table 8).

Following the above procedure, all characters in the *shin jōyō kanji* list were encoded and a database was constructed to include the alphabet codes, symbol codes, semantic codes, as well as radical and stroke codes. Examples of such coding are given in table 11 below.

Table 11: Examples of alphabet, symbol, semantic and radical-stroke character codes

character	Alphabet code	Symbol code	Semantic code	Radical and stroke code
九	PR	2PR	nine	4/5
逸	PYBHBAPCQMO	8PYB/162	escape/road	18/30/2/10/162
新	SAQLAABPOPPAB	117/75/69	stand/tree	117/75/69

8. Towards a new type of character index based on character coding

8.1 Construction and use of a new type of character index

Starting from the premise that it is necessary to develop a character index which would make search in character dictionaries more efficient and which would be suitable for learners not familiar with Chinese character writing, we developed a character code based on appropriate representations of character shapes (cf. Vorobeva

2007, 2009, p. 72). In order to develop such a new type of character index, we sorted the alphabet codes, symbol codes, semantic codes, and radical/stroke codes for all characters in the *shin jōyō kanji* list in standard dictionary order (alphabetical and numerical order). Searching for characters in such indexes should require the same amount of labour as searching for words in alphabetically ordered dictionaries to which learners from non-kanji background are accustomed.

A symbol code index and a semantic code index were developed and implemented in two character textbooks including 518 characters, *Kanji monogatari I* [漢字物語Ⅰ] (Vorobeva, 2007) and *Kanji monogatari II* [漢字物語Ⅱ] (Vorobev & Vorobeva, 2007).

Further, an alphabet code index, a symbol code index and a semantic code index were developed for the 1945 characters in the *jōyō kanji* list, selectivity coefficients were computed and compared with existing character indexes (Vorobeva, 2009).

In 2010, after the new *shin jōyō kanji* list with 2136 characters was approved, an alphabet code index, a symbol code index and a semantic code index were compiled for the new list, and a new, easier to use radical-and-stroke code index was also developed. The next sections introduce each of these new indexes.

8.2 Alphabet code index

The first part of the alphabet code index is shown in Table 12. In order to be able to use this index, one needs to memorise the 24 types of strokes and the rules governing stroke order of characters.

Table 12: First part of the alphabet code index for all characters in the *jōyō kanji* list

Alphabet coding of character	character
A	一
AA	二
AAA	三
AAABPQAAPB	耕
AAABPQPAAC	耗

Beginning learners are not yet accustomed to the constituent elements of characters. The alphabet code index is therefore expected to be easier to understand and use.

8.3 Symbol code index

Most characters are composite characters, composed of multiple constituent elements. We coded all characters in our textbook using radical and grapheme codes, compiled a database of characters with their respective radical and grapheme codes, sorted the database according to the code column, in standard dictionary ascending order (i.e. alphabetic and numerical order), and thus obtained a symbol code index. Users can thus search for characters in this index using codes based on character shape.

For example, the symbol code for the character 親 is 117/75/147. The numbers in the code are the numbers of the radical shapes into which the character can be decomposed, i.e. 立 (117), 木 (75), and 見 (147). To use the symbol code index, users refer to the table of radical numbers and grapheme codes. After some time, users generally end up remembering the numbers and codes by using them.

An excerpt from the symbol code index in *Kanji monogatari I* [漢字物語Ⅰ] (Vorobeva, 2007) and *Kanji monogatari II* [漢字物語Ⅱ] (Vorobev & Vorobeva, 2007) is shown in table 13.

Table 13: First part of the symbol code index in the textbooks *Kanji monogatari I and II*

character symbol code	**character**	**character number in *Kanji monogatari I,II***
1	一	11
1/106	百	21
1/119	来	51
1/13/46	両	238
1/132/34	夏	189

8.4 Semantic code index

Part of the semantic code index from the textbooks *Kanji monogatari I* and *II* is given in table 14. When using the semantic code index, users refer to the list of words representing the meaning labels of character constituent elements. The semantic code is obtained by listing the meaning labels of the first two constituent elements, following standard stroke order. For example, the semantic code for the character 新 is “stand/tree”, since the first two constituent elements and their corresponding meaning labels are 立 (stand), and 木 (tree). In order for the semantic code not to be too long, only the meaning labels of the first two constituent elements are used.

Table 14: First part of the semantic code index in the textbooks *Kanji monogatari I and II*

character semantic code (Russian /English)	**character**	**character number in *Kanji monogatari I,II***
Азия/сердце (Asia / heart)	悪	140
бамбук/встреча (bamboo/ meeting)	答	447
бамбук/дерево (bamboo/tree)	箱	449

8.5 Radical and stroke index

The special characteristic of this index is that users only need to remember the radicals and the basic strokes in order to use it. The first part of the radical and stroke code index is shown in Table 15.

Table 15: First part of the radical and stroke index

Radical and stroke code	**character**
1	一
1/102/17	画
1/106	百
1/132/34	夏
1/25	下
1/30/6/1/30/6/76	歌

The nine types of strokes (A, B, C, F, J, P, Q, R, W), given in Table 8 are at the same time also constituent elements appearing in the table of radicals. In the compilation of the radical and stroke based index, these strokes were treated as radicals. By analysing all characters in the *shin jōyō kanji* list, we found that more than 90% of these characters are entirely composed of elements which can be found in the traditional list of radicals. Characters that include other strokes (i.e. D, E, G, H, K, L, M, N, O, R, S, T, V, Y, Z in table 8) are relatively few, less than 10% of all characters in the *shin jōyō kanji* list. The radical and stroke code, which is mainly based on radicals, is therefore presumably easier to acquire and use than the symbol code index and the semantic code index.

Table 16 shows the frequency of use of stroke codes in the radical and stroke code index for all characters in the *shin jōyō kanji* list.

Table 16: Frequency of use of stroke codes in the radical and stroke code index for characters in the *shin jōyō kanji* list

Stroke code	**D**	**E**	**G**	**H**	**K**	**L**	**M**	**N**	**O**	**R**	**S**	**T**	**V**	**Y**	**Z**
Frequency of use	0	25	20	19	5	42	4	16	51	2	0	9	19	0	4

8.6 Comparison and evaluation of the new type of character indexes

If the new character indexes introduced in the preceding sections are learned and used, the workload necessary for search is comparable to the search in alphabetic dictionaries, resulting in more efficient character search. At the same time it can be expected that the learners gain a better insight into the structure of characters and are thus freed from rote memorisation. Table 17 gives the coefficients of selectivity (CS) of these new indexes, computed for the 1945 characters the *jōyō kanji* list. CS for each index is obtained by dividing the number of distinct characters under each code by the total number of characters in the index, and multiplying it by 100, as explained in section 6.2.

The 1945 characters in the *jōyō kanji* list were considered instead of the 2136 characters from the *shin jōyō kanji* list, in order to facilitate a comparison with existing indexes.

Table 17: Coefficients of selectivity (CS) for the new types of indexes

New type of index	**Coefficient of selectivity (%)**
Semantic code index	64.1
Alphabet code index	98.4
Symbol code index	99.4
Radical and stroke index	99.4

The results of the analysis show that the alphabet code index, the symbol code index and the radical and stroke code index all have a CS close to 100%, implying better ease of search compared with existing indexes based on character shape. In other words, in the new indexes, different characters sharing the same code are few, and many characters have a unique code. For example, in the radical and stroke code index, there are only 11 characters sharing their code with other characters (cf. table 18). CS values for the radical and stroke code index for characters in the *jōyō kanji* list and *shin jōyō kanji* list are:

jōyō kanji list: CS = (2136-11)/2136*100 = 99.5%
shin jōyō kanji list :CS = (1945-11)/1945*100 = 99.4%

Table 18: Characters from the *shin jōyō kanji* list sharing the same code in the radical and stroke code index

radical and stroke code	character 1	character 2	character 3	character 4
96	王	玉		
102	田	申	由	甲
57/2	引	弔		
9/7/28	会	伝		
30/154	員	唄		
120/102	細	紳		
83	氏	民		
1/75	末	未		
64/102	押	抽		

The CS value for the semantic code index is 64.1%, lower than the CS for the alphabetic code index, the symbol code index and the radical and stroke index. This is because the semantic code only includes the codes of the first two elements in a character, resulting in relatively many sets of characters with the same code.

In order to use character coding systems, some effort is required. For the new coding systems proposed above, it is especially important to accurately master the rules governing stroke order and the decomposition of characters into constituent elements and strokes.

The alphabet code index is probably the easiest to learn among the new indexes proposed, since users only need to memorise 24 types of basic strokes and their corresponding alphabetic codes. This index can therefore be used from the beginning stages of character instruction.

In order to use the radical and stroke index, learners need to memorise radicals, their numeric codes, and 24 basic strokes with their alphabetic codes.

In order to use the symbol index and the semantic code index, learners need to memorise radicals with their respective numerical codes, and graphemes with their alphabetic codes.

However, the symbol code index, the semantic code index, and the radical and stroke index have two merits if compared with the alphabetic code.

(1) The code is short. For example, the alphabetic code for the character 高 is SABHABGBHA, while its symbol code and its radical and stroke code (identical in both cases) is 189.

(2) In order to use the radical and stroke code, users need to learn how to structurally decompose characters and to learn about their constituent parts, which enhances comprehension and memorisation of characters.

8.7 Uses of the new index types

At the beginning stage of learning to read and write Chinese characters, the alphabetic index is probably easier to use than other indexes, since learners do not yet know the constituent elements which make up characters. However, as learning progresses, characters to be learned become more complex, being made up of more strokes, and the alphabetic code becomes longer. At the same time, learners progressively learn to recognise different constituent parts within complex characters, and are able to use other indexes, choosing the one that best suits their learning style, be it the symbol code index, the semantic code index, or the radical and stroke code index. At this point, they usually start using mainly one of these indexes.

The radical and stroke index is probably easier to learn and use than the symbol code and the semantic code index, but the latter two are probably more useful for deepening learners' understanding of character structural composition.

We used the symbol code index and the semantic code index in two textbooks we developed, *Kanji monogatari I* [漢字物語Ⅰ] (Vorobeva, 2007) and *Kanji monogatari II* [漢字物語Ⅱ] (Vorobev & Vorobeva, 2007). In the following paragraphs we explain how the use of these indexes was introduced to our learners.

In order to look up an unknown character in the textbook, learners firstly write the first two constituent elements of the character and convert them into their respective codes. At first, when they are yet not accustomed to characters, they actually write down the first elements on paper, later they learn to just imagine writing the character and convert the first two elements into their codes. They then look through either the symbol code index or the semantic code index, where codes are arranged in alpha-numeric order, and look up the character as they would in a dictionary of alphabetically arranged words, until they find the number of the character in the index. Using this number, they can then find the character inside the textbook and read the textbook explanation about the character they were searching for. Searching in these two types of indexes is equivalent to searching through an alphabetically ordered dictionary. The procedure (algorithm) used when searching the aphanumeric symbol code index is described by Vorobeva (2007, p. 25). Through use, learners get progressively accustomed to this way of looking up characters, and after some practice are able to find a character in a few seconds.

9. Selectivity of character indexes and learning burden

In the present paper, we defined the Coefficient of Selectivity as an index of efficiency for character indexes, and compared existing indexes with our newly developed index types on the basis of this coefficient. However, there is one more factor to be considered when considering the efficiency of indexes, i.e. the learning burden required from users, the special knowledge they need to acquire in order to be

able to actually use each of these indexes, such as character radicals, types of strokes and their counting, readings and meanings associated with each character, their coding etc. When discussing the efficiency of indexes, such learning burden must also be considered, as discussed in Vorobeva (2011, p. 24).

The index by total number of strokes is probably the easiest to master among the indexes presented in the previous sections. However, given its CS of 1.2%, it is clearly not an efficient index. On the other hand, the Key Words and Primitive Meanings Index (Heisig, 2001), with a CS of almost 100%, allows for very efficient search, but in order to use it, learners need to memorise approximately 2000 semantic keywords. We can therefore conclude that in future research it will be necessary to analyse the learning burden of character indexes, and further define a comprehensive index of efficiency which would reflect both selectivity and learning burden of character indexes.

10. Conclusion and further work

In the present paper we discussed the difficulties faced by learners of Japanese not familiar with Chinese characters when they use character dictionaries, presented 15 types of existing character indexes and described their characteristics. We pointed out that for users who need to look up characters in a dictionary, in addition to the generally used radical index, stroke number index and readings index, many other different types of character indexes have been developed and used, including the five step arrangement kanji table (Rosenberg, 1916), the four corner method (Wang, 1925), the phonetic key index (Shiraishi, 1971/1978), the index of katakana shapes, the index of initial stroke patterns and the index of meaning symbols (Kanō, 1998), the stroke order index (Wakao & Hattori, 1989), the index of character shapes (Sakano, Ikeda, Shinagawa, Tajima, & Tokashiki, 2009), the key words and primitive meanings index (Heisig, 1977/2001), the index by radicals (Hadamitzky & Spahn, 1981), the system of kanji indexing by patterns - SKIP (Halpern, 1988), the kanji fast finder (Matthews, 2004) and others.

In order to compare the effectiveness of different character indexes, we applied the concept of selectivity, a concept used to express the efficiency of computer data processing, to character indexes, and defined the concept of coefficient of selectivity (CS) as an index of efficiency of character indexes. We then computed CS for existing character indexes and compared their efficiency. We found that indexes based on character form or structure had a low CS in the range of 1.2% to 25.4%, probably because most indexes based on character form use only one structural element or characteristic of each character for indexing, such as the radical index using only the radical part of a character, the total stroke number index using only the number of strokes, or the initial stroke pattern index using only the form of the first stroke for indexing characters.

In order to improve the efficiency of character dictionary indexes, we considered it necessary to develop a new type of index with a hight selectivity coefficient that would be appropriate for the habits of learners not familiar with Chinese character writing. We therefore developed an alphabetic and a symbol code index, based on a code which accurately represents the form of Chinese characters, a semantic code index and a radical and stroke code index. In this paper, we described the use of these indexes, and comparatively evaluated their efficiency. We found that the alphabetic code index, the symbol code index and the radical and stroke index have a coefficient of selectivity nearing 100%, while the semantic code index only has a coefficient of selectivity of 64.1%. The reason for such a low CS is probably that it only takes into account the first two structural elements of each character, which leads to a relatively high number of characters with the same code.

We expect that learners using the above new types of indexes should be able to look up characters in dictionaries more efficiently, deepen their understanding of character structure, and be emancipated from rote learning. We implemented these new types of indexes in two introductory textbooks including 518 characters, *Kanji monogatari I* (Vorobeva, 2007) and *Kanji monogatari II* (Vorobev & Vorobeva, 2007). We plan to include these new types of indexes (an alphabetic code index, a symbol code index, a semantic code index, and a radical and stroke index) in a new textbook with 1006 characters for the initial and intermediate level that is under development, and to empirically investigate the efficiency of these indexes by surveying how learners use them in practice.

Moreover, the usefulness of character indexes should be investigated by measuring not only their coefficient of selectivity, but also the learning burden associated with them, as suggested in Vorobeva (2011, p. 24). We are therefore planning to further investigate both factors and define a comprehensive measure of efficiency for character dictionaries.

References

Akadia. (2008). *How to measure index selectivity*. Retrieved July 27, 2012, from http://www.akadia.com/services/ora_index_selectivity.html

Banno, E. [坂野永理], Ikeda, Y. [池田庸子], Shinagawa, C. [品川恭子], Tajima, K. [田嶋香織], & Tokashiki, K. [渡嘉敷恭子]. (2009). *Kanji look and learn : 512 kanji with illustrations and mnemonic hints : imeeji de oboeru 'genki' na kanji 512 : genki plus* [Kanji look and learn : イメージで覚える「げんき」な漢字 512 : genki plus]. Tokyo: The Japan Times.

Fazzioli, E. (1987). *Chinese calligraphy: from pictograph to ideogram: the history of 214 essential Chinese/Japanese characters*. New York: Abbeville Press.

Hadamitzky, W., & Spahn, M. (1981/1997). *Kanji & kana: A handbook of the Japanese writing system*. Boston: Tuttle.

Halpern, J. (1988). *New Japanese-English character dictionary*. Tokyo: Kenkyusha.

Heisig, J. (1977). *Remembering the kanji. Vol. 1*. Tokyo: Japan Publications Trading.

Henshall, K. (1988). *A guide to remembering Japanese characters*. Boston: Tuttle.

Kanō, Y. [加納喜光] (1998). *Jōyō kanji mirakuru masutā jiten* [常用漢字ミラクルマスター辞典]. Tokyo: Shogakukan [小学館].

Kod_Rozenberga [Код_Розенберга]. (2012, July 27). Retrieved July 27, 2012 from http://www.enci.ru/Код_Розенберга

Matthews, L. (2004). *Kanji fast finder - Kanji hayabiki jiten* [漢字早引き辞典]. Boston: Tuttle.

Mojikyō Kenkyūkai [文字鏡研究会]. (2002). *Pasokon yūyū kanji jutsu: Konjaku mojikyō tettei katsuyō* [パソコン悠悠漢字術:今昔文字文字鏡徹底活用]. (3rd ed.). Tokyo: Kinokuniya shoten [紀伊國屋書店].

Morohashi, T. [諸橋轍次] (1960/1984). Dai kanwa jiten [大漢和辞典]. Tokyo: Taishūkan [大修館書店].

Panasyuk, V.A. [Панасюк В.А.], & Suhanov V.F. [Суханов В.Ф.] (1983). *Bol´šoj kitajsko-russkij slovar´ - hua e da ci dian* [Большой китайско-русский словарь 華俄大辞典]. Moscow: Nauka [Наука].

Pešurov, D. A. [Пещуров Д.А.] (1891). *Kitajsko-russkij slovar´ - Po grafičeskoj sisteme* [Китайско-русский словарь. По графической системе]. Saint Petersburg: Tipografija Imperatorskoj Akademii Nauk [СПб., Типография Императорской Академии Наук].

Rosenberg, O. [ロゼンベルグ・オ] (1916). *Godan hairetsu kanjiten* [五段排列漢字典] - *Arrangement of the Chinese characters according to an alphabetical system with Japanese dictionary of 8000 characters and list of 22000 characters.* Tokyo: Kōbunsha [興文社].

Shiraishi, M. [白石光邦]. (1971/1978). *Yōsokeiteki kanji gakushū shidōhō* [要素形的漢字学習指導法]. Tokyo: Ōfūsha [桜楓社].

Stalph, J. (1989). *Grundlagen einer Grammatik der sinojapanischen Schrift*. Wiesbaden: Otto Harrassowitz.

Unicode. (2012). *Unicode 6.1.0*. Retrieved March 15, 2012 from http://www.unicode.org/versions/Unicode6.1.0/

Vasil´ev, V. P. [Васильев В. П.] (1867). *Grafičeskaâ sistema kitajskih ieroglifov. Opyt pervogo kitajsko-russkogo slovarâ* [Графическая система китайских иероглифов. Опыт первого китайско-русского словаря] [scholarly edition published online in 2010 at http://www.ci.spbu.ru/slovar/index.html].

Vorobeva, G. [ヴォロビヨワ・ガリーナ] (2007). *Kanji monogatari I* [漢字物語Ⅰ]. (2nd ed.). Bishkek: Lakprint.

Vorobev, V. [ヴォロビヨフ・ヴィクトル], & Vorobeva, G. [ヴォロビヨワ・ガリーナ] (2007). *Kanji monogatari II* [漢字物語Ⅱ]. Bishkek: Lakprint.

Vorobeva, G. [ヴォロビヨワ・ガリーナ] (2009). Sentakusei ga takai kanji sakuin no kaihatsu [選択性が高い漢字索引の開発]. *Nihongo kyōiku hōhō kenkyū kaishi* [日本語教育方法研究会誌], *16* (1), 72-73.

Vorobeva, G. [ヴォロビヨワ・ガリーナ] (2011). Kōzō bunseki to kōdoka ni motozuku kanji jitai jōhō shori shisutemu no kaihatsu [構造分析とコード化に基づく漢字字体情報処理システムの開発]. *Nihongo kyōiku* [日本語教育], (149), 16-30.

Wakao, S.[若尾俊平], & Hattori D. [服部大超]. (1989). *Kuzushi kaidoku jiten* [くずし解読字典]. Tokyo: Kashiwa Shobō [栢書房].

Wáng, Yún Wǔ [王雲五]. (1925). *Hào mǎ jiǎn zì fǎ* [號碼檢字法]. Shanghai: Shāngwù yìnshūguǎn [商務印書館].

Wáng, Yún Wǔ [王雲五] (1926). *Sì jiǎo hào mǎ jiǎn zì fǎ* [四角號碼檢字法]. Shanghai: Shāngwù yìnshūguǎn [商務印書館].

Wáng, Yún Wǔ [王雲五]. (1934). *Sì jiǎo hào mǎ jiǎnzìfǎ·fù jiǎn zì fǎ* [四角號碼檢字法·附檢字表]. Shanghai: Shāngwù yìnshūguǎn [商務印書館].

Zadoenko T. P. [Задоенко Т. П.], & Khuan, S. [Шуин Хуан] (1993). *Osnovy kitajskogo jazyka: vvodnyj kurs - Chi ch'u Han yü* [Основы китайского языка: вводный курс - 基礎漢語]. Moscow: Nauka [Наука].

RESEARCH ARTICLES
(PROJECT REPORTS)

DEVELOPMENT OF A LEARNERS' DICTIONARY OF POLYSEMOUS JAPANESE WORDS AND SOME PROPOSALS FOR LEARNERS' LEXICOGRAPHY

Shingo IMAI[*]
University of Tsukuba
imai.shingo@gmail.com

Abstract

The dictionary series *Nihongo tagigo gakushū jiten* ("*A Learner's Dictionary of Multi-sense Japanese Words*") proposes a new approach to learners' dictionaries, based on cognitive linguistics theory and on a corpus-based approach. Each entry is presented as a semantic network which follows the patterns of semantic extensions from a word's core meaning to derived meanings. A corpus, Sketch Engine, was consulted in order to select natural and frequently used examples, which were then edited to make them understandable to learners. Illustrations are also provided in an attempt to visualise the common meaning (schema) shared by the various meanings of the word.

Keywords

Learner's dictionary; monolingual Japanese dictionaries; corpus-based dictionary; semantic network; cognitive linguistics

Izvleček

Slovarska serija *Nihongo tagigo gakushū jiten* ("*Učni slovar večpomenskih japonskih besed*") prinaša nov pristop k učnim slovarjem, ki je osnovan na teoriji kognitivnega jezikoslovja in na korpusno osnovanem pristopu. Vsako geslo je predstavljeno kot semantična mreža, ki sledi vzorcem pomenskih širitev od osrednjega k drugim pomenom, ki iz tega izvirajo. Naravni in pogosti primeri rabe so bili izbrani iz korpusa s pomočjo orodja Sketch Engine in nato prirejeni, da so razumljivi uporabnikom slovarja, ki se japonščine šele učijo. Ilustracije predstavljajo skupni pomen (shemo), ki je prisoten v različnih podpomenih besede.

Ključne beside

učni slovar; enojezični slovarji japonščine; korpusno osnovan slovar; semantična mreža; kognitivno jezikoslovje

---

[*] Translated by Andrej Bekeš

1. The present state of Japanese learner's dictionaries

In recent years many teaching materials for learning Japanese have been published, easing the long-lasting shortage of such materials. In addition, use of teaching material available on computers and internet has also been increasing. Various types of grammar manuals, sentence pattern dictionaries, thesauri etc, both for teachers and for students are being published. Yet, among these, learners' dictionaries are conspicuously absent. On the other hand, a glance at the publishing situation for learning English suffices to see that both in and outside of Japan such dictionaries are being published incessantly. The quantitative, and as a consequence the qualitative gap between learner's dictionaries for learning Japanese and English is great.

Among dictionaries for learning English produced either in the United Kingdom or in the United States, there are Merriam-Webster's Learner's Dictionary, Cambridge Advanced Learner's Dictionary, Collins Cobuild Learner's Dictionary, and Oxford Advanced Learner's Dictionary. In these dictionaries, the vocabulary used for defining the meaning of lexical items is restricted to about 2000 words, which are in themselves sufficient to describe various meanings. Thus, if one learns the vocabulary used for meaning definitions, one is able to use these dictionaries. Compared to English dictionaries, similar Japanese learner's dictionaries are few. To the author's limited knowledge, there are no Japanese dictionaries comparable to the aforementioned English dictionaries aiming at intermediate to advanced level learners. Needless to say, learners of Japanese do use dictionaries, but the dictionaries they use are either monolingual Japanese dictionaries for Japanese native speakers, or domestically produced bilingual dictionaries. The aforementioned English learner's dictionaries are in principle monolingual dictionaries. On the other hand, the yet few dictionaries targeted at learners of Japanese as a foreign language are prevalently bilingual dictionaries with headword translations or headword definitions and translations.

Monolingual Japanese dictionaries for native speakers, so-called *kokugo jiten* [国語辞典] can only be used by intermediate or advanced learners. Learners usually start to use such monolingual dictionaries at the intermediate level. It goes without saying that such dictionaries have no consideration for the needs of non-native speakers and lack the information needed by learners. No monolingual Japanese dictionary for native speakers limits the vocabulary used in definition descriptions to any prescribed set. Often, definitions and explanations are more difficult to understand than the headwords they are meant to describe. Monolingual *kokugo* dictionaries often include also encyclopedic entries, which call for accurate encyclopedic information in their description, resulting in entries which are often difficult to understand. Learners of Japanese are more interested in the meaning and usage of individual words rather than in encyclopedic information, but such information on usage tends to be lacking in *kokugo* dictionaries. For example, information about the use of cases (postpositions) is not provided. Even though there are examples of use, they lack collocational information. *"Shinmeikai kokugo jiten"* is one of the few *kokugo* dictionaries which provide such kind of information, but since it is meant to be used by native speakers,

important information needed by learners of Japanese is still insufficient. Advanced learners of Japanese do not often make grammatical mistakes any more but there is often a sense of incongruity lingering around their speech and writing. Many times the reason for this is the unnaturalness of collocations.

On the other hand, bilingual dictionaries or *taiyaku jiten* [対訳辞典], unlike monolingual *kokugo* dictionaries, are accessible also to beginners. Unfortunately, most bilingual dictionaries are basically made with Japanese native speakers in mind. Therefore, all the descriptions are made to suite the native speaker and lack the information relevant to learners. There are also bilingual dictionaries published outside Japan. Such dictionaries are made to suit the speakers of a particular language, but even such dictionaries are often based on bilingual dictionaries intended for Japanese native speakers. Therefore, for the same reason as bilingual dictionaries published in Japan, they lack the information relevant to learners of Japanese.

2. The solutions suggested by the present dictionary

The dictionary series *Nihongo tagigo gakushū jiten* [日本語多義語学習辞典], to which the present author contributed the volume on adjectives and adverbs, proposes a solution based on cognitive linguistics theory that has already been employed in some English-Japanese dictionaries, namely, a synchronic description of the meaning of basic words, which are important for learners in the form of a semantic network which follows the patterns of semantic expansion from core meanings (basic meanings) to derived meanings.

In the present dictionary series, the authors did not limit the vocabulary used in definitions to any particular set. The share of verbal information in the dictionary being smaller than in other comparable dictionaries due to the limited number of headwords (only basic words) and to the ample use of illustrations, the authors consider that limiting the defining vocabulary would not have any major effect. Instead of limiting the defining vocabulary, translations of entries into 3 languages were added as a complement to the explanations in Japanese.

By introducing the cognitive linguistics point of view, the meanings of entry items were presented as derivation networks stemming from the core meaning. Therefore the order of presenting different lexical meanings is not diachronous. As far as it is possible, it presents the cognitively central (and not necessarily the most frequent) word meaning first, followed by the various meanings derived from it, all in order to facilitate the understanding of these derivational paths. Thus senseless enumeration of unconnected entry meanings can be avoided and the organic connection between different meanings can be felt more easily, contributing to a better understanding and retention of the entry's meaning.

In addition, by consistently using examples of actual use, adapted to the level of second language learners, the dictionary is conceived so as to contribute to the

learners' deeper understanding of each entry word, and, by supporting sentence patterns offered in textbooks with such examples, to foster a more natural use of Japanese. Examples are based on corpora, and a lot of care was given to selecting and editing natural, frequent examples. Nonetheless, the approach is **not corpus-driven** but **corpus-based**. While corpora were being used for consultation, examples were edited to be readable and learner-friendly.

3. Intended users of the present dictionary

While the vocabulary of the present dictionary centers on basic words which appear in textbooks for beginners, these words are actually polysemic and it is only their basic meaning that is usually introduced to beginners. Other, derived meanings not treated at the beginners' level are either introduced at the intermediate or some higher level, or, as is often the case, not treated explicitly at all in textbooks. The characteristic of this dictionary are not the descriptions of core meanings as such but how such core meanings are expanded into derived meanings. Therefore, the intended users of the present dictionary are not beginners but learners at intermediate or higher levels as well as their teachers, both native and non-native speakers. For teachers who are native speakers, derived meanings are self-evident, and they admittedly can grasp their relation to the core meaning intuitively. Yet, when teaching, it is necessary to present and explain such relations in a systematic way, and this is where the present dictionary can be of help. Further, for teachers who are not native speakers, this dictionary includes ample information such as spread of polysemy, explanations of these meanings, examples and other relevant information which can be of help to them when systematizing their own knowledge.

4. Structure and content

4.8 Selection of entries

The dictionary was complied by selecting those words in modern Japanese that are highly polysemic and therefore difficult for learners to master, and by showing word meaning networks centered on the core meanings of these basic words, in order to facilitate correct understanding and use of each word by both learners and teachers of Japanese as a second language.

The present dictionary series consists of three parts, each a separate volume, i.e., **Nouns** (121 headwords), **Adjectives and Adverbs (84 headwords)**, and **Verbs (104 headwords)**. Highly polysemous modern Japanese words were chosen as headwords. By "basic words" here the authors mean essentially words belonging to level 4 or level 3 of the former Japanese Language Proficiency Test and among them particularly those with pronounced polysemy. Further, it has to be said that words learned at the

beginning level are most often polysemous. In such cases, it is usually the core meaning which is introduced at the beginners' level, while the derived meanings are not taught at this level. Indeed, if derived meanings are used at the intermediate or higher level at all, such words are usually treated as known words, and until now derived meanings were not taught in a systematic way. Usually what happened were occasional haphazard explanations of their meaning in some reading materials. Because of this, in spite of such words being taught at the beginners' level, it was only natural that learners experienced difficulties in learning their meanings as a complete system. Even worse, it is quite possible in such circumstances that some derived meanings that are quite distant from core meanings are not perceived as such at all.

As can be seen from the above, because of the detailed explanation of polysemous basic words, the increase in the number of pages for each word resulted in a relatively limited number of entries as compared to other dictionaries. Yet if we consider that the words appearing at intermediate and higher levels tend to be less polysemous, the number of entries is not too small for the goal of the present dictionary, which is to support learning of polysemous words, , since it covers a range of vocabulary which is conductive to reaching the expected learning targets to a considerable degree. Explanation notes are quoted below to exemplify the contents and construction of the dictionary.

4.9 Structure of the entries

As can be seen in Figure 1, each entry begins with the headword (1), its difficulty level (2), the kanji (Chinese characters) used for the headword, where the character in boldface is the one that expresses its most basic meaning, while characters capped with a dot are characters which are not in the official list prescribed for regular use (Jōyō kanji). This is followed by readings (3) of the boldface characters, listed in katakana for on-yomi (Chinese loanwords) and in hiragana for kun-yomi (domestic words). Examples are given in parentheses. When the entry is a kun-yomi word, it is listed here again, along with an example. This information is followed by a network diagram (4), where arrows show how each meaning derives from the core meaning. The actual derivations are more complicated than indicated, but have been simplified where possible in keeping with the dictionary's design as a learners' dictionary. The core meaning, marked with a 0, is the meaning in modern usage that is considered the central meaning from which other meanings derive. Primary derivatives, marked with numeral (1, 2, 3 etc.) are major meanings that derive from the core meaning 0. Secondary derivatives are marked with letters of the alphabet, as in 0a, 1a, 1b, 1c, etc. They are extended meanings that stem from the core meaning or the primary derivatives. Since they represent comparatively slight differences in meaning that do not qualify as stand-alone derivatives, they are subordinated to 0, 1, 2, 3, etc.

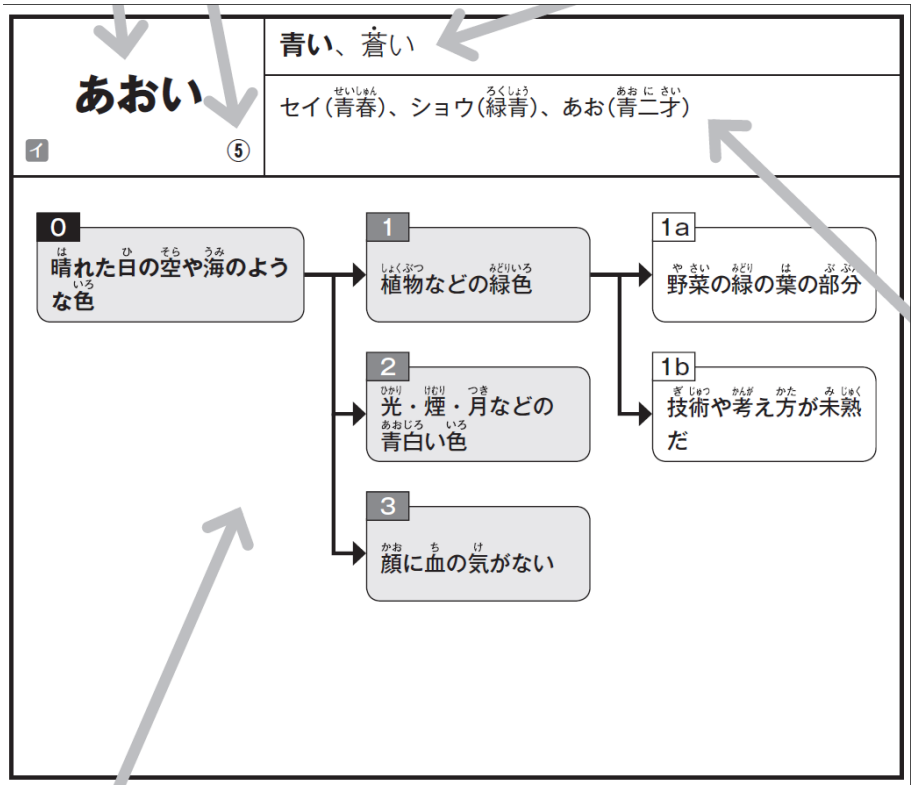


Figure 1: Structure of the headword entry

Each meaning presented in the boxes in the network diagram is then explained, translated, exemplified and illustrated as can be seen in Figure 2. The explanation (1) is given in Japanese with furigana on all Chinese characters and followed by a translation into English, Korean and Chinese. This is followed by a line (2) which describes how the meaning derives from the core meaning or the primary derivative. Example sentences (3) are provided for each meaning presented, followed by compounds, idioms, and other related expressions (4). Finally, an illustration (5) is provided to clarify the meaning. For most entries, the first example sentence is used as the basis for the illustration. These illustrations can also be used as a quick guide to the common meaning (schema) shared by the various meanings of the entry.

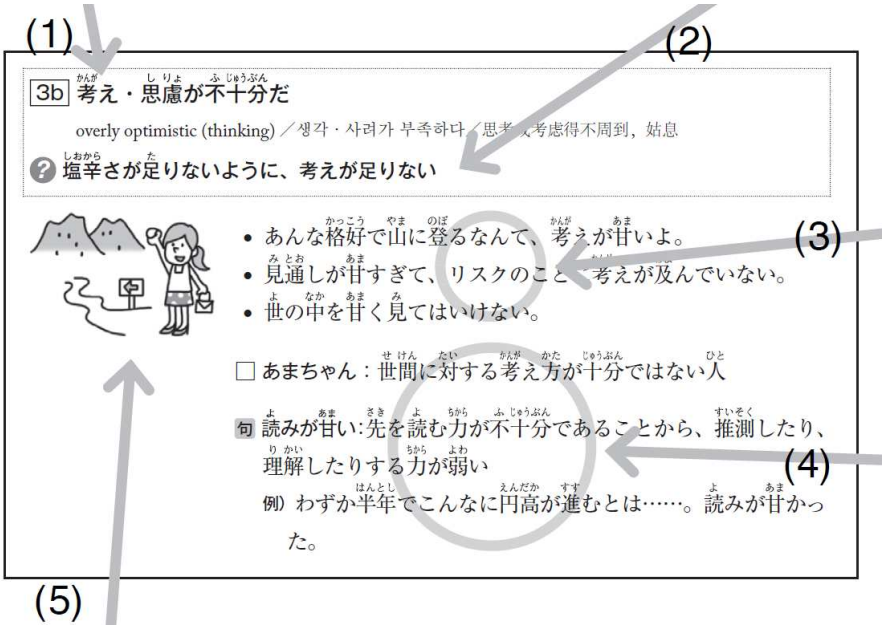


Figure 2: Meaning description with translations, illustration, examples and related expressions

When illustrations contain multiple elements, such as contrasts, the element that depicts the target meaning is generally placed on the right and/or rendered in darker shading, as can be seen in Figure 3, where the illustration conveys the target meaning "large/big (size, area)".

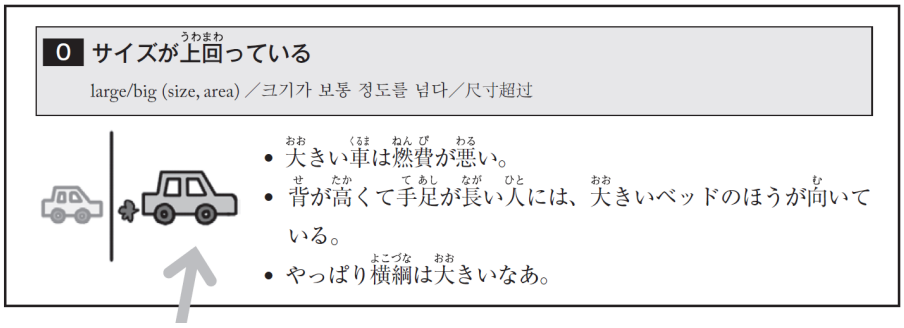


Figure 3: Illustration with multiple elements

As can be seen in Figure 4, other related expressions, such as extra compound words, idioms, transmuted words, etc. that were not covered in the meaning descriptions are enclosed in boxes marked *kanrengoku* [関連語句].

Figure 4: Box with related expressions

Distinctions in usage and other pointers on usage of the entry are given at the bottom of each entry, marked as *yōhō nōto* [用法ノート], as in Figure 5.

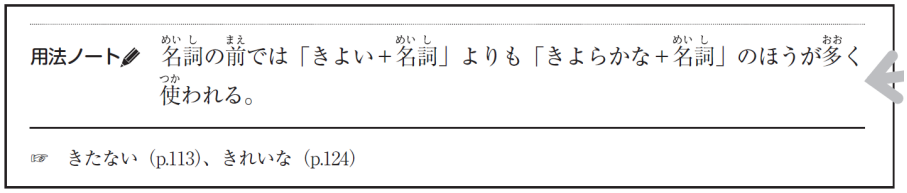


Figure 5: Usage notes and cross-references

Some entries are followed by culture notes, which provide background information on how the entry is used in the context of Japanese culture, as can be seen in Figure 6.

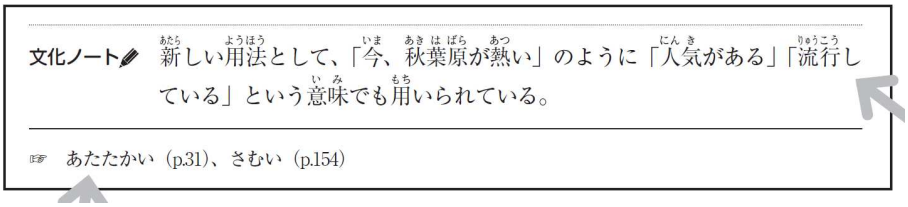


Figure 6: Culture notes and cross-references

At the bottom of each entry, there are cross-references to other entries that can be of reference, such as synonyms, antonyms, etc., which are listed with their page number, as can be seen in Figures 5 and 6.

In cases where an entry has a list of multiple Chinese characters (e.g., *atsui* - あつい: 熱い 暑い), users can refer to the example sentences to gain an understanding of the distinctions in usage between those characters. Particularly important distinctions are explained in the supplementary information sections marked with asterisks.

Intuitive symbols are used to indicate various parts of the entries: □ for compound words, 句 for idioms and proverbs, * for supplementary information, ⇔ for antonyms, ☞ for cross-references, / for words that can be used in the same pattern (e.g. 目頭を熱くする／目頭が熱くなる), ≒ for synonyms, [ ] for omissible words or particles (e.g. 幸[が]薄い), while the parts of speech used in compound words and other related expressions are marked by symbols containing the name of the part of speech in Japanese, as can be seen in Figure 7.

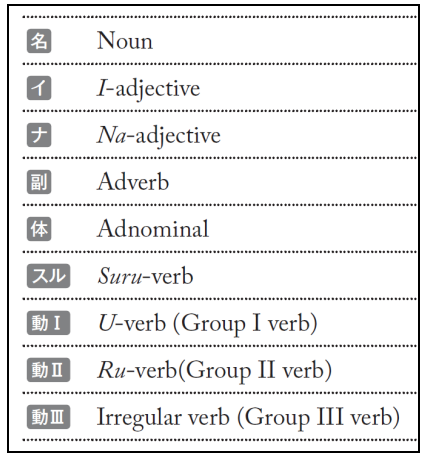


Figure 7: Symbols for parts of speech

4.10 Additional comments on the explanation notes

In this section, some additional explanations will be provided concerning items not appearing in the final version of the dictionary. Network diagrams show extensions of meaning, i.e., derivatives, based on **metaphor** (semantic extension based on similarity, e.g., *ki ga omoi* "depressed, heavy hearted", literally "with heavy spirit", when one's feelings are perceived heavy as a heavy object); **metonymy** (semantic extension based on contiguity, association, or proximity, e.g. *atatakai iro* "a warm color", where the property (warmth) of fire is expressed by its colour; **synecdoche** (extension of meaning based on subsumption or part-whole relation, e.g., *hana* "flower" standing for *sakura* "cherry blossom", where *sakura* is a kind of flower). While this analysis has been carried out during the process of writing, it is not mentioned in the completed version of the dictionary. The reason is that an explicit analysis was necessary during the preparation stage of the dictionary, while this type of information was deemed not to be necessarily easy to understand or relevant for the user. In meaning descriptions, under "relationship with superior meaning", the wording *-yoo ni* "like ..." hints at an extension based on metaphor. Also, it has to be said that derivatives should be arranged in fact "radially" and not "linearly", though for the reason of more expedient layout, they are represented "linearly". Further, the hierarchy was limited to only three levels, also because of layout expedience, even though there should be cases with deeper

reaching hierarchies. The representation is thus simplified because of layout expedience, but at the same time also more understandable to the learner than theoretically perfectly accurate but visually complex examples. The way derivatives are divided and ordered is a particular characteristic of this dictionary, which sets it apart from other dictionaries. The division and ordering are conceived so as to help the learner intuitively grasp the connections among extensions of meaning, this being the foremost characteristics of this dictionary as a learners' dictionary.

The authors tried to describe the meanings and relationships with superior meanings in as simple Japanese words and as concisely as possible, and to further enhance the understanding by adding translations in three languages.

Example sentences are ordered so that the most prototypical examples, those with meanings and uses that are most easy to understand, come first. For the selection of examples, corpora were consulted in order to make them as natural as possible and to show relevant collocations wherever possible.

A word about illustrations. Authors strived to maintain shared characteristics between illustrations, and also for shared schemata to be discernible by comparing related illustrations, thus enabling the learner to grasp the meaning extension relationships in an intuitive way. This effort was most successful with verbs. On the other hand, adjectives and adverbs were difficult to render this way. The reason for this is that the meaning of adjectives and adverbs cannot be understood without including also the meaning of the modified part, i.e., their meaning depends very much on the modified part. Because of this, illustrations of meaning must in such cases also include the modified part, and furthermore, the modified part (which could normally be expressed by a noun) tends to become more central than the adjective/adverb itself. Consequently, illustrations change along with their modified part, thus loose the shared aspects of the core meaning and meanings of derivatives. Because of this, in the case of adjectives and adverbs, illustrations should be understood as mere "illustrations" to help with understanding of the word meaning, rather than what is in cognitive linguistics called an "image schema".

5. Conclusion – suggestions for future learners' dictionaries

To conclude, based on the compilation of the present dictionary, we wish to express some thoughts regarding the shape of future learners' dictionaries, hoping that they will be of help for the compilation of such dictionaries in the future.

Firstly, the necessity of corpora. There are two types of corpora useful for compiling learners' dictionaries. One type are large scale corpora representing the usage of Japanese. There are several smaller corpora of Japanese, but only two of the size that can be directly useful for compiling dictionaries. The first one is BCCWJ

(Balanced Corpus of Contemporary Written Japanese)[1], compiled by the National Institute for Japanese Language and Linguistics (Maekawa, 2012), and the second one is the corpus JpWaC included in the Sketch Engine[2] (Srdanović-Erjavec, Erjavec & Kilgarriff, 2008). BCCWJ contains one hundred million words, with attention paid to balanced sampling. As such it is one of the largest corpora in Japan, yet it is said that for compiling English dictionaries, a one billion word corpus is necessary (Akasegawa p.c.). Compared to such size, BCCWJ still suffers from size limitation, and while its sampled data are balanced, frequencies of extracted data and collocations are still inadequate. On the other hand, the Japanese language corpus within Sketch Engine comprises 5 hundred million words, and is thus larger than BCCWJ. The problem is that it consists of data automatically collected from the web, resulting in sampled data being rather slanted. Further, there are cases of special collocations stemming all from the same URL, which means that they were produced by the same person. Because of this, some caution in using the corpus is necessary. When the present dictionary was being compiled, BCCWJ was not yet available, therefore, the Sketch Engine JpWaC had to be consulted. While Sketch Engine proved to be useful for compiling example sentences, it was inadequate as far as collocations are concerned. Ideally, a corpus for the compilation of dictionaries should be balanced, and at the same time large scale. At present the authors are compiling a one billion word corpus, by collecting data automatically from the web, while striving to make it balanced. By making the corpus freely accessible to the general public, the authors hope to contribute to the compilation of future dictionaries.

Moreover, rather than compiling dictionaries relying entirely on corpora (Corpus Driven approach), an approach based on consulting corpora (Corpus Based) seems to be preferable. For example, using examples directly from the corpora is generally inadequate because it is difficult to understand such examples without access to the social and cultural schema and the immediate textual context in which they were produced. Also, since such raw examples often contain "noise", they may not always be of help to the learner. In order for the examples not to be arbitrary or unnatural, it makes sense on the other hand to consult corpora as much as possible. Yet this does not mean that with a good corpus the compilers' work is already done. On the contrary, the task of how to select good examples, how to process them, how to extract and organize relevant information from such examples, are all an important job to be done by compilers who are intimately knowledgeable about that particular language. It is difficult to expect a dictionary to be good without such efforts by its compilers.

Another type of corpora are learner corpora. Such corpora are also recently being compiled in various places, yet at present there is no such corpus with sufficient information for the compilation of dictionaries. Learners' corpora, containing much inappropriate and mistaken usage, are difficult to tag automatically. Because they

[1] http://nlb.ninjal.ac.jp/

[2] http://ww.sketchengine.co.uk/

require manual tagging, the automatic compilation of such corpora is extremely difficult to realize. Our wish is that in the not so far future, with advances in automatic processing technology, the compilation of large scale learner corpora will also be easier to achieve.

The second important point regarding the future compilation of dictionaries is to produce them in digital form. As has already been mentioned in this paper, paper media impose a limitation on the available number of pages, therefore the realization of an ideally conceived dictionary is not always possible. Further, by increasing the number of entries, and by making the description more detailed, the size of the dictionary increases, making it more costly. The thicker the dictionary is, the more inconvenient it gets for the user. All these problems can be solved in one breath by making the dictionary in digital form. Limitations on the quantity of data, problems of portability etc. all disappear. With ample search functions provided, searching in such dictionary is considerably faster and more efficient than in paper dictionaries. Also, sound and images, movie clips etc., items that are impossible to include in a paper dictionary, can be easily added. Dictionaries from now on should be digital from the beginning, not like present digitalized dictionaries, based on paper editions. Indeed, in language classrooms paper dictionaries are no longer to be seen, learners are using dedicated electronic dictionaries, mobile phones and smart phones instead. Yet, at least as far as Japanese is concerned, many of such dictionaries are often nothing more than simplified digitalized versions of paper dictionaries in which much of the information that is traditionally included in dictionaries has been omitted. Further, as has been already mentioned before, such digitalized dictionaries are not primarily learners' dictionaries but monolingual *kokugo* Japanese dictionaries, English-Japanese bilingual dictionaries and such. Thus, while at present the access to digitalized dictionaries has become much more convenient, it is a matter of high concern that the quality of dictionaries used by learners is getting worse. What we hope for is a dictionary specifically for learners which would surpass the present paper dictionaries.

References

Imai, S. [今井 新悟] (2010). Nihongo gakushū jisho kaihatsu no kadai to yōken ni tsuite [日本語学習辞書開発の課題と要件について] ("On the tasks and conditions for developing Japanese learners' dictionaries"), *Yamaguchi kokugo kokubun* [山口国語国文], *33*, 86-96.

Imai, S. [今井 新悟] (2010). Risō no nihongo gakushūsha jisho o mezashite [理想の日本語学習者辞書を目指して] ("Towards the ideal dictionary for learners of Japanese"), *Nihongo gakushū jisho no kaihatsu to nihongo kenkyū yokōshū* [日本語学習辞書の開発と日本語研究予稿集] ("*Japanese Linguistics and the Development of Japanese Learner's Dictionaries - Proceedings*"), 1-10. Tsukuba: Tsukuba University [筑波大学].

Maekawa, K. [前川 喜久雄] (2012). Gendai kakikotoba kinkou koopasu (BCCWJ) no kouchiku to KOTONOHA keikaku no ayumi [現代書き言葉均衡コーパス(BCCWJ)の構

築とコトノハ計画の歩み] - The construction of "the Balanced Corpus of Contemporary Written Japanese (BCCWJ)" and the progress of the KOTONOHA plan. *Nihongengogakkai dai 144 kai taikai yokoushuu* [日本言語学会第 144 回大会予稿集 "Proceedings of the 144th meeting of the Linguistic Society of Japan"]), 352-357.

Moriyama, S. [森山 新], Arakawa, Y. [荒川 洋平], & Imai, S. [今井 新悟] (2009). Ninchi gengogakuteki shiten kara no nihongo gakushū jiten o kangaeru [認知言語学的視点からの日本語学習辞典を考える] ("Reflections on learners' dictionaries of Japanese from a cognitive linguistic perspective"), Paper presented at JSAA-ICJLE (Japanese Studies Association of Australia - International Conference on Japanese Language Education) Sydney, Australia: University of New South Wales.

Pardeshi, P., & National Institute for Japanese Language and Linguistics [国立国語研究所] (2006-2009). *Nihongo gakushūshayō kihondōshi yōhō handobukku no sakusei* [日本語学習者用基本動詞用法ハンドブックの作成] ("*Construction of a Handbook of Basic Verbs for Learners of Japanese*"). Retrieved from http://www.ninjal.ac.jp/research/project/b/youhoujiten/

Srdanović-Erjavec, I., Erjavec, T., & Kilgarriff, A. (2008). A web corpus and word sketches for Japanese. *Journal of Natural Language Processing - Shizen gengo shori* [自然言語処理], *15* (2), 137-159.

Dictionaries

Arakawa, Y. [荒川 洋平] (2011). *Nihongo tagigo gakushū jiten Meishi hen* [日本語多義語学習辞典 名詞編] ("A Learner's Dictionary of Multi-sense Japanese Words: Nouns"), Tokyo: ALC.

Cambridge University Press (2008). *Cambridge Advanced Learner's Dictionary. Third Edition*. Cambridge: Cambridge University Press.

Hornby, A.S., Cowie, A.P., & Gimson, A.C. (1973). *Oxford Advanced Learner's Dictionary of Current English. Third Edition.* Oxford: Oxford University Press.

Imai, S. [今井 新悟] (2011). *Nihongo tagigo gakushū jiten Keiyōshi - fukushi hen* [日本語多義語学習辞典 形容詞・副詞編] ("A Learner's Dictionary of Multi-sense Japanese Words: Adjectives, Adverbs"), Tokyo: ALC.

Moriyama, S. [森山 新] (2012). *Nihongo tagigo gakushū jiten Dōshi hen* [日本語多義語学習辞典 動詞編] ("A Learner's Dictionary of Multi-sense Japanese Words: Verbs"), Tokyo: ALC.

Perrault. S.J. (ed.) (2008). *Merriam-Webster's Advanced Learner's English Dictionary*. Springfield, MA: Merriam-Webster.

Sinclair, J. (ed.) (1987). *Collins Cobuild English Language Dictionary*. London & Glasgow: Collins.

Yamada, T. [山田 忠雄], Shibata, T. [柴田 武], Sakai, K. [酒井 憲二], Kuramochi, Y. [倉持 保男], Yamada, A. [山田 明雄], Uwano, Z. [上野 善道], Ijima, M. [井島 正博], Sasahara, H. [笹原 宏之] (eds.) (2012). *Shinmeikai kokugo jiten dai 7han* [新明解国語辞典第 7 版]. Tokyo: Sanseido [三省堂].

DEVELOPMENTS OF READING TUTOR, A READING SUPPORT SYSTEM FOR JAPANESE LANGUAGE LEARNERS

Yoshiko KAWAMURA[*]
Tokyo International University
kawamura@tiu.ac.jp

Abstract

The present paper gives an overview of the tools and materials included in the Japanese language reading tutorial system Reading Tutor and the multilingual lexicographical project Reading Tutor Web Dictionary. This is followed by a discussion of possible uses of Reading Tutor and the Web Dictionary in Japanese language instruction and for supporting autonomous language learning. The paper further presents one particular use of these tools and resources in the development of learning material for foreign candidates taking the Japanese national examination for certifying care workers. We conclude with suggestions for effective guidance in fostering autonomous vocabulary learning.

Keywords

reading support; multilingual dictionary; autonomous learning; language for special purposes; language learning support tools

Izvleček

Članek predstavlja orodja in gradiva v sistemu za podporo branju v japonščini Reading Tutor in v večjezičnem slovarskem projektu Reading Tutor Web Dictionary ter njihovo možno uporabo tako za poučevanje japonščine kot tudi za podporo samostojnemu jezikovnemu učenju. Nadalje predstavlja konkreten primer uporabe teh orodij in virov za izdelavo učnega gradiva za tuje kandidate, ki se pripravljajo na japonski državni izpit za zdravstvene delavce. V zaključku predlaga nekaj pedagoških pristopov za učinkovito podporo samostojnemu učenju besedišča.

Ključne besede

bralna podpora; večjezični slovar; samostojno učenje; jezik stroke; orodja za podporo jezikovnemu učenju

[*] Translated by Kristina Hmeljak Sangawa

1. Introduction

The Japanese language reading tutorial system Reading Tutor has been used in many ways by Japanese learners and teachers around the world since its appearance on the web in 1999. It has been accessed more than 2 million times since its inception, and it is presently being accessed 1500 times a day on average. In 2003 a companion project was launched to add multiple languages to the Reading Tutor dictionary tools. This paper is organised in the following way. Section 2 describes the reading support system Reading Tutor, section 3 proposes ways in which this tool can be used to foster autonomous learning, and section 4 presents the multilingual lexicographical project Reading Tutor Web Dictionary and its possible uses In section 5, we present a case study of the use of these resources for the development of learning materials for foreign candidates taking the Japanese national examination for certified care workers, and section 6 concludes with suggestions for effective guidance and learner support geared at fostering autonomous vocabulary learning in the age of the internet.

2. The Japanese language reading tutorial system Reading Tutor

Reading Tutor (http://language.tiu.ac.jp/) is an Internet site combining reading materials and various tools to support reading and learning Japanese as a second language. As seen in Table 1, the uses of Reading Tutor are graphically displayed on the front page of this site. Reading Tutor is freely available online and can be accessed and used any time from anywhere in the world.

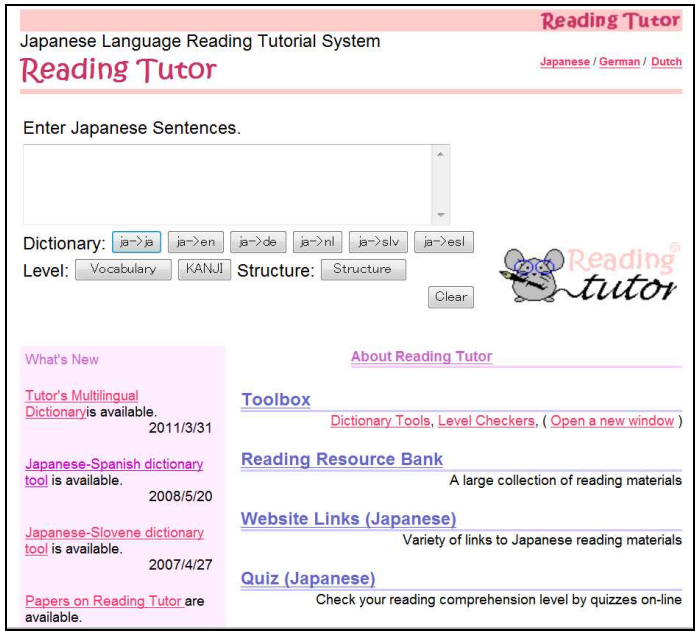


Figure 1: Reading Tutor's top page

Reading Tutor includes the following tools and resources which were developed to support Japanese language reading and learning. In order to use the tools, users input (type or paste) text into the upper box, and click on the button of the tool they wish to use.

2.1 Dictionary tool

The following dictionaries are available, as can be seen from the button acronyms: a monolingual dictionary with Japanese definitions of the Japanese headwords (ja->ja), a Japanese-English (ja->en), Japanese-German (ja->de), Japanese-Dutch (ja->nl), Japanese-Slovene (ja->slv) and a Japanese-Spanish (ja->esl) dictionary. This tool links all words appearing in the input text to the chosen dictionary and displays the text alongside dictionary information on a new page. When the user clicks any word to look up dictionary information, relevant dictionary information appears on the right side, and clicked words appear at the bottom of the left frame, enabling the user to save a list of unknown words for later revision.

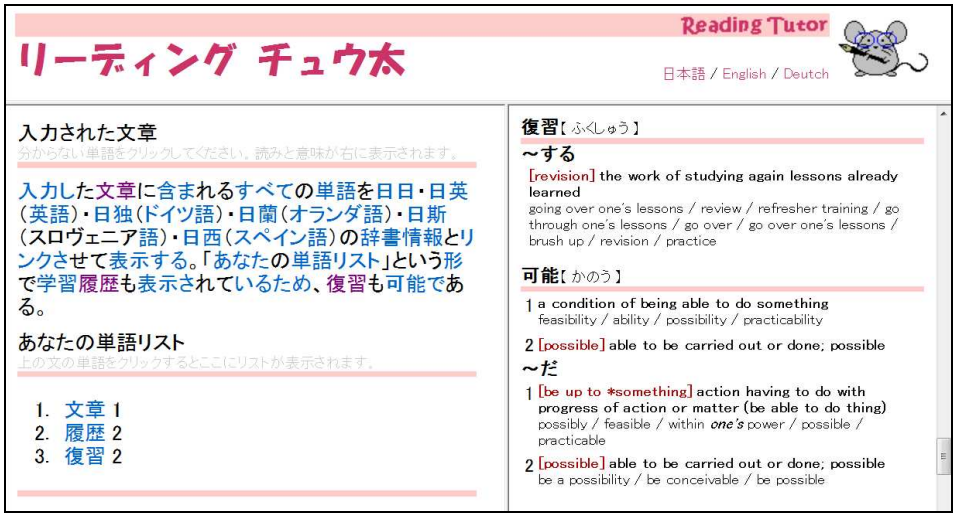


Figure 2: Reading Tutor dictionary tool output window

Figure 2 shows an example of the dictionary tool's output, where the user has selected the Japanese-English dictionary and clicked once on the word 文章 and twice each on the words 履歴 and 復習. The last clicked word is shown at the top of the right-hand frame.

2.2 Level Checker (Vocabulary Checker and Character Checker)

These two tools analyse the text, determine the difficulty level of each word or each character in the text according to the old version of the Japanese Language

Proficiency Test Content Specifications (JF & AIEJ, 2002; hereafter abbreviated to JLPT), and display the text showing words or characters in the colour of the JLPT level they belong to, alongside a list of all words or characters appearing in the text, divided by JLPT levels, as displayed in Figure 3 (see next section).

2.3 Grammatical patterns tool

This tool automatically detects grammatical patterns and functional words in the text, links them to (a part of) information in the dictionary of grammatical patterns *Nihongo bunkei ziten* [日本語文型辞典] (Group JAMASII, 1998) and displays them in a layout similar to the dictionary tool. The tool can be accessed by pressing the button [structure]. The site also offers the reading resources listed below.

2.4 Reading resource bank

This repository contains reading material that can be immediately used for studying Japanese either in class or for self-study. Each text is marked for level of difficulty, calculated according to the vocabulary it contains, making it easy to choose the material that is most appropriate for the reader.

2.5 Collection of website links

The link page presents carefully selected sites that may be useful to learners of Japanese as a second language grouped by subject into news, Japanese culture, Japanese language learning and Japanese language teaching.

2.6 Grammar quiz

The site includes grammar questions at levels 1 and 2 of JLPT, where users can check their own grammatical competence. This section includes both an automatic answer checker and grammar explanations.

Tools and materials which make up Reading Tutor thus serve different purposes and can be useful both for Japanese language instruction and for supporting autonomous learning, as explained in detail in Tutor’s introductory manual (Kawamura, 2009).

3. Editing teaching materials using Reading Tutor

3.1 Verifying learners' proficiency level

By combining material from the Reading Resources Bank with the Vocabulary Checker, teachers can easily gauge their students' level. When learners read a passage and click on unknown words to check the reading or meaning, these words are recorded in a section named "Your word list". Teachers can then retrieve this list, feed it into the Vocabulary Checker, and find the JLPT level the unknown words belong to. The same process can be used for any other reading material not included in the Reading Resources Bank, simply by inputting the text into the dictionary tool first and then using it as reading material. One advantage of this process is that it allows students to check their own level objectively.

3.2 Developing reading materials appropriate for the students' level

The Vocabulary Checker can also be used to develop reading materials which are appropriate to the students' level. When the user inputs a text into the Vocabulary Checker, the text is processed and appears as in Figure 3 below, with each word in the text colour-coded for its level in JLPT. The text in Figure 3 is an example of a processed text, containing the entry 雷 [*kaminari* - "thunder"] from the site *Key Aspects of Japan* (Sugiura & Gillespie, 2012).

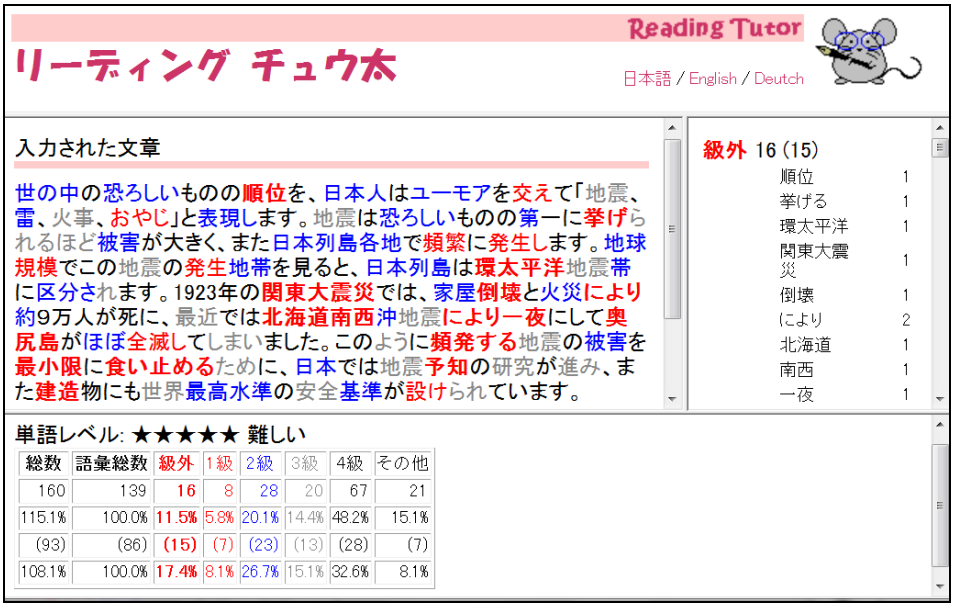


Figure 3: Output of Reading Tutor Vocabulary Level Checker

As can be seen in Figure 3 all words in the processed text are colour-coded according to their level in JLPT: level 4 words (the easiest) are displayed in black, level 3 (upper beginners) in grey, level 2 (intermediate) in blue, level 1 (advanced) in red, and words which are not included in the JLPT are displayed in bold red characters. Teachers can thus read through the colour-coded text and rewrite the parts not appropriate to the students' level. For example, if a teacher is preparing reading material for intermediate students, the words in red could be rewritten. This process can be then repeated with the rewritten text to check for any remaining difficult words, thus ensuring the text has been edited appropriately for the level of the students.

3.3 Developing supplementary learning materials

Reading materials for language courses are usually accompanied by supplementary material such as word lists, comprehension questions, cloze tests, grammar tests, Chinese character tests, texts with phonetic guides (also called *furigana*, text in smaller type added above or below Chinese characters to show pronunciation), and other types of material. Some of these materials - word lists, Chinese character quizzes, phonetic guides - can be easily prepared using the tools offered on the Reading Tutor web site.

a) Word lists: To prepare a word list for a reading passage, teachers can feed the passage into the Vocabulary Checker and obtain a list of all words contained. If needed, they can then feed this list into one of the dictionary tools to obtain a list of words with their readings and translations.

b) Kanji quiz: The colour-coded output of the Vocabulary Checker can be used to prepare kanji quizzes. The colour-coded text may be useful in helping student visually grasp their reading level.

c) Text with phonetic guides: A text with phonetic guides can easily be edited using the *furigana* function of the Reading Tutor Web Dictionary, as explained in section 4. A text with phonetic guides can be used as preparation for a kanji quiz, for practicing reading aloud and other exercises.

3.4 Integrating autonomous learning into reading classes

Reading classes can be conducted more effectively by combining them with autonomous learning, individual preparation and review. This is especially effective in classes of students with very different proficiency levels, or mixed classes with students from countries using kanji script and those where kanji is not used. If the text to be studied is made available on the web, students can prepare and review their reading using the dictionary tool. They can also use these tools to prepare for writing or reading aloud tests. However, they should be warned that all tools use an automatic parser and that errors may occur.

4. The multilingual Japanese dictionary *Reading Tutor Web Dictionary*

To answer the demand for different dictionaries from learners of Japanese around the world, a project for the creation of a multilingual dictionary was started in 2003. The result of this project is offered online as the *Reading Tutor Web Dictionary1*. In this project, a monolingual dictionary of Japanese headwords with Japanese definitions and examples is being translated into more than 20 languages by dictionary teams around the world. Each entry is made available online as soon as it is edited by the editing team. The compilation of the Japanese-Russian and Japanese-Vietnamese dictionaries have been completed for words in the JLPT.

The *Reading Tutor Web Dictionary* includes a tool for analysing any input and glossing it with entries from any of the dictionaries it includes, a tool for adding phonetic guides (*furigana*) above all Chinese characters, and an example search tool.

All tools can be accessed from the *Reading Tutor Web Dictionary* top page, as shown in Figure 4.

Figure 4: The Reading Tutor Web Dictionary top page

The tools on this page can be used in the same way as Reading Tutor. When a user types or pastes any text into the text box at the top of the page, and presses the Dictionary button, the text is automatically analysed and glossed with entries from the selected dictionary or dictionaries. By default, the dictionary of the language set as the preferred language in the user's web browser is automatically selected. In addition, the user can choose to add any number of dictionaries by ticking the boxes in the dictionary list below the text box. For example, by ticking the boxes next to the words

[1] http://chuta.jp/

Japanese, English and Slovenian, as in Figure 4 above, the user obtains glosses from the Japanese-Japanese, Japanese-English and Japanese-Slovenian dictionaries at the same time, as shown in Figure 5.

Figure 5. *Reading Tutor Web Dictionary's* text output

Figure 5 shows the output obtained by inputting the Japanese version of this paper's first paragraph into the Reading Tutor Web Dictionary and selecting Japanese, English and Slovene. Each word in the processed text in the left frame is linked to the dictionary entries listed in the right frame. When the user clicks on any of the linked words on the left, the corresponding dictionary entry in the frame on the right side automatically scrolls up and appears at the top of the right hand frame. At the same time, the same word appears in the bottom part of the left frame, under the heading あなたの単語リスト [*anata no tango risuto*] ("Your word list"). Figure 5 is a picture of the user's monitor after the user has clicked on the words 公開 [*kōkai* "opening"], 関係 [*kankei* "relationship"] and 超える [*koeru* "to get over"]; the last word clicked is displayed at the top of the right hand frame. As can be seen in this screenshot, each meaning described and translated in the dictionary entry is also accompanied by examples, so that users can check not only the meaning or translation of each word, but also its typical usage. The dictionary entry for the verb *koeru* has been completed in the Japanese-Slovenian dictionary and is therefore displayed with examples and their Slovenian translations.

In the case of English, Korean, Chinese, Indonesian and Tagalog, all entries of the JLPT word list have been translated into these languages in the form of a mini-dictionary where entries have only the word's translation equivalents without examples, so that even for entries of the main multilingual dictionary where the

examples have not been edited yet, the word's translations appear in the glossary. In the Figure 5 example, entries from the Japanese-English mini-dictionary appear in the right frame.

All words listed under "Your word list" are linked to the glossary on the right, so that users can easily review them after reading.

Examples in the dictionary entries are also preprocessed, so that the user only needs to click on any example to see it with a glossary of all the words it contains on its right side.

The *furigana* function can be used by ticking the "Furigana" checkbox at the bottom of the Reading Tutor Web Dictionary top page. If the checkbox is ticked, the output in the left frame is displayed with phonetic guides above each word, as can be seen in Figure 6.

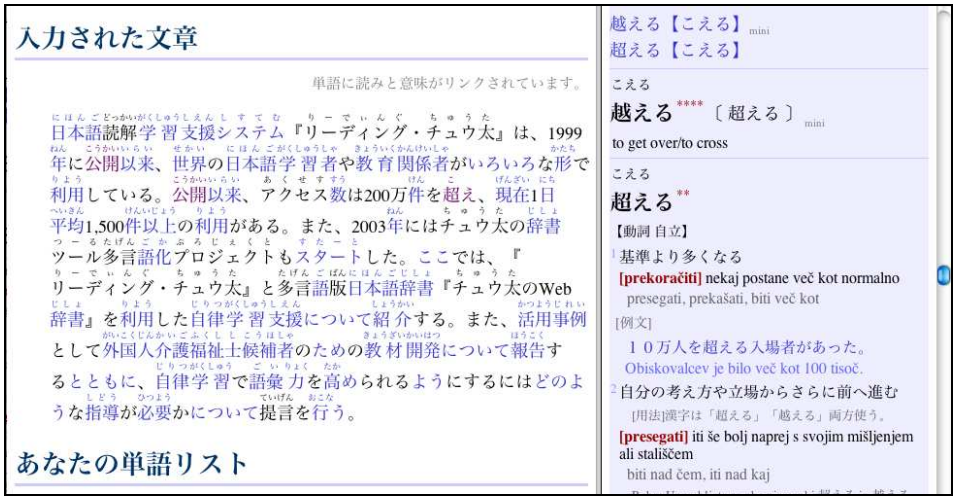


Figure 6: *Reading Tutor Web Dictionary*'s text output with phonetic guides (furigana)

4.1 Example search, fuzzy searches

Users can also search for single dictionary entries and examples using the "Word / Example Sentence Search" box at the bottom of the web dictionary top page, as shown below in Figure 7. Here it is possible to use the special character * (an asterisk) for fuzzy searches. When the user inputs a word without any special characters, the system displays the entry of the headword that exactly matches the word. By adding an asterisk to the end of a character string in the search box, the user can find entries for all headwords beginning with that character string (for example, しか* returns しかし,しかた etc.), and by adding an asterisk to the beginning of the word, find all headwords ending with that string (for example, *出す returns 出す, but also 思い出す, 言い出す, 飛び出す etc.).

Moreover, by adding an asterisk to both sides of the search string, the user can see all occurrences of that string both within headwords and in the examples, thus finding words that are present in the examples, even if they are not yet included as dictionary entries.

Figure 7: Word/Example sentence search box on *Reading Tutor Web Dictionary*'s top page

In Figure 7, the string として [*toshite* "as"] is enclosed in asterisks to search for all examples containing this character strings. Figure 8 below shows the result of this search.

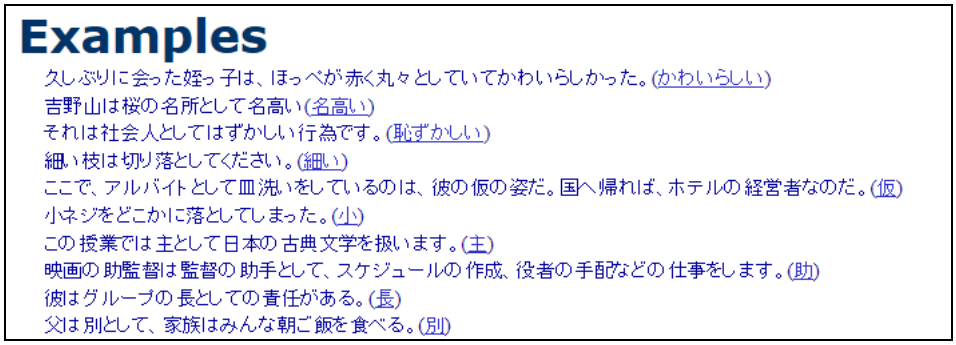


Figure 8: Part of the output obtained by searching for *として* using the Word/Example sentence search tool

Examples containing the search string are displayed as shown above, followed by the headword they belong to, in brackets. If the Japanese part of the entry has been translated into other languages, these translations can be displayed by clicking on the headword shown within brackets.

This tool retrieves all matches for any character string, without further linguistic analysis. As can be seen in Figure 8, searching for the string *toshite*, which means "as is" if used by itself, results in an output screen also containing examples where *toshite* is part of another word, such as *otoshite* ("dropping"). Finding all occurrences of the string *toshite* used as a compound particle can be a useful exercise in reading. Such a list can also be used in class, for an exercise in distinguishing different uses of a word or pattern.

5. Case study: Using Reading Tutor's tools to develop learning materials for foreign candidates to the Japanese national examination for certified care workers

This section presents a case study in which the Reading Tutor tools were used to develop learning materials to support autonomous study. From 2009 to 2011, a set of learning materials was developed to help foreign candidates studying for the Japanese national examination for care workers. These learners are prospective health care workers who have come from Indonesia and the Philippines to Japan on the basis of Japan's Economic Partnership Agreements with these countries. These candidates are expected to pass the Japanese national examination for certified care workers within 4 years of their arrival to Japan, while working as trainees in hospitals or other institutions. They are only allowed to take the examination once. They usually come to Japan with little or no knowledge of the Japanese language, but cannot afford to spend much time in language classes because of their full time employment as nursing trainees. We therefore decided to develop learning materials to support autonomous study, using Reading Tutor's tools to develop a vocabulary list, a character list and a dictionary of nursing terminology (Kawamura & Nomura, 2010, Kawamura et al., 2011)

5.1 Construction of a dictionary of nursing terminology

As a preliminary step in the development of learning materials, we surveyed the vocabulary used in past national examinations for care workers, using the Vocabulary Checker. As a result of this survey, the examination questions were found to include 3799 words not listed in the JLPT vocabulary list. We then compiled a Japanese-Indonesian-Tagalog-English glossary of all words in the JLPT and of all words appearing at least twice in the past examination questions surveyed. Each Japanese headword in the glossary usually has one translated equivalent for each language, but the dictionary editing system also allows for more than one translation to be added to each word. The glossary was incorporated into the Reading Tutor Web Dictionary, making it immediately available to foreign candidates of the Japanese national examination for care workers.

5.2 Development of an instructional vocabulary list for candidates to the national examination for care workers

In order to compile an instructional vocabulary list to help candidates to the national examination for care workers, vocabulary frequently appearing in the examination was surveyed using Vocabulary Checker.

All vocabulary used in the 16 national examinations from the 3rd to the 18th was surveyed, and 1146 words were found to appear at least 16 times in these examinations. By excluding vocabulary of level 3 and 4 (the easiest) of the Japanese

Language Proficiency Test, we obtained a list of 808 words, which we named the “808 Nursing Wordlist” (介護単語 808 [*Kaigo tango 808*]).

	見出し語	読み	品詞	やさしい日本語	英語	インドネシア語	メモ
□□	愛	あい	名詞	人やものを大切に思う気持ち	love/affection	cinta/mencintai	
□□	相手	あいて	名詞	何かを一緒にする人	partner/opponent	pasangan/lawan	
□□	明らか	あきらか	な形	はっきりしている	clear	terang	
□□	預かる//預ける	あずかる//あずける	動詞	頼まれて管理や世話をする//管理や世話をしてもらう	to keep//to leave sth with sb	dititipi (titipan)//menitipkan	
□□	アセスメント	あせすめんと	名詞	何かを始める前に調べること	assessment	penilaian	
□□	与える	あたえる	動詞	あげる	to give	memberikan	
□□	～当たり	あたり	接尾	～にどれだけあるか	per	per/setiap	
□□	悪化（する）	あっか	動詞	悪くなる（こと）	degeneration/getting worse	memburuk	

Figure 9: Sample of the **“**808 Nursing Wordlist**”** (介護単語 808 [*Kaigo tango 808*]) for candidates to the national examination for care workers

As can be seen in Figure 9, each entry in the 808 Nursing Wordlist is composed of a Japanese headword, its reading (in hiragana), a definition in simple Japanese, and one or more English and Indonesian equivalents. The Japanese explanations use words from the high frequency level words of Japanese.. The English and Indonesian equivalents were compiled using the Reading Tutor Web Dictionary. A column with checkboxes was added on the left side for learners to be able to monitor their own progress.

5.3 Compilation of supplementary material to support autonomous learning

Since candidates to the national examination for care workers come from linguistic backgrounds where Chinese characters are not used, the large number of characters to be newly learned is a serious problem. This needs to be taken into consideration when planning efficient ways to learn the nursing terminology included in the 808 word list described above. We therefore compiled a supplementary set of learning materials aimed at mastering Chinese characters, the “Nursing Character List” (介護漢字 [*Kaigo kanji*]). An excerpt from this list is shown in Figure 10.

	漢字・単語	読み	品詞	やさしい日本語	英語	インドネシア語	メモ
□	定	さだめる・テイ	動詞	決める	to decide	dapat menetapkan	
□	定める	さだめる	動詞	決める	to decide/to appoint	dapat menetapkan	
□	定期	ていき	名詞	決まった期間	regular	periodik	
□	定年	ていねん	名詞	会社などをやめることになっている年齢	retiring age	usia pensiun	
□	指定（する）	していする	動詞	物や場所をそれと決める（こと）	appointment/to appoint	penunjukan	
□	設定（する）	せってい	動詞	決める（こと）	setup/to set up	penetapan/menetapkan	
□	測定（する）	そくてい	動詞	測る（こと）	measurement/to measure	ketetapan	
□	特定（する）	とくてい	動詞	それだと決める（こと）	identification/specific/to identify	spesifik	

Figure 10: Sample of the “Nursing Character List” (介護漢字 [*Kaigo kanji*]) for candidates to the national examination for care workers

In order to compile this material, the 808 Nursing Wordlist was fed into the Kanji Checker and all characters used in the 808 Nursing Wordlist were ranked and listed according to frequency, beginning with the most frequent character. As can be seen in Figure 10, each character is listed in a highlighted (grey) line containing its possible readings, the part of speech of the word the character represents when used by itself, an explanation of its meaning in simple Japanese, translations into English and Indonesian, and some space for user’s notes. This line with basic information about a single character is followed by all 808 Nursing Wordlist entries which include the character. These words are arranged according to the position of the character in the word, in order to help users focus their attention on word formation and to strengthen their ability to infer a word’s meaning from the characters in the compound. Words which begin with the headword character are displayed first, followed by words ending with the character, and finally by all other words containing the character.

The structure of the entries, showing the part of speech of the word represented by a character when used by itself and in a list of compound words ordered by the location of the character within the word, is conducive to a better understanding of word formation and word networks, as can be seen in the following examples.

Verb-like characters in compound words are often accompanied by characters for the verb’s object on their right side, or characters for verb modifiers on their left side. For example, when the character 定, which can be used to represent the root of the verb *sadameru* [定める “to decide”], is used in the first part of a compound word, such as in *teiki* [定期 “fixed period, fixed term”], *teinen* [定年 “retirement age”], the second part of the compound word usually refers to the object of this verb (in the case of 定 - *sadameru* - “to decide, to fix”, the second character in the compound refers to what is being decided or fixed: *-ki* 期 “period”, *-nen* 年 “year, age”). On the other hand, when the character is used as the last part of a compound word, such as in *shitei* [指定

"indicate" + "decide" → "designate, specify, appoint"], *settei* [設定 "provide, prepare" + "decide" → "establish, set up"], *sokutei* [測定 "measure, fathom" + "decide" → "measure"], *tokutei* [特定 "special" + "measure" → "specify"], the first part of the compound word refers to the manner in which the action described by the verb is carried out (in the above cases, the first character in the compound refers to how something is being decided).

Noun-like characters in compound words can act as modifiers to the following characters when appearing at the beginning of a compound, or can be modified by the characters preceding them when appearing at the end of a compound. For example, when the character 人, which can be used to represent the word *hito* [人 "person"], is used as the first part of a compound word, such as in *jinsei* [人生 "person" + "life" → "human life"], *jinkou* [人工 "person" + "construction" → "man-made, artificial"], the prefix 人 modifies the following morpheme, meaning "of man", "human". On the other hand, when the character is used as the last part of a compound word, such as in *kojin* [個人 "single" + "person" → "individual"], *roujin* [老人 "old" + "person" → "old person"], the morpheme written with this character (*jin*) is modified by the first part of the compound word, meaning "some kind of person".

The Nursing Character List also includes an explanation on how Chinese character compounds are formed. Vocabulary learning can be easier when the mechanisms for composing Chinese character compounds is understood. By observing these rules, learners acquire the ability to guess the meaning of unknown compounds, which may lead to better reading proficiency.

The 808 Nursing Wordlist and the Nursing Characters List are available for download at http://chuta.jp/Archive/, for learners anywhere.

5.4 Building a workbook with controlled vocabulary for the national examination for care workers

The following guidelines were followed to develop a workbook for the national examination for care workers.

a. vocabulary is limited to JLPT level 3 and 4 and the 808 Nursing Wordlist;

b. expressions not included in these lists are only used when they are deemed too important to be reworded using controlled vocabulary, and are explained in notes;

c. words related to Japanese culture or systems are also explained in notes;

d. explanations in notes are written using basic vocabulary and words in the 808 Nursing Wordlist;

e. sentences are as short as possible; bulleted lists are also used if necessary.

Learning materials for beginning students are usually compiled with simplified vocabulary, but considering that the materials are meant to prepare students for the

state examination, we decided that words frequently appearing in the examination should not be rewritten using simpler expressions, and therefore adopted the guidelines outlined above.

To aid the editing team in compiling the workbooks according to the above guidelines, we needed a tool that would enable editors to check immediately whether the explanations and exercises they were writing contained any vocabulary outside the scope of the selected wordlists. By using the Vocabulary Checker within the Reading Tutor site (section 2.2), we developed a tool which we named "Nursing wordlist Checker" and have made it publicly available at (http://basil.is.konan-u.ac.jp/chuta/level/).

The Nursing wordlist Checker analyses any input text using the morphological analyser MeCab (Kudo, 2006), splits the text into words, checks it against the wordlists mentioned above and shows the same text with each word coloured using the following colour scheme:

- **black**: basic words (included in JLPT level 3 and level 4) and function words;
- **green**: words in the 808 Nursing Wordlist;
- **blue**: words from JLPT level 2 which are not included in the 808 Nursing Wordlist;
- **red**: all other words.

Editors used this tool while compiling the workbooks to check whether the words they were using were included in the specified vocabulary lists. When a keyword they needed to use was not included in the lists, they added a gloss which explains its meaning with words from the list. All words from the 808 Nursing Wordlist which appeared in the text are marked with bold typeface, so that learners can easily discern important words and will find their explanation in the accompanying 808 Nursing Wordlist. All texts in the workbooks are marked with *furigana* (pronunciation guides), so that learners can easily find them in the Reading Tutor Web Dictionary.

These were the guidelines and tools used to compile five workbooks in simplified Japanese to support autonomous learning for the new curriculum of the Japanese national examination for care workers (Kawamura, 2011). While the above paragraphs describes a case study used to develop learning materials for prospective care workers, the same procedure could also be used to develop materials for other specialised areas.

6. Japanese teaching in the age of the internet

With the spread of the internet, learners of Japanese around the globe can now access diverse information about and in Japanese, without restrictions on time and place. At the same time, the motives of learners approaching Japanese are also diversifying. In parallel with these changes, the modes of Japanese language education

also need to change from textbook-centered classroom teaching, to a mode of language education which is conducive to autonomous learning. In this context, three approaches are crucial in supporting autonomous learning: 1) reading instruction which includes both speed-reading and intensive reading; 2) vocabulary training concentrating on word usage; and 3) character training which fosters the ability to infer the meaning of unknown words and characters. Let us consider each one in turn.

6.1 Reading instruction including both fast and intensive reading

Learners need to develop the ability to skim or scan written documents in order to find only the necessary information within a long text. At the same time, when they find a stretch of text with the information they are looking for, they need to be able to understand it accurately. With the explosion of information available now, it has become even more important to train learners to read both fast and accurately.

6.2 Vocabulary training concentrating on word usage

Each word in a language refers to meaning which expands in multiple directions, but the semantic features of words are different in each language. Vocabulary and reading instruction therefore needs to guide learners to notice not only idioms and fixed expressions, but also the semantic area covered by each word. One way of achieving this is by using a dictionary. For example, using the Reading Tutor Web Dictionary, where different meanings and usages of each word are described, translated and exemplified, learners can appreciate the different ways in which each word can be used by trying to translate each example into their mother tongue. Such an activity can help them realise the need to understand words in context, and thus contribute to their linguistic awareness.

6.3 Character training fosters the ability to infer the meaning of unknown words and characters

The number of words and characters learners must memorise when learning Japanese, especially for those coming from non-kanji backgrounds, is daunting. As described in section 5, learners need instruction regarding compound formation for words of Chinese origin. A sound knowledge of the basic rules by which such words are formed can help learners enhance their ability to guess the meaning of unknown words. Trying to infer the meaning of a new word both from its context and also by analysing the characters with which it is composed and their position within the word, before looking it up in a dictionary, may help learners deepen their vocabulary knowledge over time.

The three approaches outlined above aim at fostering autonomous learning. In this age when information in and regarding Japanese can be accessed online from anywhere in the world, Japanese language education can greatly profit from the advantages of the

Internet. We hope that some of new modes of instruction introduced here will foster learners' independence and help them advance steadily and efficiently in their learning efforts.

Acknowledgments

The present paper is a revised and extended version of presentations delivered at The University of Tsukuba, Japan Women's University, The Korean Association of Japanese Language Teaching, and The Institute of East Asian Studies at Thammasat University. The development of learning support tools and learning materials described in this paper was made possible by the collaboration of HOBARA Rei (The University of Tokyo), KITAMURA Tatsuya (Konan University) and all the members of the Reading Tutor project team. I would like to extend my sincere thanks to all of them.

References

Group JAMASII [グループ・ジャマシイ] (Eds.) (1998). *Nihongo bunkei ziten* [日本語文型辞典]. Tokyo: Kurosio Publishers [くろしお出版].

Japan Foundation [国際交流基金], & Association of International Education, Japan [日本国際教育協会]. (2002). *Nihongo nōryoku shiken Shutsudai kijun (kaiteiban)* [日本語能力試験出題基準【改訂版】] - *Japanese Language Proficiency Test: Test Content Specifications (Revised Edition)*. Tokyo: Bonjinsha [凡人社].

Kawamura, Y. [川村よし子] (2009). *Chūta no tora no maki --- Nihongo kyōiku no tame no intānetto katsuyōjutsu* [チュウ太の虎の巻―日本語教育のためのインターネット活用術] ("*Practical uses of the internet for Japanese language education*"). Tokyo: Kurosio [くろしお出版].

Kawamura, Y. [川村よし子], & Nomura, A. [野村愛] (2010). Kaigo no tame no mini jisho o kumiireta jisho tsūru no kaihatsu [介護のためのミニ辞書を組み入れた辞書ツールの開発] ("Development of a dictionary tool including a mini-dictionary for health work"). *Nihongo kyōiku hōhō kenkyūkaishi - Japanese Language Education Methods* [日本語教育方法研究会誌], *17* (1), 22-23.

Kawamura, Y. [川村よし子], Nomura, A. [野村愛], Natō, A. [名藤杏子], Kaneniwa, K. [金庭久美子], Saiki, M. [斉木美紀], Kitamura, T. [北村達也] (2011). Kaigo fukushishi kōhosei no tame no kokka shiken ni muketa kyōzai no kaihatsu [介護福祉士候補生のための国家試験に向けた教材の開発]. In *Nihon kyōiku kōgaku gakkai 27 kai zenkoku taikai kōen ronbunshū* [日本教育工学会第 27 回全国大会講演論文集], 625-626.

Kawamura, Y. [村よし子] ed. (2011). *Yasashii nihongoban kaigo fukushishi shin curriculum gakushū workbook* [やさしい日本語版 介護福祉士新カリキュラム学習ワークブック], 5 vols., Shizuoka: Shizuoka Prefecture [静岡県].

Kudō, T. [工藤拓]. (2012). *MeCab : Yet Another Part-of-Speech and Morphological Analyzer*, Retrieved from: http://mecab.sourceforge.net/.

Sugiura, Y. [杉浦洋一], & Gillespie, J.K. (2012). *Key Aspects of Japan - Nichi-ei taiyaku Nihon bunka kiiwaado jiten* [日英対訳日本文化キーワード事典]. Retrieved from: http://www.japanlink.co.jp/ka/.

JAPANESE LEARNING SUPPORT SYSTEMS: HINOKI PROJECT REPORT

Bor HODOŠČEK
Tokyo Institute of Technology,
Graduate School of Decision Science and Technology,
Department of Human System Science
hodoscek.b.aa@m.titech.ac.jp

Kikuko NISHINA
Professor Emeritus,
Tokyo Institute of Technology
knishina@m06.itscom.net

Abstract

In this report, we introduce the Hinoki project, which set out to develop web-based Computer-Assisted Language Learning (CALL) systems for Japanese language learners more than a decade ago. Utilizing Natural Language Processing technologies and other linguistic resources, the project has come to encompass three systems, two corpora and many other resources. Beginning with the reading assistance system Asunaro, we describe the construction of Asunaro's multilingual dictionary and its dependency grammar-based approach to reading assistance. The second system, Natsume, is a writing assistance system that uses large-scale corpora to provide an easy to use collocation search feature that is interesting for its inclusion of the concept of genre. The final system, Nutmeg, is an extension of Natsume and the Natane learner corpus. It provides automatic correction of learners errors in compositions by using Natsume for its large corpus and genre-aware collocation data and Natane for its data on learner errors.

Keywords

CALL; reading assistance system; writing assistance system; scientific and technical Japanese corpus; learner corpus; genre

Izvleček

V poročilu predstavljamo projekt Hinoki, ki je bil zastavljen pred več kot desetimi leti za izdelavo spletnih sistemov za računalniško podprto učenje japonščine kot tujega jezika. Z uporabo jezikovnih tehnologij in drugih jezikovnih virov so bili v okviru projekta razviti trije sistemi, dva korpusa in veliko drugih virov. V nadaljevanju predstavljamo sistem Asunaro za podporo branju, izgradnjo njegovega večjezičnega slovarja in pristop k podpori branju, ki sloni na odvisnostni slovnici; sistem za podporo pisanju Natsume s preprostim vmesnikom za iskanje žanrsko določenih kolokacij v obsežnih korpusih; ter sistem Nutmeg za samodejno popravljanje napak. Nutmeg je nadgradnja sistema Natsume in učnega korpusa Natane, ponuja samodejno

popravljanje napak med samim pisanjem z uporabo žanrsko določenih kolokacijskih informacij iz obsežnih korpusov preko sistema Natsume in informacij o napakah piscev, ki se učijo japonščine kot tujega jezika, iz korpusa Natane.

Ključne besede

računalniško podprto učenje jezika; sistem za podporo branju; sistem za podporo pisanju; korpus znanstvene in tehnične japonščine; učni korpus; žanr

1. Preface

According to a 2009 report from the Japan Foundation, there are over three and a half million people learning Japanese outside Japan.[1] Fortunately, access to good general educational materials has become easier with the advent of the Internet. However, the situation for learners with specialized language needs, such as those who are pursuing a degree at a Japanese institution of higher education, has unfortunately not improved as much.

The Japan Student Services Organization (JASSO) reports that there are 138,075 international students in Japan; another report by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) Agency for Cultural Affairs reports that there are 40,799 international students studying Japanese in Japan.[2] For these students who are pursuing specialized study at institutions of higher education in Japan, the following are just some of the skills they will have to master:

- read textbooks
- write reports and papers
- listen to lectures and take notes
- present at conferences or seminars

Because it is hard to tailor the Japanese language class to meet the specialized language needs of each learner's field of specialization, an alternative is needed. One way of approaching this problem is to provide Computer-Assisted Language Learning (CALL) systems for use online. CALL systems can supplement the language learning provided to learners and assist them in studying material from their field of specialization. The construction of such self-learning, individualized learning systems

[1] Detailed statistics are available in the Japan Foundation's 2009 "Survey Report on Japanese Language Education Abroad", available at https://www.jpf.go.jp/j/japanese/survey/result/dl/survey_2009/gaiyo2009.pdf

[2] While JASSO provides numbers for international students in their 2011 report available at http://www.jasso.go.jp/statistics/intl_student/data11_e.html, they do not provide information on the number of students who are required to take Japanese classes. MEXT offers an independent report with slightly different numbers, available at http://www.bunka.go.jp/kokugo_nihongo/jittaichousa/h23/gaikoku_6_03.html, that does contain the number of Japanese language learners.

has been the goal of the Hinoki project. The following report describes three systems and several linguistic resources, the results of pursuing this goal for over a decade.

1.1 Report Overview

After providing an overview of the Hinoki project, Chapter 2 describes the linguistic resources in use by the project. Chapter 3 introduces the Asunaro reading support system, which features courseware designed for science and engineering students, as well as a multilingual dictionary that includes several commonly underrepresented Asian languages. Chapter 4 describes Natsume, a writing assistance system that is backed by large-scale corpora and provides an easy to use search interface for collocations. Chapter 5 introduces the search system for our Natane learner corpus, which has applications to second language acquisition research and machine-learning applications for automatic learner error detection and correction. Chapter 6 introduces Nutmeg, an automatic error correction system for learner's writing. Finally, Chapter 7 concludes this report by offering a summary and our perspectives for future work.

2. Linguistic Resources

The Hinoki project relies heavily on linguistic resources, though it is also a producer of such. Linguistic resources used in the project are native and learner corpora, as well as dictionaries. To meet the goals of the project, some linguistic resources had to be developed: multi-lingual dictionaries, purpose-specific corpora, as well as learner corpora.

In the earlier systems, emphasis was put on native resources, as they enable a Data-Driven Learning approach to learning Japanese.

However, to really know where and why learners make mistakes, a learner corpus is also essential and is where more recent efforts have been focused on.

2.1 Native Resources

As part of the *Nihongo kōpasu* ("Japanese Corpus") project led by the National Institute for Japanese Language and Linguistics (NINJAL) for 4 years, the main goal of our group was to explore the ways in which the project's new Balanced Corpus of Contemporary Written Japanese (BCCWJ) could be applied to Japanese language education. As we are focused on finding ways to assist Japanese language learners in writing academic reports and papers, it was necessary to compile another corpus containing this genre, in addition to using the BCCWJ. For several other reasons explained below, the Japanese version of Wikipedia was also used.

2.1.1 BCCWJ

The National Institute for Japanese Language and Linguistics (NINJAL) created the Balanced Corpus of Contemporary Written Japanese (BCCWJ) in the span of five years between 2006 and the end of 2011 (Maekawa, 2007b, 2007a, 2012). The objective of the project was to compile a tagged corpus of contemporary written Japanese that had sufficient scale and coverage of sub-varieties of written language to offer a representative sample of Japanese written language. Such a language resource had not previously existed for Japanese, and its creation was seen as important for the future development of any research with a need for representative Japanese language data, including official government language policy.

The BCCWJ consists of the Publication, Library and Special-Purposes sub-corpora, each of them accounting for roughly one-third the size of the BCCWJ. The Publication sub-corpus includes books, magazines, and newspapers and is sampled from all published material in Japan between 2001 and 2005. The Library sub-corpus includes books sampled from several library holdings within the Greater Tokyo Metropolitan area. The Special-Purposes sub-corpus differs from the other two in that it should not be considered a representative sample of written Japanese, but rather serve as useful comparison material for the others.

2.1.2 Scientific and Technical Japanese Corpus (STJC)

Unfortunately, the data needs of providing writing and reading assistance in an academic context are not fully satisfied by the BCCWJ. While some sub-corpora are close in subject matter (topic) and writing style (register), the lack of inclusion of genuine research papers from academic journals precludes their ability to serve as a representative sample of written science and engineering discourse. It was thus necessary to build a new corpus that contained a representative and authentic sample of academic writing. The new corpus was named the Scientific and Technical Japanese Corpus (STJC), and consists of papers from several scientific and technical journals written in Japanese. The following criteria were used when choosing which journals to collect papers from:

1. The journal must specialize in some scientific or engineering field.
2. The journal is published by a society with at least a thousand members.
3. The journal has reasonable review standards.
4. The journal allowed us to use the text from the papers as example sentences in our system.

Currently, papers are included from the Journal of Natural Language Processing, the Journal of the Japan Society of Civil Engineers, the Journal of Nippon Medical School, the Journal of the Chemical Society of Japan, the Journal of Environment and Natural Resources Engineering, as well as the Journal of the Institute of Electrical Engineers of Japan.

2.1.3 Wikipedia

The decision to include the Japanese version of Wikipedia[3] was made for several reasons. For many tasks, the quantity of data provided by the BCCWJ is sufficient (Maekawa, 2011). However, due to the nature of the data used in Natsume, which includes triplet combinations of nouns, case particles and verbs, the amount of extractable data for any but the most common expressions quickly becomes insufficient. Additionally, other NLP technologies deployed in the Natsume system, such as getassoc[4], are more precise at scales of data in the range of Wikipedia. Another requirement of the project is that text data from corpora should be legally available to be displayed online. The permissive license of Wikipedia allows all us to show all sentences as example sentences in Natsume.

One unfortunate side effect of including Wikipedia is that for many less frequent collocations, the only information on them is available in Wikipedia, making any genre comparisons impossible. Another demerit of including Wikipedia is the inclusion of grammatical mistakes and comparatively long sentences, which average at around 51 characters compared with an average of 35 for newspapers (see Table 1).

Table 1: Character counts and average sentence length for all corpora.

Corpus	Subcorpora	Characters	Average sentence length
STJC		6,108,143	58.46
Wikipedia		372,901,202	51.10
BCCWJ	Books	53,801,124	37.59
	Yahoo! Q&A	9,763,298	30.69
	Diet minutes	8,712,108	75.06
	Textbooks	1,818,571	28.44
	White papers	8,443,965	58.25
	Yahoo! Blogs	5,246,121	23.26
	Magazines	455,634	31.41
	Newspapers	1,188,355	34.90
Total		**468,438,521**	**48,04**

[3] Currently a snapshot from 2008. Wikipedia data dumps are available from http://dumps.wikimedia.org/.

[4] Available from http://getassoc.cs.nii.ac.jp/.

2.2 Natane – Learner Corpus

2.2.1 Introduction

Natane is a Japanese language learner corpus annotated with learner errors. The main benefit of learner corpora in the context of writing assistance, when compared to native corpora, is that they enable insights into the kinds of errors learners make. For example, in the case of Natane, comparing learner error tendencies based on their first language might guide customizations to lesson plans based on the learner's first language.

Compared to native corpora containing the writings of native speakers, learner corpora are often smaller in size and variety. This is due to the difficulty of obtaining learner writing, which in most cases is elicited for the construction of the corpus and not collected from readily-available sources as in the construction of the BCCWJ. Another common differentiator is the inclusion of error annotations and background information on the learners who produced the material used.

The end goal of the construction of this corpus is the construction of well-formed and sufficient machine-learnable data for automatic writing error correction. It should be noted that while relatively simpler things like the construction of a spellchecker, co-occurrence checker, or writing style checker are possible, features that hinge on an understanding of semantics and discourse are hard to make practical even in state-of-the-art NLP systems.

As an ongoing joint project with several Japanese language teachers, the collection and annotation of the corpus initially proceeded along the following stages:

1. Collection of learner essays and their transcription.
2. Pilot annotation of learner errors using Excel (Cao & Nishina, 2010; Cao, Kuroda, Yagi, & Nishina, 2010).
3. Analysis of pilot annotation and definition of final error classification framework (Cao, Kuroda, Yagi, & Nishina, 2011; Cao, Yagi, & Nishina, 2012).
4. Use of the multipurpose annotation tool Slate for error tagging.

2.2.2 Collection

The essays were collected from undergraduate and graduate students as well as students attending Japanese language schools. All essays were written to a specific topic, though not all topics are the same. Each learner's age, nationality, university level, first language, major and Japanese language learning experience, as well as other background information, were recorded with the essay. Additionally, learners signed a waiver authorizing the anonymized usage of their essay in our project.

Although more than 5000 sentences have been collected, currently only around 3500 have been annotated[5]. In its present state, Natane consists of 285 essays obtained from 192 learners, totaling 205,520 characters. From a total of 9,041 annotations, there are 6,789 learner errors. The distribution of learners by their first language is biased towards Mandarin Chinese speakers, who account for more than half of learners and essays. The remaining languages are predominantly from Asia.

Table 2: Distribution of essays by first language.

First language	Male	Female	Unknown Sex	Total
Mandarin Chinese	62	64	26	152
Marathi	6	23	7	36
Vietnamese	18	9	0	27
Korean	24	3	7	34
Spanish	2	0	0	2
Malay	8	0	0	8
Slovenian	7	0	0	7
Hungarian	1	0	0	1
Thai	1	0	0	1
Unknown	5	0	12	17
Total	**133**	**90**	**62**	**285**

2.2.3 Pilot Annotation

While error classification frameworks for languages such as English and French already exist (Díaz-Negrillo & Fernández-Domínguez, 2006; Granger, 2003; L'Haire & Faltin, 2003), there were no preexisting comprehensive error annotation scheme or descriptive framework for Japanese language learner errors. Because of the lack of such a framework, the project decided to construct one itself, drawing from previous research as well as the annotator's teaching experience (Cao & Nishina, 2010). During the pilot annotation process, it became clear that there were two kinds of error annotations. The first were ordinary, unambiguous errors and the second kind were errors where the annotator felt the particular language usage was unnatural. Ordinary errors include deviations from standard orthography, syntactic function (voice, tense, aspect, modality), conjugation, and subject-predicate incongruity. They are typically easy to annotate and occur frequently. Unnatural errors include word choice, addition or omission of text units (phrase, paragraph, etc.), and are typically less frequent and harder to annotate, leading to lower agreement between annotators.

[5] An up-to-date breakdown of data included in Natane, including the nationalities of learners, is available at http://hinoki.ryu.titech.ac.jp/natane/stats.

2.2.4 Error Classification Framework

The feedback gained from the pilot annotation process was crucial for refinements in the error classification framework (Cao, Kuroda, Yagi, & Nishina, 2011). The resulting error annotation framework is hierarchical, able to take into account different viewpoints regarding learner errors, as well as enable the systematic annotation of such errors (Yagi & Suzuki, 2012).

The hierarchy consists of at most four levels, with higher levels corresponding with courser, more abstract categories, and branches out in three principal dimensions:

1. Error level – the linguistic level of the error (i.e. phoneme, word, phrase, ..., discourse; the word tag is further classified into word classes like noun, verb, etc.)
2. Error category – type and form of error
   - type: addition, omission, word order, deviation from standard orthography, etc.
   - form: conjunction, conjugation, collocation, (Japanese letter) script
3. Error source – reason or background for error (i.e. annotator's subjective opinion on source of error: register and style mismatch, coherence, first language interference, etc.)

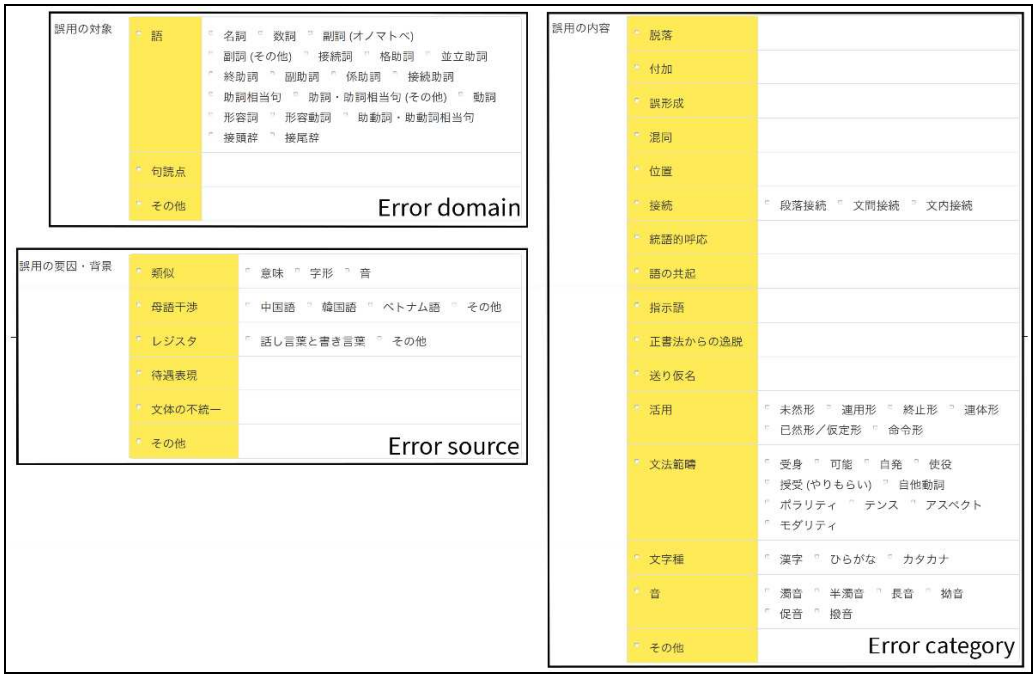


Figure 1: The hierarchical error classification framework used in Natane

2.2.5 Error Annotation Process with Slate

After the error classification framework was decided on, the choice had to be made between continuing to use Excel to annotate the corpus or finding another solution. Though Excel's free-form nature served the formative stage of the annotation process, significant drawbacks related to its ad-hoc usage became clear. The choice was then made to use the web browser-based Slate corpus annotation and management system[6], as it offers the following advantages over Excel: higher data integrity and greater data diversity (Kaplan, Iida, Nishina, & Tokunaga, 2012). Slate decreases the chance for inconsistent annotation by eliminating the chance for errors with respect to formatting differences between annotators and misplacement of annotations into the wrong table cell, among other problems. Using Slate also increases the diversity of possible annotations, by enabling more than one annotation per segment (sentence) as well as annotations that overlap or span multiple sentences. Previously the format of the Excel table limited the amount of possible error annotations to one per sentence. Slate also provides an overhead view of the hierarchical error classification framework that - coupled with an interface that allows the user to see all annotations at a glance - enables efficient and speedy annotation.

As there was considerable data included in the existing Excel tables, it was not re-annotated but rather converted for inclusion into Slate. All new annotations are being recorded using Slate. Three teachers specializing in Japanese language education at different universities separately annotated all essays using the Slate corpus annotation and management system.

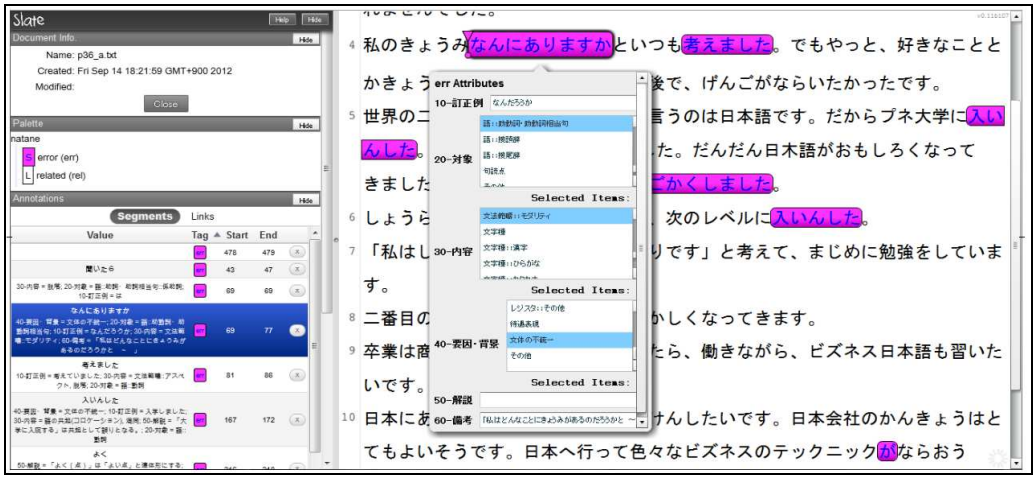


Figure 2: An example of composition errors annotated with Slate; marked areas represent errors, with the left pane providing detailed information including all error annotations.

[6] More information is available at Slate's homepage: http://www.cl.cs.titech.ac.jp/slate/.

2.3 Conclusion

The Hinoki project depends on the existence of many large-scale corpora, most of which are already available to the research community. For more specialized needs, such as the inclusion of a representative sample of scientific and technical Japanese, no corpora existed, so one had to be constructed. The available Japanese language learner corpora are still few, although recent developments have increased the number available: Learner's Language Corpus of Japanese[7], Teramura corpus[8], NINJAL's learners corpus[9], JC Corpus[10] are just some of the corpora available now. The existing major differences between Natane and these learner corpora is that they are more focused on the annotation of grammatical errors and thus have a less comprehensive error classification framework than the one used in Natane.

Though not mentioned in this chapter, without the availability of high-quality Natural Language Processing tools for Japanese, it would be hard to impossible to make use of much of these linguistic resources. The specific tools used in each system are detailed in the explanations of each system separately.

3. Asunaro: Multilingual Reading Assistance System

3.1 Introduction

The first system developed under the Hinoki project was the Asunaro multilingual reading assistance system (Nishina, Okumura, Yagi, et al., 2002; Nishina, Okumura, Abekawa, et al., 2004). Development of Asunaro began in 1999 and the system was first released online in 2002.

At its inception, Asunaro was unique in that it integrated a multilingual reading and learning environment into one online system accessible to anyone with an Internet connection. At the time, most systems targeted English language learners, while Asunaro incorporates several Asian languages. This was important because the number of international students from neighboring Asian countries studying at universities in Japan is greater than that of students from English-speaking countries.

The main goal of the system was to help Japanese learners read and understand academic material in Japanese. The main target of the system is Japanese language learners enrolled in Japanese universities majoring in the fields of science and engineering. Many of them are expected to be able to read academic papers and textbooks in their field, but it is often difficult to provide for their specialized learning

[7] http://cblle.tufs.ac.jp/llc/ja/

[8] http://teramuradb.ninjal.ac.jp/

[9] http://jpforlife.jp/taiyakudb.html

[10] http://www34.atwiki.jp/jccorpus/pages/21.html

needs in university Japanese language classes. The use of Asunaro was seen as a way to enable personalized learning for those learners.

3.2 Main Features

Users accessing the Asunaro system are presented with the main screen containing a text box into which they can paste or directly enter Japanese language text for analysis. The main screen is split into three areas consisting of the user input area in the top left, the translation and example sentence area in the top right, and a detailed word and phrase view of single sentences at the bottom.

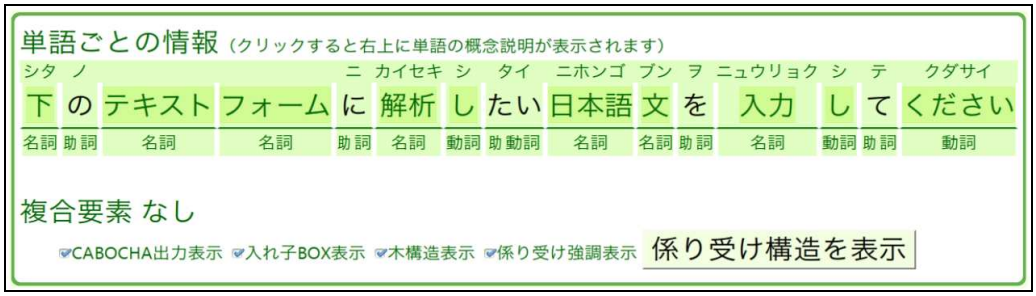


Figure 3: Bottom area containing morphologically analyzed user input with readings and word class information provided by MeCab (Kudo, 2012).

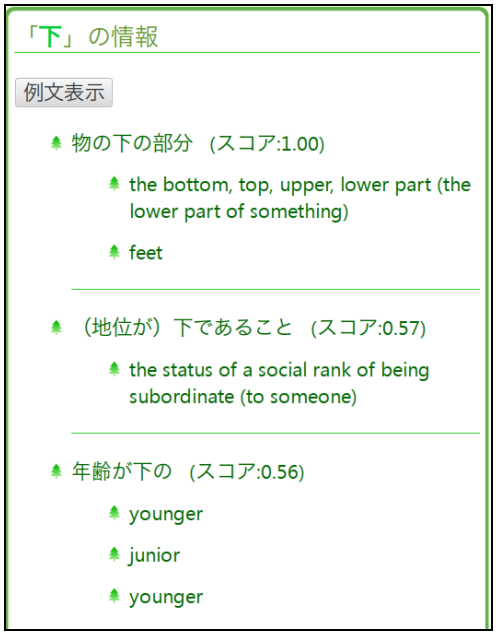


Figure 4: Top right area containing translations and example sentences of user selected words or phrases.

Users click words or phrases[11] in the bottom area to update translations in the top right area. Translations appear in order of importance, based on the application of meaning disambiguation using the surrounding word context.

Finally, clicking on the arrow at the beginning of each sentence takes the user to the secondary screen where they can see the dependency structure of the sentence.

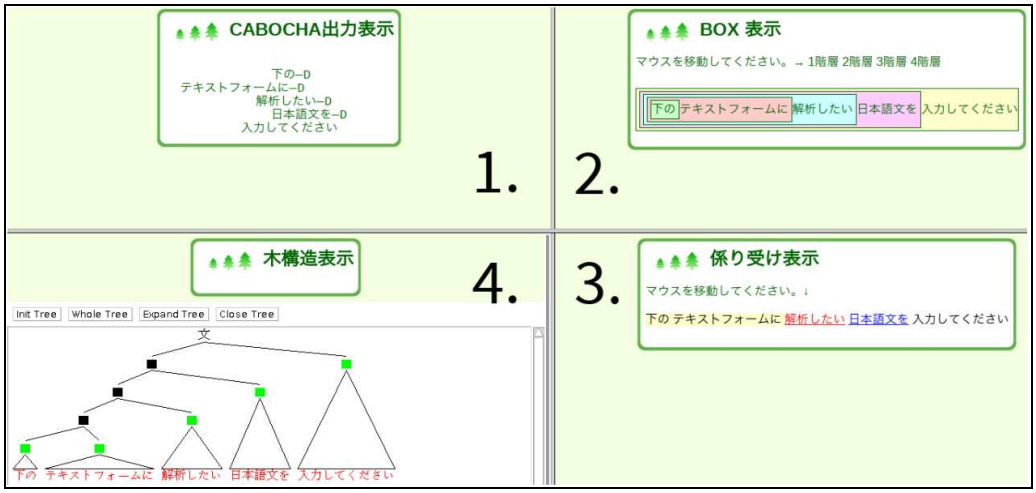


Figure 5: Different views of dependency structure in secondary screen. (Clockwise from top left)
1. raw output from the Japanese dependency parser CaboCha (Kudo & Matsumoto, 2012),
2. embedded dependency structure ("ireko") display,
3. mouse-over dependency link-like display, and
4. tree structure representation of sentence.

3.3 Courseware

However, the usage outlined above is in many ways too difficult for non-advanced learners. For beginning-to-intermediate learners who are studying in the fields of science and engineering, the provided courseware is more appropriate. Learning a language through reading is best when the material read is learner-level appropriate. Asunaro makes use of a textbook (Nishina, 2001) which is written specifically for intermediate level undergraduate science and engineering students. The main goal of the courseware is to help science and engineering students achieve proficiency in technical communication to be able to read papers and discuss research in seminars. All courseware in Asunaro was checked for parsing mistakes and manually corrected.

[11] Idioms and phrases like *te wo tunagu* and *kao ga hiroi* are automatically recognized as such and also marked as phrases.

Additionally, the courseware contains an audio playback feature so users can listen to the courseware material while learning to read it.

3.4 Multilingual Dictionary

Although electronic Japanese-English dictionaries have been available since the beginning of the 1990s, for many Asian languages such as Malay, Thai or even Chinese, no electronic dictionary was available at the time of Asunaro's inception. As more than half of all international students in Japan come from other Asian countries, support of languages other than English was seen as a high priority. The EDR Electronic Dictionary is used for its translations between English and Japanese, as well as its concept ID, which links every Japanese word to a concept.[12] Enabling translations of Japanese words into languages other than English required the construction of a new multilingual dictionary that would map words from the target language to EDR's concept ID. This differed from many similar systems at the time that used English as the intermediary language. Excluding the Japanese and English entries from the EDR dictionary, the multilingual dictionary contains around 25,000 entries for Chinese, and around 5,000 each for Thai, Indonesian and Malay.

Another unique feature of the system was that it provided a common language-independent framework for handling compound expressions. This is important as compounds and phrasal units are language-dependent and must be handled on a per-language basis. For Japanese, phrasal units and compounds are detected using CaboCha and the EDR electronic dictionary.

3.5 Related Work

Reading Tutor, which is a reading assistance system widely used in Japan and abroad, also contains a multilingual dictionary (Kawamura, Kitamura, & Hobara, 2012, 2000). Additionally, Reading Tutor's "Kyozai Banku" (Kawamura & Kitamura, 2001) is a similar effort to the courseware feature in Asunaro to provide leveled reading material.

Rikai.com is a popular website that provides hiragana readings and English translations of online text[13]

3.6 Conclusion

Asunaro was constructed to assist students from the fields of science and engineering to read and understand technical Japanese. For beginning- to intermediate-

[12] More information on the EDR Electronic Dictionary is available at https://www2.nict.go.jp/out-promotion/techtransfer/EDR/index.html.

[13] Rikai.com is accesible from http://www.rikai.com.

level learners, it includes courseware aimed at assisting them to eventually be able to read authentic texts from their field. For advanced learners, the copy and paste nature of the system allows them to focus on learning just the sentences they do not yet fully understand. It uses the EDR Electronic Dictionary as a basis for constructing a larger multilingual dictionary, presenting learners with glosses into their native language: English, Chinese, Malay, Thai and Indonesian.

4. Natsume: Writing Assistance System

4.1 Introduction

Natsume is an online writing assistance system that began operating in 2009 (Hodošček, in press, 2012; Abekawa, Hodošček, & Nishina, 2011). The initial focus during the development of Natsume was to enable users to not just be able to search for collocations, but also be able to convince themselves of the correct usage of a collocation in several ways. Thus, Natsume was to not just provide raw collocation information, but was to enable users to look for similar collocations and compare collocational tendencies between different genres.

While Asunaro assists international students in reading, Natsume focuses on assisting them in writing technical Japanese. For example, writing reports or papers at universities can be hard if the students cannot differentiate between what words or expressions are spoken and what are written Japanese. As a study and writing aid, the use of conventional (non-corpus-based) electronic dictionaries is prevalent among international students. However, these dictionaries seldom contain information on a word's usage with respect to written and spoken language. Natsume, by virtue of having access to corpora from various genres, contains information that can be used to determine if a word is appropriate for spoken or written Japanese.

When writing in a second language, it is often the case that one knows the meaning of a noun or verb, but does not know what verb goes together with what noun. Conventional dictionaries often contain only a limited amount of information on frequently co-occurring patterns of words. These frequently co-occurring patterns of words are called collocations and are important because they offer more contextual information about a word than what is found in conventional dictionaries. Moreover, knowledge of collocations has been shown to be essential to achieving high second language proficiency (Pawley & Syder, 1983).

Users can use the system to find collocations of a word, check the correct use of a word or collocation by looking at example sentences, and compare observed frequencies in various genres. This follows the philosophy of data-driven language learning by giving users access to authentic information which they can then use as the basis for any decisions with respect to writing and word choice.

Natsume's current target users are intermediate to advanced learners of Japanese, as well as Japanese native speakers.

4.2 Main Features

The interface can be divided into three views:

1. Collocation view – where users search for the collocate words of any noun, adjective or verb.
2. Genre comparison view – looking at the genre frequency distribution of a collocation reveals that collocation's genre tendencies.
3. Example sentence view – authentic examples enable the learner to see how the collocation is used at the sentence level.

Users must select the particular collocation pattern they want to search for and a matching noun, verb or adjective into the search box to start the search. Searching for a word will present several lists, grouped by case particle and sorted by frequency, of the searched word's collocates. The sorting scheme is user selectable and one can choose from the default frequency, Dice's coefficient, t score, Jaccard similarity coefficient, Log-likelihood ratio, Chi-square coefficient, and Mutual-Information score for different types of collocations. The color bars at the right of every collocate indicate the relative frequency (or score) of the collocation in all corpora. Additionally, users can search for and compare two or more similar patterns at the same time to help decide on which one is more suited for them. Using this feature, users can additionally resort on any input word, which makes it easy to see at a glance which words collocate with which input words.

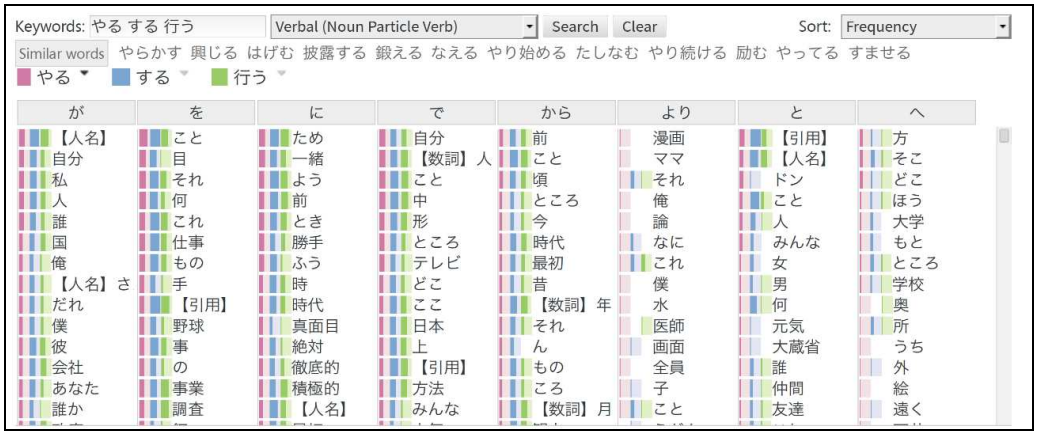


Figure 6: Main interface containing word search input area, similar words feature and collocates of the three input words. Frequency information for each verb is uniquely color coded.

When the user is interested in seeing more information on a particular collocation triplet, clicking on the collocate will load the genre comparison view to the bottom of the main collocation view. The behavior of the click can be set to one of:

- particle/conjugation expansion – can be used to compare among different grammatical uses of collocates
- synonym expansion – can be used to automatically compare among similar collocates
- no expansion (default) – standard view, only provides genre information of selected collocation
- click expansion – can be used to manually compare genre information of collocates

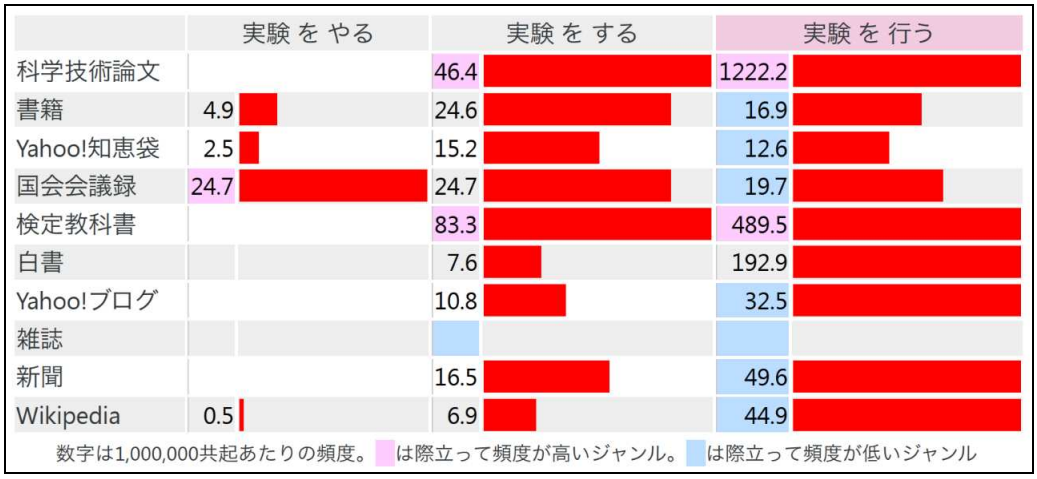


Figure 7: Comparing three collocates of /jikken/ “experiment” taking the /wo/ case particle: /yaru/ “to do” (colloquial), /suru/ “to do”, and /okonau/ “to conduct, carry out”.

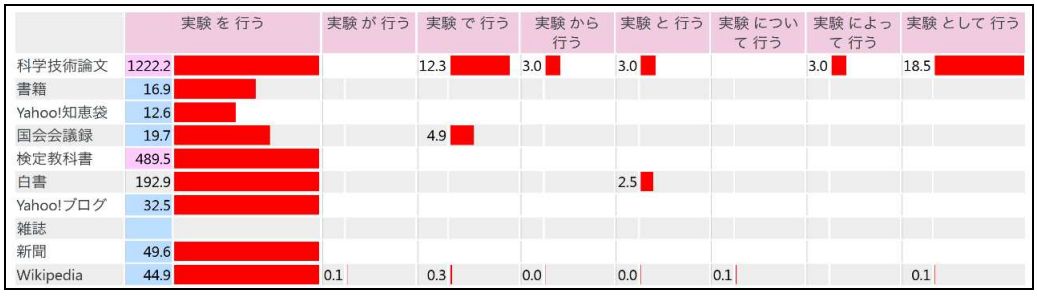


Figure 8: Comparing genre frequencies between different patterns including /jikken/ and /okonau/ using the case particle and conjugation expansion feature.

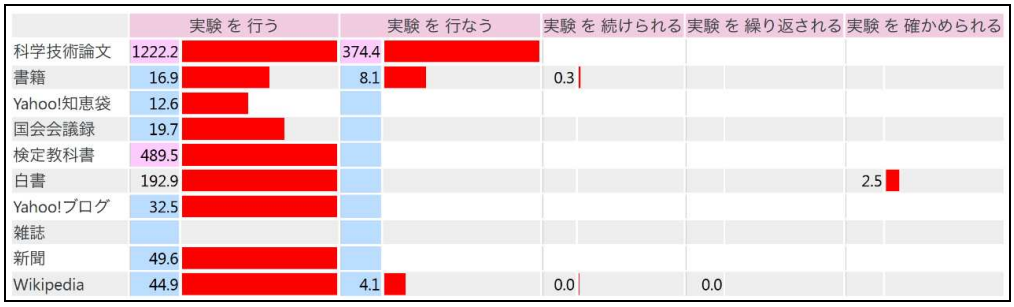


Figure 9: Comparing genre frequencies between similar collocates of /jikken/ using the similarity expansion feature.

In the genre comparison view, users can visually compare a collocations usage across different genres. The frequency numbers visible in the genre comparison interface are the relative frequency of occurrence of a collocation per 100,000 collocations. This is done to ensure the frequencies are comparable even if the corpus sizes differ, as is the case here. Additionally, Natsume uses the chi-square test to color-code genres as blue if the frequency of occurrence is significantly larger than the average across all genres, and pink if the frequency is significantly lower. Genres that are not color coded do not significantly differ from the mean.

When even more information is desired, the user can bring up the example sentence view which shows example sentences from the selected corpora. Sentences are displayed randomly by genre up to a limit of six sentences per genre. One can judge if a collocation is suitable for one's writing context by comparing its frequency across genres, its differences with similar collocations, and the actual usage as seen in example sentences from different corpora.

Figure 10: Examples sentences of /jikken wo okonau/.

4.3 Related Work

In parallel to the construction of the BCCWJ, NINJAL commissioned the construction of two search systems, one that is freely available and offers basic KWIC search features, called Shonagon, and another subscription-based one that allows searching with regular expressions over short and long unit words, called Chunagon[14]. Another system that shares Natsume's focus on Japanese language education is NINJAL-LWP[15], a lexical profiler for a subset of the BCCWJ (Pardeshi, 2012). It contains features similar to Natsume, but differentiates itself by providing many different kinds of collocations.

Perhaps the most sophisticated collocation query system for Japanese is the Sketch Engine, a "Corpus Query System incorporating word sketches, one-page, automatic, corpus-derived summary of a word's grammatical and collocational behaviour" (Sketch Engine, 2012; Kilgarriff, Rychly, Smrz, & Tugwell, 2004). The Sketch Engine supports multiple languages including Japanese through a 400 million token web-based corpus (JpWaC) that was first released in 2008. More than 50 collocational and grammatical relations are in use in the word sketch grammar (Srdanović-Erjavec, Erjavec, & Kilgarriff, 2008). The Sketch Engine also contains a unique word comparison feature, called word sketch difference, which is in some aspects similar to searching for several words at the same time using Natsume, though it is also more sophisticated.

4.4 Collocation Data Extraction

The Japanese dependency analyzer CaboCha was used to extract the dependency structure of all sentences in the corpora, from which noun, particle and verb or adjective dependent patterns were extracted. Post-processing was performed on verbs to differentiate passive (/iwareru/, passive of "say") or potential (/ieru/ "be able to say")[16] with causative (/iwasu/ "to make talk") voice usage, as well as combine verbal compounds into single units (/kaki + hajimeru/ "begin to write"). Nouns were post-processed to normalize numbers, dates and personal names.

[14] Available at http://www.kotonoha.gr.jp/shonagon/ and https://chunagon.ninjal.ac.jp/, respectively.

[15] NINJAL-LWP is accessible from http://nlb.ninjal.ac.jp/.

[16] Passive and potential usage is not always discriminated by the underlying MeCab morphological analyzer and IPA electronic dictionary.

Table 3: Collocation token and type count per genre.

Genre	NPV tokens	NPV types	NPAdj tokens	NPAdj types	AdjN tokens	AdjN types
STJC	323,182	164,803	16,936	8,755	20,153	9,794
Wikipedia	17,935,354	6,818,612	614,75	203,547	950,112	338,833
Books	2,837,802	1,547,893	117,309	59,793	236,534	121,21
Yahoo! Q&A	394,228	241,74	32,848	18,005	35,508	20,328
Diet minutes	404,819	193,774	17,555	8,318	30,092	13,809
Textbooks	96,007	66,72	3,185	2,225	6,925	4,857
White papers	393,881	165,535	15,978	6,934	27,567	11,016
Yahoo! Blogs	184,122	136,134	10,418	7,678	18,759	13,877
Magazines	20,144	17,132	922	797	1,91	1,766
Newspapers	60,447	49,683	2,006	1,665	3,652	3,179
Total	**22,649,986**	/	**831,907**	/	**1,331,212**	/

4.5 Genre

The defining feature of Natsume is the ability to differentiate between expressions that are suitable for writing in an academic context and those that are not. Consider the following example from the STJC corpus:

- /Taisha ni yori hasseishita nisankatanso wa mizu ni yōkaishi, .../ "The CO2 produced from metabolism dissolved in water, ... "[17]

Comparing the expression /nisankatanso ga mizu ni yōkaisuru/ "CO2 dissolves in water", taken from the example below, with the expression /satō ga mizu ni tokeru/ "sugar dissolves in water", it is clear that the former is written in an academic, technical style, while the latter is of a more informal, spoken variety. For learners without the native language intuition needed to arrive at the same conclusion, Natsume provides a data-driven way of helping them to take a first step towards gaining this kind of intuition.

4.6 Conclusion and Future Work

Natsume is primarily a system to assist a specific part of the writing process: finding the right words for a particular writing context, which in this case is technical

[17] Excerpt from Terajima, R, Shimada, S., Oyama, T., & Kawasaki, S. (2009) "Fundamental Study of Siliceous Biogrout for Eco-Friendly Soil Improvement", Doboku Gakkai Ronbunshuu C, Vol. 65 No. 1, p. 120-130.

Japanese. Currently, example sentences in Natsume are displayed randomly. Tailoring the example sentence view to display examples appropriate to the learner's proficiency level is a future goal of the project (Hodošček, Abekawa, Murota, & Nishina, 2012).

5. Natane: Error Search System

5.1 Introduction

Having introduced the corpus Natane in Chapter 2, this chapter focuses on the search capabilities of Natane and how they might help the researcher or Japanese language teacher in analyzing their own students' errors or identify particular areas where learners have tendencies to make errors.

5.2 Interface

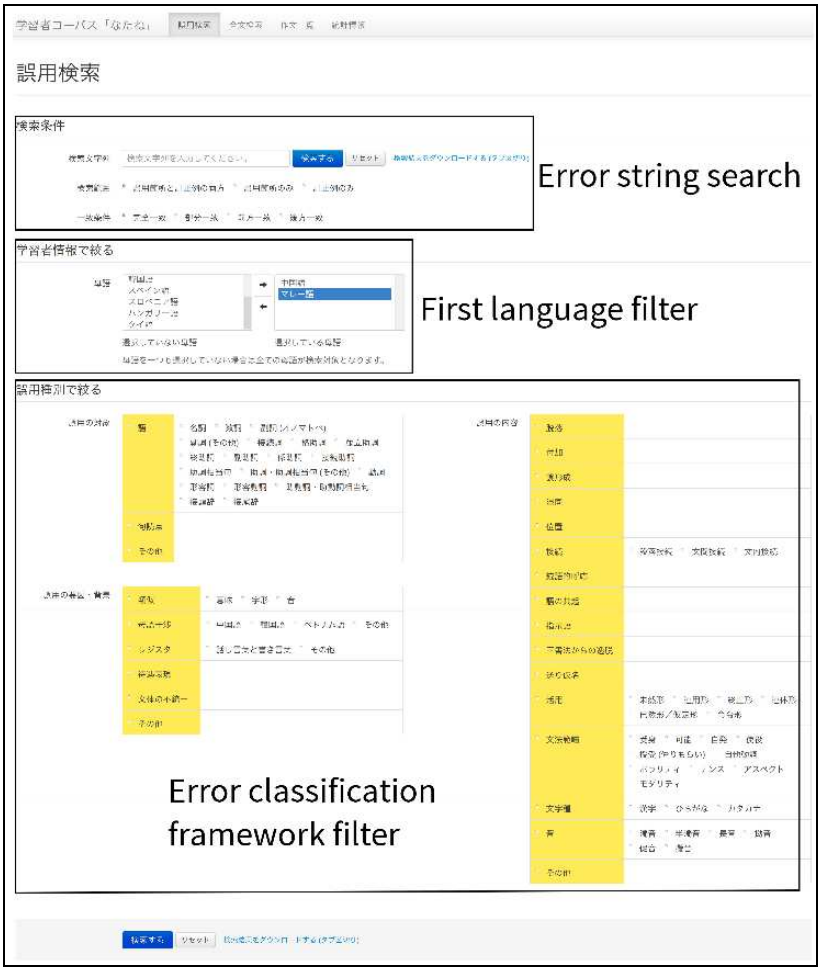


Figure 11: Natane interface: search for learner errors and filter based on first language and specific error types.

Figure 12: Searching for /yaru/ “to do” will return all learner errors containing the word. Here the correct way of writing the second sentence is to replace /yaru/ with its polite version /suru/.

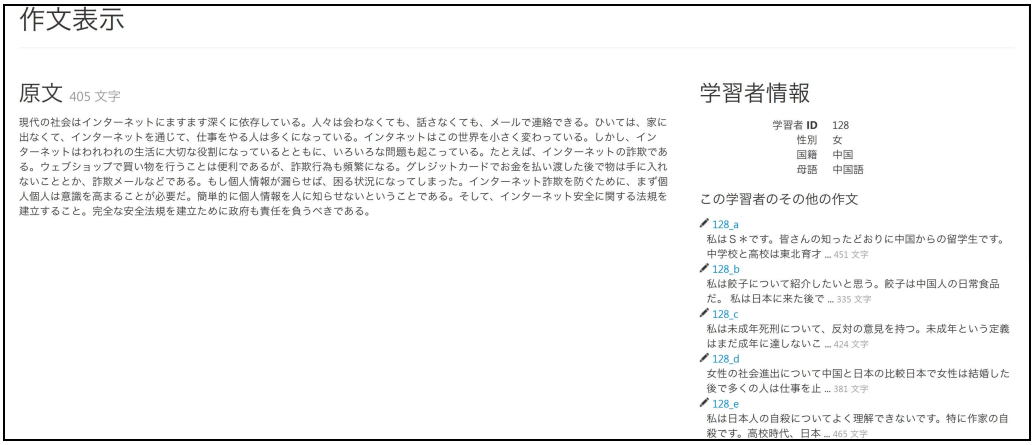


Figure 13: Situating the previous error in the learner essay.

誤用一覧

15 件の誤用タグが付与されています。

作文 ID	誤用 ID	前文脈	誤用箇所	後文脈	誤用の対象	誤用の内容	誤用の要因・背景	訂正例	解説
128_f	25084	現代の社会はインターネットにますます	深くに	依存している。人々は会わなくても、話さなくても、メールで連絡できる。ひいては、家	語::形容詞	付加, 混同, 誤形成		深く	形容詞の副詞的用法の誤り、
128_f	25085	現代の社会はインターネットにますます深くに	依存している	。人々は会わなくても、話さなくても、メールで連絡できる。ひいては、家に出なくて、	語::助動詞・助動詞相当句, 語::動詞	文法範疇::アスペクト		依存することになってきた	
128_f	25086	すます深くに依存している。人々は会わなくても、話さなくても、メールで連絡できる。	ひいては	、家に出なくて、インターネットを通じて、仕事をやる人は多くになっている。インタネ	語::副詞	混同		さらには	
128_f	25087	依存している。人々は会わなくても、話さなくても、メールで連絡できる。ひいては、家	に	出なくて、インターネットを通じて、仕事をやる人は多くになっている。インタネットは	語::助詞・助詞相当句::格助詞	混同		を	
128_f	25088	的に個人情報を人に知らせないということである。そして、インターネット安全に関する	法規を建立する	こと。完全な安全法規を建立ために政府も責任を負うべきである。				法規を確立する	
128_f	25089	お金を払い渡した後で物は手に入れないこととか、詐欺メールなどである。もし個人情報	が	漏らせば、困る状況になってしまった。インターネット詐欺を防ぐために、まず個人個人	語::助詞・助詞相当句::格助詞	混同		を	
128_f	25090	いということである。そして、インターネット安全に関する法規を建立すること。完全な	安全法規を建立ため	に政府も責任を負うべきである。	語::動詞	語の共起(コロケーション), 脱落, 誤形成	類似::意味	安全法規を確立するため	
128_f	25091	しまった。インターネット詐欺を防ぐために、まず個人個人は意識を高まることが必要だ	。簡単的に	個人情報を人に知らせないということである。そして、インターネット安全に関する法規	語::副詞	混同, 誤形成, 付加		簡単に	
128_f	25092	ても、メールで連絡できる。ひいては、家に出なくて、インターネットを通じて、仕事を	やる	人は多くになっている。インタネットはこの世界を小さく変わっている。しかし、インタ	語::動詞	混同	レジスタ::話し言葉と書き言葉	する	
128_f	25093	、困る状況になってしまった。インターネット詐欺を防ぐために、まず個人個人は意識を	高まる	ことが必要だ。簡単的に個人情報を人に知らせないということである。そして、インター	語::動詞	文法範疇::ヴォイス::自他動詞		高める	
128_f	25094	人に知らせないということである。そして、インターネット安全に関する法規を建立する	こと。	完全な安全法規を建立ために政府も責任を負うべきである。	語::助動詞・助動詞相当句	脱落	文体の不統一, レジスタ::その他	ことである。	
128_f	25095	欺である。ウェブショップで買い物を行うことは便利であるが、詐欺行為も頻繁になる。	グレジットカード	でお金を払い渡した後で物は手に入れないこととか、詐欺メールなどである。もし個人情	語::名詞	文字種::カタカナ, 音::濁音	類似::音	クレジットカード	
128_f	25096	ーネットを通じて、仕事をやる人は多くになっている。インタネットはこの世界を小さく	変わっている	。しかし、インターネットはわれわれの生活に大切な役割になっているとともに、いろい	語::助動詞・助動詞相当句, 語::動詞	文法範疇::アスペクト	類似::意味	している	
128_f	25097	存している。人々は会わなくても、話さなくても、メールで連絡できる。ひいては、家に	出なくて	、インターネットを通じて、仕事をやる人は多くになっている。インタネットはこの世界	語::助動詞・助動詞相当句, 語::動詞	接続::文内接続, 位置, 混同, 誤形成, 付加	類似::意味	出ることなく	
128_f	25098	は手に入れないこととか、詐欺メールなどである。もし個人情報が漏らせば、困る状況に	なってしまった	。インターネット詐欺を防ぐために、まず個人個人は意識を高まることが必要だ。簡単的	語::助動詞・助動詞相当句	文法範疇::アスペクト		なってしまう	

Figure 14: Viewing all learner errors in a given essay.

Searching for errors relating to the verb /yaru/ returns two errors. One example is the sentence /intānetto wo tsūjite shigoto wo yaru hito wa ōku natte iru/ “the number of people working on the Internet is increasing”, where the usage of /yaru/ is wrong because it is the colloquial form of /suru/.

5.3 Conclusion and Future Work

Natane is a learner corpus that has many potential uses, though we envision two main types of usages, one by Japanese language educators and the other by NLP researchers.

In this chapter, we described the search interface for the Natane corpus, which is targeted at the former. Japanese language educators can make use of Natane to find examples of learner errors. Also, the data provided is useful for analyzing error

tendencies due to first language interference, as well as for observing the language acquisition process.

The latter usage is primarily aimed at applications in NLP and machine learning, where Natane can be used to construct novel error correction systems. An example of one such system is introduced in the next chapter.

With the existence of several Japanese language learner corpora that all make use of different error classification frameworks, a movement towards a common standard is, perhaps, the most pressing issue.

6. Nutmeg: Writing Assistance and Automatic Error Correction System

6.1 Introduction

Natsume, while useful for finding collocations, does not automatically correct the learner's writing. In an evaluation of Natsume, it became clear that for every collocation the learner checked using Natsume, there were many more that went unchecked (Hodošček, Abekawa, Bekeš, & Nishina, 2011). The next obvious step was to develop a system that checks learners' writing and provides feedback on any errors they may have made. This writing assistance system was named Nutmeg and provides basic feedback for learners' writing using automatic error identification (Yagi, Hodošček, & Nishina, 2012). The system is unique in that it does this from two sources: native and learner corpora.

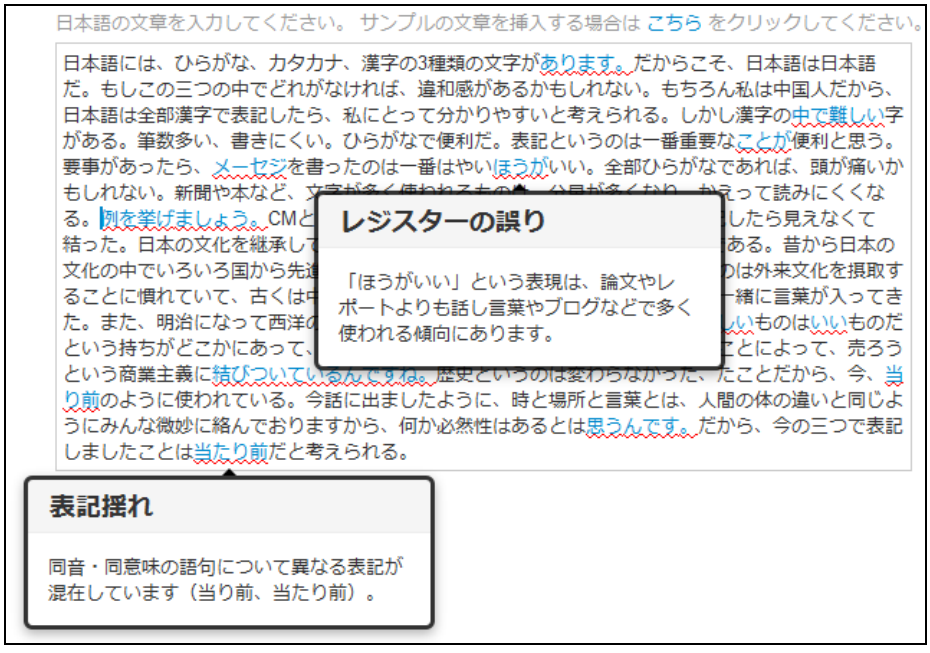


Figure 15: Nutmeg interface showing two correction suggestions.

6.2 Related Work

Compared to other existing Japanese language automatic composition correction systems, Nutmeg strives to incorporate both native and learner corpora in its correction model. An example of a more narrow application of a similar system is Chantokun. Developed at the Nara Advanced Institute of Science and Technology (NAIST), Chantokun[18] is a system that detects and corrects case particle misuse based on corrected Japanese language sentences from the Lang-8 website[19], a language-exchange social networking website where users with different first languages correct each other's writing (Mizumoto & Komachi, 2012).

An example of a system that focuses on the native corpus side of automatic error correction is the Japanese proofreading system Tomarigi[20] (Oono & Inazumi, 2011). Another example that uses the dependency structure of a sentence to revise complex sentences into easier to understand ones is the jcorrect tool (Oosaki, 2006).

6.3 Error Correction Method

In general, Nutmeg uses the native BCCWJ and STJC corpora as "correct data", whereas it uses learner errors from Natane as "incorrect data". Thus expressions that are tagged as errors in Natane become candidates for automatic correction.

Natane contains 386 orthographic errors. One way of detecting outright orthographic errors is if they go unrecognized in morphological analysis. Additionally, most such errors are found within two letters of a word. A word including an error is replaced with the corresponding word in the native word list. For example, suppose a learner were to mistaken the word /messeeji/ "message" as /meeseji/. If there is no prior learner error of the same word in Natane, then a morphological analysis reveals that it is an unknown word. The unknown word can then be matched to similar words contained in the morphological dictionary. Finally, the correct orthography can be presented to the learner.

Though the language contained in the BCCWJ and STJC should, in principle, be considered correct, this does not preclude the use of native corpora as instruments in identifying learner errors. One example is making use of the various genres available in Natsume, through which it is possible to correct collocation usage from the genre perspective. For example, using data from Natsume it is possible to automate the process of checking if a collocation is appropriate for an academic report as outlined in Chapter 4. If a learner uses the collocation /jikken wo yaru/ in an academic report, the system will be able to identify the inappropriate usage and offer the replacement

[18] See http://cl.naist.jp/chantokun/ for more information.

[19] Accessible from https://lang-8.com/.

[20] More information and download links available at http://www.pawel.jp/outline_of_tools/tomarigi/.

collocation /jikken wo okonau/ as a correction. This is possible because of the existence of relatively incompatible genres, such as those that lean towards a formal writing style (STJC and White papers) and those that lean towards an informal or spoken writing style (Yahoo! Blogs, Yahoo! Q&A and Diet minutes) (Hodošček & Nishina, 2011). The corpora in Natsume can thus be divided into so-called positive and negative genres and the relative frequencies of collocations in those genres can be tested using the chi-square test. When an expression like /jikken wo yaru/ is used we can determine that it is incorrect because its frequency in the negative genres is significantly high, while its frequency in positive genres is significantly low. A replacement collocation could be found by searching through similar collocations and testing them in the same manner or by using WordNet to expand the available search space (Bond et al., 2009; Isahara et al., 2012).

6.4 Conclusion and Future Work

The aim of Nutmeg is to become a compositional tool that is able to automatically warn learners of potential mistakes as they are making them. The two types of data backing Nutmeg's error correction facilities are native and learner corpora corresponding to the data provided by Natsume and Natane, respectively.

Though the available size of native corpora is much greater than that of learner corpora, an avenue for improvement to collocation error correction is to provide candidate replacement expressions for learner errors, perhaps using WordNet. More effort must be put into obtaining or constructing other specific-purpose corpora if other writing genres, such as business writing, are to be considered.

It is also clear that Natane should be expanded in scale in order to conduct a more comprehensive quantitative evaluation. The implementation of an automatic error correction system must be treated cautiously, because the results of automatic error correction depend on the annotations being objective. This is especially difficult for learner errors at the semantic or discourse levels, as it is here that annotators subjectivity most easily comes into play. Thus, as a first step, easier items such as orthographic errors should be considered (Yagi, Hodošček, & Nishina, 2012).

7. Conclusion and Future Work

In the span of just over a decade, the Hinoki project has produced the Asunaro, Natsume and Nutmeg systems as well as the Natane learner corpus. As the project is led by linguists, language teachers, computer engineers and educational engineering researchers, it has been able to synthesize ideas from these disciplines together into several multi-viewpoint CALL systems.

The construction of Asunaro resulted in the construction of a novel electronic multilingual dictionary that contains several often underrepresented Asian languages.

Asunaro also applied state of the art NLP research to provide a practical dependency grammar-based reading assistance system.

Natsume was developed as a corpus-backed collocation search tool that allows users to find new collocations that fit their writing style by enabling them to check the correctness of Japanese collocations they are not confident about. An immediate goal of the development of Natsume is the addition of new types of collocations to the search interface. Another goal is being pursued in ongoing work to channel Natsume's knowledge of genres and collocations into Nutmeg for use in automatic error correction. Finally, the extension of available native corpora to other learner-specific purposes, such as the writing of emails or business writing is also being considered.

The development of Natane has resulted in a unique Japanese learner corpus and an accompanying search system. It has applications for both language researchers and educators, as well as NLP applications. The future direction of Natane is closely aligned with that of Nutmeg, the usage of which will hopefully contribute to the development and further validation of Natane's error classification framework.

Nutmeg is an extension of both Natsume and Natane into automatic error correcting for learner writing. From the development of Natane it became clear that simpler orthographic and syntactic factors are easier to objectively annotate than semantic and discourse factors, which are more prone to subjective decision making on the part of the annotator. This subjective decision making also leads to greater difficulty in automating error correction at a reasonable precision. There is thus a need for a greater volume of annotations, that are objectively classified in the error classification framework of Natane. This is essential in order to realize more sophisticated error correction and composition assistance.

Finally, an effort should be made to move from the localized lexical writing assistance seen in Natsume and Nutmeg towards a more comprehensive discourse-level composition assistance. For this purpose, more inter-system collaboration with other projects is needed.

References

Abekawa, T., Hodošček, B., & Nishina, K. (2011). Go no kyōki wo kōritsuteki ni kensaku dekiru nihongo sakubun shien shisutemu Natsume no shōkai [Introduction to efficient collocation search in Japanese writing assistance system Natsume]. (Vol. 17, pp. 595–598). Proceedings of The 17th Annual Meeting of The Association for Natural Language Processing. Toyama: The Association for Natural Language Processing.

Bond, F., Isahara, H., Fujita, S., Uchimoto, K., Kuribayashi, T., & Kanzaki, K. (2009, August). Enhancing the Japanese WordNet. (pp. 1–8). ACL-IJCNLP 2009. The 7th workshop on Asian Language Resources. Singapore.

Cao, H., & Nishina, K. (2010). Establishment of error classification framework for error database. *Journal of Japanese language education methods*, *18*(1), 38–39.

Cao, H., Kuroda, F., Yagi, Y., & Nishina, K. (2011, August). Analysis of error classfication framework for learner corpus. (Vol. 2, pp. 520–521). International Conference on Japanese Language Education 2011. Tianjin, China.

Cao, H., Kuroda, F., Yagi, Y., Suzuki, T., & Nishina, K. (2010, July). Gakushūsha sakubun shien shisutemu no tame no goyō dētabēsu sakusei: dōshi no goyō bunseki wo chūshin ni [Constrction of learner error database for composition assistance: Focus on error analysis of verbs]. (Vol. 2, pp. 1571–1579). International Conference on Japanese Language Education 2010. Taipei, Taiwan.

Cao, H., Yagi, Y., & Nishina, F. K. K. (2012, August). Construction of learner corpus Natane and possible application. (pp. 1–4). 5th international conference on Computer Assisted Systems for Teaching & Learning Japanese (CASTEL/J). Nagoya. Retrieved from http://2012castelj.kshinagawa.com/proceedings/Poster/Poster5_Cao.pdf

Díaz-Negrillo, A., & Fernández-Domínguez, J. (2006). Error tagging systems for learner corpora. Revista española de lingüística aplicada, *19*, 83–102. Retrieved from http://dialnet.unirioja.es/descarga/articulo/2198610.pdf

Granger, S. (2003). Error-tagged learner corpora and CALL: a promising synergy. *The Computer Assisted Language Instruction Consortium (CALICO) Journal*, *20*(3), 465–480.

Hodošček, B. (2012). Sakubun sien to rejisutā [Writing assistance and register]. In K. Nishina, M. Kamada, H. Cao, T. Utashiro & T. Muraoka (Eds.), *Nihongo gakusyūsien no kōtiku: gengokyōiku kōpasu sisutemu kaihatu [Constructing Japanese Language Learning: Language education, corpus and system development]* (3, pp. 275–287). Tokyo: Bonjinsha.

Hodošček, B. (in press). Kōpasu no shūshū, setsumei, janru; jikken to bunseki [Corpus collection, explanation and genre; Experiment and analysis]. In Y. Sunakawa (Ed.), *Kōza nihongo kōpasu [Japanese corpus textbook series]* (Chap. 5, Vol. 5). Asakura Publishing Co., Ltd.

Hodošček, B., & Nishina, K. (2011, August). On the treatment of register in writing assistance systems. (Vol. 2, pp. 522– 523). International Conference on Japanese Language Education 2011. Tianjin, China.

Hodošček, B., Abekawa, T., Bekeš, A., & Nishina, K. (2011). Assisting co-occurrence production in report writing: Evaluation of writing assistance tool Natsume. *Journal of Technical Japanese Education*, 13, 33–40.

Hodošček, B., Abekawa, T., Murota, M., & Nishina, K. (2012, August). Readability of example sentences in writing assistance tool Natsume. (pp. 1–4). 5th international conference on Computer Assisted Systems for Teaching & Learning Japanese (CASTEL/J). Nagoya. Retrieved from http://2012castelj.kshinagawa.com/proceedings/Poster/Poster8_BorHodoscek.pdf

Isahara, H., Bond, F., Kanzaki, K., Uchimoto, K., Kuroda, K., Kuribayashi, T., Torisawa, K. (2012). Japanese WordNet. Retrieved December 10, 2012, from http://nlpwww.nict.go.jp/wn-ja/index.en.html

Kaplan, D., Iida, R., Nishina, K., & Tokunaga, T. (2012). Slate - a tool for creating and maintaining annotated corpora. *Journal for Language Technology and Computational Linguistics*, *26*(2), 89–101. Retrieved from http://www.cl.cs.titech.ac.jp/publication/673.pdf

Kasahara, S. (2012). Chantokun —tōkeiteki nihongo kōsei— [Chantokun —statistical correction of Japanese—]. Retrieved December 10, 2012, from http://cl.naist.jp/chantokun/

Kawamura, Y., & Kitamura, T. (2001, March). Development of a Japanese reading resource bank using the Internet. *Current report on Japanese-language education around the globe, 6*, 241–255. Retrieved from http://ci.nii.ac.jp/naid/110001046390/en/

Kawamura, Y., Kitamura, T., & Hobara, R. (2000, August). Development of a reading tutorial system for JSL and JFL learners using the EDR Electronic Japanese-English Dictionary. *Japan journal of educational technology, 24*, 7–12. Retrieved from http://ci.nii.ac.jp/naid/10008760744/en/

Kawamura, Y., Kitamura, T., & Hobara, R. (2012). Japanese language reading tutorial system. Retrieved December 10, 2012, from http://language.tiu.ac.jp/

Kilgarri ▪ , A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of EURALEX*.

Kudo, T. (2012). MeCab: Yet Another Japanese Dependency Structure Analyzer. Retrieved December 10, 2012, from https://code.google.com/p/mecab/

Kudo, T., & Matsumoto, Y. (2012). CaboCha: Yet Another Japanese Dependency Structure Analyzer. Retrieved December 10, 2012, from https://code.google.com/p/cabocha/

L'Haire, S., & Faltin, A. (2003). Error diagnosis in the FreeText project. *The Computer Assisted Language Instruction Consortium (CALICO) Journal, 20*(3), 481–495. Retrieved from https://calico.org/a-290-Error%20Diagnosis%20in%20the%20FreeText%20Project.html

Maebo, K. (2012). A survey of register labelling in Japanese dictionaries - Towards the labelling of words in dictionaries for learners of Japanese. *Acta Linguistica Asiatica, 3*(3), 9-26.

Maekawa, K. (2007a, March 1–3). Design of a balanced corpus of contemporary written Japanese. In *Proceedings of the symposium on Large-scale Knowledge Resources (LKR2007)* (pp. 55–58). Tokyo Institute of Technology. Tokyo, Japan.

Maekawa, K. (2007b). KOTONOHA and BCCWJ: Development of a Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the first international conference on Korean language, literature, and culture* (Vol. 2, pp. 158–177). Corpora and Language Research. Seoul.

Maekawa, K. (2011, October). Linguistics-oriented language resource development at the National Institute for Japanese Language and Linguistics. In *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)* (pp. 1–6). doi:10.1109/ICSDA.2011.6085971

Mizumoto, T., & Komachi, M. (2012, February). Robust NLP for Real-world Data: 3. Why is Japanese so Hard to Learn?—A Preliminary Investigation on Realistic Japanese Learners' Corpus and Application of Natural Language Processing to Japanese Language Learning and Education—. *IPSJ Magazine, 53*(3), 217–223.

Nishina, K. (2001). *Yasashi kagaku gijutsu nihongo dokkai nyūmon (kaitei-ban) [A gentle introduction to scientific and technical Japanese reading comprehension (revised edition)]*. International Student Center, Tokyo Institute of Technology.

Nishina, K., Okumura, M., Abekawa, T., Yagi, Y., Bilac, S., & Fu, L. (2004, March 8–9). Asunaro CALL system: combining multilingual with multimedia. In *International symposium on Large-scale Knowledge Resources LKR 2004* (pp. 69– 72). Tokyo Institute of Technology. Tokyo, Japan.

Nishina, K., Okumura, M., Yagi, Y., Totugi, N., Sawaya, T., Fu, L., … Abekawa, T. (2002). Kōbun hyōji to tagengo intāfēsu wo sonaeta nihongo dokkai gakushū shien shisutemu no kaihatsu [Development of a reading assistance system for Japanese language containing grammar display and multilingual interface features]. In *Proceedings of the 8th conference of the Association of Natural Language Processing* (pp. 228–231). Association of Natural Language Processing.

Oono, H., & Inazumi, H. (2011, March). Support tool of Japanese document proofreading and polish : Tomarigi : overview of efforts to support and labor-saving, for the labor of

correction. *Research report of JSET Conferences*, *2011*(1), 325– 332. Retrieved from http://ci.nii.ac.jp/naid/10029781745/en/

Oosaki, H. (2006). Tips for technical writing. Retrieved December 10, 2012, from http://www.ispl.jp/~oosaki/research/ tips-jcorrect/

Pardeshi, P. (2012). Compilation of Japanese Basic Verb Usage Handbook for JFL Learners: A Project Report. *Acta Linguistica Asiatica*, *2*(2), 37-63.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency, In *Language and Communication* (pp. 191–226). London: Longman.

Sketch Engine: SketchEngine. (2012). Retrieved December 10, 2012, from http://www.sketchengine.co.uk/

Srdanović-Erjavec, I., Erjavec, T., & Kilgarriff, A. (2008). A web corpus and word sketches for Japanese. *Information and Media Technologies, 3*(3), 529–551.

Yagi, Y., & Suzuki, T. (2012). Gakushūsha sakubun kōpasu no kōchiku to goyō no bunseki [Construction of learner corpus and error analysis]. In K. Nishina, M. Kamada, H. Cao, T. Utashiro & T. Muraoka (Eds.), *Nihongo gakushūshien no kōchiku: gengokyōiku kōpasu shisutemu kaihatsu [Constructing Japanese Language Learning: Language education, corpus and system development]* (3, pp. 249–274). Tokyo: Bonjinsha.

Yagi, Y., Hodošček, B., & Nishina, K. (2012, March). BCCWJ to gakushūsha sakubun kōpasu o riyōshita nihongo sakubun shien [Japanese writing assistance using the BCCWJ and a learner corpus]. In *Dai ikkai kōpasu nihongogaku wākushoppu yokōshū [Proceeding of the first workshop on Japanese corpus linguistics]*. Dai ikkai nihongo kōpasu wākushoppu [First workshop on Japanese corpus linguistics]. Tokyo.

Acknowledgments

We thank Irena Srdanović for help with a revision. We also thank the editors for their useful comments and suggestions.

JASLO: INTEGRATION OF A JAPANESE-SLOVENE BILINGUAL DICTIONARY WITH A CORPUS SEARCH SYSTEM

Kristina HMELJAK SANGAWA
University of Ljubljana
kristina.hmeljak@ff.uni-lj.si

Tomaž ERJAVEC
Jožef Stefan Institute
tomaz.erjavec@ijs.si

Abstract

The paper presents a set of integrated on-line language resources targeted at Japanese language learners, primarily those whose mother tongue is Slovene. The resources consist of the on-line Japanese-Slovene learners' dictionary jaSlo and two corpora, a 1 million word Japanese-Slovene parallel corpus and a 300 million word corpus of web pages, where each word and sentence is marked by its difficulty level; this corpus is furthermore available as a set of five distinct corpora, each one containing sentences of the particular level. The corpora are available for exploration through NoSketch Engine, the open source version of the commercial state-of-the-art corpus analysis software Sketch Engine. The dictionary is available for Web searching, and dictionary entries have direct links to examples from the corpora, thus offering a wider picture of a) possible translations in concrete contextualised examples, and b) monolingual Japanese usage examples of different difficulty levels to support language learning.

Keywords

bilingual lexicography; corpus search; parallel corpus; readability level

Izvleček

Članek predstavlja japonsko-slovenski slovar jaSlo, spletni slovar za slovensko govoreče učence japonščine, in vključitev primerov iz dveh korpusov s pomočjo odprto-kodnega korpusnega iskalnika NoSketch Engine. Korpusa sta jaSlo (milijon besed), vzporedni korpus japonskih in slovenskih besedil, ki je bil zgrajen za ta namen in vsebuje večinoma literarna, spletna in akademska besedila, ter JpWaC-L (300 milijonov besed), korpus spletnih besedil, razdeljenih v povedi, ki so rangirane po težavnostnih stopnjah. S pregledno povezavo korpusnih primerov in slovarskih iztočnic v dvojezičnem slovarju za učence japonščine kot tujega jezika, ponuja sistem uporabnikom prijazen dostop k slovarskim podatkom, tj. reprezentativnim prevodnim ustreznicam, in korpusnim podatkom, ki ponujajo a) širšo sliko možnih prevodnih ustreznic v konkretnih primerih s sobesedilom in b) enojezične primere rabe japonskih besed v povedih različnih težavnostnih stopenj, za podporo jezikovnemu učenju. Članek predlaga možne rabe tega gradiva pri učenju japonščine in se zaključi s smernicami za prihodnje delo.

Ključne besede

dvojezično slovaropisje; korpusno iskanje; vzporedni korpus; stopnja berljivosti

1. Introduction - background to the project

Bilingual dictionaries are one of the most basic tools needed by learners of foreign languages, especially at the beginning and intermediate stages of learning, when they are not yet able to use monolingual resources effectively. However, dictionary compilation is also a very labour-intensive and time-consuming enterprise, requiring considerable financial and human resources that are often not available for smaller language pairs.

The Japanese-Slovene dictionary jaSlo being compiled at the University of Ljubljana is an example of such a low-cost bilingual lexicographical project targeted at a few hundred users, which strives to make efficient use of available resources to balance its limitations stemming from the limited number of users it targets. The dictionary is moreover being compiled for a language pair without any previous lexicographical tradition, and with very little comparative linguistic research or translated texts to build upon.

The first stages of the project involved collaborative compilation, encoding conversion, enrichment with third-party resources and web deployment (Erjavec, Hmeljak Sangawa, & Srdanović, 2006).

To facilitate the editing of Japanese-Slovene dictionary entries for this under-researched language pair, a parallel corpus was compiled to complement the use of intuition and of sets of bilingual dictionaries (such as Japanese-English and English-Slovene dictionaries) when editing new entries, and to check the accuracy and validity of translations in the earlier dictionary version. At the same time, a web-derived corpus of Japanese was developed in a separate project (Srdanović, Erjavec, & Kilgarriff, 2008).

A first attempt at adding usage examples from the monolingual and the parallel corpus mentioned above was described previously (Hmeljak Sangawa, Erjavec, & Kawamura, 2009) and was followed by other interface enhancements following a usability study (Hmeljak Sangawa & Erjavec, 2010).

1.1 Corpus-based lexicography

Monolingual dictionaries have long made use of collections of attested examples of usage to select the list of lemmas to be included and to describe them, in some cases prescriptively, citing only expressions used by canonical authors, such as in the *Vocabolario dell' Accademia della Crusca* (1612) or the *Diccionario de Autoridades de la Real Academia Española* (1726-1739), in other cases descriptively, striving to cover as comprehensively as possible attested usages of words, such as in Samuel Johnson's *A Dictionary of the English Language* (1755), the *Oxford English Dictionary* (1884-1928) or Jacob and Wilhelm Grimm's *Deutsches Wörterbuch* (1854-).

With the advent of automatically searchable electronic corpora, corpus use in lexicography acquired a new dimension. Beginning with pioneering works such as the *Trésor de la langue française* (Imbs et al., 1971-1994) and the Collins Cobuild project (Sinclair, 1987), the use of electronic corpora has nowadays become standard practice in monolingual lexicography, making use of increasingly large-scale corpora to support the accuracy and increase the speed of dictionary compilation both in corpus-based and corpus-driven dictionaries (Rundell & Kilgarriff, 2011).

Some reports mention the use of monolingual corpora to support the editing of one of the two languages in a bilingual dictionary, for example to verify the naturalness of collocations or to compare the semantic prosody of both source and target language in bilingual dictionaries (Ferraresi, Bernardini, Picci, & Baroni, 2008; Srdanović, 2012; Šorli, 2012), to provide typical L2 examples in uni-directional bilingual dictionaries (Adamska-Sałaciak, 2006), or to find usage examples and verify regional variants of one of the two languages covered by the dictionary (Kilgarriff, Pomikálek, Jakubíček, & Whitelock, 2012).

The extraction of terminology from parallel corpora also has a long tradition in the field of natural language processing (Church & Gale, 1991; Wu & Xia, 1994). However, while automatic terminology extraction from parallel corpora is a well-developed area of research in the fields of machine translation and automatic language processing, it is not standard practice in the production of dictionaries for human users.

Parallel and comparable corpora have also been used by translators since before the advent of electronic corpora, to complement bilingual dictionaries. Their use has been advocated by translator trainers (Zanettin, 2002; Bernardini & Castagnoli, 2008) and translation theorists (Baker, 1995).

In lexicographic theory, the use of parallel corpora in bilingual dictionary-making was proposed almost two decades ago (Hartmann, 1994; Hartmann, 1996), and later again (Corréard, 2005; Krishnamurty, 2005), but as noted recently (Salkie, 2008), reports of bilingual dictionaries based on parallel corpora are rare.

One of the earliest reports presents some pioneering work for the compilation of a Canadian French-English dictionary, a language pair with one of the first large-scale parallel corpora (Roberts, 1996; Roberts & Cormier, 1999). Citron & Widmann (2006) report on HarperCollin's use of an in-house English-French aligned corpus of translated literature to improve existing dictionary translations in a dictionary targeted at the most demanding users. Some recent work on French-Slovene lexicography (Perko & Mezeg, 2012) compares existing dictionary entries with data from a parallel corpus, highlighting the usefulness of parallel corpus data for finding translational equivalents, predictable/unpredictable collocations and multiword discourse markers, and the limitations of such corpora stemming from their availability and size, and for their inclusion of context-bound or even wrong translations. However, bilingual lexicography in general does not seem to have made yet much systematic use of parallel corpora.

The need for the automatisation of bilingual dictionary compilation for lesser used languages where dictionary publication does not pay off the publisher's investment has recently been noted by Héja and Takács (2012), who propose a model of an automatically generated bilingual proto-dictionary and present an example of an automatically generated English-Hungarian dictionary that might be used not only by lexicographers but also by end users.

In this line of thought, our project also proposes the use of a parallel corpus to complement a bilingual dictionary, targeted both at the dictionary editors and its users.

The following sections present the latest developments of this project: a new user interface with interlinked but separate access to dictionary entries and corpus examples, an augmented parallel corpus, and a new interface to both monolingual and bilingual corpus examples. Section 3 presents possible uses of these resources for learning Japanese as a second language, and section 4 concludes with plans for further work.

2. Resources for Slovene-speaking learners of Japanese

Three types of resources are offered on the same site and interlinked for ease of use. The first component of the site is a bilingual Japanese-Slovene dictionary targeted at beginning and intermediate Slovene-speaking learners of Japanese. The other two resources, a web-derived corpus of Japanese examples of usage marked by difficulty level, and a Japanese-Slovene parallel corpus, can be accessed through a common querying system.

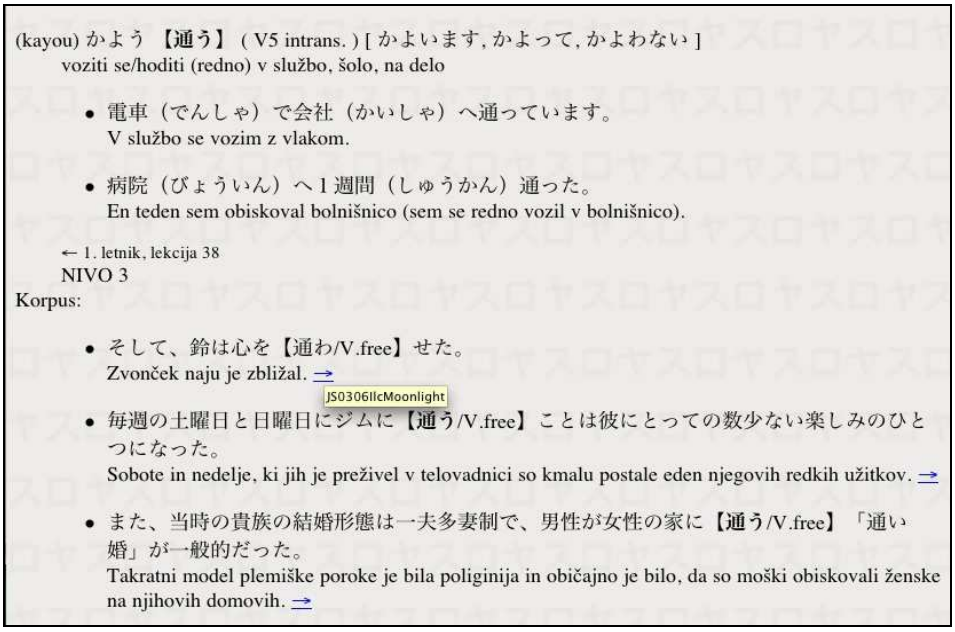
2.1 The Japanese-Slovene dictionary jaSlo

The dictionary was compiled by combining Japanese-Slovene glossaries developed at the Department of Asian and African Studies at the University of Ljubljana to be used in beginning and intermediate language courses, then checked against the complete word list of the Japanese Language Proficiency Test (JF & AIEJ, 2004) to add JLPT vocabulary not yet present in the glosses, resulting in ca. 10,000 Japanese lemmas with approximately 25,000 Slovene translational equivalents. The dictionary was then converted into a TEI-compliant XML format and released online at http://nl.ijs.si/jaslo/, as described by Erjavec, Hmeljak Sangawa, and Srdanović (2003).

The database was later revised and enlarged both manually, verifying and correcting entries, adding usage examples and missing translational equivalents, and also automatically, adding Latin alphabet transcriptions of all headwords, difficulty levels according to the JLPT vocabulary list (from level 4 - very easy, to level 1 - very difficult), and normalising part-of-speech labels, as described by Erjavec et al. (2006).

The dictionary was later further enlarged with translated examples extracted from a purpose-built Japanese-Slovene parallel corpus (Hmeljak Sangawa & Erjavec, 2008), which is described in more detail in the following section of this article. Examples were extracted for all headwords found in the corpus, obtaining new examples for 4648 of the 9891 headwords. In the case of frequent words which had tens of examples, the shortest six examples were selected, since sentence length is a robust indicator of readability.

The corpus itself had been manually validated during compilation, and we could therefore be relatively confident of the translation quality and appropriate alignment of the extracted sentences in general, but manual validation of each extracted and appended sentence was not possible due to time constraints. The corpus-extracted examples were therefore graphically separated from the rest of the entry and marked with the label *Korpus*, in order to warn users that the corpus-extracted sentences were not purposely selected or revised example sentences, but rather naturally occurring examples of usage. In such translations, the headword is not always translated with one of the translation equivalents given in the dictionary lemma itself, or even translated at all. In the corpus-extracted examples, the entry headword was highlighted by means of square brackets and bold type, and a small arrow at the end of each example provided a link to data regarding the source text. The name of the file from which the example was taken could be summoned up by mouse-over to function as an indication of text type. An example of such an entry with corpora examples can be seen in Figure 1.


(kayou) かよう 【通う】 (V5 intrans.) [かよいます, かよって, かよわない]
voziti se/hoditi (redno) v službo, šolo, na delo

- 電車（でんしゃ）で会社（かいしゃ）へ通っています。
  V službo se vozim z vlakom.
- 病院（びょういん）へ 1 週間（しゅうかん）通った。
  En teden sem obiskoval bolnišnico (sem se redno vozil v bolnišnico).

← 1. letnik, lekcija 38
NIVO 3
Korpus:

- そして、鈴は心を【通わ/V.free】せた。
  Zvonček naju je zbližal. →
  JS0306IIcMoonlight
- 毎週の土曜日と日曜日にジムに【通う/V.free】ことは彼にとっての数少ない楽しみのひとつになった。
  Sobote in nedelje, ki jih je preživel v telovadnici so kmalu postale eden njegovih redkih užitkov. →
- また、当時の貴族の結婚形態は一夫多妻制で、男性が女性の家に【通う/V.free】「通い婚」が一般的だった。
  Takratni model plemiške poroke je bila poliginija in običajno je bilo, da so moški obiskovali ženske na njihovih domovih. →

Figure 1: Example of a jaSlo dictionary entry with corpus examples in the 2009 version

The addition of examples to half of the dictionary entries had the obvious advantage of providing additional usage information and possible new translation candidates to a middle-sized dictionary, but the mechanical addition of corpus examples directly to the dictionary entries also had some drawbacks. One problem was that users might not realise that the corpus excerpts were not necessarily the most typical examples of Japanese usage nor the most central translations of the given headword. A survey of 80 headwords with automatically appended examples revealed that examples for 8% of the lemmas included useful new translational equivalents, but 2% included context dependent or unnecessarily divergent translations that might be misleading for beginning users, and as much as 8% of the examples were assigned to the wrong dictionary entry because of lemmatisation errors that could confuse inexperienced users.

We therefore decided to separate the dictionary from the parallel corpus in the new dictionary interface, and linked each dictionary entry to an automatically generated corpus query which opens in a new browser window, thus clearly separating the edited dictionary entry from the automatically generated concordances of corpus lines. This should hopefully help users differentiate between edited entries and examples (a source of information that dictionary users seldom question), and examples from authentic texts, where users are more likely to expect idiosyncratic expressions and possible deviations from conventional usage. This is similar to the approach adopted by Breen (2004), who linked a large Japanese-English dictionary with examples in a corpus of parallel Japanese-English sentences, noting that this also had the advantage of decoupling the maintenance of the dictionary file from that of the corpus.

The same format was adopted to link all dictionary entries to examples in a web-derived corpus of Japanese, created previously for a separate project (Srdanović, Erjavec, & Kilgarriff, 2008) and later split into five sub-corpora of graded difficulty, as described in section 2.3.

Figure 2 shows the same headword showed in Figure 1, but within the new interface, with links to parallel and graded corpus examples. By clicking on any of the numbers in the bottom two lines, the user has direct access to concordances of the headword in all linked corpora, described in the following two sections.

Figure 2: Example of a jaSlo dictionary entry with links to corpus examples in the 2012 version

2.2 The Japanese-Slovene parallel corpus jaSlo

After the publication of the third version of the dictionary in 2006, a parallel corpus was built from some parallel texts that had accumulated as a by-product of academic activities: student coursework (Japanese texts on society and popular culture translated into Slovene, Slovene texts on tourism translated into Japanese) and lecture handouts (texts by visiting professors from Japanese universities on the history, literature, geography and society of Japan, translated into Slovene by staff at the University of Ljubljana). The corpus was built to serve both as a source of possible translational equivalents for the dictionary compilers, and as a source of examples for dictionary users. However, since most of these texts were too difficult for beginning and intermediate learners, we also added two sets of more readable texts: excerpts of Japanese novels recently translated into Slovene, and localised pages obtained from multilingual web portals, mostly texts originally written in other languages (English, French, Russian etc.) and translated both into Japanese and into Slovene, given the lack of direct translations from Japanese to Slovene and vice-versa. The Japanese novels were digitised, while the web material was manually checked for translation quality, discarding sub-standard texts and non-corresponding parts. This first version of the corpus was composed of multilingual web pages (46.3%), revised student coursework (24.5%), literary fiction (15.7%) and translated lecture handouts (13.5%).

All texts were normalised into plain UTF-8 text files, aligned at sentence level, and the alignments manually validated. It was then lemmatised using Chasen (Matsumoto, Takaoka, & Asahara, 2007) for the Japanese part and “ToTaLe” (Erjavec et al., 2005) for the Slovene part of the corpus, obtaining a sentence-level aligned corpus of 7914 translation units, corresponding to 226,220 Japanese morphemes and 171,261 Slovene words, as described previously (Hmeljak Sangawa, Erjavec, &

Kawamura, 2009). Examples of word usage were automatically extracted from this corpus and appended directly to the corresponding dictionary entries.

An analysis of the examples extracted from this corpus for a sample of dictionary entries revealed that examples from light literature were overall the easiest and therefore the most usable as dictionary examples, when compared with examples from the other sub-corpora, especially if compared to the sub-corpus of academic prose containing particularly complex sentences with specialised vocabulary. In the second phase of corpus-building we therefore enlarged the corpus focusing mainly on literary texts. Going through the same steps as described above, we added excerpts from 14 novels of 10 Japanese contemporary authors as well as two other types of texts, mainly because of their availability in electronic format: a small collection of personal correspondence and other miscellanea translated by the first author and her colleagues, and the Japanese and Slovene translations of the New Testament. The latter amounts to more than one third of the complete corpus in size, and was added because of its availability and because the alignment could be done automatically with minimal manual validation, since all sentences are already coded using the same system in all languages into which the Bible is translated. Biblical text is admittedly not ideal reading material for beginning or intermediate learners of Japanese as a foreign language, but we included these texts into the corpus nonetheless, since the corpus interface allows for the selection (or exclusion) of texts to be included in the concordance according to their genre label, making it easy for users to exclude biblical text when they need easier examples, and allowing for its inclusion when they need as many examples as possible.

The present, 2nd version of the parallel corpus thus contains texts from the previous version, including multilingual web pages, revised student coursework, literary fiction and lecture handouts, and the newly added selection of literary fiction and the New Testament. The size of the parallel corpus and its sub-corpora is given in table 1.

Table 1: The size of the parallel corpus jaSlo and of its subcorpora

	no. of documents	**no. of Japanese tokens**	**no. of Slovene tokens**
literary	24	295,969	220,427
biblical	25	284,189	188,159
web-derived	34	98,276	59,921
coursework	28	42,607	32,796
academic	9	31,337	23,376
personal	12	10,741	7,716
Total	**132**	**763,119**	**532,395**

The corpus, encoded in TEI P5 (TEI, 2011), was then converted to a format suitable for concordancers, in particular CUWI (Erjavec, in print) based on the open source corpus workbench CWB (Christ, 1994) and the open-source system NoSketchEngine (Rychly, 2007). The corpus is made available through these two powerful concordancers on the nl.ijs.si server.

Figure 3 shows the concordance obtained via NoSketchEngine when searching for the verb *kayou* (the same as in Figure 1 and 2) in the parallel corpus jaSlo. The list on the left side shows the codes of the documents containing the word composed of an acronym indicating the direction of translation (JS for translations from Japanese to Slovene, EJS for translations from English to Japanese and Slovene, SJ for translations from Slovene to Japanese, etc.), and a word from the title or the author of the document. Clicking on these document codes brings up a window with source information including author and translator names (when known), the title of the document, its year and mode of publishing, as shown in figure 4. The second column contains the Japanese sentences containing the word, and the third column contains their translation into Slovene.

Figure 3: Example of a concordance from the parallel Japanese-Slovene corpus jaSlo

Figure 4: Display of source information for one of the documents in the corpus

For each entry in the Japanese-Slovene dictionary, a link to its concordance in the parallel corpus was added at the end of the entry (as seen in Figure 2), in order to bring the corpus examples as close to the dictionary user as possible, but without obstructing the dictionary itself.

2.3 The Japanese web corpus jpWaC-L and its difficulty-level sub-corpora

The third resource on the jaSlo site is jpWaC-L, a web corpus for learners of Japanese as a foreign language. It was derived from jpWaC, a 400 million word corpus of Japanese texts (Srdanović, Erjavec, & Kilgarriff, 2008) constructed by crawling the web using the methods proposed by Sharoff (2006) and by Baroni and Kilgarriff (2006). The jpWaC corpus is large, cleaned of text duplicates, lemmatised and part-of-speech tagged, and as such an ideal source of word usage examples.

Given its size, examples could be found for all lemmas in our dictionary, but examples for basic vocabulary were too many and in most cases too difficult for beginning learners. We therefore marked sentences in the corpus by five difficulty levels, and also made five sub-corpora of jpWaC-L, each one corresponding to one difficulty level (Hmeljak Sangawa, Erjavec, & Kawamura, 2009).

We first annotated each word in the corpus with its difficulty level according to the Japanese Language Proficiency Test specifications (JF & AIEJ, 2004), ranging from 4 (easiest words) to 1 (most difficult words), and assigned level 0 to words not appearing in the JLPT list. We then identified in the corpus well-formed and relatively simple sentences. This was achieved by the following set of heuristics, obtained empirically by repeated tests and evaluation:

1) no duplicate sentences (only one occurrence of a sentence was retained);
2) between 5 and 25 tokens in length (to exclude short fragments and long complex sentences);
3) containing less than 20% of punctuation marks an numerals;

4) containing not more than 20% words at level 0 (to avoid too much difficult vocabulary or proper names);
5) not containing words written with non-Japanese characters;
6) not containing opening or closing quotes or parentheses (to avoid errors of segmentation);
7) not beginning with punctuation (to avoid improperly segmented fragments);
8) ending in a full stop, the Japanese character kuten, 。(to include only full sentences);
9) containing at least one predicate, i.e. a verb or an adjective.

This process identified about 3 million sentences, amounting to approximately 50 million text tokens. These sentences were then further subdivided to exemplify words at each of the JLPT levels, selecting sentences which do not contain words from a more difficult level, and containing at least 10% words belonging to the targeted difficulty level. Each sentence was marked with its difficulty level, from 4 (with the easiest words) to 1 (with the most difficult words), while the easy sentences containing vocabulary outside the scope of the JLPT list were given level 0. The remaining sentences in jpWaC-L, i.e. those not appropriate for language learners are given level -1.

As mentioned, we also extracted all the sentences of the 4-0 difficulty levels and made from them separate (sub)corpora, named jpWaC-L4 to jpWaC-L0. These corpora do not contain connected text, but are suitable for looking at individual sentences of a given difficulty level - as they are much smaller than the complete jpWaC-L, complex queries take much less time.

The size of the complete corpus and of the subcorpora is given in Table 2.

Table 2: Size and composition of jpWaC-L and its 5 sub-corpora of graded difficulty level

Corpus	Size (in tokens)	%
jpWaC	409,030,315	
jpWaC_L	51,341,958	100
jpWaC_L0	43,763,041	85.24
jpWaC_L1	1,629,340	3.17
jpWaC_L2	4,608,635	8.98
jpWaC_L3	1,039,984	2.03
jpWaC_L4	300,958	0.59

This (or, rather, a very similar) corpus of sentences marked for difficulty level was made available in 2008 on the same portal as the dictionary jaSlo, but with its own search interface, separated from the dictionary search window.

In the new noSketchEngine dictionary interface, links to examples in each difficulty-level sub-corpus (if there are any) and in the complete jpWaC-L are added at the end of each entry, alongside links to the parallel corpus jaSlo, in order to facilitate access to examples during dictionary use, as can be seen in Figure 2. Since jpWaC-L contains examples of use for most dictionary headwords, most entries in the dictionary have links to jpWaC-L0 and to the sub-corpora of the same or higher difficulty level as the headword.

3. Possible uses of the resources for learners of Japanese as a foreign language

While dictionary entries provide explicit information on each headword's meaning (by means of the most typical and intuitive translations), on its morphology and syntax (by listing parts of speech and inflected verb forms) and stylistic or pragmatic restrictions on usage (by means of usage labels), corpus examples can also fulfil many functions.

First, the corpora described above can be used as a standalone resource to look up the translation(s) (in the parallel corpus) or usage (in both corpora) of words not yet included in the dictionary.

Second, they can be used to find or confirm particular aspects of word usage that are not described in detail in the dictionary entry, including additional translational equivalents, morphological forms, syntactic structures, and pragmatic, stylistic or idiomatic restrictions on word usage.

The parallel corpus jaSlo can be useful for finding translational equivalents in both directions, particularly for encoding purposes, given the present lack of a Slovene-Japanese dictionary. Moreover, translational equivalents appearing together with their context of use can help users choose the right translation both in terms of exact shade of meaning and in terms of stylistic and pragmatic appropriateness. Japanese is particularly rich in synonyms which differ mainly in terms of levels of formality and politeness, and selecting the most appropriate word among several possible candidates is always challenging for learners, who could therefore profit from corpus examples.

Pragmatic aspects of word usage are particularly difficult to describe explicitly in dictionary entries, and may be learnt more easily through exposure to a sufficient number of examples. By observing and analysing concordances for words such as the discourse marker やはり, which has no exact translational equivalent in Slovene, users can infer their pragmatic and discursive role.

Other aspects of word usage can be found in both corpora. Learners at the beginning and intermediate level often have difficulties with verb and adjective conjugation and with syntactic structures, especially if these differ from those of their translational equivalents in the learners' mother tongue, such as in the case of Japanese

adjectives expressing feelings; the adjective 寒い (*samui* "cold"), for example, can be translated by an adjective (*hladen* or *mrzel*), but also a verb (*zebsti*) or a noun (*mraz*). Example sentences at selected levels of difficulty can help users learn, confirm and reinforce such patterns of usage.

4. Conclusions and directions for further work

In the previous sections we presented three interlinked on-line resources for Slovene learners of Japanese: a Japanese-Slovene dictionary, a Japanese-Slovene parallel corpus, and a corpus of web-derived examples at different difficulty levels, and discussed their possible uses in the context of learning Japanese as a foreign language.

Plans for future work include the enhancement of both the dictionary and the parallel corpus, which are conceived as open-ended projects. The dictionary lemma list is presently based on the JLPT vocabulary list which lacks recent vocabulary, frequent loanwords and culturally-bound terms. In the next revision of the dictionary we plan to enhance jaSlo's lemma list by checking it against the new instructional vocabulary list recently created at the University of Tsukuba on the basis of a corpus of Japanese language textbooks and of a section of the Balanced Corpus of Contemporary Written Japanese (Sunakawa, Lee, & Takahara, 2012). We also plan to analyse the dictionary server's log files of unsuccessful searches to check for words users have looked up and have not found in the dictionary.

Another area in which the system could be improved is the linking of dictionary entries with corpus examples, firstly on the level of lemmatisation in the corpus, by separating more systematically examples including only a single headword from examples including the same word in a compound, phrase, or multi-word unit, and link the appropriate examples to the relative subentries.

Finally, empirical evaluations of dictionary use, including log analyses, user surveys and user observation, are also being planned in order to keep tuning the dictionary to its users.

References

Adamska-Sałaciak, A. (2006). Translation of dictionary examples - Notoriously unreliable? In E. Corino, C. Marello, & C. Onesti (Eds.), *Proceedings of the Twelfth EURALEX International Congress, Torino, Italia, September 6th - 9th, 2006* (pp. 493-501). Alessandria: Edizioni dell'Orso.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target 7*(2), 223-243.

Baroni, M. & Kilgarriff, A. (2006). Large linguistically-processed web corpora for multiple languages. In *Proceedings of the Eleventh Conference of the European Chapter of the*

Association for Computational Linguistics (pp. 87-90). Stroudsburg: Association for Computational Linguistics.

Bernardini, S. & Castagnoli, S. (2008). Corpora for translator education and translation practice. In E. Yuste-Rodrigo (Ed.), *Topics in language resources for translation and localisation* (pp. 39-55). Amsterdam / Philadelphia: Benjamins.

Breen, J. (2004). *JMdict: a Japanese-multilingual dictionary*. In G. Sérasset (Ed.), *Proceedings of the workshop on multilingual linguistic resources* (pp. 71-79). Stroudsburg, PA, USA: Association for Computational Linguistics.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *Proceedings of the Conference in Computational Lexicography, COMPLEX'94* (pp. 23–32). Budapest: Hungarian Academy of Sciences.

Church, K. & Gale, W. (1991). Identifying Word Correspondences in Parallel Texts. In P. Price (Ed.), *Proceedings, DARPA Speech and Natural Language Workshop* (pp. 152-157). San Mateo, CA: Morgan Kaufmann.

Citron, S. & Widmann, T. (2006). A bilingual corpus for lexicographers. In E. Corino, C. Marello, & C. Onesti (Eds.), *Proceedings of XII EURALEX International Congress* (pp. 251-255). Alessandria: Edizioni dell'Orso.

Corréard, M.-H.. (2006). Bilingual lexicography. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (2nd ed., vol.1, pp. 787-796). Amsterdam: Elsevier.

Erjavec, T. (in print): Vzporedni korpus SPOOK: označevanje, zapis in iskanje. In Š. Vintar (Ed.), *Slovenski prevodi skozi korpusno prizmo*. Ljubljana: Znanstvena založba Filozofske fakultete.

Erjavec, T., Hmeljak Sangawa, K., & Srdanović, I. (2003). An XML TEI encoding of a Japanese-Slovene learners' dictionary. In V. Rajkovič (Ed.), *Information Society 2003 Proceedings Volume B* (pp. 20-26). Ljubljana: Institut Jožef Stefan.

Erjavec, T., Ignat, C., Pouliquen, B., & Steinberger, R. (2005). Massive multi-lingual corpus compilation: Acquis Communautaire and ToTaLe. In *Proceedings of the 2nd Language & Technology Conference, April 21-23, 2005* (pp. 32-36). Poznań: Wydawnictwo Poznańskie.

Erjavec, T., Hmeljak Sangawa, K., & Srdanović, I. (2006). jaSlo, a Japanese-Slovene Learners' Dictionary: Methods for Dictionary Enhancement. In E. Corino, C. Marello, & C. Onesti (Eds.), *Proceedings of the Twelfth EURALEX International Congress, Torino, Italia, September 6th - 9th, 2006* (pp. 611-616). Alessandria: Edizioni dell'Orso.

Ferraresi, A., Bernardini, S., Picci, G., & Baroni, M. (2008). Web corpora for bilingual lexicography: A pilot study of English/French collocation extraction and translation. In *The International Symposium on Using Corpora in Contrastive and Translation Studies 25th -- 27th September 2008, Zhejiang University, China.* Retrieved from http://www.sis.zju.edu.cn/sis/sisht/dlwy/UCCTS2008papers/UCCTS%20Ferraresi_et_al.pdf

Geyken, A. & Lemnitzer, L. (2012). Using Google books unigrams to improve the update of large monolingual reference dictionaries. In R. V. Fjeld, & J. M. Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress* (pp. 362-366). Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.

Hartmann, R.R.K. (1994). The use of parallel text corpora in the generation of translation equivalents for bilingual lexicography. In W. Martin, et al. (Eds.), *Euralex 1994 Proceedings* (pp. 291-297). Amsterdam: Vrije Universiteit.

Hartmann, R.R.K. (1996). Contrastive textology and corpus linguistics: On the value of parallel texts. *Language Sciences 18*(3-4), 947-957.

Héja, E. & Takács, D. (2012). An online dictionary browser for automatically generated bilingual dictionaries. In R. V. Fjeld, & J. M. Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress* (pp. 468--477). Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.

Hmeljak Sangawa, K. & Erjavec, T. (2008). 学習者用日本語辞書のための対訳例文獲得 [Gakushūshayō nihongojisho no tame no taiyaku reibun kakutoku]. In *Proceedings of the Workshop on Natural Language Processing for Education, co-located with the 14th Annual Meeting of The Association for Natural Language Processing, 21 March 2008, University of Tokyo* (pp. 19-22). Tokyo: The Association for Natural Language Processing.

Hmeljak Sangawa, K., Erjavec, T., & Kawamura, Y. (2009). Automated collection of Japanese word usage examples from a parallel and a monolingual corpus. In S. Granger, & M. Paquot (Eds.), *eLexicography in the 21st century : new challenges, new applications : Proceedings of eLex 2009* (pp. 137-147). Louvain: Presses Universitaires de Louvain.

Hmeljak-Sangawa, K. & Erjavec, T. (2010). The Japanese-Slovene dictionary jaSlo: Its development, enhancement and use, *Studia Kognitywne = Études Cognitives 10*, 211-224.

Imbs, P., et al. (Eds.). (1971-1994). *Trésor de la langue française*. (16 vols.) Paris: CNRS - Gallimard.

Japan Foundation, & Association of International Education Japan. (2004). *Japanese Language Proficiency Test Content Specifications* (Revised ed.). Tokyo: Bonjinsha.

Kilgarriff, A., Pomikálek, J., Jakubíček, M., & Whitelock, P. (2012). Setting up for corpus lexicography. In: R. V. Fjeld, & J. M. Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress* (pp. 778-785) Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.

Krishnamurty, R. (2006). Corpus lexicography. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (2nd ed., Vol. 3, pp. 250-254). Amsterdam: Elsevier.

Matsumoto, Y., Takaoka, K., & Asahara, M. (2007). *Chasen - Japanese Morphological Analyzer. v. 2.4.0*. [http://chasen-legacy.sourceforge.jp/]

Perko, G. & Mezeg, A. (2012). Uporaba francosko-slovenskega vzporednega korpusa pri slovarski analizi nekaterih mejnih področij idiomatike. In M. Šorli (Ed.), *Dvojezična korpusna leksikografija. Slovenščina v kontrastu: novi izzivi, novi obeti* (pp. 12-34). Ljubljana: Trojina.

Roberts, R. (1996). Parallel-text analysis and bilingual lexicography. In *Papers presented at AILA 1996*. Retrieved from http://www.dico.uottawa.ca/articles-fr.htm

Roberts, R. & Cormier, M. (1999). *L'analyse des corpus pour l'élaboration du Dictionnaire canadien bilingue.* Retrieved from http://www.dico.uottawa.ca/articles/paris99.zip

Rundell, M. & Kilgarriff, A. (2011). Automating the creation of dictionaries: Where will it all end? In F. Meunier, et al. (Eds.), *A Taste for Corpora: In honour of Sylviane Granger* (pp. 257-281). Amsterdam: John Benjamins.

Rychlý, P. (2007). Manatee/Bonito, a modular corpus manager. In *Proceedings of 1st workshop on recent advances in Slavonic natural language processing* (pp. 65-70). Brno: Masaryk University. 65-70.

Salkie, R. (2008). How can lexicographers use a translation corpus? In *The international symposium on using corpora in contrastive and translation studies 25th -- 27th September 2008, Zhejiang University, China*.Retrieved from http://www.sis.zju.edu.cn/sis/sisht/dlwy/UCCTS2008papers/UCCTS%20Salkie.pdf

Sharoff, S. (2006). Open-source corpora: Using the net to fish for linguistic data, *International Journal of Corpus Linguistics, 11*(4), 435-462.

Sinclair, J. (Ed.). (1987). *Looking up: An Account of the Cobuild Project in Lexical Computing*. London: Collins ELT.

Srdanović, I. (2012). Dvojezična korpusna leksikografija in japonski jezik: model za izdelavo japonsko-slovenskega slovarja kolokacij. In Šorli, M. (Ed.), *Dvojezična korpusna leksikografija. Slovenščina v kontrastu: novi izzivi, novi obeti* (pp. 117-133). Ljubljana: Trojina.

Srdanović, I., Erjavec, T., & Kilgarriff, A. (2008). A web corpus and word sketches for Japanese, *Journal of Natural Language Processing* - 自然言語処理, *15*(2), 137-159.

Sunakawa, Y., Lee, J.-H., & Takahara, M. (2012). The Construction of a Database to Support the Compilation of Japanese Learners' Dictionaries, *Acta Linguistica Asiatica 2*(2), 97-115. Retrieved from http://revije.ff.uni-lj.si/ala/article/view/174

Šorli, M. (2012). Semantična prozodija v teoriji in praksi - korpusni pristop k proučevanju pragmatičnega pomena: primer slovenščine in angleščine. In M. Šorli (Ed.), *Dvojezična korpusna leksikografija. Slovenščina v kontrastu: novi izzivi, novi obeti* (pp. 90-116). Ljubljana: Trojina.

TEI Consortium. (2011). *TEI P5: Guidelines for Electronic Text Encoding and Interchange: Version 1.9.1*. Retrieved from http://www.tei-c.org/Guidelines/P5/

Wu, D. & Xia, X. (1994). Learning an English-Chinese lexicon from a parallel corpus. In *AMATA-94: Proceedings of the First Conference of the Association for Machine Translation in the Americas* (pp. 206-213). Columbia: AMT.

Zanettin, F. (2002). Corpora in translation practice. In E. Yuste-Rodrigo (Ed.), *Language resources for translation work and research - LREC workshop #8* (pp. 10-14). Retrieved from http://www.lrec-conf.org/proceedings/lrec2002/pdf/ws8.pdf