University of Ljubljana
FACULTY OF ARTS

Acta Linguistica Asiatica

Volume 2, Number 2, October 2012

Lexicography of Japanese as a Second/Foreign Language

ACTA LINGUISTICA ASIATICA
Volume 2, Number 2, October 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/
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

It is my pleasure to introduce this thematic issue dedicated to the lexicography of Japanese as a second or foreign language, the first thematic issue in Acta Linguistica Asiatica since its inception.

Japanese has an outstandingly long and rich lexicographical tradition, but there have been relatively few dictionaries of Japanese targeted at learners of Japanese as a foreign or second language until the end of the twentieth century. With the growth of Japanese language teaching and learning around the world, the rapid development of very large scale linguistic resources and language processing technologies for Japanese, a new generation of aggregated, collectively developed or crowd-sourced resources evolving in the context of the social web, a shift from static paper to constantly developing electronic resources, the spread of internet access on hand-held devices, and new approaches to the use of language reference resources stemming from these developments, dictionaries and other reference resources for learners, teachers and users of Japanese as a foreign/second language are being developed and used in new ways in different user communities. However, information about such developments often does not reach researchers, lexicographers, dictionary users and language teachers in other user communities or research spheres. This special issues wishes to contribute to the spread of such information by presenting some recent developments in this growing field.

Having received a very lively response to our call for papers, not all papers selected for publishing could fit into this issue, and part of them will be included in the December issue of ALA, which is also going to be dedicated to Japanese lexicography.

The first round of papers included in this issue presents a varied cross-section of current JFL lexicographical work and research. All papers in this issue point out the relative scarcity of appropriate reference works for learners of Japanese as a foreign language, especially when compared to lexicographical resources for Japanese native speakers, and each of the endeavours presented here confronts this lack with its own original approach. Reflecting the paradigm shift in Japanese language research, where corpus research is again playing a central role, most papers presented here take advantage of the bounty of newly available corpora and web data, most prominent among which is the Balanced Corpus of Contemporary Written Japanese developed by the National Institute for Japanese Language and Linguistics in Tokyo, and which is used by **Mogi**, **Pardeshi et al.** and **Sunakawa et al.** in their lexicographical research and projects, while **Blin** taps data for his research from the web, another increasingly important linguistic resource.

The first two papers offer two perspectives on existing Japanese dictionaries. **Tom Gally** in his paper *Kokugo Dictionaries as Tools for Learners: Problems and Potential* points out the drawbacks of currently available Japanese dictionaries from the perspective of learners of Japanese as a foreign language, but at the same time offers a very detailed and convincing explanation of the merits of monolingual Japanese dictionaries for native speakers (*kokugo* dictionaries), such as their comprehensiveness, detailedness and quantity of contextual information, when compared to bilingual dictionaries, which make them a potentially useful resource even for an audience they are not targeting - foreign language learners. His detailed explanation of possible uses and potential hurdles and pitfalls learners may encounter in using them, is not only accurate and informative, but also of immediate practical value for language teachers and lexicographers.

Toshinobu Mogi, in his paper *Towards the Lexicographic Description of the Grammatical Behaviour of Japanese Loanwords: A Case Study*, investigates the lexicographic description of loanwords in Japanese reference works and notes how information offered by currently available

dictionaries, especially regarding the grammatical aspects of loanword use, is not sufficient for learners of Japanese as a foreign language. After pointing our the deficiencies of current dictionary descriptions and noting how dictionaries sense divisions do not reflect the frequency of different senses in actual use, as reflected in a large-scale representative general corpus of Japanese, he uses a fascinatingly detailed analysis of the behaviour of a Japanese loanword verb to describe a corpus-based method of lexical description, based on the correspondence between usage forms and senses, which could be used for the compilation of Japanese learners' dictionaries meant for the reception and production of Japanese.

The second part of this special issue is composed of four reports on particular aspects of ongoing lexicographical work targeted at learners of Japanese as a foreign language.

Prashant Pardeshi, **Shingo Imai**, **Kazuyuki Kiryu**, **Sangmok Lee**, **Shiro Akasegawa** and **Yasunari Imamura** in their paper *Compilation of Japanese Basic Verb Usage Handbook for JFL Learners: A Project Report*, after pointing out - as other authors in this issue - the lack of a detailed and pedagogically sound lexicographical description of Japanese basic vocabulary for foreign learners, propose a corpus-based on-line system which incorporates insights from cognitive grammar, contrastive studies and second language acquisition research to solve this problem. They present their current implementation of such a system, which includes audio-visual material and translations into Chinese, Korean and Marathi. The system also uses natural language processing techniques to support lexicographers who need to process daunting amounts of corpus data in order to produce detailed lexical descriptions based on actual use.

The next article by **Marcella Maria Mariotti** and **Alessandro Mantelli,** *ITADICT Project and Japanese Language Learning*, focus on the learner's perspective. They present a collaborative project in which Italian learners of Japanese compiled an on-line Japanese-Italian dictionary using a purposely developed on-line dictionary editing system, under the supervision of a small group of teachers. One practical and obvious outcome of the project is a Japanese-Italian freely accessible lexical database, but the authors also highlight the pedagogical value of such an approach, which stimulates students' motivation for learning, hones their ICT skills, makes them more aware of the structure and usability of existing lexicographic and language learning resources, and helps them learn to cooperate on a shared task and exchange peer support.

The third project report by **Raoul Blin**, *Automatic Addition of Genre Information in a Japanese Dictionary*, focuses on the labelling of lexical genre, an aspect of word usage which is not satisfactorily presented in current Japanese dictionaries, despite its importance for foreign language learners when using dictionaries for production tasks. The article describes a procedure for automatic labelling of genre by means of a statistical analysis of internet-derived genre-specific corpora. The automatisation of the process simplifies its later reiteration, thus making it possible to observe lexical genre development over time.

The final paper in this issue is a report on *The Construction of a Database to Support the Compilation of Japanese Learners' Dictionaries*, by **Yuriko Sunakawa**, **Jae-ho Lee** and **Mari Takahara**. Motivated by the lack of Japanese bilingual learners' dictionaries for speakers of most languages in the world, the authors engaged in the development of a database of detailed corpus-based descriptions of the vocabulary needed by learners of Japanese from intermediate to advanced level. By freely offering online the basic data needed for bilingual dictionary compilation, they are building the basis from which editors in under-resourced language areas will be able to compile richer and more up-to-date contents even with limited human and financial resources. This project is certainly going to greatly contribute to the solution of existing problems in Japanese learners' lexicography.

Kristina Hmeljak Sangawa

RESEARCH ARTICLES

KOKUGO DICTIONARIES AS TOOLS FOR LEARNERS: PROBLEMS AND POTENTIAL

Tom GALLY
The University of Tokyo
cwpgally@mail.ecc.u-tokyo.ac.jp

Abstract

For second-language learners, monolingual dictionaries can be useful tools because they often provide more detailed explanations of meanings and more extensive vocabulary coverage than bilingual dictionaries do. While learners of English have access to many monolingual dictionaries designed specifically to meet their needs, learners of Japanese must make do with Kokugo dictionaries, that is, monolingual dictionaries intended for native Japanese speakers. This paper, after briefly describing Kokugo dictionaries in general, analyzes a typical entry from such a dictionary to illustrate the advantages and challenges of the use of Kokugo dictionaries by learners of Japanese.

Keywords

monolingual; Japanese; kokugo; dictionary; learners

Izvleček

Enojezični slovarji so lahko koristno orodje pri učenju tujega jezika, saj pogosto ponujajo bolj podrobne razlage pomenov in pokrivajo bolj obsežno besedišče kot pa dvojezični slovarji. Medtem ko imajo učenci angleščine kot tujega jezika na voljo veliko enojezičnih slovarjev, ki so bili izdelani prav za njihove potrebe, pa morajo učenci japonščine uporabljati enojezične slovarje japonščine, imenovane Kokugo, t.j. slovarje, ki so namenjeni govorcem japonščine kot maternega jezika. Pričujoči članek - po kratki splošni predstavitvi slovarjev Kokugo - skozi analizo slovarskega članka iz takega slovarja oriše prednosti in izzive rabe slovarjev Kokugo za učenje japonščine kot tujega jezika.

Ključne besede

enojezični; japonščina; kokugo; slovar; učenci

1. Introduction

In the past few decades, learners of English as a second language have benefited from the publication and rapid development of many monolingual dictionaries designed specifically to meet their needs. These dictionaries, which include *Oxford Advanced Learner's Dictionary, Collins COBUILD Advanced Learner's English Dictionary*, and similar volumes from Cambridge, Longman, Macmillan, and Merriam-Webster, have incorporated many learner-friendly features, including a controlled defining vocabulary, greater attention to collocations and idioms both as headwords and in definitions and examples, extensive use of corpora for meaning explication and example selection, and new macro- and microstructure designs. (For more information on these dictionaries, see Cowie, 1999, and Béjoint, 2010, pp. 163–200.) The rapid innovations in these dictionaries have been driven not only by advances in lexicography and corpus linguistics but also by the huge global market for English-learning materials, making learner's dictionaries, despite the large investment necessary for their creation, a potentially lucrative source of income for publishers.

Learners of Japanese, however, have not been nearly as fortunate, as there are no monolingual dictionaries of Japanese currently available that meet the needs of intermediate and advanced learners.[1] Learners fluent in English, Chinese, or Korean, which have reasonably good bilingual dictionaries with Japanese, might not suffer significantly from this lack, but speakers of most other languages are at a severe disadvantage when trying to learn Japanese. Furthermore, at least in the case of English, most of the bilingual dictionaries used by learners of Japanese were in fact written for native Japanese speakers and thus lack many features needed by second-language learners, including explanatory definitions for difficult-to-translate headwords, verb-conjugation categories and other grammatical information, and usage notes. Perhaps the greatest drawback of bilingual dictionaries published for fluent Japanese speakers, when considered from the learner's perspective, is the omission of headwords that Japanese users are not likely to seek when using a bilingual dictionary into another language, including slang, dialect, archaisms, variants, and proper names.

Some of these drawbacks of bilingual dictionaries of Japanese can be overcome through the use of a type of dictionary often overlooked in Japanese-language education: monolingual dictionaries of Japanese aimed at native speakers of the

[1] The monolingual *Dictionary of Basic Japanese Usage for Foreigners* was published by Japan's Agency for Cultural Affairs in 1971. This dictionary incorporated features useful for learners, including explanatory definitions written in relatively simple language, many example sentences, and full conjugation information for verbs and adjectives. Although a second edition appeared in 1975 and a third with about 4,500 headwords in 1990, the dictionary is no longer in print, let alone available in digital form. Two companion volumes, listed in library catalogs but not consulted for this study, were *Dictionary of Chinese Characters for Foreigners* [外国人のための漢字辞典 *Gaikokujin no Tame no Kanji Jiten*] (1966) and *A Specialized Scientific Dictionary for the Foreigners: Physical Science* [外国人のための専門用語辞典〈自然科学系〉 *Gaikokujin no Tame no Senmon Yōgo Jiten: Shizen Kagaku Kei*] (1966), are also out of print.

language. Usually called *kokugo jisho* [国語辞書] or *kokugo jiten* [国語辞典], these dictionaries are readily available in both paper and digital versions from commercial Japanese publishers, and they have many advantages over bilingual dictionaries: their definitions are often explanations of the headword's meaning, rather than mere synonyms; they indicate conjugation categories of verbs; and, while their inclusion of slang and other nonstandard language is sometimes limited, they do contain a wider range of vocabulary than most bilingual dictionaries. Because these dictionaries were written for native speakers of Japanese, however, they present significant hurdles to learners, particularly in the comprehensibility of their definitions and examples. This paper therefore examines the typical features of these monolingual Japanese dictionaries—called Kokugo dictionaries here—and discusses the advantages and disadvantages of those features for people learning Japanese as a second language.

2. Contemporary Kokugo Dictionaries

A wide range of Kokugo dictionaries are currently available for native speakers, from the 14-volume *Nihon Kokugo Daijiten,* a comprehensive historical dictionary of the language from the earliest recorded times to the present, to small, inexpensive dictionaries sold in 100-yen stores that are intended mainly to provide the meanings and orthography of "hard" or often misunderstood words.[2] Although dictionaries all along this spectrum can be used profitably by learners, this paper will concentrate on two categories of Kokugo dictionaries that are likely to be most useful: midsized dictionaries that focus on contemporary general vocabulary, and comprehensive dictionaries that also include historical vocabulary and encyclopedic entries.

Among the many midsized dictionaries aimed at general users are *Shinmeikai Kokugo Jiten, Iwanami Kokugo Jiten, Sanseidō Kokugo Jiten,* and *Meikyō Kokugo Jiten.* These dictionaries typically claim to have about 70,000 entries, and their paper editions have between about 1400 and 1900 pages, including front and back matter. Their headwords, senses, and examples primarily reflect the modern Japanese language, and they contain few encyclopedic entries. (The second edition of *Meikyō Kokugo Jiten,* for example, contains brief entries for *Nihon* "Japan" and *Chūgoku* "Chugoku region; China" but none for *Tōkyō* or *Amerika.*)

One-volume comprehensive dictionaries are printed in a larger format and contain more pages, usually around 3000, and claim to have around 230,000 entries. Three comprehensive dictionaries that, as of 2012, have been updated recently are *Kōjien, Daijirin,* and *Daijisen.* (Many similar dictionaries have been published in the past century, but most are no longer being updated.) In addition to the contemporary

[2] Okimori, Kurashima, Katō, & Makino (1996) contains a full list and descriptions of Kokugo and other dictionaries published in Japan up through the mid-1990s. Some of the many books in Japanese about the history and characteristics of Kokugo dictionaries are Kurashima (1997), Ishiyama (2007), and Kurashima (2010).

vocabulary covered by the midsized dictionaries, the comprehensive dictionaries also contain archaic headwords and senses, and citations are often taken from classical or canonical literary works. They also contain many proper names and technical words that are missing from the midsized dictionaries. Perhaps the most important difference among these dictionaries for learners is the sense order: while *Kōjien* orders the multiple senses of a headword with the earliest or most basic meanings first, *Daijirin* and *Daijisen* give the most common contemporary meanings first.

As of 2012, all of the dictionaries named above are available in paper form. Most are also available in digital formats, which might include cd- and/or dvd-roms, portable electronic dictionaries, free and/or subscription-based Web sites, and smartphone, tablet, and personal computer applications. Data on dictionary sales in Japan are held closely by publishers, but anecdotal evidence, including observations of the dictionaries used by university students and the space allocated to paper dictionaries in bookstores, suggests that the era of paper dictionaries is coming to an end. While digital versions do offer some distinct advantages to students, including faster lookup times, intra- and inter-dictionary links, and, on some devices, handwritten input, the actual content of digital Kokugo dictionaries is so far largely identical to that of their paper versions. For this reason, and because the rapid progress of digital and network technology makes it difficult to predict how Kokugo dictionaries might be delivered to users in coming years, this paper will focus only on the content of dictionary entries, not their medium of presentation.

3. A Typical Entry in a Midsized Dictionary

To see the advantages and challenges of the use of Kokugo dictionaries by learners of Japanese, let us examine in detail an entry for a word that an intermediate or advanced learner might want to look up in a dictionary: the verb *satoru*. This word was chosen because one of its two main senses is used in general contexts in the contemporary language while the other is limited to a particular cultural domain. The entry for *satoru* from the second print edition of the midsized *Meikyō Kokugo Jiten* (2010) appears below. This is followed by a detailed explanation of the entry's components and the implications of each component for a learner accessing the entry. In the explanations, the romanization of each component is given in italics for the convenience of readers.

さと・る【悟る（゛覚る）】〖他五〗❶ものの本質や意味などを（直感的に）はっきりと理解する。また、隠されていたことなどをはっきりと認識する。「車の重大さを—」「敵に—・られないように注意せよ」❷仏教で、心の迷いを去って永遠の真理を会得する。悟りを開く。「仏法の真理を—」[可能] 悟れる [名] 悟り

3.1 Headword

The headword is listed in kana order based on the pronunciation of its unmarked imperfective form, さとる *satoru*, not by its usual orthographic representations (悟る or, less commonly, 覚る). Thus the preceding word in the paper dictionary is さとり *satori* and the following word is サドル *sadoru.* For learners using paper Kokugo dictionaries, this pronunciation-based listing can be frustrating, as often a word one wishes to look up appears in a text at least partly in kanji, rather than entirely in kana; if one does not know the reading of the kanji, one cannot find it easily in a Kokugo dictionary.[3] This problem is usually alleviated with electronic dictionaries, which, depending on the hardware and software, allow kanji-containing words to be looked up using cut-and-paste, stylus or finger input, optical character recognition, or selection of kanji components (multiradical lookup).

In *Meikyō,* the boundary between the verb stem さと *sato-* and suffix る *-ru* is indicated by a *nakaguro,* or black dot (・); the same symbol is used in this dictionary to separate the stems and suffixes of adjectives. *Meikyō* also uses a hyphen (–) to separate the parts of compound words; the word サドンデス *sadondesu* "sudden death", for example, appears as a headword as サドン–デス. Neither the black dot nor the hyphen would appear in those words in a regular text. These markers, which are normally omitted from bilingual dictionaries, can provide useful clues to learners about the morphemic structure and etymology of headwords.

One of the challenges for learners using most dictionaries of Japanese, including all currently available Kokugo dictionaries, is that words can be looked up only by their canonical, unmarked form. If a reader encounters a conjugated form of the verb *satoru,* such as the potential *satoreru* or the negative passive participle *satorerarenakute,* and wants to find the meaning of the word in a dictionary, he or she must be able to deduce that the plain imperfective form is *satoru.* A fairly high level of grammatical knowledge is therefore necessary before a learner can use such dictionaries effectively.[4]

3.2 Orthography

Because this verb can be written not only in kana but also with kanji, the two usual kanji representations follow the headword in brackets: **悟る（〝覚る）**. The lack of any marking or further bracketing of the first version, **悟る**, indicates that this is a standard

[3] Some printed Kokugo dictionaries, including *Iwanami Kokugo Jiten* and *Daijirin,* have indexes of kanji and "hard-to-read" kanji combinations (*jukugo*), but those indexes exclude many word forms that learners would need to look up. A compete kanji and kanji-compound index to the second edition of the comprehensive dictionary *Daijirin* was published in 1997 as a separate volume (*Kanjibiki Gyakuhiki Daijirin*), but its bulk makes it unwieldy for casual use.

[4] An exception is Jim Breen's WWWJDIC, a free online Japanese-English dictionary. Searches for most conjugated or declined forms of words lead to the standard headword forms.

written form of the verb. The second version, 覚る, is both enclosed in curved parentheses, indicating that it is a nonstandard form, and marked with the symbol ▷, indicating that, while the kanji 覚 appears on the *Jōyō Kanji* (常用漢字) list of characters designated by the government for everyday use, *sato-* is not an officially designated reading for that character. Other symbols are used in this dictionary to indicate when kanji do not appear on the *Jōyō* list at all, when a reading is in an annex to the *Jōyō Kanji* list, and when a combination of characters has a special reading.

This detailed information about the status of different written forms of words can be useful to learners for at least two reasons. When a reader learns from a dictionary that the written form of a word he or she has encountered in a text is nonstandard, the reader can often infer something about the text's provenance: it might predate the government's postwar orthographic standards, it might not have been subjected to the rigorous editing applied to newspapers and some other publications, or it might reflect the author's individual preferences or literary sensibility. The orthographic labeling also helps the learner decide what form to use when writing in Japanese; a person composing a university report or a job application letter, for example, might decide to use the standard forms even if he or she prefers the nonstandard forms.

3.3 Part-of-Speech Information

The next item in the entry, 他五, consists of two abbreviations of verbal categories. The character 他 indicates that the headword is a transitive verb (他動詞 *tadōshi*), while the character 五 shows that it follows the *godan* (五段) conjugation pattern. For other headwords, this information might be 名, for 名詞 (*meishi*, "noun"); 形, for 形容詞 (*keiyōshi*, "adjective"); 形動, for 形容動詞 (*keiyō dōshi*, "adjectival verb"); 代, for 代名詞 (*daimeishi* , "pronoun"); etc.

This grammatical information, especially about verb categories, is usually omitted from bilingual dictionaries aimed at native speakers of Japanese. Learners opening a Kokugo dictionary for the first time, however, are likely to be confused by them, as the abbreviations might refer to grammatical categories that the learners know by very different names. *Godan* conjugation verbs, for example, are often called "consonant-stem" verbs in textbooks of Japanese written in English, and understanding the term *godan* and similar expressions requires familiarity with Japanese grammar as it has been taught in Japanese schools. Kokugo dictionaries also often indicate the categories of verbs for the literary language (文語 *bungo*), which many students do not need to learn. In order to get the most out of this section of Kokugo dictionary entries, therefore, students would have to make a conscious effort to learn the abbreviations and their meanings in the context of traditional Japanese school grammar.

3.4 Definitions

This entry for *satoru* has two senses, marked with the numbers ❶ and ❷. Within each sense is a definition followed by an example or two.

The definition of the first sense is ものの本質や意味などを（直感的に）はっきりと理解する。また、隠されていたことなどをはっきりと認識する。 Mono no honshitsu ya imi nado o (chokkanteki ni) hakkiri to rikai suru. Mata, kakusarete ita koto nado o hakkiri to ninshiki suru. This might be translated as "To understand clearly (intuitively) the essence, meaning, etc. of something. Or, to recognize clearly something that is hidden."

The definition of the second sense is 仏教で、心の迷いを去って永遠の真理を会得する。悟りを開く。 *Bukkyō de, kokoro no mayoi o satte eien no shinri o etoku suru. Satori o hiraku*, which might be translated as "In Buddhism, to cast away confusions of the soul and to obtain eternal truth; to achieve *satori* [enlightenment]".

It is here, in the definitions, that Kokugo dictionaries offer the greatest potential for learners of Japanese but also present the greatest challenges. To see why, compare the above translation of the first sense with the corresponding definitions in three Japanese-English dictionaries (sense specifiers in Japanese have been omitted):

> perceive ((that)); realize ((that)) (Shogakukan Progressive Japanese-English Dictionary, 4th ed.)

> 1 see; notice; perceive; discern; guess; sense; wake to…; be alive to…; be aware of…; get wind of…
> 2 understand; comprehend; apprehend; realize. (*Kenkyusha's New Japanese-English Dictionary*, 5th ed.)

> (1) to perceive; to sense; to discern; (2) to understand; to comprehend; to realize (*Jim Breen's WWWJDIC*)

While the bilingual dictionaries offer only short glosses, the *Meikyō* definition gives a full explanation, incorporating semantic elements that would be difficult to discern from the English translations. These include *honshitsu ya imi nado* "essence, meaning, etc.", as examples of things that might be the objects of the verb, and *chokkanteki ni* "intuitively" and *hakkiri to* "clearly", as modifiers of the verb *rikai suru* "to understand". The second half of the first definition can also aid the learner's understanding by indicating that the object of the verb might be something that has been hidden.

For the second sense, where bilingual dictionaries give some variation on "be spiritually awakened; attain enlightenment" (*Progressive*), the Kokugo dictionary provides helpful additional information, particularly the explicit reference to Buddhism and the explanation *kokoro no mayoi o satte* "cast away confusions of the soul". A reader unfamiliar with Buddhism and seeing the word *satoru* used in the Buddhist

sense for the first time is likely to understand the concept much better after reading the *Meikyō* definition than the English glosses.

Most definitions in Kokugo dictionaries are similarly explanatory, and as such they should be more useful to learners than mere glosses. The problem, of course, is that learners must be able to read and understand the explanations. Because the dictionaries are written for native speakers of Japanese of normal educational attainment, the definitions assume that users have a fairly wide vocabulary of Japanese and know at least the *Jōyō Kanji*. While the first definition of *satoru* might be understandable to higher-intermediate learners, the second definition, in particular the transitive use of the verb 去る *saru* "to get rid of something undesirable", will be more difficult to grasp. Most learners would have to look up at least several words contained in the definition, a time-consuming task that, while not uneducational in its own right, would be distracting if the leaners' immediate purpose is to understand what *satoru* means in a particular text. It is this issue—the difficulty in understanding definitions—that presents the greatest hurdle to the effective use of Kokugo dictionaries by learners of Japanese.

3.5 Examples

The entry for *satoru* contains three short examples. For the first sense, the examples are 車の重大さを— *kuruma no jūdaisa o* [*satoru*] "realize the importance of automobiles" and 敵に—・られないように注意せよ *Teki ni* [*sato*]*rarenai yō ni chūi se yo* "Be careful not to be noticed by the enemy." For the second sense, the example is 仏法の真理を— *Buppō no shinri o* [*satoru*] "realize the truth of Buddhism". (Like many Kokugo dictionaries, *Meikyō* replaces the headword or headword stem with a dash (—) in examples, presumably to save space.) While brief, these examples can provide useful information to learners: the use of *satoru* with the object particle *o* in the first and third examples reinforces the fact that this is a transitive verb; the use of *jūdaisa* "importance" and *shinri* "truth" as the verb's object shows that it often takes an abstract noun as an object; the second example shows that, when used in the passive, the verb can take a personal noun as its subject. All of these insights, of course, assume that the reader knows the readings and understands the meanings of the other words in the examples; if the examples contain unknown vocabulary, it can be a difficult, time-consuming task to figure out the meanings of examples, especially with paper dictionaries.

As with most examples in midsized Kokugo dictionaries, the origins of these phrases and sentences are not indicated. In general, such dictionaries seem to use a combination of verbatim corpus examples, modified corpus examples, and examples invented by the lexicographers to illustrate meanings and show typical collocations.[5]

[5] The examples (as well as definitions) in one widely used Kokugo dictionary, *Shinmeikai Kokugo Jiten*, have been the target of criticism, praise, and affection because of their specificity, opinionatedness, and

Comprehensive dictionaries also contain many examples with citations, usually from literary, often classical, sources. Of the six examples for three senses of *satoru* in the sixth edition of *Kōjien* (2008), for example, two are from the 11th-century *Tale of Genji* and one from the 13th-century *Tale of the Heike*. The third edition of *Daijirin* (2006) has an example from 20th-century literature, the 1906 novel *Kusamakura* (English title: *The Three-Cornered World*) by Sōseki Natsume. This heavy use of examples from classical and canonical literature—a result of the dictionaries' historical association with the scholastic field *kokugoka* (国語科), that is, the study of Japan's national language and literature—contrasts sharply with the examples in recent English-English learners' dictionaries, which rely largely on citations taken or adapted from corpora that cover a wide variety of contemporary spoken and written sources.

As a first impression, the examples in midsized Kokugo dictionaries are likely to seem more accessible to learners than those in the larger comprehensive dictionaries. Often, however, mixed in with the comprehensive dictionaries' literary examples are contemporary phrasal examples similar to those in midsized dictionaries; if learners know how to distinguish the contemporary from the classical examples, they can obtain as much benefit from the examples in the comprehensive dictionaries as from those in the midsized dictionaries.

3.6 Other Information

The entry for *satoru* in *Meikyō* ends with two derived forms of the headword, 悟れる *satoreru* and 悟り *satori*; the abbreviation 可能 *kanō* indicates that the former is the potential form of the verb and the abbreviation 名 *mei* that the latter is the nominal form. Other information that might be included in entries in this and other Kokugo dictionaries include the historical kana spelling of the headword, an abbreviation indicating the headword's pitch accent pattern,[6] etymologies, grammar and usage notes, and information about synonyms and antonyms. In addition to the entries themselves, many dictionaries also contain explanatory columns on meaning and usage, illustrations, appendices on various topics, and other supplements designed to give "added value" to the dictionaries and make them more attractive to consumers.

---

occasional humor (see, for example, Nishiyama et al., 1992, and Akasegawa, 1996). The eccentricity of *Shinmeikai*'s examples is probably a disadvantage for learners, but its examples have advantages as well, particularly their larger number compared with other dictionaries of the same size and the inclusion of brief glosses to explain the meaning of idiomatic expressions. For example, the entry for the noun *sanaka* contains the examples 夏の—〔＝暑い盛り〕 *natsu no* [*sanaka*] (= *atsui sakari*) 'the *sanaka* of summer (= the hottest period)' and 冬の—〔＝最も寒い時〕 *fuyu no* [*sanaka*] (= *mottomo samui toki*) 'the *sanaka* of winter (= the coldest time)'. The brief explanations, though intended for native speakers, can be useful for learners as well. No other Kokugo dictionary offers such extensive glossing of examples.

[6] Of the dictionaries mentioned in this paper, only *Shinmeikai* and *Daijirin* give accent information. When asked by the author about the widespread omission of this important element of Japanese pronunciation from Japanese dictionaries, several Japanese lexicographers have pointed out that pitch accent in Japanese varies widely by dialect and that native Japanese speakers rarely seek that information in dictionaries.

4. An Imperfect Yet Still-Valuable Tool

Compared with English-English learners' dictionaries, which incorporate a wide range of useful innovations made possible by recent advances in corpus linguistics and lexicographic theory and practice, Kokugo dictionaries are not nearly as useful tools for students of Japanese. The vocabulary of definitions and examples is not controlled sufficiently for learners, the grammatical information is inadequate for people who have not yet mastered Japanese grammar, and the word-lookup system, especially with paper versions, can be frustrating for students still learning kanji. These inadequacies are, of course, due to the dictionaries' omission of features that are usually not needed by native Japanese speakers.

Nevertheless, Kokugo dictionaries do offer some advantages over bilingual dictionaries. Their explanatory definitions can often be enlightening to a learner who has been unable to figure out what a word means based on brief bilingual glosses. The information on orthography, grammar, and usage is more detailed than that in many bilingual dictionaries, especially those aimed at native Japanese speakers. The larger comprehensive dictionaries define many archaic, slang, dialect, and technical words that might be difficult to find elsewhere, and their compact explanations of Japanese people, places, and things can greatly assist readers trying to learn not only the Japanese language but also the history and culture of Japan. Perhaps most importantly, like all monolingual dictionaries, Kokugo dictionaries offer an immersive experience in the language: beginning students, who might struggle even to find headwords listed in the yet-to-be-mastered kana order, are at least exposed to the Japanese writing system while trying to use the dictionaries, while more advanced learners will have a chance to guess at the meanings of words that they are likely to encounter again in the future. Combined with the explanatory definitions, this immersion also helps to protect learners from one of the greatest dangers of bilingual dictionary use: assuming that a word given as the translation of a headword in a dictionary can be used to translate that same word in any context.

Of course, Japanese learners would benefit even more from learners' dictionaries similar to those available for English. Unfortunately, however, no such dictionaries exist today, and the author has not heard of any under preparation. The reason is not hard to guess: the number of learners of Japanese is tiny compared to that of English, and commercial publishers, already suffering from a drop-off in sales of their conventional dictionaries, cannot make the huge, long-term investment necessary to design and produce an entirely new type of dictionary. Until the Japanese government, or perhaps a private foundation, decides to support such a project in the future, it seems likely that Japanese learners who want the advantages of a monolingual dictionary will continue to need to use Kokugo dictionaries, imperfect as they may be.

References

Books

Akasegawa, G. [赤瀬川原平]. *Shinkai-san no nazo* [新解さんの謎]. Tōkyō: Bungeishunjū [文藝春秋].

Béjoint, H. (2010). *The lexicography of English.* Oxford, UK: Oxford University Press.

Cowie, A. P. (1999). *English dictionaries for foreign learners: A history.* Oxford, UK: Oxford University Press.

Ishiyama, M. [石山茂利夫]. (2007). *Kokugo jisho: Dare mo shiranai shusshō no himitsu* [国語辞書 誰も知らない出生の秘密]. Tōkyō: Sōshisha [草思社].

Kurashima, N. [倉島長正]. (1997). *'Kokugo' to 'kokugo jiten' no jidai, jō: Sono rekishi* [「国語」と「国語辞典」の時代・上 —その歴史—]. Tōkyō: Shōgakukan [小学館].

Kurashima, N. [倉島長正]. (2010). *Kokugo jisho 100-nen: Nihongo o tsukamaeyō to kutō shita hitobito no monogatari* [国語辞書 100 年 日本語をつかめようと苦闘した人々の物語. Tōkyō: Ohfu [おうふう].

Nishiyama, S., & QQQ no Kai [西山里美と QQQ の会]. (1992). Jisho ga konna ni omoshirokute ii kashira: Sanseidō "Shimeikai kokugo jiten" shukan ni ateta santsū no tegami [辞書がこんなに面白くていいかしら 三省堂『新明解国語辞典』主幹に宛てた三通の手紙]. Tōkyō: JICC.

Okimori, T. [沖森卓也], Kurashima, T. [倉島節尚], Katō, T. [加藤知己], & Makino, T. [牧野武則]. (1996). *Nihon jisho jiten* [日本辞書辞典]. Tōkyō: Ohfu [おうふう].

Dictionaries

Daijirin [大辞林] (3rd ed.). (2006). Tōkyō: Sanseidō [三省堂].

Daijisen [大辞泉] (1st ed., revised). (1998). Tōkyō: Shōgakukan [小学館].

Dictionary of basic Japanese usage for foreigners [外国人のための基本語用例辞典 Gaikokujin no tame no kihongo yōrei jiten] (3rd ed.). (1990). Tōkyō: Printing Bureau, Ministry of Finance [大蔵省印刷局 Ōkurashō Insatsukyoku].

Iwanami kokugo jiten [岩波国語辞典] (7th ed.). (2011). Tōkyō: Iwanami Shoten [岩波書店].

Jim Breen's WWWJDIC. Retrieved from http://www.csse.monash.edu.au/~jwb/cgi-bin/wwwjdic.cgi?1C

Kanjibiki gyakuhiki daijirin [漢字引き・逆引き大辞林]. (1997). Tōkyō: Sanseidō [三省堂].

Kenkyusha's new Japanese-English dictionary [研究社新和英大辞典 Kenkyūsha shin Wa-Ei daijiten] (5th ed.). (2003). Tōkyō: Kenkyūsha [研究社].

Kōjien [広辞苑] (6th ed.). (2008). Tōkyō: Iwanami Shoten [岩波書店].

Meikyō kokugo jiten [明鏡国語辞典] (2nd ed.). (2010). Tōkyō: Taishūkan [大修館].

Nihon kokugo daijiten [日本国語大辞典] (2nd ed.). (2001). Tōkyō: Shōgakukan [小学館].

Sanseidō kokugo jiten [三省堂国語辞典] (6th ed.). (2007). Tōkyō: Sanseidō [三省堂].

Shogakukan progressive Japanese-English dictionary [小学館プログレッシブ和英中辞典 Shōgakukan puroguresshibu Wa-Ei chūjiten] (4th ed.). (2011). Tōkyō: Shōgakukan [小学館].

Shinmeikai kokugo jiten [新明解国語辞典] (7th ed.). (2011). Tōkyō: Sanseidō [三省堂].

TOWARDS THE LEXICOGRAPHIC DESCRIPTION OF THE GRAMMATICAL BEHAVIOUR OF JAPANESE LOANWORDS: A CASE STUDY

Toshinobu MOGI
Naruto University of Education
tmogi@naruto-u.ac.jp

Abstract

The present papers offers a case study of a Japanese loanword verb, with the aim of contributing to corpus-based research on Japanese loanwords and of providing a foundation for the compilation of a dictionary of grammatical patterns of loanwords for learners of Japanese as a foreign language. The case study presents an analysis of actual usage of loanword *suru*-verbs in the large-scale Balanced Corpus of Contemporary Written Japanese, which is followed by a detailed analysis of all examples of the polysemous verb *katto-suru*. It is thereby shown how corpora can help in describing loanwords by matching a word's meaning with its patterns of usage, and how such a description can be useful to learners of Japanese as a foreign language.

Keywords

loanwords; gairaigo; corpus linguistics; sentence pattern; Japanese language teaching

Izvleček

Pričujoči članek predstavlja študijo primera japonskega glagola tujega izvora, z namenom prispevati h korpusno-osnovanim raziskavam japonskih tujk in obenem predlagati osnovne smernice za sestavo slovarja slovničnih vzorcev japonskih tujk za učence japonščine kot tujega jezika. Predstavljena je analiza rabe samostalniških glagolov na *-suru* v velikem uravnoteženem korpusu sodobne pisne japonščine (BCCWJ), čemur sledi podrobna analiza vseh primerov večpomenskega glagola *katto-suru*. S tem je prikazano, kako lahko uporaba korpusa pripomore k opisu tujk z vzporejanjem podpomenov in vzorcev rabe posameznih besed ter kako lahko tak opis koristi učencem japonščine kot tujega jezika.

Ključne besede

tujke; gairaigo; korpusno jezikoslovje; stavčni vzorec; učenje japonščine

1. Introduction

The case study presented in this paper is a corpus-based contribution to Japanese loanword research which aims at preparing the conceptual framework for the compilation of a dictionary of grammatical patterns for learners of Japanese.

Section 2 reviews problems raised in previous linguistic and pedagogical research on loanwords, highlighting the need for research on Japanese loanwords which takes into account their grammatical behaviour. The following sections present a case study of nominal verbs (*suru* verbs) of foreign origin. Section 3 gives an account of the use of nominal verbs of foreign origin in a very large corpus of contemporary written Japanese, while section 4 offers an analysis of a polysemous verb, *katto-suru*, based on corpus examples. We thereby show how a corpus-based detailed description of a word, aligning the word's meanings with the word's syntactical behaviour, produces a description which is useful for learners of Japanese as a foreign language.

2. On the need for research on the grammar of loanwords

In Japanese lexicology, Japanese words are traditionally categorised according to their origin (*goshu* 語種 in Japanese), i.e. according to how the word came to be used in Japanese. The lexicon of contemporary Japanese could be broadly divided into *native words* (*koyūgo* 固有語) which existed originally in Japanese, and *borrowed words* (*shakuyōgo* 借用語) which were borrowed from other languages, but it is traditionally categorised further, making a three-fold distinction between domestic words (*wago* 和語), words borrowed from Chinese (*kango* 漢語) and words borrowed from other (mainly European) languages (*gairaigo* 外来語). The distinction can be summarised as follows:

(1) Categories of Japanese words according to their origin (*goshu* 語種):
a. domestic words: *wago* 和語
b. borrowed words: *kango* 漢語 (of Chinese origin),
gairaigo 外来語 (from other languages)

This paper is only concerned with words of foreign origin borrowed from languages other than Chinese (*gairaigo*). Hereafter, the term *loanword* shall only be used in this restricted meaning, as a term for words of foreign, but not Chinese origin. If compared with research on domestic words and words of Chinese origin, research on loanwords (*gairaigo*) is lagging far behind, as has been noticed in previous research (e.g. Ishino 1996, Kim 2011). There are only very few analyses of the meaning of loanwords and the difference in usage with respect to domestic words and words of Chinese origin, while the grammatical behaviour of loanwords - how loanwords are used within sentences - has hardly been studied at all.

On the other hand, research on Japanese language teaching (e.g. Sawada, 1993; Nakayama et al., 2008) has often stressed that loanwords (alternatively termed also

katakana-words) pose considerable difficulties to foreign learners of Japanese and that their teaching and explanation has not been adequately addressed. In fact, loanwords which appear in textbooks of Japanese as a foreign language are usually only introduced with an example and a one-word gloss (a synonym) to indicate meaning, while their grammatical behaviour is generally not discussed at all.

Loanwords share some general grammatical properties with words borrowed from Chinese, namely, that they can be used as nouns (in their most basic form), or as verbs (if the light-verb *suru* is appended), or as adjectives (with the addition of the copula *da* or of the particle *na*). However, these rules cannot be applied to all loanwords, as can be seen in (2), and learners need to check and memorise the grammatical and syntactical properties of each word individually. For each loanword, they need to check whether it can be used as a verb, whether such a verb is transitive or intransitive and in what syntactical pattern it can be used, etc. Consider the loanwords listed in (2): while words in list (2) a. can be used as verbs with the addition of *suru*, words in list (2) b. cannot.

(2) a. *adobaisu-suru* (“to advise”), *imeeji-suru* (“to imagine”), *katto-suru* (“to cut”), *rirakkusu-suru* (“to relax”);

b. **kureemu-suru* (*“to do claim/complaint”), **doriimu-suru* (*“to do dream”), **shotto-suru* (*“to do shot”), **panikku-suru* (*“to do panic”)

However, in existing dictionaries of verb patterns such as Koizumi et al. (1989) or collocation dictionaries such as Himeno (2004), which are used by teachers and learners of Japanese as a foreign language, loanword verbs such as those given in (2) are not included. Descriptive research on loanword verbs, adjectives and nouns, especially research examining their syntactic behaviour, is therefore needed both from a general linguistic point of view, in order to have a better description of this part of the Japanese lexicon, as well as from an applied point of view, to obtain basic data from which applied linguistic research such as lexicography or Japanese language teaching could greatly profit.

The final aim of such research on the grammatical behaviour of loanwords is to accumulate accurate descriptions of individual loanwords, which are eventually to be edited into a dictionary of loanword grammatical patterns. The present paper makes a first step in this direction by presenting a case study of the verb *katto-suru,* analysing its semantic and grammatical characteristics which can be extracted from a corpus, and showing how such information can be ordered and presented to learners of Japanese.

3. Loanword *suru*-verbs in BCCWJ

The following analysis is based on data from the 2009 monitor data version of the Balanced Corpus of Contemporary Written Japanese (hereafter abbreviated to BCCWJ). BCCWJ is a large-scale corpus consisting of texts from different media, including books, white papers, the online Q&A web-site *Yahoo!Chiebukuro* and

Minutes of the Japanese Diet (parliamentary proceedings), amounting to ca. 45 million words.

A corpus-wide search for loanword *suru*-verbs yielded 18,094 tokens corresponding to 1,421 types.[1]

Table 1 shows the list of loanword *suru*-verbs with the first 30 frequency ranks, alongside their description in three dictionaries and two other works: the largest existing dictionary of Japanese loanwords, the *Concise katakana-go jiten* (4th ed., with ca. 48,000 lemmas), two medium-size general Japanese language dictionaries, *Meikyō kokugo jiten* (ca. 70,000 lemmas) and *Iwanami kokugo jiten* (7th ed., ca. 65,000 words), a textbook of Japanese loanwords (Sasaki, 2001), and a research paper on basic loanwords in the context of teaching Japanese as a foreign language (Sawada, 1993).

Table 1: The 30 most common loanword *suru*-verbs in BCCWJ and their description in five reference works[2]

No.	Lemma	Lemma (romanized)	No. of examples	Concise[4] 2010	Iwanami kokugo[7] 2009	Meikyō kokugo 2002	Sasaki 2001	Sawada 1993
1	クリック	*kurikku*	1,385	Δ	○ (tr.)	○ (tr.)	×	×
2	チェック	*chekku*	815	○	○ (tr. & intr.)	○ (tr.)	○ (tr.)	○
3	スタート	*sutaato*	553	○	○ (intr.)	○ (intr.)	○ (intr.)	○
4	インストール	*insutooru*	338	○	○ (tr.)	○ (tr.)	×	×
5	コピー	*kopii*	334	○	○ (tr.)	○ (tr.)	○ (tr.)	○
6	コントロール	*kontorooru*	313	○	○ (tr.)	○ (tr.)	○ (tr.)	○
7	メール	*meeru*	304	Δ	Δ	Δ	×	×
8	カバー	*kabaa*	284	○	○ (tr.)	○ (tr.)	○ (tr.)	○
9	クリア	*kuria*	261	○	○ (tr.)	○ (tr.)	○ (tr.)	○

[1] The search was performed using the whole-text search system package Himawari, which is included in the BCCWJ monitor data set, searching with the condition "katakana + *suru* in all inflected forms", which does not include "katakana + *ru*" nor "katakana + *dekiru*". Types were counted according to the following criteria. (1) Orthographic variants such as *ba/va* (バ/ヴァ) or variants with long/short vowels are counted as separate types. (2) Compound words such as *katto&peesuto-suru* or *pasukatto-suru* are counted as individual, separate types. (3) Words of mixed origin, such as *dotakyan-suru* are also included. (4) Words which are generally accompanied by the particle *wo* when used with the verb *suru* (such as *sakkaa-suru*) are also included.

[2] In Table 1, the symbol ○ indicates that the reference work contains a description of the word's usage as a *suru*-verb, possibly including the distinction transitive (tr.) /intransitive (intr.); the symbol Δ indicates that the reference work contains the word in question, but only in a usage other than as a *suru*-verb; the symbol × indicates that the word is not listed as a lemma.

No.	Lemma	Lemma (romanized)	No. of examples	Concise[4] 2010	Iwanami kokugo[7] 2009	Meikyō kokugo 2002	Sasaki 2001	Sawada 1993
10	カット	*katto*	252	○	○ (tr.)	○ (tr.)	○ (tr.)	○
11	イメージ	*imeeji*	243	○	○ (tr.)	○ (tr.)	○ (tr.)	○
12	ダウンロード	*daunroodo*	240	Δ	○ (tr.)	○ (tr.)	×	×
13	セット	*setto*	220	○	○ (tr.)	○ (tr.)	○ (tr.)	○
14	プレゼント	*purezento*	206	○	○ (tr.)	○ (tr.)	○ (tr.)	○
15	リード	*riido*	206	○	○ (tr. & intr.)	○ (tr.)	○ (tr. & intr.)	○
16	アピール	*apiiru*	205	○	○ (tr. & intr.)	○ (tr. & intr.)	○ (tr.)	○
17	キス	*kisu*	202	○	○ (intr.)	○ (intr.)	×	○
18	チャレンジ	*charenji*	197	○	○ (tr. & intr.)	○ (intr.)	○ (intr.)	○
19	アクセス	*akusesu*	194	Δ	○ (intr.)	○ (intr.)	×	×
20	ヒット	*hitto*	174	○	○ (intr.)	○ (intr.)	○ (intr.)	○
21	リラックス	*rirakkusu*	169	○	○ (intr.)	○ (intr.)	○ (intr.)	○
22	アップ	*appu*	167	○	○ (tr. & intr.)	○ (tr. & intr.)	○ (tr. & intr.)	○
23	キャンセル	*kyanseru*	167	○	○ (tr.)	○ (tr.)	○ (tr.)	×
24	サポート	*sapooto*	164	○	○ (tr.)	○ (tr.)	○ (tr.)	○
25	ノック	*nokku*	150	○	○ (tr.)	○ (intr.)	×	○
26	メモ	*memo*	142	○	○ (tr.)	○ (tr.)	○ (tr.)	○
27	エスカレート	*esukareeto*	134	○	○ (intr.)	○ (intr.)	○ (intr.)	○
28	デザイン	*dezain*	132	○	○ (tr.)	○ (tr.)	×	○
29	リンク	*rinku*	126	○	○ (tr.)	○ (tr.)	○ (tr.)	×
30	シフト	*shifuto*	123	○	○ (intr.)	○ (intr.)	×	○
31	デビュー	*debyuu*	123	○	○ (intr.)	○ (intr.)	×	○
32	バックアップ	*bakkuappu*	123	○	○ (tr.)	○ (tr.)	○ (tr.)	○

The 20 most frequent verbs were the following.

(3) *kurikku-suru* (“to click”), *chekku-suru* (“to check”), *sutaato-suru* (“to start”), *insutooru-suru* (“to install”), *kopii-suru* (“to copy”), *kontorooru-suru* (“to control”), *meeru-suru* (“to mail”), *kabaa-suru* (“to cover”), *kuria-suru* (“to clear”), *katto-suru* (“to cut”), *imeeji-suru* (“to imagine”), *daunroodo-suru* (“to download”), *setto-suru* (“to set”), *purezento-suru* (“to give as a present”), *riido-suru* (“to lead”), *apiiru-suru* (“to appeal”), *kisu-suru* (“to kiss”), *charenji-suru* (“to challenge”), *akusesu-suru* (“to access”), *hitto-suru* (“to hit”)

Among these, the underlined verbs are mostly used as technical terms related to computers. The reason for the verb *kurikku-suru* to have such a prominent frequency is that this word is used repeatedly in explanations of operation instructions in software user manuals. Approximately 900 of the examples found come from such manuals. Similarly, approximately 90% of the occurrences of *insutooru-suru*, *meeru-suru* and *daunroodo-suru* come from the site *Yahoo!Chiebukuro*.

If we compare the list of the 20 most frequent loanwords in (3) and the list of "basic loanwords" proposed by Sawada (1993), we find that only the underlined words in (3) are not included in Sawada's list. Leaving aside computer terminology, which is highly susceptible to changes in time, we therefore find that most basic loanwords were found in this corpus search.

In the following section, the verb *katto-suru* is analysed in detail. The verb *katto-suru* was chosen because of its polysemy and because it is a general verb, and its use is not restricted to one particular subject area.

4. A case study: analysis of the verb *katto-suru*

Subsection 4.1 presents a semantic analysis of the verb *katto-suru*, based on examples from the BCCWJ, while subsection 4.2 offers an analysis of the sentence-final forms and other elements co-occurring with this verb.

4.1 Semantic characteristics

The loanword *katto* is derived from the English word *cut*. The loanword dictionary *Concise katakana-go jiten* (4th ed.) defines it as follows. The definition given below is partly abridged.

(4) カット [cut]　～する
1. 切ること，削除，省略。
2. 洋裁の裁断。
3. 〖テニス・卓球など〗打球に後退回転を与えること。
4. 〖バスケットなど〗相手のボールを横合いから奪い取ること。
5. 〖野球〗野手が他の野手の送球を途中で捕球すること。
6. 〖トランプ〗札を切ること。親が札を切り混ぜた後，子がその札を２分し，いかさまのないことを確認するもの。
7. 小型のさし絵，写真。
8. 映画やテレビで，ショットを編集作業により，適当な長さに切ったもの。また，監督や演出家が，１つのショットの撮影を終える時にする合図。
9. 〖美容〗髪を切り，形を整えること。
10. 〖ダンス〗上げた足をもう一方の足にたたきつけるようにして，すばやく立ち足を変えるもの。基本ステップの１。
11. 宝石を多面体に削ること。

This could be translated as follows.

(4) *katto* [cut] *~suru*

1. cutting, deletion, omission.
2. cutting cloth in dressmaking.
3. {tennis, table-tennis etc.} giving a backwards spin to a batted ball.
4. {basketball etc.} snatching the opponent's ball from aside.
5. {baseball} a fielder's getting another fielder's ball.
6. {card games} cutting the cards; the non-dealer's splitting the pack of cards into two parts after the dealer has shuffled the cards, to avoid cheating.
7. small illustration, photograph.
8. in film or television, a take that is cut to appropriate length during editing; or, the signal given by the film director or an actor at the end of a take.
9. {hairdressing} cutting and arranging hair.
10. {dance} quick change of supporting foot by flinging the moving foot onto the other foot; one of the basic steps.
11. shaping a rough gemstone into a faceted shape.

This is a very detailed description, including various uses of the word as a technical term. However, there is no explanation regarding which of these meanings is commonly used, nor any indication as to whether the word can be used as a verb (in the form *katto-suru*) in all the meanings listed.

To investigate these points, we collected examples of *katto-suru* from BCCWJ and ordered them according to their meaning. The result of this analysis is presented in Table 2, including frequency information.

Table 2: Semantic analysis of *katto-suru* examples

Meaning		**No. of examples**	
[1] cut	[food] *wo katto suru* - to reduce into small pieces, to divide [peel/rind/husk] *wo katto suru* - to peel, to pare, to skin	46	98
	[long thin things] *wo katto suru* - to shorten [thin things] *wo katto suru* - to make smaller	52	
[2] cut hair	[hair] *wo katto suru* - to cut and arrange	33	33
[3] cut down, reduce	[images/text/items] *wo katto suru* - to delete, to abridge	45	103
	[money/quantity] *wo katto suru* - to reduce, to cut down	58	
[4] cut off, block, obstruct	[ultraviolet rays/light] *wo katto suru* - to cut off, to block	12	14
	[a ball/ a pass] *wo katto suru* - to take, to intercept	2	
[5] other meanings		4	4
	Total:		**252**

Senses [1] to [4] were created on the basis of corpus examples, but also taking into account definitions in existing dictionaries. The common basic meaning of *katto-suru* could be described as “to make a cut into something, making it into another (smaller) shape”.

Examples of usage for each sense described in Table 2 are given below. The codes in brackets refer to the sample ID in BCCWJ.

(5) sense [1]: to reduce into small pieces, to divide, to make shorter, smaller

a. 野菜なども使う分だけ皮をむいて使う大きさにカットし、密閉容器に入れていくといい。

Yasai nado mo tsukau bun dake kawa o muite tsukau ookisa ni katto shi, mippei youki ni irete iku to ii.

“You should also peel just the vegetables you will need, cut them to the size you need and put them into airtight containers.” (PB25_00290)

b. 皮をナイフで薄くカットします。果肉は食用とします。

Kawa o naifu de usuku katto shimasu. Kaniku wa shokuyou to shimasu.

“Thinly cut off the husk. The fruit pulp will be used for dish preparation.” (LBm5_00033)

c. 編み終わりは、ひもの余分をカットし、内側にボンドでとめる。

Amiowari wa, himo no yobun o katto shi, uchigawa ni bondo de tomeru.

“When the knitting is finished, cut off the excess thread and paste it inside.” (LBs5_00032)

d. 先ほど受験用写真をきれいにカットするコツを質問した者です。

Sakihodo juken’you shashin o kirei ni katto suru kotsu o shitsumon shita mono desu.

“I am the one who just asked about tricks for cleanly cutting/trimming pictures for exams.” (OC10_00401)

(6) sense [2]: to cut and arrange hair

避暑地のお嬢さんらしく、帽子に隠れた髪は、学校にいる時より短くカットされている。

Hishochi no ojouchan rashiku, boushi ni kakureta kami wa, gakkou ni iru toki yori mijikaku katto sarete iru.

“Her hair, tucked away under her hat, is cut shorter than when she is at school, as would befit a young lady in a summer resort.” (LBr9_00274)

(7) sense [3]: to delete, to abridge, to reduce, to cut down

a. 読んでみると、議事録からは『あー』とか『うー』はカットされている。

Yonde miru to, gijiroku kara wa “aa” to ka “uu” wa katto sarete iru.

“On reading it, one finds that *aahs* and *uhms* have been cut out of the minutes.” (PB26_00141)

b. 公務員の数は半分になり、給与は３割カット、退職金は半減、または全額カットされる。

Koumuin no kazu wa hanbun ni nari, kyuuyo wa sanwari katto, taishokukin wa hangen, mata wa zengaku katto sareru.

"The number of public servants is halved, allowances reduced by 30%, retirement money reduced by half or completely abolished." (PB53_00657)

(8) sense [4]: to cut off, to block, to intercept

a. 前述の通り遮光板で上方の光が綺麗にカットされるため、マルチリフレクターより暗く感じる。

Zenjutsu no toori, shakouban de jouhou no hikari ga kirei ni katto sareru tame, maruchirefurekutaa yori kuraku kanjiru.

"As stated earlier, the light above is completely cut off by the gobo, so that one feels it darker than with a multi-reflector." (OC06_03304)

b. ディフェンスしてて相手の蹴ったボールを胸でカットした、とかならまだしも。

Difensu shite te aite no ketta booru o mune de katto shita, toka nara madashimo.

"I could get it if he had got the other's ball with his chest for defence." (OC14_03488)

(9) other uses:

a. ＜斜めに切るように打つ＞　バレーボールのボールなどを打つときにボールの下部を手のひらでカットするようにするのと、

[to strike with a diagonal slash] *Bareebooru no booru nado o utsu toki ni booru no kabu o tenohira de katto suru youni suru no to,*

"Striking the lower part of the ball with one's palm of the hand in volleyball and the like, and..." (PB4n_00054)

b. ＜解除する＞　中古車で、リミッターがカットされている車は売られているものなのでしょうか？

[to unlock, to deactivate] *Chuukosha de, rimittaa ga katto sareteiru kuruma wa urareteiru mono na no deshouka?*

"In the case of used cars, are cars with deactivated speed limiters being sold?" (OC06_00923)

From a semantic point of view, sense [2] "to cut hair" could be merged into sense [1], since it coincides with the sense "to shorten long thin things". However, since examples categorised as sense [2] contain mediative expressions (expressions of semantic indirectness), as will be explained later, these examples were put in a separate sense.

Sense [3] is generally presented in dictionaries as "to delete/abridge a part of a text or sum of money", but in fact it may express the deletion of something in its completeness, as it can collocate with expressions such as *zenbu* (全部 "all") or

zengaku (全額 "the whole sum") as in example (7b). However, these cases may also be interpreted as "considering a larger unit, making the larger unit smaller".

It should be noted that most dictionaries present the use of *katto-suru* as a specialised term in sports, such as senses 3. and 4. in the dictionary description quoted at (4), but that examples of this use such as (8b) and (9a) were actually very rare in BCCWJ.

4.2 Syntactical characteristics

This section explores the syntactical characteristics of the verb *katto-suru*, beginning with co-occurrence patterns, for each of the senses presented in 4.1.

4.2.1 Co-occurrence patterns

Let us first consider the case particles and adverbs occurring in sentences where the predicate is *katto-suru*. Table 3 presents data for all patterns occurring in at least 5 examples.

Table 3: Co-occurrence patterns of *katto-suru*

Sense	No. of examples	Case particles				Adverbial expressions	
		o **(object)**	***de*** **(instr.)**	***de*** **(location)**	***kara*** **(source)**	**Adverbs of result**	**Adverbs of quantity**
[1] cut	98	66	9		1	32	8
[2] cut hair	33	11	3	6		11	
[3] reduce	103	52			5	4	16
[4] block	14	12	2				
[5] other	4	2					
Total	**252**	**143**	**14**	**6**	**6**	**47**	**24**

The direct object of the transitive verb *katto-suru* marked by particle *o*, including cases where the object is topicalised and marked by particle *wa*, appears in the same sentence in 143 of 252 cases (56.7%). There are 54 further cases when the patient, which is usually marked by particle *o*, appears as the subject of a passive sentence marked by particle *ga*, and 10 cases where the patient is the head of a noun-modifying clause and therefore not accompanied by any particle. An accurate count of the examples where the patient (usually accompanied by particle *o*) is really absent therefore yields 45 examples altogether (17.9%). Since 20 of these are examples of sense [2], we can conclude that one of the characteristics of sense [2] is that it tends not to co-occur with the verb's direct object, although it is sometimes difficult to decide whether an example such as (10) should be considered to be a case of transitive sentence where the object *kami o* (髪を "hair *o*") is omitted, or a case of intransitive

sentence. In this analysis, such sentences were considered to be cases of transitive sentences.

(10) 私はいつも美容院でカットしている。
Watashi wa itsumo biyouin de katto-shite iru.
"I always have [my hair] cut at the hairdressers." / "I also have a cut at the hairdressers.")

Other particles that characteristically co-occur are instrumental particle *de* with sense [1] (e.g. *hasami de katto-suru* "cut with scissors"), locative particle *de* with sense [2] (e.g. *biyouin de katto-suru* "to cut / have a haircut at the hairdressers"), source particle kara with sense [3] (e.g. *chingin kara katto-suru* "to cut from wages").

If we now consider adverbial expressions, we find that sense [1] and [2] are often accompanied by adverbs which express the result of cutting, such as "...*o mijikaku* ("short") / *hitokuchidai ni* ("to a mouthful") / *suki na katachi ni* ("to one's preferred shape") *katto-suru*", while sense [3] is often accompanied by adverbs of quantity or degree, such as "... *o sukoshi* ("a little") / *ichibu* ("in part") / *zengaku* ("completely, for the whole sum") *katto-suru*".

On the basis of this analysis of co-occurrences, table 4 summarises typical patterns for each sense, taking into account patterns which occurred in approximately 10% or more examples.

Table 4: Senses and patterns of *katto-suru*

Sense		Arguments	Patterns
[1] cut	to process / peel / separate	[food] *o* ([tool] *de*)	*o* case (+ instrumental *de* case) (+ expression of result)
	to shorten / make smaller	[thin and long / thin object] *o* ([tool] *de*)	
[2] cut hair	to cut and arrange hair	([hair] *o*) ([place] *de*)	(*o* case) (+ locative *de* case) (+ expression of result)
[3] reduce	to delete / abridge	[images / text / items] *o*	*o* case (+ expression of quantity / degree)
	to reduce the quantity	[money / quantity] *o*	
[4] block	block, obstruct	[ultraviolet rays / light] *o* ([tool] *de*)	*o* case (+ instrumental *de* case)
	take, intercept	[a ball / a pass] *o* ([bodypart] *de*)	

As could be seen above, by analysing corpus data it is possible to present detailed information regarding cases, arguments and adverbs which tend to co-occur with a particular verb, and the semantic category of nouns which tend to appear in these

arguments. In the case of *katto-suru*, it was shown that all senses appear in transitive uses of the verb, but that each sense appears in its own particular pattern.

4.2.2 Sentence-final forms

Let us now see the characteristic sentence-final forms such as voice markers and auxiliary verbs which appear after the verb *katto-suru*. Table 5 presents forms which have appeared in at least 5 examples.

Table 5: Sentence-final forms of *katto-suru*

Sense	**No. of examples**	***-rare-* (passive)**	***-te iru* (progressive/ resultative/ havitual)**	***-te shimau* (perfect)**	***-te morau* (benefactive)**	***-te iku* (future continuation)**	***-you* (volitive)**
[1] cut	98	9	4	1		2	
[2] cut hair	33	2	8		6	1	
[3] reduce	103	39	16	6	1	4	5
[4] block	14	2		1			
[5] other	4	2	2				
Total	**252**	**54**	**30**	**8**	**7**	**7**	**5**

A very conspicuous point is that sense [3] tends to occur in the passive form much more than other senses (72.2% of all passive sentences pertain to this sense). This tendency reflects the fact that the situation of "reducing / cutting down surplus parts of a text, image or sum of money" is depicted from the point of view of the (possibly unwitting) receiver more often than the other senses. This is also corroborated by the fact that as much as 6 examples of *-te shimau*, an auxiliary verb expressing regret, are used in this sense in the passive form. Conversely, the volitive form *-you*, which is used when viewing the act from the opposite point of view, such as the administration or management, is used very rarely and almost exclusively in parliament proceedings and economic texts.

Sense [2], on the other hand, often occurs with benefactive verbs (*-te morau* "receive" or *-te kureru* "give (to me)"), expressing gratitude for the received action. *Katto-suru* used in sense [2] exhibits the characteristic of semantic indirectness (mediativeness) (*kaizaisei* 介在性, Sato, 2005), in the sense that it can be used in examples such as (10) even when it was a hairdresser or someone else (and not the subject of the sentence) who actually cut the hair, at the subject's request. The syntactic construction with an auxiliary benefactive verb could therefore be considered as a syntactic reference to the agent who acted upon request.

If we now consider the form *-te iru*, we find as many as 22 examples where the form expresses the resulting state of an action (16 of these are in the passive form),

while most remaining examples refer to sense [2]: 6 examples of repetitive action / habit, 1 example of progressing action. The tendency of sense [2] to appear in forms expressing repetition can be seen as stemming from the fact that the action of "cutting one's hair" is something done habitually.

Finally, all examples of the form *-te iku*, regardless of the verb sense, express gradual development or progression of the action described and do not exhibit any particularity regarding sense.

4.3 Analysis

As could be seen in the above analysis, usage examples of *katto-suru* taken from the corpus BCCWJ can show not only the characteristics of nouns which can typically co-occur as objects, but also the characteristics of co-occurring expressions and predicate forms which are typical for each sense.

For example, with regard to sense [2] "cut one's hair", the following syntactical characteristics have been observed.

(11) a. There are cases in which the object with particle o is not expressed, since the object "hair" is taken for granted.

b. There is sometimes an argument with locative case particle *de*, such as "*biyouin de*" ("at the hairdresser's").

c. The verb is often accompanied by expressions of result, such as "*mijikaku*" ("short").

d. The verb is often in the form *-te iru*, expressing repetitive action or habit.

e. It can appear in sentences expressing semantic indirectness (mediativeness), in which reference to the agent is made by means of auxiliary benefactive verbs.

The fact that we can observe such an equivalence between the senses of a verb and its syntactical characteristics means that we can predict - to a certain extent - the meaning and subsense of a word from the form of its sentence, and vice versa the form of a sentence from the meaning (or subsense) of a word.

In the context of teaching Japanese as a foreign language, providing learners with a description of a verb which includes not only its lexical meaning, but also its patterns of usage as presented in table 4, would help them in receptive and productive tasks, since it would provide them with the information they need to tell, for example, which sense of *katto-suru* is meant in the sentence they read, judging from the co-occurring words and patterns, or to predict, for example, which words and patterns can be used with *katto-suru* when they want to use this verb in a particular sense.

5. Conclusion

The results of the case study presented in this paper indicate the importance of analysing loanwords from a grammatical perspective, investigating their behaviour within sentences and not only their meaning.

Judging from the many differences in the description of loanwords in monolingual Japanese dictionaries, as presented in Table 1, it is clear that much is still unknown regarding loanwords in contemporary Japanese. The first task that awaits us is to build up a comprehensive description of basic loanwords, on the basis of corpus data. As has been shown in the present paper, the results of such a description would be very useful to the teaching of Japanese as a foreign language.

At the same time, while preparing detailed analyses of individual words, a descriptive lexicographical framework aimed at foreign learners needs to be developed, indicating what information needs to be included in a description of grammatical patterns. Moreover, even a corpus of the size of BCCWJ may sometimes not offer enough examples for fine-grained distinctions of meaning. The solution of this problem is another task that awaits us.

References

Himeno, M. (2004). *Kenkyūsha Nihongo Hyōgen Katsuyō Jiten*. Kenkyūsha.

Ishino, H. (1996). Jiten ni okeru Gairaigo no Gogi Kijutsu. In *Gengogakurin 1995-1996*. Tokyo: Sanseidō. 273-286.

Kim, E. (2011). 20-seiki Kōhan no Shinbun Goi ni okeru Gairaigo no Kihongoka. (Shift of the Loanwords to Basic Words in the Japanese Newspaper Vocabulary in the Second Half of the 20th Century.) *Handai Nihongo Kenkyū*, Monograph 3. Toyonaka: Osaka University.

Kitahara, Y. (2002). *Meikyō Kokugo Jiten*. Tokyo: Taisyūkan.

Koizumi, T., Funaki, M., Honda, K. Nitta,Y. & Tsukamoto, H. (1989). *Nihongo Kihon Dōshi Yōhō Jiten*. Tokyo: Taisyūkan.

Nakayama, E., Jinnouchi, M., Kiryu, R., & Miyake, N. (2008). Nihongo kyōiku ni okeru “Katakanago Kyōiku” no Atsukawarekata. (Teaching Japanese as a Foreign Language: Katakana and its Implementation in the Syllabus.) *Nihongo Kyōiku*, 138, 83-91.

Nishio, M., Iwabuchi, E., & Mizutani, S. (2009). *Iwanami Kokugo Jiten*. 7th ed. Tokyo: Iwanami Shoten.

Sanseidō Henshūjo. (2010) *Concise Katakanago Jiten*. 4th ed. Tokyo: Sanseidō.

Sasaki, M. (2001). *Yoku Tsukau Katakanago*. Tokyo: Alc.

Sato, T. (2005). *Jidōshi-bun to Tadōshi-bun no Imiron*. Tokyo: Kasamashoin.

Sawada, T. (1993). Nihongo Kyōiku no tameno Kihon Gairaigo ni tsuite. (Loanwords Usage in Japanese: The Fundamental Points for the Japanese Language Teaching.) *Bulletin of Nara University of Education. Cultural and Social Science*, 42(1), 225-239.

RESEARCH ARTICLES
(PROJECT REPORTS)

COMPILATION OF JAPANESE BASIC VERB USAGE HANDBOOK FOR JFL LEARNERS: A PROJECT REPORT

Prashant PARDESHI
National Institute for Japanese Language and Linguistics (NINJAL)
prashantpardeshi@gmail.com

Shingo IMAI	Tsukuba University
Kazuyuki KIRYU	Mimasaka University
Sangmok LEE	Kyushu University
Shiro AKASEGAWA	Lago Institute of Language
Yasunari IMAMURA	National Institute for Japanese Language and Linguistics (NINJAL)

Abstract

In this article we introduce a collaborative research project entitled "*Nihongogakushuushayou kihondoushi youhouhandbook no sakusei* (Compilation of Japanese Basic Verb Usage Handbook for Japanese as Foreign Language (JFL) Learners)" carried out at the National Institute for Japanese Language and Linguistics (NINJAL) and report on the progress of its research product, namely, a prototype of a basic verb usage handbook (referred to as "handbook" below). The handbook differs in many ways from the conventional printed dictionaries or electronic dictionaries available at present. First, the handbook is compiled online and will be made available on internet for free access. Secondly, the handbook is corpus-based: the contents of the entry are written taking into consideration the actual use of the headword using the BCCWJ corpus. Also, it contains illustrative examples of particular meanings culled from the BCCWJ corpus as well as those coined by the entry-writers. Third, the framework used in the description of semantic issues (polysemy network, cognitive mechanism underlying semantic extensions and semantic relationships among various meanings, etc.) is cognitive grammar, which adopts a prototype approach. Fourth, it includes audio-visual contents (such as audio files and animations/video clips etc.) for effective understanding, acquisition and retention of various meanings of a polysemous verb. Fifth, the handbook is bilingual (Japanese-Chinese, Japanese-Korean and Japanese-Marathi) and incorporates insights of contrastive studies and second language acquisition. The handbook is an attempt to share cutting edge research insights of various branches of linguistics with Japanese language pedagogy. It is hoped that the handbook will prove to be useful for JFL learners as well as Japanese language teachers across the globe.

Keywords

basic verbs; corpus-based; cognitive grammar; audio-visual contents; bilingual dictionary; multilingual dictionary

Izvleček

Članek predstavlja skupinski raziskovalni projekt z naslovom "*Nihongogakushuushayou kihondoushi youhouhandbook no sakusei* (Izdelava priročnika o rabi japonskih osnovnih glagolov za učence japonščine kot tujega jezika)", ki poteka na Državnem inštitutu za japonski jezik in jezikoslovje (National Institute for Japanese Language and Linguistics - NINJAL), ter poroča o trenutnem stanju raziskovalnega izida, t.j. prototipa priročnika o rabi osnovnih glagolov (v nadaljevanju "priročnik"). Priročnik se v marsičem razlikuje od običajnih tiskanih in elektronskih slovarjev, ki so trenutno dosegljivi. Prva značilnost je ta, da se priročnik ureja preko spleta in bo prosto dostopno objavljen na spletu. Druga je ta, da je priročnik osnovan na korpusih: pri redakciji gesel se upošteva dejanska raba iztočnic v korpusu BCCWJ, priročnik pa vsebuje tako primere rabe posameznih podpomenov, ki se črpajo iz korpusa BCCWJ, kot tudi primere, ki jih sestavijo redaktorji. Tretja značilnost je ta, da se semantični vidiki (pomenske mreže, kognitivni mehanizmi, ki botrujejo pomenskim širitvam, ter pomenske povezave med posameznimi podpomeni, ipd.) opisujejo v okviru kognitivne slovnice s prototipnim pristopom. Četrta značilnost je ta, da vključuje zvočne in slikovne vsebine (zvočne posnetke, animacije, videoposnetke ipd.) kot pomoč pri učinkovitem razumevanju, učenju in pomnjenju različnih pomenov večpomenskih glagolov. Peta značilnost je ta, da je priročnik dvojezičen (japonsko-kitajski, japonsko-korejski in japonsko-maratski) in vključuje spoznanja protistavnega jezikoslovja in vede o učenju tujih jezikov. Priročnik je poskus zlitja najnovejših raziskovalnih spoznanj različnih vej jezikoslovja z didaktiko japonskega jezika. Upamo, da bo priročnik koristil tako učencem kot učiteljem japonščine po celem svetu.

Ključne beside

osnovni glagoli; korpusno osnovan; kognitivna slovnica; zvočno-slikovne vsebine; dvojezični slovar; večjezični slovar

1. Introduction

Verbs as predicators are one of the crucial components determining the skeleton of a sentence, which serves as a basic unit of communication. For improving communication skills in Japanese it is imperative for JFL (Japanese as foreign language) learners to master various usages of basic verbs used frequently in day-to-day communication in a systematic way. At the National Institute for Japanese Language and Linguistics (NINJAL), a collaborative research project entitled "*Nihongogakushuushayou kihondoushi youhouhandbook no sakusei* (Compilation of Japanese Basic Verb Usage Handbook for Japanese as Foreign Language (JFL) Learners)" is being carried out (project leader: Prashant Pardeshi, timeline: October 2009-September 2012). The aim of the project is to develop a prototype for the compilation of a handbook of usage of basic verbs in Japanese frequently used in day-to-day conversation by integrating state-of-the-art insights from various related fields such as Cognitive Linguistics, Corpus Linguistics, Japanese Linguistics, Japanese Language Pedagogy, Contrastive Linguistics, and Linguistic Typology. The envisaged end product is a set of small-scale bi-lingual handbooks such as Japanese-Chinese, Japanese-Korean and Japanese-Marathi, compiled adopting the prototype developed in the project. We believe that such a bilingual handbook of usage of Japanese basic verbs

would be of great help for JFL learners in their effort to acquire the Japanese language systematically and efficiently.

The handbooks under compilation differ from existing dictionaries in various respects such as compilation policy, scope and contents of description and the writing and editing process. In this article we report on the progress of the project and salient features of its envisaged research output, namely, a prototype of a bilingual Japanese basic verb usage handbook (referred to as handbook below). The structure of this article is as follows. In section 2 we provide the outline of the handbook project and a overview of the salient features of the handbook under preparation. Against this backdrop, in section 3 we exemplify the organization of each entry with the help of a concrete example – the verb *hashiru* "to run" – and describe the (tentative) methodology of description. One of the salient features of the handbook is that it is corpus-based. In section 4, we describe the tools/interfaces developed for retrieving information necessary for writing an entry from the corpora of correct use of Japanese and of the errors of JFL learners. Further, the compilation and editing work of the handbook is carried out online using a web-based editing tool. In section 5, we describe the multilingual editing tool developed in this project. This tool allows us to transcend the barriers of space and time. Furthermore, we are developing audio-visual contents in order to foster understanding of various meanings of polysemous verbs. In section 6 we introduce those contents. Finally, in section 7 we discuss future prospects.

2. Overview of the handbook project and salient features of the handbook

2.1 Overview of the handbook project

We believe that systematic learning of polysemous basic verbs including features such as the semantic behaviour (semantic extensions of a verb and interrelationship among its various meanings, related words such as synonyms, antonyms etc., proverbs/idioms involving the verb in question etc.), grammatical/syntactic behaviour (voice and polarity bias, aspectual and modal characteristics, co-occurrence restrictions, modifiers/adverbial elements, ungrammatical/unnatural usages, etc.), argument structure (case frame), genre/register bias, etc. is necessary in order to master communication skills in Japanese. Further, it is also necessary to know where and how the Japanese language (target language: L2) is similar to or different from the user's mother tongue (source language: L1). In view of this, the aim of the project is to develop a prototype for the compilation of a handbook of usage of Japanese basic verbs by integrating state-of-the-art insights from various related fields of theoretical and applied linguistics for the JFL learner. At present, 58 scholars from various parts of the globe are participating in this project. Out of these 58 scholars, 42 are native speakers

of Japanese while 16 are non-native researchers working on Japanese language for a long period of time[1].

Since the primary goal of the project is qualitative, viz. developing a prototype of a bilingual basic verb usage handbook, we decided to restrict the quantity (number) of verbs and focus on highly polysemous basic verbs which pose a great challenge for JFL learners. Concretely speaking, we focus on the following 11 verbs: verbs of spatial motion (vertical motion: *agaru* "go/move up", *ageru* "cause to go/move up", *sagaru* "go/move down", *sageru* "cause to go/move down", and horizontal motion: *hashiru* "run"), and verbs of temporary or permanent transfer of possession (*ageru* "to give something to someone as a present/gift", *morau* "to receive something from someone as a present/gift", *uru* "to sell", *kau* "to buy", *kasu* "to lend" and *kariru* "to borrow"). All of these verbs are highly polysemous: for example, in our handbook there are 19 meanings/senses for *agaru* "go/move up", 22 for *ageru* "cause to go/move up", and 11 for *hashiru* "run". In section 3, we describe the policy and method of description of an entry through the example of the entry for *hashiru* "run" in our handbook.

2.2 Salient features of the handbook

The handbook under preparation is in electronic online form and the target users of the handbook are envisioned to be advanced JFL learners and native as well as non-native teachers of Japanese. In addition to the dictionary-like usage for looking up the meaning and examples illustrating various meanings of a verb, the handbook serves as a reference grammar also containing many salient features such as explanation of cognitive mechanisms underlying semantics extensions, notes on grammatical and non-grammatical usages, pragmatics or context-related explanations, tips from the contrastive perspective (comparison with the L1 of the JFL learner), "real" examples from the corpus, visual contents such as image-schema (static, abstract line drawings as well as concrete animations and video-clips), and audio-contents such as accent pattern and sound-files for all illustrative examples. Further, the descriptions and "coined" examples are all based on the actual use of the verb as "objectively" gleaned through the corpus data.

Out of all these salient features, two features can be considered as "discriminatory" ones that set apart the present handbook from the bi-lingual dictionaries available at present: (i) corpus-based approach: drawing on a corpus of "correct use" of Japanese native speakers and one of "erroneous use" of JFL learners in addition to the intuitions of scholars for the composition of entries and (ii) incorporating the insights of cognitive linguistics and contrastive linguistics.

For the corpus of "correct use" of Japanese native speakers we used the BCCWJ corpus (Maekawa, 2012) developed by the National Institute for Japanese Language

[1] For further details visit the project HP: http://www.ninjal.ac.jp/research/project/b/youhoujiten/.

and Linguistics (NINJAL) and developed an interface called NINJAL-LWP for the BCCWJ corpus (NLB) to cull the information necessary for writing a entry. For "erroneous uses" of JFL learners we used the data from Teramura (1990) and developed a interface to retrieve relevant information from it (see section 4 for details). The prototype of the handbook under preparation incorporates examples from BCCWJ corpus culled with the help of NLB and thus offers both coined as well as real examples side-by-side (see the tentative design in Figure 1).

For incorporating the insights of cognitive linguistics we have incorporated visual contents such as image-schema (static, abstract line drawings as well as concrete animations and video-clips), and audio-contents such as accent pattern and sound-files for all illustrative examples taking full advantage of the web-based nature of the handbook. As for incorporating insights of contrastive linguistics, in addition to grammatical similarities and differences between Japanese and JFL's native language we have provided extra-grammatical information such as notes on pragmatics and cultural factors.

The handbook is compiled/edited using a web-based editing tool connecting scholars in Japan, China and India. Such a handbook differs in many respects from contemporary bilingual dictionaries and therefore we purposely call it a bilingual handbook. In the following sections prominent salient features of the handbook are discussed.

3. The organization of an entry and the (tentative) methodology of description

3.1 Organization of an entry

The organization of an entry/headword is explained below with the help of the concrete example of the verb *hashiru* (to run). Following this, the methodology of description is mentioned. However it should be borne in mind that the statement pertaining to the methodology of description is tentative and subject to change.

〔アクセント：Accent〕LHL

〔活用：Conjugation〕hasir- Group I

〔語義一覧：List of senses/meanings〕

1. 人、動物などが、（足を交互にすばやく動かして）速やかに前進する (a person or an animal moves quickly ahead (by quickly moving its legs alternatively))
2. 乗り物が速く動く (a vehicle moves fast)
3. 乗り物が運行する (transportation operates)

4. 目的の場所へ急いで移動する (to move to the destination hurriedly)
5. 目的のために動き回る (to run around for some purpose)
6. 逃げる。自分の立場から逃げてある側につく (to run away, to flee from one's own side and join another side)
7. 好ましくない傾向に傾く (incline towards an undesired trend)
8. 速くさっと見る (take a quick look)
9. 感覚、現象などが一瞬にして現れる（現れて消える）(sudden appearance [and disappearance] of a feeling or phenomenon)
10. 道、川、亀裂などがある方向に延びている、通じている (extension or continuation of a road or a river or a crack etc. in a particular direction)
11. 活動する。実績を上げる (to work, to achieve results)

The details of the sense 1 are described below. Owing to space restrictions, other senses are not discussed here.

〔語義：Sense/meaning〕

人、動物などが、（足を交互にすばやく動かして）速やかに前進する
a person or an animal moves quickly ahead (by quickly moving its legs alternatively).

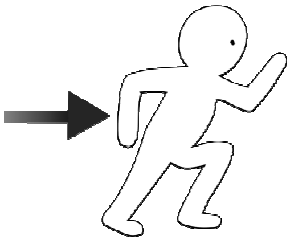
〔表記：Orthographical form〕

走（はし）る

〔自他：Transitivity〕

自動詞 (Intransitive)

〔イメージ：Image〕



〔構文フレーム：Construction frame〕

- 基本フレーム：<人・動物> が走る (Basic frame: <person/animal> NOM runs)
- オプショナル要素 (Optional elements/adjuncts)
  （起点）から (source) kara, （着点）まで (goal) made
  （場所１／位置）を (location 1/position) wo,（距離１）を (distance 1) wo

（場所２）で (location 2) de, （道具）で (instrument) de, （速さ）で (speed) de,
（距離２） (distance 2), （目的）で (purpose) de, （時間１）で (time 1) de,
（様態） (manner), （時間２）(time 2)

〔共起例：Collocations〕

＜人・動物＞が <person/animal> ga	① 人 (person)：私 (I), ー（さん） Mr./Mrs./Ms. X, 彼 (he), 子供 (child), 選手 (player) ② 動物 (animal)：馬 (horse), 猫 (cat), ネズミ (mouse)
（起点）から (source/starting point) kara	① 建物 (building)：駅 (station), 家 (house) ② 場所 (place)：東京 (Tokyo), 箱根 (Hakone), （人／もの）のところ (from the location of a person or an object)
（場所１／位置）を (place 1/position) wo	① 場所 (place)：公園 (park), 屋内 (in-house), 校庭 (school ground), 砂浜 (beach), ～沿い (along something), 歩道 (walkway), 山道 (mountain trail/pass), コース (course), 廊下 (corridor), 水の上 (on or above the water surface), 闇の中 (in the dark), 暖かい日差しの中 (in the warm sun) ② 位置 (position)：目の前 (in front of one's eyes), 先頭 (ahead), トップ (top), はるか前方 (way ahead in the forward direction)
（距離１）を (distance 1) wo	マラソン (Marathon), 42.195km, ハーフマラソン (half marathon), 長距離 (long distance), 短距離 (short distance)
（場所２）で (location 2) de	公園 (park), 屋内 (indoor), 校庭 (school ground), 砂浜 (beach)
（道具）で (instrument) de	ジョギングシューズ (jogging shoes), 裸足 (bare foot)
（速さ）で (speed) de	全速力(with full speed), 時速 50km (50 km/hour)
（距離２） (distance 2)	100 メートル (100 meters), 50 メートル (50 meters)
（目的）で (purpose) de	国体 (national tournament), オリンピック (Olympic), レース (race)
（時間１）で (time 1) de	1 時間 (one hour), 100 メートルを 11 秒 (100 meters in 11 sec)
（様態） (manner)	ゆっくり (slowly), 速く (fast), 一目散に (as fast as one's legs can/could carry one), 勢いよく (fiercely), 息せき切って (breathlessly), トロトロ (feebly), ビュンビュン (zippingly)
（時間２） (time 2)	１時間 (one hour), 10 分 (10 minutes)

〔非共起例：Wrong collocations〕

（様態）(manner)　（誤）(inappropriate/incorrect)　遅く (slowly)

〔例文・作例：examples/coined examples〕

- 大学の中を新しい靴でゆっくりと 10km 走った。((I) slowly ran 10 km at the university wearing new shoes.)
- 犬が公園の中を向こうからこちらへ走ってくる。(The dog runs across the park from the other side to here.)
- 駅伝で東京から箱根まで走る。(To run from Tokyo to Hakone in the ekiden race.)
- 駅まで大通りを走っていく。(To run to the station along the boulevard street.)
- 家のまわりをゆっくり 20 分ほど走った。((I) ran slowly around my house for 20 minutes.)

〔例文・コーパス：examples/from corpus: not translated into the target language〕

- まるで競争しているみたいな勢いで廊下を走るとは。（ベティ・ニールズ作、和香ちか子訳『幸せへの航海』, 2004）
- ちなみに、お巡りさんは歩道を走っています。(Yahoo!知恵袋, 2005, マナー、冠婚葬祭)
- かなり本格的に走る人たちばかりで、半分ぐらいは外国人の感じもしますが、日本人であれ外国人であれ、こんなに大勢の人が走るのは健康になりたいためでしょうか。(阿久悠『詩小説』, 2000. 9 文学)

〔個別の誤用情報：Information on errors pertaining to specific use〕

(1) 語義 1 で到達地点の「に」をとることはできない。到達地点の「に」を用いた場合、語義 4 の解釈になる。語義 1 で到達地点の「に」をとるときは「走っていく」「走りこむ」などと方向を表す動詞を伴った複合動詞にする必要がある。(In the case of sense/meaning 1, the goal location cannot be marked with the particle “*ni*”. If the goal location is marked with the particle “*ni*”, the meaning changes to sense/meaning number 4. In order to use the particle “*ni*” in the case of sense/meaning 1, it is necessary to use a complex predicate such as “*hashitte iku*” or “*hashirikomu*” which contain a verb implying direction.

（誤: ungrammatical use）駅に全速力で走った。（語義 4 の解釈になる。）
（正: grammatical use）駅まで全速力で走った。
（正: grammatical use）駅に全速力で走っていった。

(2)（様態）の「はやく」は速度を表す「速く」であり、時期を表す「早く」は用いない。The adverb “*hayaku*” is the one that expresses “*speed*” and not the one that expresses “*an early time/period/season*”.

（誤: ungrammatical use）あの選手はとても早く走る。
（正: grammatical use）あの選手はとても速く走る。

— 中略 abbreviated —

〔文法：Grammar〕

語義 sense	走らせる	走ろう	走っている	走った
	使役 causative form	意思 volitional form	進行 progressive form	過去 past form
1	○	○	○	○
2	○	×	○	○
3	○	×	○	△
4	○	△	○	○
5	○	○	○	○
6	△	△	△	○
7	○	×	○	○
8	○	×走らせよう	○走らせている	○走らせた
9	×	×	×	○
10	△	×	○（状態）	×
11	△	△	○（状態）	△

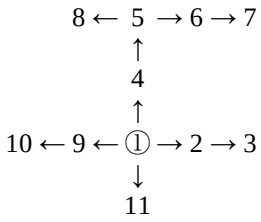
〔複合語： Compounds〕

▸走り回る ▸走り去る ▸走り通す ▸走り込む ▸走り抜く ▸走り抜ける
▸走り過ぎる ▸走り高跳び ▸走り幅跳び ▸突っ走る ▸ひた走る ▸小走り
▸ひとっ走り ▸使い走り ▸走り書き ▸走り読み ▸口走る ▸先走る
▸才気走る ▸石（いわ）走る ▸血走る ▸ご馳走

〔慣用句・ことわざ: Idioms/Proverbs〕

▸ペン（筆）が走る ▸虫酸が走る

〔語義ネットワーク: Semantic network〕



〔関連語（ワードファミリー）: Related words (word family)〕

- 同義語 Synonyms：▸駆（駈）ける ▸駆け足 ▸馳せる ▸ダッシュする
- 類義語 Near-synonyms：▸歩く ▸通る ▸動く ▸進む ▸行く ▸飛ぶ

3.2 The methodology of description: the content and the intent

〔Accent〕

In the case of accent, H stands for high and L stands for low pitch accent. However, for conveying accent information, the audio medium is more effective than the visual and we provide audio files to convey accent in addition to the visual representation.

〔Conjugation〕

The stem of the verb and its conjugation pattern is provided. As for the conjugation pattern, the classification widely used in Japanese language education (Group I, II and III) is adopted.

〔List of meanings/senses〕

The basic meaning is presented first and derived meanings follow as distinct senses. The basic meaning is also known as the central meaning and in a polysemous word it is considered as the most basic sense/meaning. This meaning is more concrete, more frequent and corresponds to what is known as the prototypical sense. The order of senses/meanings in the list of senses/meanings is decided taking into consideration the semantic closeness or remoteness of the sense in question to the central meaning. However basically this relationship is not linear there is some inevitable arbitrariness in determination of the order of meaning/senses. A semantic network diagram (described below) is also presented in order to show relationships among meanings graphically.

〔Meaning/Sense〕

The meaning/sense is explained in simple, easily understood terms. Some key words are intentionally used in order to make clear the relationships among the explanations. Such a strategy will also help to foster the understanding of connections in the semantic network. Also, the explanation is devised in such a way that the semantic congruence between the constructional meaning suggested by the construction frame discussed later and the core arguments and adjuncts would be easier to comprehend.

〔Orthographic representation〕

The orthographic representation in Kanji (Chinese) characters is provided with kana reading.

〔Transitivity〕

The transitivity of the verb in question is given. Depending on the meaning/sense, the transitivity may vary. However, the transitivity given here is that of the basic/central meaning.

〔Image〕

Providing a pictorial image of the meaning/sense helps in facilitating understanding of the meaning/sense in question. Image plays an important role especially in the derived/extended meanings/senses. Images are modeled on image schema proposed in the theory of cognitive linguistics. However we adopted more concrete images as compared to theoretical image schema. Further, in the case of image, unlike image

schema, emphasis is given to ease of understanding rather than theoretical precision. For the image, still pictures, animation as well as video clips are used (see section 6 for details).

〔Construction frame〕

The construction frame is shown in the form of a two tier structure: obligatory core arguments and optional adjuncts. However, as shown below, in some cases judgment between the two is difficult. For example, the verb *kaku* "to write" is a two-place predicate taking two core arguments, however in a construction like <person 1> write <a letter> to <person 2> it behaves like a 3-place predicate. In such cases, in the construction grammar approach (cf. Goldberg (1995), the construction containing 3 arguments is assumed. One falls in a dilemma on the issue of whether the 3-place construction should be incorporated in the description of a dictionary entry for the verb *kaku* "write". This is because, if one proceeds with adopting the construction-centered explanation, one needs to include extremely eccentric constructions as well, resulting in dramatically swelling the length of the description. Even if one adopts such a description policy, the issue of deciding whether the phrase <person 2> *ni* should be treated as an argument or as an adjunct remains unsolved. Viewed from the meaning/sense of the verb it is an adjunct while viewed from the point of a construction it is an argument. At present this issue is left to the decision of the entry writer and editor, however, by referring to the frequency count, this issue can be resolved to a certain extent.

〔Collocations〕

Collocations are shown for both arguments and adjuncts. This is because collocations differ from one sense to another as well as from one case particle to another. Collocations are ordered in the sequence of collocation frequency deduced using the BCCWJ corpus browsing tool called NINJAL-LWP for BCCWJ (NLB for short). As a statistical index expressing the strength of a collocation, a score called "Mutual Information (MI)" score is available, however the MI score tends to cull expressions involving high degree of idiomaticity, so we decided to use raw frequency as a criterion for the purpose of listing collocations. Arranging collocation based on the raw frequency deduced from NLB ensures objectivity and authenticity. However, on the other hand, owing to the limitation on the size of the corpus (65 million words in NLB, 100 million words in BCCWJ) there is no guarantee that all the collocations needed to be listed in the dictionary are culled without any leakage. Therefore, some collocations which do not appear in the NLB, but which are thought to be necessary for learners are added. This measure, to a large extent, depends on the experience of the editor. In future, if the size of the corpus is increased, it is expected that the selection of collocations on the basis of the frequency criterion would become easier. For this purpose, the Tsukuba WEB Corpus (TWC) with a projected ten times the entries of BCCWJ is under preparation.

〔Wrong collocations〕

Here collocations which are prone to lead to wrong usage are described.

〔Examples: coined examples〕

For each meaning/sense we provided more than 3 coined examples. In order to avoid examples ending only with dictionary form (plain style, non-past), we have made a deliberate attempt to coin examples involving variation of tense, aspect, voice, modality etc. Such a move also helps to enhance naturalness of examples. Quite often we have even used complex sentences as well.

〔Examples: from corpus〕

We have provided examples culled from the BCCWJ corpus as well. The purpose of providing examples from corpus is to provide examples that are natural in the context of situation in question. However, on the other hand there is the criticism that such examples are difficult for non-natives to comprehend. The same observation has been made during the process of compilation of this handbook as well. It has been pointed out that real examples from a corpus are hard to comprehend unless one has sufficient knowledge of socio-cultural background. It became clear in our handbook that translation of such examples into another language is a big obstacle. Especially, considering the typically High Context Communication (Hall, 1976) nature of Japanese, it is easy to imagine that the problems of real examples would be much graver than in English. Whether to stick to real-examples only or to allow coined examples for the point of view of second language education is a complex issue with no satisfactory solution. At present, taking merits of both, we have decided to include natural examples as well as tailored examples. However, since the translation of natural examples is an extremely difficult task, we have decided not to translate the corpus examples.

〔Information on wrong usage: in the case of specific meanings〕

Mistakes that learners tend to make often are described under this heading. For information on wrong usage by JFL learners, various databases including Teramura database (http://teramuradb.ninjal.ac.jp/) are used. However, since these corpora are developed individually, the size of each of them is rather small and it is difficult to deduce general patterns of mistakes from them. Under such circumstances we have to heavily rely on the teaching experience of the editor. The following are examples from learners' corpora:

Spoken language corpus:

> 発話対照データベース、生活対照データベース (taiwa taishou detaabeesu, seikatsu taishou deetabeesu)
>
> 日本語学習者会話データベース(nihongo gakushuusha kaiwa deetabeesu)
>
> 日本語学習者会話ストラテジーデータ (nihongo gakushuusha kaiwa sutoratejiideeta)
>
> KY コーパス (KY koopasu)
>
> タグ付き KY コーパスと検索ツール(tagutsuki KY koopasu to kensaku tsuuru)
>
> BTS による多言語話し言葉コーパス (BTS ni yoru tagengo hanashikotoba koopasu)
>
> インタビュー形式による日本語会話データベース（上村コーパス）(intabyuu keishiki ni yoru nihongo kaiwa deetabeesu (Uemura koopasu))

Written language corpus:

寺村誤用例集データベース (Teramura goyou reishuu deetabeesu)

日本語学習者言語コーパス (nihongo gakushuusha gengo koopasu)

作文対訳 DB (sakubun taiyaku DB)

自然言語処理の技術を利用したタグ付き学習者作文コーパス (shizengengoshori no gijutsu wo riyou shita tagutsuki gakushuusha sakubun koopasu)

日本・韓国・台湾の大学生による日本語意見文データベース (nihon/kannkoku.taiwan no daigakusei ni yoru nihongoikenbun deetabeesu)

JLPTUFS 作文コーパス(JLPTUFS sakubun koopasu)

In addition to the above list, there are many corpora which are either not made public or are accessible to only few individuals. For the effective use of intellectual resources, it is desired that an organization like NINJAL take the lead in the development of a platform like CHILDES (Child Language Data Exchange System) which allows accumulation of data in a common platform.

〔Grammar〕

Here we have shown the behavior of the verb with respect to grammatical categories like aspect, voice, tense etc. A conclusion is still not reached on whether to include categories like direct passive, indirect passive, imperative form, other sentence-final expressions. Further, whether to make judgments on grammaticality of such categories based on intuitions of individuals or on the basis of corpus frequency is also not yet decided. For making judgments on grammaticality (especially the subtle ones, shown by triangle sign) on the basis of corpus frequency, the size of the BCCWJ corpus seems not to be sufficient.

〔Compounds〕

Compound words are too large in number and hence it is impractical to include all of them. If so, again one has to decide on the basis either of intuition or of corpus frequency in order to decide potential candidates that should be listed. We would like to make use of the corpus for this and at present are using frequency as a criterion for listing compound words.

〔Idioms and proverbs〕

Idioms and proverbs consist of elements which are tightly bound together and the meaning of the whole cannot be guessed from the combination of the meanings of the parts. In other words, it can be said that semantic transparency is low in the case of idioms and proverbs. However, the transparency is a gradient concept and the decision of collocation or proverb is bound to be arbitrary. One yardstick for this decision can be MI (Mutual Information) score. The higher the degree of idiomaticity the greater the MI score (see section 4.1.2).

〔Semantic network〕

The relationships among meanings/senses are visually shown with the help of a radial category network diagram. The basic or central meaning is the one that is known in cognitive linguistics as the prototypical meaning. The relationships among meanings/senses are visually shown with the help of a radial category network diagram. The basic or central meaning is the one that is known in cognitive linguistics as the prototypical meaning. Derivations from it are arranged in a way to be understood intuitively. These semantic derivations themselves are products of linguistic research. Many cognitive linguists are also involved in this project. However, there is no guarantee that the semantic derivations are determined on the basis of a single meaning. Also the sequence of diachronic change and synchronic relationship often do not match. In view of these considerstions, while insights from cognitive linguistics form the basis of description, often changes have been made in favour of intuitive understanding. There are places where accuracy of description from the point of cognitive linguistics conflicts with intuitive understanding. In such cases we have preferred educational considerations such as ease of understanding for teachers and learners.

As for the network, showing just the connection is not enough. The strength of the connection should also be shown. We are thinking of showing the strength or weakness of the connections visually in terms of the thickness of the line or the distance between the senses so as to foster understanding in a visual and intuitive way.

〔Related words (word family)〕

At present, we have listed words with almost the same meaning and synonyms as related words. Listing of antonyms is also under consideration. We are thinking of presenting the word family in the form of a radial category network, if possible.

4. Developing tools for corpora of correct usage and wrong usage

One of the important policies we adopted to create this handbook is to make good use of available corpora. To compile a corpus-based handbook or dictionary, the existence of tools which enable dictionary writers to use corpora adequately and efficiently in the process of dictionary making is indispensable. In this project we chose the Balanced Corpus of Contemporary Written Japanese (BCCWJ) as a corpus of correct use by Japanese natives and the *Gaikokujin gakushuusha no nihongo goyoureishuu* (*Collection of errors of JFL learners*, 1990), compiled by Hideo Teramura and his colleageus, as a corpus of wrong usages of JFL learners. We developed search tools for each of these corpora. In the following two subsections, we will describe the features and functions of both tools.

4.1 NINJAL-LWP for BCCWJ (NLB)

NINJAL-LWP for BCCWJ (NLB, http://nlb.ninjal.ac.jp) is an online search tool for the BCCWJ, jointly developed by the National Institute of Japanese Language and

Linguistics (NINJAL) and Lago Institute of Language (LIL). The basic unit of this system is LagoWordProfiler (LWP), which LIL has developed for dictionary writing and editing. LWP has been successfully utilized in several projects of English-Japanese, Japanese-English dictionary making.

Figure 1: The headword Window of NLB

BCCWJ is the first balanced corpus of the Japanese language, developed by NINJAL, and its final version was made public at the end of 2011. It is a large corpus of more than 100 million words, the size of which is comparable to the British National Corpus. The main component of the corpus consists of random samples from books, newspapers, magazines using rigid statistical methods to establish representativeness. Nine additional sub-corpora are provided for special purposes, including web text, which shows different usage patterns from those of text of the print media (Maekawa, 2012).

4.1.1 Lexical profiling

The most important feature of NLB is its introduction of the lexical profiling methodology. Lexical profiling is now a standard method for making corpus-based dictionaries because it satisfies the requirements for using corpora in dictionary making. A concordancer used to be a standard tool in the earliest corpus lexicography. On the COBUILD Project, which made extensive use of corpora for the first time, the writing staff wrote headword entries by analyzing concordance lines from a concordancer (Sinclair, 1987). Concordance lines enable the dictionary writer to analyze individual words in real context. However, the larger the number of lines, the more difficult it is to grasp the whole variety of linguistic phenomena. To solve this difficulty, lexicographers realized the importance of summarizing linguistic phenomena comprehensively by use of abstraction (lemmatization, POS tagging, and chunking) and statistical measures (the MI score, the T score, etc.). In this process,

lexical profiling as a new approach gradually developed (Church et al., 1991). At the end of the 1990s, a practical lexical profiling tool called *Word Sketch* appeared (Kilgarriff & Rundell, 2002). This software was first used for compiling *Macmillan English Dictionary for Advanced Learners*, and then it developed into the integrated system *Sketch Engine*, which is now used in many dictionary projects.

Lexical profiling has two important requirements. The first is comprehensiveness. Linguistic research, in general, focuses on a particular linguistic behavior and adopts an approach that examines individual instances carefully and thoroughly. On the other hand, what dictionary making requires is to examine each headword's overall behavior. A dictionary writer needs to grasp a headword's behavior as comprehensively as possible. When implementing a search tool, which patterns to extract and how to classify those extracted patterns are vital keys to ensure comprehensiveness.

The other key is time efficiency. This is essential in dictionary making. The number of headwords in a dictionary range from several thousand to one hundred thousand. To make best use of a corpus when writing a large number of headwords, an environment that enables dictionary writers to use a corpus efficiently is indispensable. Key factors to realize this environment include search speed and a user interface.

4.1.2 Lexical profiling in NLB

So how does NLB satisfy the requirements of lexical profiling? As to comprehensiveness, NLB deals with the orthographical variety of the Japanese language. Japanese is usually written in three types of characters: *hiragana*, *katakana* and *kanji*. This means a word could be written in at least three ways. The noun *hito*, which means *a person*, can be written as ひと in hiragana, or ヒト in katakana, or 人 in kanji, with different connotations. In the case of compound verbs, things are complicated by the fact that some verbs have two or more kanji candidates with slightly different meanings. The compound verb 取り上げる (toriageru), which means *pick up* or *adopt*, can also be written as 採り上げる. Including a variation of kana suffixes, more than ten orthographical forms for トリアゲル are possible. From the point of view of comprehensiveness, it is, in many cases, more appropriate to group two or more orthographical variants into the most typical orthographical form than to give each form a headword status. NLB deals with this issue by incorporating the idea of representative orthographical form. In the previous example of 取り上げる, more than ten orthographical forms are all grouped into the

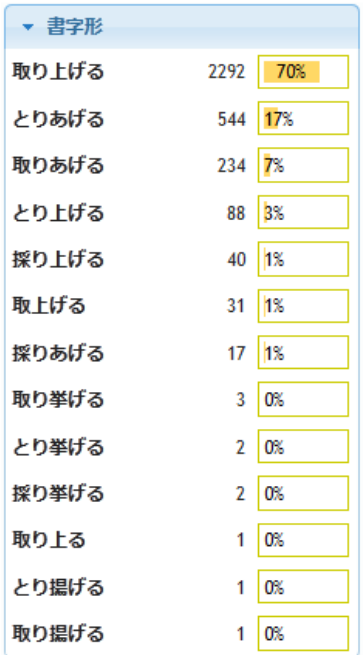


Figure 2: Orthographical forms for *toriageru*

representative form 取り上げる, which consists of a headword entry. Figure 2 shows the frequency distribution of orthographical forms for 取り上げる in BCCWJ.

In order to maximize time efficiency, NLB has a user interface that allows the user to examine grammatical patterns, collocations, and examples from the corpus in the same window (See Figure 1). On *Sketch Engine*, which we mentioned earlier, a screen transition occurs every time the user looks for examples for each collocation. A user interface with frequent screen transitions is problematic from the point of view of time efficiency. With the recent spread of large screen displays, it is not so difficult as before to introduce a user interface with a minimum of screen transitions. Although user interfaces for corpus search tools have not been given much attention until recently, its importance is expected to increase as the size of corpora increases and more sophisticated search functions are implemented. Search speed is another important factor closely related to time efficiency. NLB shows the results of collocations and examples almost instantly by optimizing the structure of the database.

Another important feature of NLB is its function to sort collocations by raw frequency and other statistic measures such as the MI-score and the logDice score. Figure 3 shows collocations of N を買う(*N wo kau*, to buy *N*). In the upper part of the figure, collocations are ordered by raw frequency, and in the lower part, by MI score. The MI score has a tendency to be unreliably high among low-frequency collocations. To avoid this reliability issue, NLB provides a filter function to remove low-frequency collocations. In the lower part of Figure 3, low-frequency collocations of less than five instances are excluded from the list. You can see idiomatic expressions like 顰蹙を買う(upset someone), 歓心を買う(seek someone's favor), 失笑を買う(make someone laugh at you) are top of the list. Sorting collocations by multiple statistic measures is an extremely useful function.

...を買う 1264件

コロケーション	頻度	MI	N-S
ものを買う	177	3.92	-2.29
本を買う	119	5.72	-0.30
【一般】を買う	118	2.12	-0.26
車を買う	99	6.38	0.16
物を買う	96	6.33	0.57
株を買う	87	8.42	-1.11
切符を買う	79	10.77	-0.48
家を買う	74	4.43	-0.23
のを買う	71	1.03	-1.99
土地を買う	66	6.84	-0.46
それを買う	59	2.64	-0.91
服を買う	56	7.85	-0.63
品を買う	50	6.70	0.12
商品を買う	49	6.27	0.32
券を買う	46	8.73	0.50
顰蹙を買う	46	13.01	-0.48

Page 1 /13 100

...を買う 180件

コロケーション	頻度	MI	N-S
顰蹙を買う	46	13.01	-0.48
歓心を買う	18	13.01	-0.40
不興を買う	25	12.69	-0.04
失笑を買う	9	11.69	-0.22
不評を買う	10	11.69	-0.02
反感を買う	39	10.97	0.14
切符を買う	79	10.77	-0.48
パンプスを買う	6	10.21	0.02
馬券を買う	21	9.78	0.14
土産を買う	45	9.70	0.21
宝くじを買う	12	9.69	-0.40
おみやげを買う	9	9.52	-0.04
一役を買う	5	9.46	0.09
怨みを買う	34	9.42	-1.27
ウーロン茶を買う	5	9.21	0.11
チケットを買う	23	9.03	0.16

Page 1 /2 100

Figure 3: Collocations of *N wo kau*

NLB also facilitates creating examples with dictionary-making-oriented functionality. On the example panel (the right-most panel of Figure 1), examples for a collocation are shown in ascending order of their character counts. This helps the dictionary writer to use corpus examples for reference easily and effectively. Each corpus example is color-coded according to the sub-corpus it belongs to, which enables the writer to know where each example comes from quickly. In addition, the writer can examine the context of a corpus example just by clicking its source information label.

As we have seen, NLB provides an ideal environment for Japanese dictionary making, by dealing with the wide variety of orthographical forms in Japanese, and offering a user-friendly interface.

4.2 The Teramura Wrong Usage Database

Gaikokujin gakushuusha no nihongo goyoureishuu (*Collection of errors of JFL learners*) is a report compiled by Teramura Hideo and his team in the late 1990s, after they collected and classified misuse samples from compositions written by overseas students from 24 countries. The total of the misuse samples amounts to 6,300, with misuse labels attached to misuse positions. Other information includes learner's nationality and composition type.

The online version of this report, Teramura Wrong Usage Database provides a search function. The user can search misuse examples by combining conditions (a type of misuse, a learner's nationality, a composition type, etc.) Figure 4 shows the "search

from misuse type" function. Misuse types are shown in a tree structure, effectively informing the user of how many misuse instances there are for each type on any combination of nationalities and composition types.

Figure 4: Teramura Wrong Usage Database

Most conventional Japanese dictionaries for native speakers and foreign learners, including ones with a learning or teaching purpose, only show correct usages; very few show wrong usages. This tool enables us to include useful wrong usage information for learners such as wrong collocations in a definition entry.

5. Crossing the barriers of space and time: An online multi-lingual editing tool

Compiling a dictionary requires a lot of time and human resources. It is usually the case that there is an editor-in-chief who directs lexicographers in charge of writing up entries. The editor-in-chief proofreads the entries that the lexicographers have written, and corresponds with them as often as necessary. Proofreading may be done by different proofreaders and the editor-in-chief manages the editorial activity. This process usually takes a long time, and is not ideal if time for the compilation is limited. Another drawback of this traditional system is that lexicographers will usually have no chance to examine entries that the other lexicographers write.

To overcome these problems, we have developed a web-based editing system so that the editors, lexicographers and proofreaders can have access to the entry data for editing, reviewing and proofreading processes.

To develop the current online editor system, our experience in compiling A Dictionary of Basic Verbs in Japanese for Marathi, the outcome of Prashant et al. (2007)'s project is fully exploited. Under a limited budget, we made use of free applications to achieve our goal: a wiki system to store the entry data in XML format. Wiki is a system for collaborative editing online and has a repository system, under which all older versions of wiki pages are stored. By comparing the current version with one of the older versions, editors can tell what have been changed, deleted and/or added in the latest version. In this new system, we take advantage of the repository feature of wiki.

In the current system, the lexicographers write entries in Japanese first. Then the Japanese entries are translated into four foreign languages (Marathi, Korean, Chinese and English) by translators. At this stage some additional information will be added that is related to cultural and linguistic differences between Japanese and the target language.

The following sub-sections give a brief outline of the online editorial system.

5.1 An outline of the online editorial system

5.1.1 Some features of the online editor

The online editor developed for this project has the following features:

- Data are input in a textbox area on the editor and stored in an XML structure.
- The data input in the editor are reflected in a preview function to check how they look in the HTML format instantaneously.
- Employing Yahoo API, it is possible to assign furigana, the phonetic transcription of kanji, in a format that may be convertible into other formats like HTML.
- The lexicographers can read the entries that are written by the others online and post a comment, which will be shared by all editors.

5.1.2 Online editor as a plug-in of Dokuwiki

The editor is not a standalone application but is developed as a plug-in for Dokuwiki. Dokuwiki is a Unicode-based wiki application and does not require a binary database system like SQL because data pages are saved in text files. Each entry is organized in an XML format and stored as a Dokuwiki page. Since the file is a text file, it can be directly used as an XML file for data-processing.

The lexicographers first login to the Dokuwiki homepage as in Figure 5.

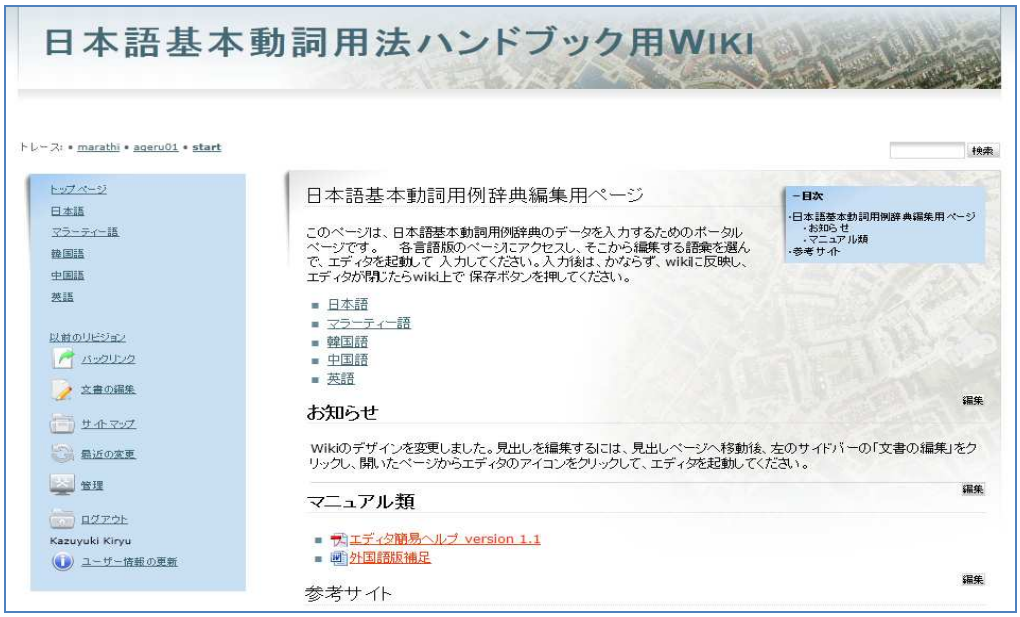


Figure 5: The homepage of the editorial system on Dokuwiki

5.1.3 Starting the online editor

After logging in, lexicographers choose the language, and then select one of the entries in the list to edit it. The Wiki page shows the XML data of the entry, but it is not directly edited. They start the plug-in online editor. On starting up, the editor retrieves the XML data from the Wiki page. The view of the entry data is formatted in an Explorer view, with a tree structure displayed on the left pane and each sub-data displayed on the right pane, as in Figure 6.

Figure 6: A full view of the online editor

Figure 7 shows the view when one of the items is selected and its editing area is displayed on the right pane.

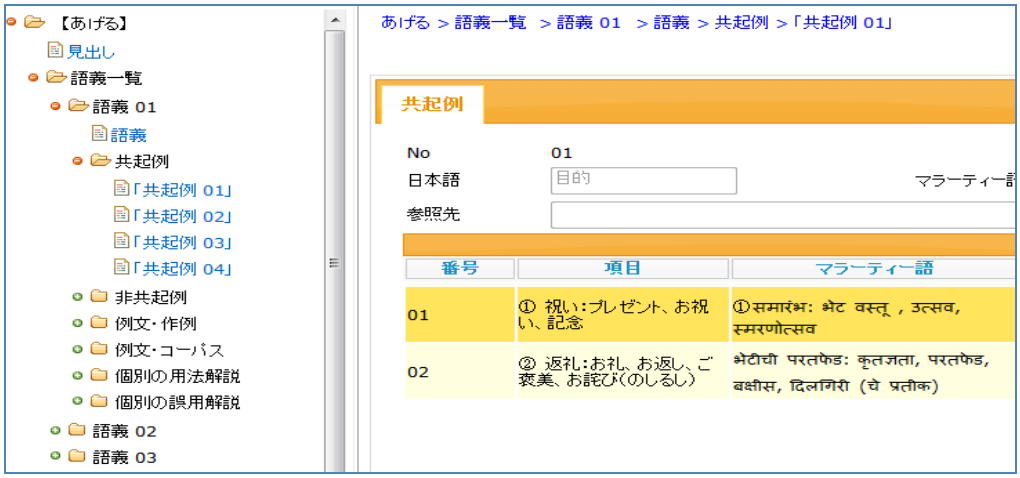


Figure 7: The editing pane for Collocation 01（共起例 01） is open on the right page

5.1.4 The preview function

The editor has a preview function. There are two types of preview: the entire view of the entry and the partial view of an item of the entry. The preview is generated via XSLT as an HTML page. An image of the full-scaled preview in Marathi is shown in Figure 8, and an image of the partial view is shown in Figure 9. Since it is a bilingual version, both the Japanese data and the respective Marathi data are shown. In the bilingual version, as shown in Figure 8, an additional piece of information from a contrastive point of view（対照情報）is also provided when necessary. This information will not be included in the Japanese version.

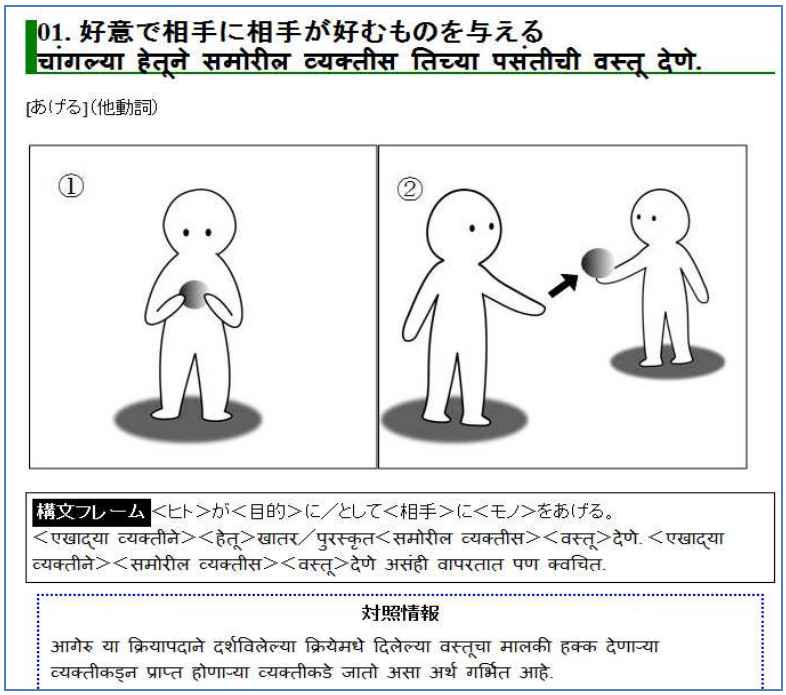


Figure 8: The full-scaled preview of the Marathi translation of *ageru*

The layout design of the preview in Figures 8 and 9 is not intended to be final, but to be temporary just for convenience. The final layout design will be developed differently and be applied to generate the final product from the same XML data.

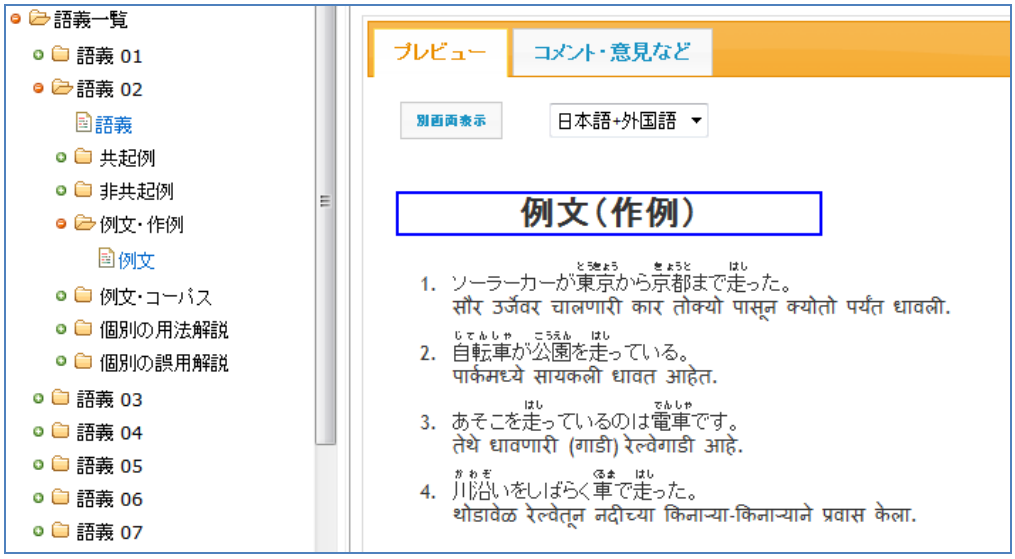


Figure 9: Built-in partial preview of portion of examples

5.1.5 Comparison of different versions

Dokuwiki's revision control makes it possible to compare the latest version with any older version. When two versions are compared, differences will be displayed. This is one of the major merits in using Dokuwiki for entry data management.

5.1.6 Posting comments and improving the quality of description

The editor has a function of posting a comment on the data, while editing or reviewing. Comments are sent to all the editorial members to share the information by email. The comments can also be viewed on the editor and follow-up comments posted. Through this process, the editorial members can exchange ideas and opinions about entry data so that the lexicographer in charge can improve the quality of the descriptions and examples.

6. Audio-visual contents

Taking into consideration the cognition and memorization process in learning new words and meanings, we have incorporated audio-visual contents in the handbook. We believe audio-visual contents facilitate the comprehension and memorization of various meanings of a verb. A brief discussion of the audio-visual contents is provided below.

6.1 Audio-visual contents

In the present handbook, before presenting specific meanings of a polysemous verb, we first provide an abstract image schema which represents the core, shared meaning of the verb in question. Following this, a radial semantic network of the various meanings of a polysemous verb is provided. These two visual contents set the stage for zooming into a specific meaning. On moving to a specific meaning, we provide an animated illustration of the representative example of that meaning. The animated illustrations are a set of still hand-drawn pictures which are connected in such a way that they depict the semantic scenario as it unfolds in time. Audio contents are also added to the animated illustrations. In addition to the abstract image-schema animated illustration depicting a specific meaning, we also provide video-clips as well. A brief description of these three audio-visual contents is given below.

6.1.1 Image schema

The verbs included in the present handbook are all fairly polysemous. For example, the entry of *agaru* "rise/go up" in our handbook has as many as 19 meanings. In the cognitive linguistics perspective all these meanings are considered to share a core or prototypical meaning which is illustrated with the help of an abstract line diagram which is widely referred to as an image schema. For example, in the case of

the verb *agaru* “rise/move up” “motion of an entity in a upward direction” is taken to be the core meaning and it is illustrated with the help of an image schema shown in figure 10.

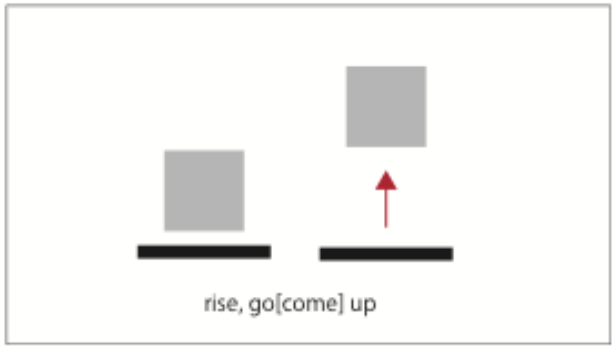


Figure 10: Image schema for the verb *agaru* “rise/move up”

All the meanings are derived from this prototypical meaning through semantic extensions of various types. Image schema would be useful for learners to understand the core or prototypical meaning of a polysemous verb and also to appreciate the connection with specific meanings.

6.1.2 Animated illustrations

As mentioned earlier, we have included animated illustrations in order to facilitate comprehension and retention of specific meanings. From previous research (Dwyer, 1978; Lin & Dwyer, 2004; Dwyer, 2007; Chou & Hsiao, 2010), it has become increasingly clear that the static visual instruction serves as a powerful learning strategy and improves information acquisition and retrieval capabilities. Using these insights, rather than using multimedia contents, we use multiple hand-drawn animations and we show them in a sequence along with audio contents synchronized with them as shown in Figure 11.

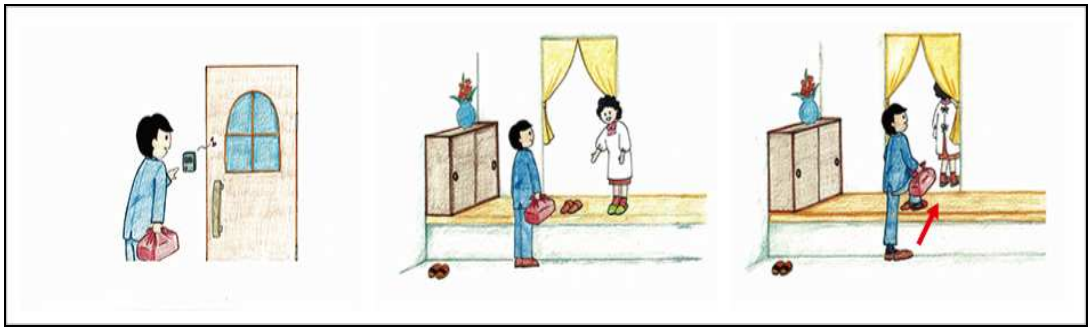


Figure 11: Animated illustrations for *ie ni agaru* “to visit someone”

6.1.3 Video clips

Contextbased information such as deixis (e.g. the use of auxiliary verb indicating the location of the speaker in expressions such as *agatte iku/kuru* (come/go up)), or the resultant state conveyed by *-te iru-* form as in the expressions such as *hata-ga agatte iru* (the flag is raised), or the perfectly fried, crisp *tenpura* as depicted by the adverb *karatto* and the like can be effectively conveyed using a video clip. Wherever necessary and feasible, we have tried to provide video-clips to foster understanding of subtle meanings. Figure 12 below illustrates the expression *kaidan o agatte iku* (go up climbing the stairs) wherein the scenario is shot from the backside of the person climbing the staircase to induce the viewers "viewpoint" in the interpretation of the scene.

Figure 12: Video clip depicting *kaidan o agatte iku* (go up climbing the stairs)

7. Future prospects

From the foregoing discussion it should be clear that the handbook in preparation differs in many respects from bilingual dictionaries available now. The content of the entries is based on information gleaned from corpora and is augmented with insights from various sub-fields of linguistics, especially cognitive linguistics and contrastive linguistics. Further, the handbook includes audio-visual contents in order to improve information acquisition and retrieval capabilities.

The handbook will be made available for free access on internet around April 2013. After getting feedback from JFL learners and teachers of Japanese in various parts of the globe, we plan to make improvements both in content as well as presentation. We also plan to increase the number of headwords and the target languages beyond English, Chinese, Korean and Marathi, if collaborators are willing to volunteer their services. Finally, we strongly believe that the output of the handbook project will make a substantial contribution not only to Japanese language research and Japanese language pedagogy but also to corpus linguistics, contrastive linguistics and linguistics in general.

References

Church, K., Gale, W., Hanks, P. & Hindle, D. (1991). Using Statistics in Lexical Analysis. In: Sernik, E. (Ed.) *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, 115-164. New Jersey: Psychology Press.

Chou, P., & Hsiao, H. (2010). The Effect of Static Visual Instruction on Students' Online Learning: A Pilot Study. *Interdisciplinary Journal of Information, Knowledge, and Management.* 5. [Available online: http://www.ijikm.org/Volume5/IJIKMv5p073-081Chou456.pdf]

Dwyer, F. M. (1978). *Strategies for improving visual learning*. State College, PA: Learning Services.

Dwyer, F. M. (2007). The program of systematic evaluation (PSE): Evaluating the effects of multimedia instruction 1965-2007. Educational Technology, XLVII(5), 41-45.

Goldberg, A. F. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Hall, E. T. (1976). *Beyond Culture*. Anchor books.

Kilgarriff, A., & Rundell, M. (2002). Lexical Profiling Software and its Lexicographic Applications: A Case Study. In: Braasch, A. & Povlsen, C. (Eds.) *Proceedings of the Tenth EURALEX Congress*, 807-819. Copenhagen.

Lin, C., & Dwyer, F. M. (2004). Effect of varied animated enhancement strategies in facilitating achievement of different educational objectives. *International Journal of Instructional Media*, 31(2), 185-198.

Maekawa, K. (2012). Gendai kakikotoba kinkou koopasu (BCCWJ) no kouchiku to KOTONOHA keikaku no ayumi (The construction of "the Balanced Corpus of Contemporary Written Japanese (BCCWJ)" and the progress of the KOTONOHA plan). *Nihongengogakkai dai 144 kai taikai yokoushuu* (Proceedings of the 144th meeting of the Linguistic Society of Japan), 352-357.

Pardeshi, P., & Akasegawa, S. (2011). BCCWJ wo katsuyou shita kihondoushi handobukku sakusei: koopasu buraujingu shisutemu NINJAL-LWP no tokuchou to kinou (Compilation of basic verbs handbook using the BCCWJ corpus: Salient features and functions of the corpus browsing system NINJAL-LWP). *Gendai kakikotoba kinkou koopasu (BCCWJ) kansei kinen kouenkai yokoushuu* (*The proceedings of the symposium commemorating the completion of "the Balanced Corpus of Contemporary Written Japanese (BCCWJ)"*), 205-216. Tokyo: National Institute for Japanese Language and Linguistics.

Pardeshi, P., & Kiryu, K. (2007). "Nihongo-Maraathiigo kihondoushiyouhoujiten" sakusei purojekuto: indo ni okeru nihongo kyouiku no kisozukuri ni mukete ("Japanese-Marathi Basic Verb Dictionary" Compilation Project: A step toward the construction of the foundation of Japanese language education in India). *Koube daigaku ryuugakusei sentaa kiyou* (*Bulletin of Kobe University International Student Center*) 13: 87-102.

Sinclair, J. (1987). *Collins Cobuild Dictionary English Language Dictionary*. London: Harper Collins Publishers.

Sinclair, J. (Ed.) (1987). *Looking Up, An account of the COBUILD Project in lexical computing*. London: Collins ELT.

Teramura, H. (1990). *Gaikokujin gakushuusha no nihongo goyoureishuu* (*Collection of errors of JFL learners*). Report of grant-in-aid study [available online at: http://teramuradb.ninjal.ac.jp/teramura.goyoureishu.pdf].

ITADICT PROJECT AND JAPANESE LANGUAGE LEARNING

Marcella MARIOTTI
Ca' Foscari University of Venice
mariotti@unive.it

Alessandro MANTELLI
Ca' Foscari University of Venice
mantrex@gmail.com

Abstract

This article aims to show how the Nuclear disaster in Fukushima (3 March 2011) affected Japanese Language teaching and learning in Italy, focusing on the ITADICT Project (Marcella Mariotti, project leader, Clemente Beghi, research fellow and Alessandro Mantelli, programmer). The project intends to develop the first Japanese-Italian online database, involving more than 60 students of the Japanese language interested in lexicographic research and online learning strategies and tools. A secondary undertaking of ITADICT is its Latin alphabet transliteration of Japanese words using the Hepburn system of romanization. ITADICT is inspired by the EDICT Japanese-English database developed by the Electronic Dictionary Research and Development Group established in 2000 within the Faculty of Information Technology at Monash University. The Japanese-Italian database is evolving within the Department of Asian and North African Studies at Ca' Foscari University of Venice, the largest in the country and one of the main teaching centres of Japanese in Europe in terms of the number of students dedicated to it (more than 1800) and number of Japanese language teaching hours (1002h at B.A. level, and 387h at M.A. level).

In this paper we describe how and why the project has been carried out and what the expectations are for its future development.

Keywords

ITADICT; Japanese-Italian database; lexicography; Japanese language; online database; collaborative editing; Japanese language learning

Izvleček

Pričujoči članek predstavlja projekt ITADICT (vodja projekta Marcella Mariotti, sodelujoči raziskovalec Clemente Beghi, programer Alessandro Mantelli) in vpliv nuklearne katastrofe v Fukushimi 3. marca 2011 na učenje japonščine v Italiji. Cilj projekta je razvoj prve spletne japonsko-italijanske baze podatkov, pri njem pa sodeluje več kot 60 študentov japonščine, ki jih zanima slovaropisje in učne strategije ter orodja na spletu. Drugi cilj projekta ITADICT je prečrkovanje japonskih besed v latinico, po sistemu Hepburn. Projekt je zastavljen po vzoru japonsko-angleške podatkovne baze EDICT, ki jo je razvila skupina Electronic Dictionary Research and Development Group (skupina za raziskovanje in razvoj elektronskih slovarjev), ki je bila ustanovljena leta 2000 na Fakulteti za informacijsko tehnologijo na Univerzi Monash. Japonsko-italijanska baza podatkov se razvija na Oddelku za azijske in severno-afriške študije na Univerzi Ca'Foscari v Benetkah, eden od glavnih centrov za učenje japonščine v Evropi in

največji v Italiji po številu študentov (1800) in številu učnih ur japonščine (1002 na prvi stopnji in 387 na drugi stopnji študija).

Članek predstavlja ozadje in način izpeljave projekta ter načrte za prihodnji razvoj.

Ključne besede

ITADICT; japonsko-italijanska baza podatkov; leksikografija; japonski jezik; spletna baza podatkov; sodelovalno urejanje; učenje japonščine

1. ITADICT Project and Japanese Language Learning

The ITADICT Project (http://virgo.unive.it/itadict/eng/about), is aimed at the creation of a freely accessible Japanese-Italian database, and is expressly inspired by Jim Breen's JMdict/EDICT Project that initiated in 1991 at Monash University.

The database was started separately by both Marcella Mariotti (Ca' Foscari University of Venice) and Clemente Beghi (Ca' Foscari University of Venice) between 2007 and 2008 as part of their research. At the time Beghi was a Ph.D student at Cambridge University doing research on Esoteric Buddhist Iconography, so he edited Buddhist and, for other reasons, Floral terms. In the mean time, Mariotti was a JSPS post-doc researcher at International Christian University (Tokyo), where she needed an Italian translation and transliteration in Latin alphabet of all the words present in her Hypermedia Dictionary of Japanese Grammar BunpoHyDict (Mariotti 2008), so this was her starting point for editing more than 3000 words in the database.

They are both grateful to Jim Breen (Monash University) who brought their research interests together.

ITADICT became one unified voluntary project coordinated by Marcella Mariotti, at the end of 2010, when Beghi and Mariotti were both teaching Japanese Language at the Department of East Asian Studies (now Department of Asian and North African Studies) at Ca' Foscari University of Venice. In one and a half years, they involved more than 60 of their students, who became an integral part of the project, actively translating terms from Japanese into Italian and inserting them in the ITADICT EDITOR later developed by Alessandro Mantelli.

2. Translating students: Why involving them?

The strategic role of pleasure in the long-term acquisition processes of a foreign language has been stressed by neurolinguistics and researchers such as Danesi (2003), Schumann (2006) and Balboni (2002). Moreover, the more social networks, eLearning and mLearning sites and applications spread around on PCs and smartphones, the more students are fascinated, aware of and concerned with the object of their studies,

proving that "learning is like a utility - like water or electricity - that flows in a network or a grip that we tap into when we want". (Downes 2007)

According to a survey conducted in 2009 (Ferrari, 2010) students at Ca' Foscari often approach Japanese Language studies not intentionally, as a conscious part of a wider life-plan they have, but more as a way to foster their curiosity, to feel closer to "the real thing": the original language of loved novels, movies, animated movies, manga, *dorama*, inspired sutras or martial arts.

Maybe due to a disenchantment felt by students of Japanese in Italy in the new century, following the economical crisis of Japan, this emotional motivation is quite specific to the learners of the net-generation (Mariotti, forthcoming, Miyake 2012), while far removed from those of the Nineties.

The above may explain why, as soon as Mariotti introduced ITADICT project to her students as part of a presentation about fansubbing and language learning (Mariotti, 2011), many of them were willing to participate. Their motivations were diverse: mainly they were interested in creating a tool for translating Japanese into Italian using a mouse-over dictionary[1], and in learning strategies to use the online research tools on Japanese language sites to conduct lexicographical research (Mariotti 2012[2]).

A secondary motivation followed. As of 2012, Ca' Foscari students need to complete a period of internship before they can graduate. In 2011 this internship was "warmly suggested", and students received 5 or 6 University Credits for it. Particularly because of the nuclear incident in Fukushima following the tragic 3.11 earthquake and tsunami in North Eastern Japan, Ca' Foscari students were discouraged from applying for an internship in Japan as they had usually done. As a result, 42 out of the 64 students collaborating on ITADICT were 2011 prospective graduates who did not know where to complete their internship and, intrigued by the ITADICT project, chose to do so by taking part in the project.

3. What is ITADICT?

As mentioned, ITADICT was born out of different needs, largely mirroring those of the original Japanese-English EDICT project, which started in 1991 from MOKE (Mark's Own Kanji Editor), a word processor with integrated Japanese-English Dictionary (Breen 2010). ITADICT project focused on creating a file for a Japanese-Italian database that could be used by third party software to easily read and translate Japanese texts (e.g. Rikaichan – popup dictionary tool for Firefox browser-, Japan

[1] E.g. http://www.polarcloud.com/rikaichan/

[2] Online anonymous survey *Why ITADICT*?, addressed to the 66 students and collaborators who worked on ITADICT (Sept. 2012).

Goggles -iPhone app to translate words from live camera-, Kotoba/Imiwa? – iPhone dictionary-app to manage Japanese-Other Languages databases-, and more).

In 1999 the EDICT project, which had been limited by very a simple dictionary structure, evolved into the more complex JMDict Project JMDict (Japanese-Multilingual Dictionary) Project (managed by the EDRG Electronic Dictionary Research Group). (Breen 2004)

JMDict employs an XML structure to support a much richer dictionary entry format including multiple kanji surface forms and readings. The original EDICT format is generated from this project as a legacy format mainly for older software packages. An expanded "EDICT2" format is also generated which more closely follows the XML content. For our purposes we started using the simpler "Traditional" EDICT file where there is only 1 kanji form and 1 reading per entry/line in plain text, with less marking of different semantic fields than in the newer EDICT2:

KANJI [KANA] /(PoS tag) gloss/gloss/...

The file had about 160.000 entries (with one line per entry), where most common entries had a (P) mark for "priority" at the end of the line. Breen's online pages describe the process used to determine the priority of a term, mainly marked after a) Alexandre Girardi's (NAIST-MULTITEL) match-analysis between EDICT entries, the 1994-1998 corpus of Mainichi Shinbun, and b) the 10,000 common words in the collection *Ichimango goi bunruishū* (Senmon kyouiku Publishing 1998). Although, as Breen underlines:

> While the priority markings accurately reflect the status of entries with regard to the various sources, they must be seen as only providing a crude indication of how common a word or expression actually is in Japanese. The "(P)" markings in the EDICT and EDICT2 files appear to identify a useful subset of "common" words, but there are clearly some marked entries which are not very common, and there are clearly unmarked entries which are in common use, particularly in the spoken language. (Breen, 2010)

3.1 Latin alphabet transliteration according to the Hepburn system

Since our purpose was to allow as many people as possible to approach the Japanese language and to enable, let us say, a primary school teacher to say some words to a Japanese-native-speaker child at school even without knowing *kanji* or *kana*, we added a new characteristic to Breen's EDICT format: the Latin alphabet transliteration of the *hiragana* readings in brackets. ITADICT line, then, appeared as follows:

KANJI [KANA, latin] /(PoS tag)/gloss/gloss/...
おい付く [おいつく, oitsuku] / (v5k,vi) raggiungere/uguagliare/arrivare al livello di/(P)/

This decision was followed by a heated debate, and since the database was developed inside university academia, and the whole process itself was part of a teaching/learning project, we adopted the Hepburn system of transliteration and did not rely on automatic transliteration tools (e.g. Romaji Translator), but rather chose to manually transcribe each entry.

3.2 Use of monolingual JA-JA dictionary

With the intent to not only produce an accurate Japanese-Italian database for the general user, but also to offer our students a professionalizing experience and autonomous learning strategies, we encouraged them to refer to online and offline monolingual Japanese-Japanese dictionaries and discussion groups (e.g. Yahoo's Chiebukuro or kotoba.ne.jp), above and beyond utilizing them to only translate from English. This was intended to avoid "false friends" as well, which are quite numerous in English and Italian, such as 春分 *shunbun*, translated into English as "vernal equinox" and mistakenly translated into the Italian "equinozio d'inverno" (winter equinox), instead of "equinozio di primavera" (spring equinox). Further explanations about checking entered translations through the ITADICT Editor are given in section 5 of this article, dedicated to ITADICT Editor.

4. How was the project organized?

4.1 Repartition of the "traditional" EDICT file

The work with our students started with the extraction of 18626 priority words (P) from the "traditional" EDICT file, resulting in a 2.8 MB .txt file that was split into 18 smaller files of 1000 entries each. Students volunteering were assigned 250 entries each, while internship students were assigned 1000 entries each. The former had a very flexible deadline, while the latter had to complete the internship in 3 months.

Assigned entries were sent to the students as an .rtf file e-mail attachment, with lines numbered from 1 to 1000 for each "priority" file. (Figure 1)

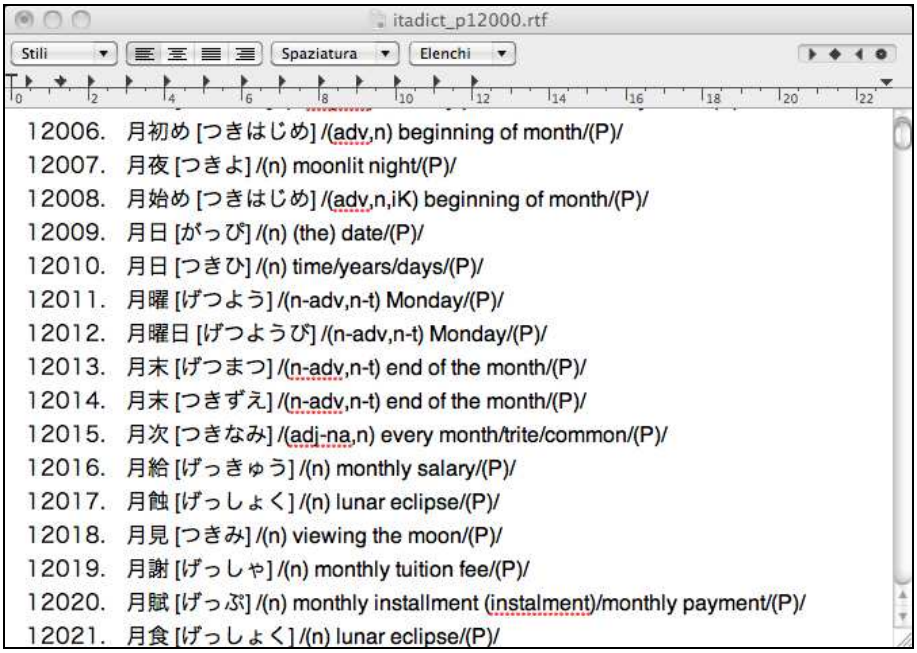
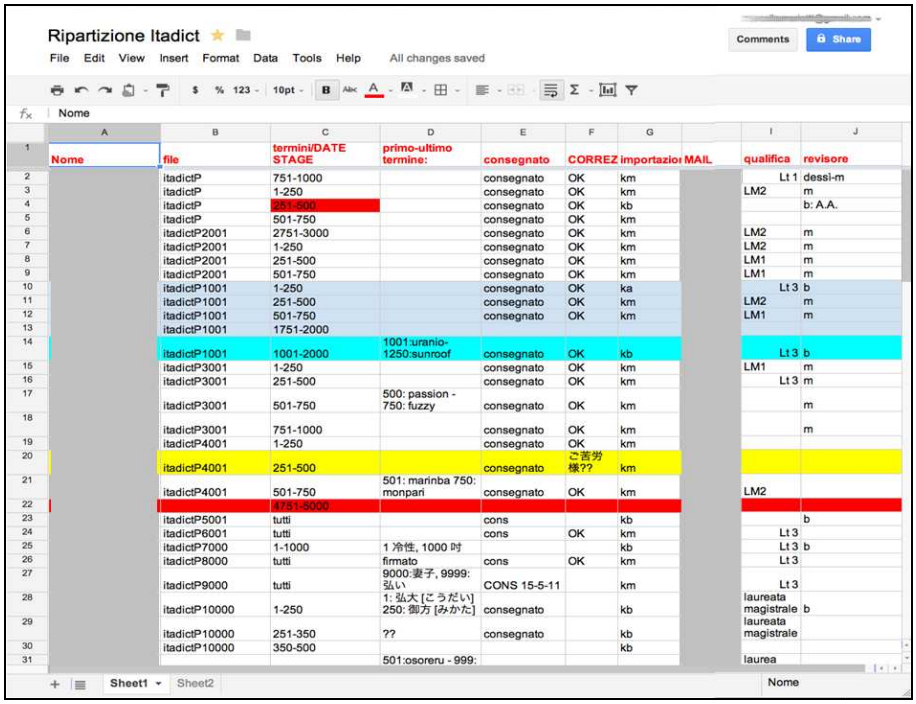


Figure 1: Partitioned .rtf file of (P)riority words with numbered lines

The overall exchange of files was managed on a shared online google spreadsheet called *Ripartizione ITADICT* created by Mariotti on November 3, 2010 (Figure 2).

The spreadsheet included the following information:

- student's name and surname,
- assigned file or portion of file,
- deadline of the work,
- first/last line to translate,
- delivered date,
- reviewed status,
- supervisor of the (later) import in the online new EDITOR developed by Mantelli,
- private e-mail (upon written agreement, so as to be able to contact the student-translator even in the future),
- grade (undergraduate or graduate),
- supervisor of the delivered entries.



Ripartizione Itadict

Nome	file	termini/DATE STAGE	primo-ultimo termine:	consegnato	CORREZ	importazior	MAIL	qualifica	revisore
	itadictP	751-1000		consegnato	OK	km		Lt 1	dessi-m
	itadictP	1-250		consegnato	OK	km		LM2	m
	itadictP	251-500		consegnato	OK	kb			b: A.A.
	itadictP	501-750		consegnato	OK	km			
	itadictP2001	2751-3000		consegnato	OK	km		LM2	m
	itadictP2001	1-250		consegnato	OK	km		LM2	m
	itadictP2001	251-500		consegnato	OK	km		LM1	m
	itadictP2001	501-750		consegnato	OK	km		LM1	m
	itadictP1001	1-250		consegnato	OK	ka		Lt 3	b
	itadictP1001	251-500		consegnato	OK	km		LM2	m
	itadictP1001	501-750		consegnato	OK	km		LM1	m
	itadictP1001	1751-2000							
	itadictP1001	1001-2000	1001:uranio-1250:sunroof	consegnato	OK	kb		Lt 3	b
	itadictP3001	1-250		consegnato	OK	km		LM1	m
	itadictP3001	251-500		consegnato	OK	km		Lt 3	m
	itadictP3001	501-750	500: passion - 750: fuzzy	consegnato	OK	km			m
	itadictP3001	751-1000		consegnato	OK	km			m
	itadictP4001	1-250		consegnato	OK	km			
	itadictP4001	251-500		consegnato	ご苦労様??	km			
	itadictP4001	501-750	501: marinba 750: monpari	consegnato	OK	km		LM2	
		4751-5000							
	itadictP5001	tutti		cons		kb			b
	itadictP6001	tutti		cons	OK	km		Lt 3	
	itadictP7000	1-1000	1 冷性, 1000 吋			kb		Lt 3	b
	itadictP8000	tutti	firmato	cons	OK	km		Lt 3	
	itadictP9000	tutti	9000:妻子, 9999: 弘い	CONS 15-5-11		km		Lt 3	
	itadictP10000	1-250	1: 弘大 [こうだい] 250: 御方 [みかた]	consegnato		kb		laureata magistrale	b
	itadictP10000	251-350	??	consegnato		kb		laureata magistrale	
	itadictP10000	350-500				kb			
			501:osoreru - 999:					laurea	

Sheet1 Sheet2 Nome

Figure 2: Ripartizione ITADICT's spreadsheet

Some students participated more than one time; in this case, e-mail and personal data were inserted only the first time they registered on the form.

4.2 Online discussion group

Since there were very important rules to follow (e.g. the entry syntax, according to which semantic fields had to be divided by a slash without space before or after it), and some questions (e.g. whether to add more specific information about artists' names or not, or which ISO had to be used for source language of *katakana* words) were to be decided by involving all the contributors, on October 23, 2010 we opened a forum (googlegroup Itadict), where more experienced students directly trained by Mariotti could help newcomers, and where the community could discuss how to manage problematic matters (neologisms, technical terms, fields of pertinence...), share useful links and collect/answer queries.

As of September 15, 2012, the dedicated googlegroup *ITADICT* collected 464 messages from its 86 members, who participated in 155 discussions. The first conversations, up to March 26, 2011 were mainly about:

1. how to correctly open the .rtf files,
2. how to find and fill in the documents for internship registration,

3. how to manage the repetition of similar entries with diverging pronunciation such as the example below, a problem solved by the most recent format JMDict (J apanese Multilingual Dictionary by Breen).

   a. 付いている [ついている] /(exp,uk) to be lucky/to be in luck/(P)/
   b. 付いている [ついてる] /(exp,uk) to be lucky/to be in luck/(P)/
   c. 付いてる [ついている] /(exp,uk) to be lucky/to be in luck/(P)/
   d. 付いてる [ついてる] /(exp,uk) to be lucky/to be in luck/(P)/

4. how/where to look up unknown terms
5. how to transliterate a *kanji-kana* compound word: separately or not? Adding source language or not? How?... (e.g. イベント処理 [イベントしょり, *ibento shori*] /(n) elaborazione degli eventi, event processing {comp} (en: event (+ shori))
6. how to transliterate words into Latin characters following Hepburn Style.
7. … and more.

We finally agreed to combine and summarize all collected answers into one "guideline document" (edited by Silvia Rivadossi and Marina Monego in March 2011): this enhanced and sped up the workflow. But since the number of students involved was exponentially increasing, the need for an online editor became evident. Revised instructions were posted by Piercarlo Tommasi in November 2011, after the Editor was developed.

5. The ITADICT Editor and the front-end user interface

Alessandro Mantelli, consolidated system engineer and adjunct professor of informatics at Ca' Foscari University of Venice, volunteered to conceive and develop an appropriate web editor (ITADICT Editor) that would enable us easily to:

a) maintain the original *kanji/kana* - English database and "Part of Speech" classification[3]
b) import data from all the translation entry files produced until then,
c) monitor collaborators' work by login and password furnished upon request,
d) check and where necessary modify each entered translation,
e) monitor and coordinate the trainees showing:

[3] English Part of Speech labels were not translated into Italian, but rather left in English so that third party applications already conforming to Breen's Jmdict format would be able to process the ITADICT database as well. (E.g. Japanese Verb Conjugator Katsuyo)

a. name of the translator of the entry
b. time of the entry
c. modification history
d. number of entries translated by each student
e. entries left to be translated

f) filter the database by translator's name (so as to monitor the work done) and other useful fields (see. "how to search online")

g) further broaden the database

h) share the ITADICT search engine online with the world

i) export the database of Italian translations and of Latin transcriptions in order to merge it into third party applications (e.g. Rikaichan)

The following screenshots show each of the above features.

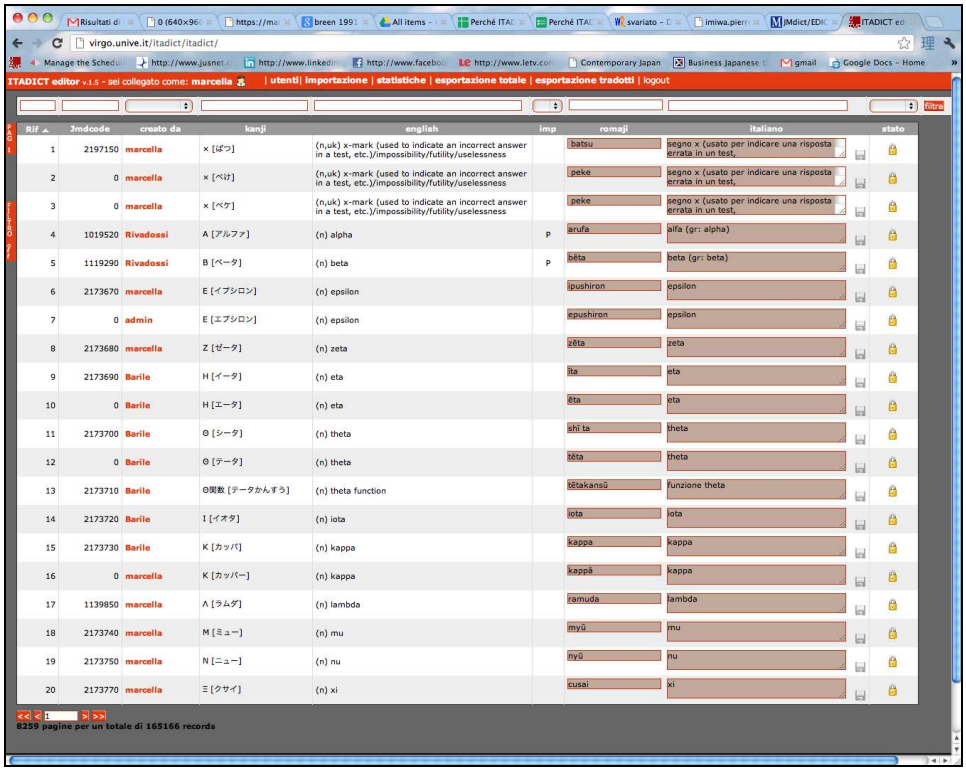


Figure 3: ITADICT Editor page

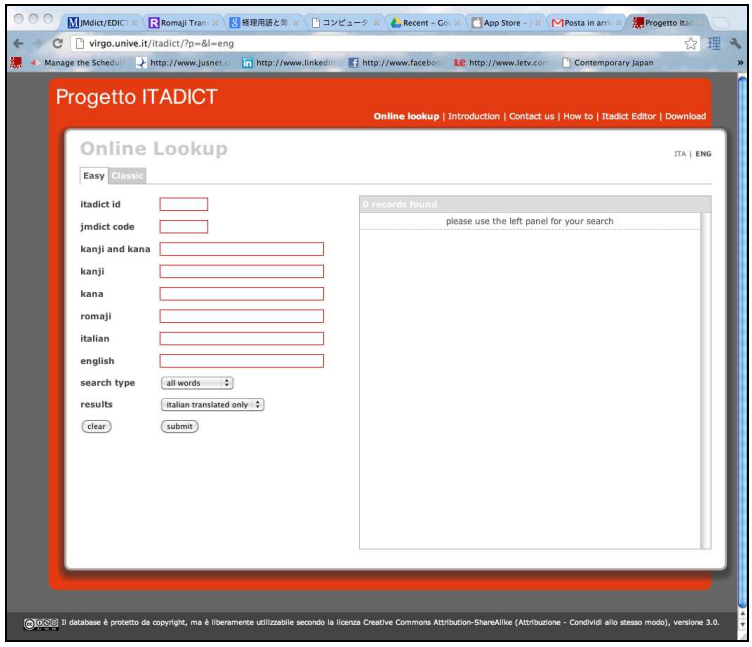


Figure 4: ITADICT online search interface (by code, by kanji and kana, by kana, by romaji, by Italian, by English, filters: only terms translated into Italian or All terms).

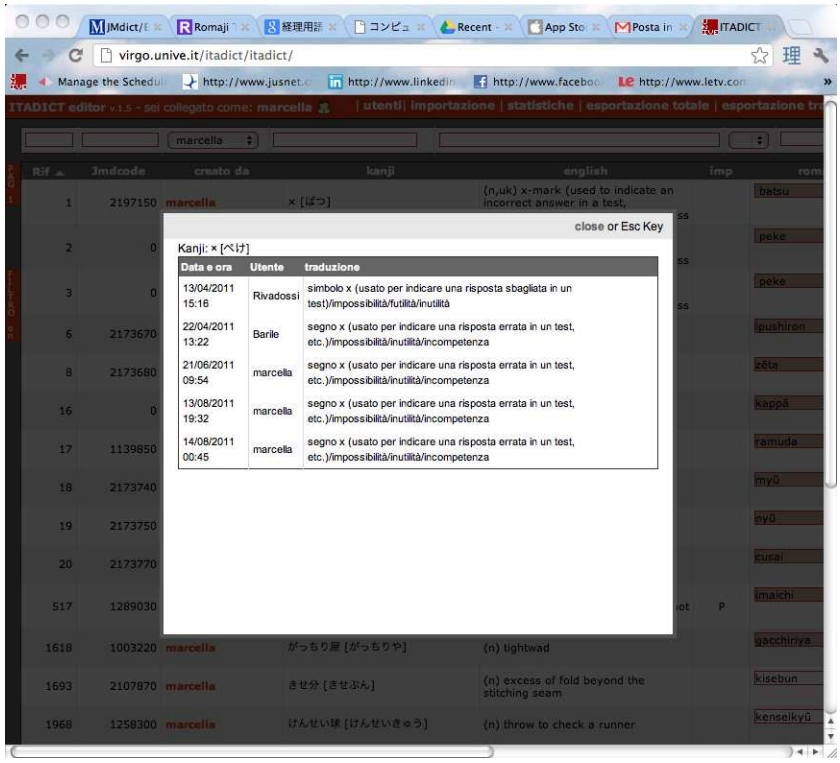


Figure 5: History of each inserted term (date, time/ user/ translation history)

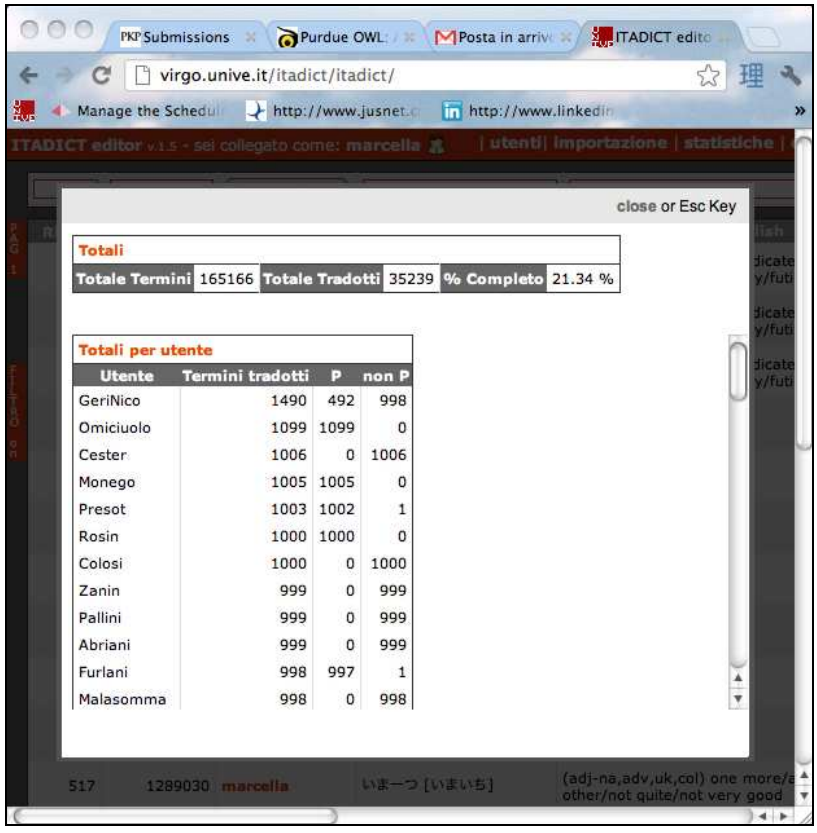


Figure 6: Statistics (Total terms, translated, % completed, terms by translator -priority/non priority)

6. Technical considerations

6.1 Php language to override static html

From a data management point of view, one of the main ITADICT project goal is to provide a scalable development architecture as well as an easy way to enter and edit dictionary data.

Considering the limitations of a traditional html-based approach that lacks a database structure and advanced conditional procedures, ITADICT was developed using a client server model architecture.

A client server application is a piece of software that runs on a client computer and makes requests to a remote server. In ITADICT, the client is the browser in which the web url is invoked such as Firefox or Chrome. From a client perspective, nothing changes compared with a traditional html page: the client reads pages from a server (a remote computer where pages are stored) and displays them. If there are javascript procedures, they are executed as well. From a server point of view, however, there are two components that cause significant changes in the process of creating the page: an

advanced and widely used server side language called *php*, and a database (data organizer system) called *mysql*. These two components are normally invoked before the http service (a piece of software installed in the server that provides the web page[4]). In ITADICT the http service used is called *apache*[5] and is the most widely used web service[6].

Thanks to a server side language such as *php*, *static* limits of *html* can be overriden. Programmers can use a high level language to produce software and call on the data in the database through direct *sql* invocation. The page is parsed by the *http* server and displayed in the browser.

ITADICT database contains all the dictionary data organized in a sql table schema. Vocabulary data can be entered using a dedicated editor whose access is granted by login and password. Kanji data is based on the original EDICT dictionary and registered in the database in binary and base 64 format[7].

6.2 ITADICT basic architecture

Client – Server data communication follows two approaches:

1. Traditional or synchronous method: Data is sent to the server. Page content is rebuilt and presented to the browser.
2. Ajax or Asynchronous Method: Data is sent to the server. Only the needed portion of the page is recreated and passed asynchronously (in the background) to the page, to substitute for old content in real time without interfering with the display and behavior of the existing page.

The client-server approach enables us to separate the interface from the data layer, allowing us to create different interfaces to be connected to the same database in the future.

The asynchronous method is especially used in the front-end (publicly available) user interface where the interaction with the server is smooth and dynamic. In place of the traditional browser loading status, an animated icon is displayed when data communication between client and server occurs.

As data management is separated from the client interface, it is possible to create different ITADICT interfaces such as mobile applications that connect to the same database.

---

[4] Without an http service or web server no page can be displayed on a browser. Every page of the World Wide Web is provided by an http service.

[5] http://projects.apache.org/projects/http_server.html

[6] Netcraft survey July 2012: http://news.netcraft.com/archives/2012/07/03/july-2012-web-server-survey.html

[7] Traditional varchar format has been avoided to avoid misleading utf8 conversions.

The following diagram represents ITADICT's basic architecture.

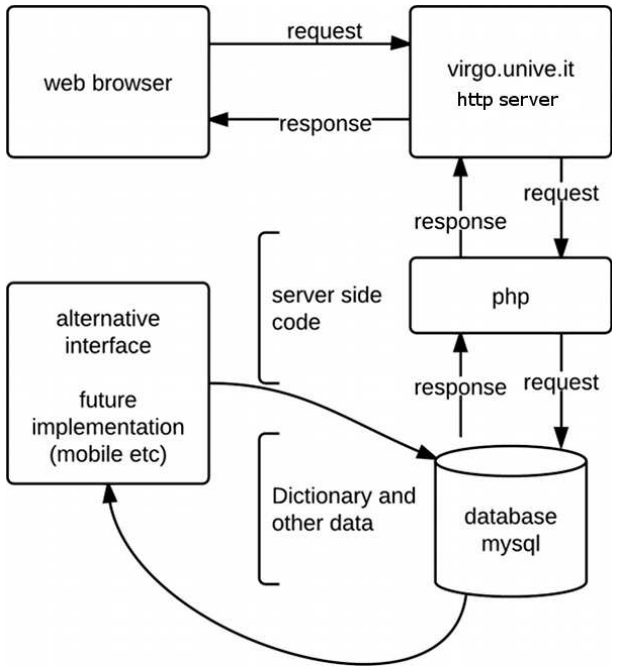


Figure 7: ITADICT's basic architecture

6.3 Future system implementation

ITADICT is a system still in development. Currently we are moving from the original EDICT data source to the more advanced JMdict data format. The database structure itself is going to change and new features - such as adding custom terms - are going to be implemented shortly in the editor.

From a front-end point of view, we are planning to expand search functions and include an svg render engine for the displayed kanji.

7. What did ITADICT internship offer to the students?

Working on ITADICT (.rtf file or Editor) helped students acquire several skills, beside learning new English and Japanese words, such as:

- how to use shared documents,
- how to use online Japanese searching engines,
- how to use a monolingual Japanese-Japanese dictionary,

- how to autonomously translate terms not present in Japanese-Italian dictionaries,
- how to thoroughly adhere to strict editing rules,
- how to ask for help from colleagues and via online translation forums (e.g. Biblit, Langit, WordReference and more),
- how to use online discussion forums,
- how to open a new e-mail account,
- how to correctly transliterate Japanese into Latin alphabet according to the Hepburn system.

Besides these acquired skills, one of the major satisfactions offered by the ITADICT internship is the opportunity to actively participate in the creation of a language tool the students themselves will use in the future.

The database of the translated terms can be used online as well as downloaded and merged into third party applications through a dedicated link[8].

8. Conclusions and future work

Up to September 2012 about 35239 entries have been transliterated into Latin characters (*romaji*) translated and (most but not all) revised. The most common 18.681 entries (marked with a "priority" P in the database) have been completed and, as mentioned above, we are now proceeding to match the "traditional Edict" format we started from, with the newer JMdict format. This is meant primarily to avoid misleading readings such as the pair mentioned above, where ついてる (*tsuiteru)* is actually not the transcription of the entry 付いている (tsuite iru):

付いている [ついてる, tsuite ru] /(exp,uk) to be lucky/to be in luck/(P)/

We are planning to reach 50.000 terms by the beginning of 2013 and then reach our final goal of 180.000 terms hopefully within two more years. Special projects are underway, such as the development of lists of specialized words (nanotechnology, leather processing industry, textile industry, furniture industry[9]) and a more user friendly layout of the open search engine reachable at virgo.unive.it/itadict.

The exported database can be used with open source dictionary software and reading aid software the same way as Edict files can. The file is in Unicode/ISO-10646 coding with UTF-8 encapsulation.

[8] http://virgo.unive.it/itadict/?p=download&l=eng

[9] As part of a European Social Fund project selected from Ca' Foscari University to be presented to Veneto Region. Title: "Language tools to support the internationalization of Veneto Region's companies: terminology in the textile, leather, chemical-environmental, securities and furniture industries" (Project Leader: Marcella Mariotti).

The authors welcome new collaborators: anyone interested in the project can contact us using the online format: we will be happy to further improve our project.

List of collaborators from whom we received the disclosure agreement.

Nicola Angaran
Nicolò Anesa
Alice Aniello
Andrea Belluomini
Antonio Benasaglio Berlucchi
Alice Berto
Giulia Bianco
Stefano Boggia
Rita Bovina
Paola Celentano
Marco Colosi
Matteo Contrini
Elia Dal Corso
Massimo Dalla Pria
Mauro Dalle Prane
Giorgia Dessì
Giacomo Orseolo Ferro
Valentina Gastaldello
Niccolò Geri
Andrea Giolai
Alessandra Grillo
Alberto La Spada
Martina Malasomma
Lara Marinozzi
Elisa Martini
Marina Monego
Alessia Omiciuolo
Elena Ominetti
Chiara Pallini
Giulia Perin
Giovanni Presot
Silvia Rivadossi
Elena Tessari
Jacopo Tiezzi
Federico Tombari
Aurora Torchia
Cinzia Zanin

References

Balboni, P. E. (2002). *Le sfide di Babele*. UTET.

Breen, J. (2004). JMDict: a Japanese-multilingual dictionary. In: *Coling 2004 workshop on multilingual linguistic resources*, Geneva, Switzerland, pp. 71-78.

Breen, J. (2010). JMdict/EDICT Japanese/English Dictionary Project http://www.csse.monash.edu.au/~jwb/edict_doc.html#IREF05

Danesi, M. (2003). Second Language Teaching, A View from the Right Side of the Brain. Dordrecht: Kluwer.

Downes, S. (2007). *Trends and Impacts of E-Learning 2.0.* Keynote presentation delivered at the International Conference on Open Course Ware and e-Learning, Taipei, Taiwan.

Ferrari, M. (2010). *La didattica del giapponese a studenti universitari italiani attraverso la canzone. Analisi di un ciclo di tre lezioni sperimentali sulla canzone all'Università Ca' Foscari di Venezia*. (Japanese language teaschin to italian university students. Analysis of

three experimental classes at Ca' Foscari Univerity of Venice) Master Thesis (tutors: Mariotti, M. & Balboni, P.). Ca' Foscari Unviersity of Venice.

Mariotti, M. (2011). "Fansubbing *e didattica: vecchie e nuove sfide e Tecnologie*" (Fansubbing and language learning/teaching: old and new challenges and technologies", paper presented at the workshop *La traduzione dalle lingue orientali: limiti, specificità e prospettive*, (Translating from East-Asian Languages: limits, specificities and perspective, Nicoletta Pesaro coordinator), Ca' Foscari University of Venice, 15 Febbruary 2011.

Mariotti, M. (Forthcoming). *Introduzione alla lingua giapponese* (Introduction to Japanese Language), Carocci.

Mariotti, M. (2011). BunpoHyDict: a Hypermedia Dictionary of Japanese Grammar and its Development. In: Association of Japanese Language Teachers in Europe (AJE), *Japanese Language Education in Europe (Yoropa Nihongo Kyoiku)*, 15 , pp. 179-188.

Miyake, T. (2012). "Japanese Studies as 'subculture'", paper presented at the conference *"Manga Worlds": Subcultures, Japan, Japanology*, Kyoto International Manga Museum, 3 June 2012.

Schumann, J. H., et al. (2006). *The Neurobiology of Learning*. Routledge.

Quoted links

Breen's *Lexicographical* Details:
http://www.csse.monash.edu.au/~jwb/edict_doc.html#IREF05

BunpoHyDict (under construction):
www.bunpohydict.com

Chiebukuro:
http://chiebukuro.yahoo.co.jp/

Goo Japan:
goo.ne.jp

Introduction to EDICT (2005):
http://www.csse.monash.edu.au/~jwb/edict_doc_old.html

Introduction to EDICT within the JMdict Project (2012):
http://www.csse.monash.edu.au/~jwb/edict.html

ITADICT database download:
http://virgo.unive.it/itadict/?p=download&l=eng

ITADICT:
http://virgo.unive.it/itadict

Japan Goggles:
http://japangoggles.lucsens.com/

Kotoba.ne.jp:
http://www.kotoba.jp/

Kotoba/Imiwa?:
http://imiwa.pierrephi.net/

Online survey *Why ITADICT?":*
https://docs.google.com/spreadsheet/ccc?key=0Anfm7ZNoCqUCdHhxVVJLU3FjZXNBWl9iU3RQRUNHREE#gid=0

Rikaichan:
http://www.polarcloud.com/rikaichan/

Romaji Translator:
http://www.romaji.org/

The “traditional” EDICT:
http://www.csse.monash.edu.au/~jwb/edict_doc.html

The Jmdict Project (2012):
http://www.csse.monash.edu.au/~jwb/jmdict.html

AUTOMATIC ADDITION OF GENRE INFORMATION IN A JAPANESE DICTIONARY

Raoul BLIN
EHESS - School for Advanced Studies in the Social Sciences
blin@ehess.fr

Abstract

This article presents the method used for the automatic addition of genre information to the Japanese entries in a Japanese-French dictionary. The dictionary is intended for a wide audience, ranging from learners of Japanese as a second language to researchers. The genre characterization is based on the statistical analysis of corpora representing different genres. We will discuss the selection of genres and corpora, the tool and method of analysis, the difficulties encountered during this analysis and their solutions.

Keywords

lexicography; corpus; genre; Japanese

Izvleček

Članek predstavlja metodo za samodejno dodajanje informacij o žanru k japonskim iztočnicam v japonsko-francoskem slovarju, ki je namenjen tako učencem japonščine kot tujega jezika kot tudi raziskovalcem. Žanrski opis je osnovan na statistični analizi korpusov različnih žanrov. Članek opisuje izbiro žanrov in korpusov, orodja in metode za analizo, težave pri analizi in rešitve zanje.

Ključne besede

slovaropisje, korpus, žanr, japonščina

1. Introduction

In the Japanese language, general lexicons and dictionaries (whether monolingual or multilingual) published to date provide very little information about the genres of the various entries. Furthermore, dictionaries that address genre generally focus on only one theme (for example the law) and cannot be used to compare different genres.

However, information regarding the genre of words is extremely important in language production since it enables the writer to choose vocabulary appropriate to the

genre of the discourse or text in question. In terms of translation, such information is useful to choose the appropriate word in the target language. In addition, it will also enable the reader to define the style of a text.

Defining and manually annotating genre information for all lexical entries in a dictionary is very costly, both in terms of man-hours and financially. It is also fraught with many methodological difficulties.

The first difficulty is to provide clear criteria that can be used by lexicographers to classify the words. In a strictly scientific approach, if the criteria were qualitative, they would require evaluation and testing. A few lexicographers may use them to classify a sample of entries, after which the rate of agreement between lexicographers would be calculated. The criteria must be revised and retested until the rate of agreement is sufficiently high. This is a long and laborious process.

The second difficulty is to assign the lemma to one genre. Except for highly specialised vocabulary, words may appear in a variety of genres. Therefore, merely classifying a word as belonging to one genre to the exclusion of other genres would not be a suitable classification method. Rather, the word's classification should be graduated in terms of each genre under consideration. Once again, the evaluation procedure mentioned above becomes necessary.

The third difficulty concerns human knowledge limitations. The classification method presupposes that the lexicographer has a good knowledge of all the genres under consideration. Such lexicographers are certainly very rare. Teams of two or more researchers from different specialties are necessary. Therefore, the evaluation process mentioned above does not cover the rate of agreement between lexicographers working alone, but between teams of lexicographers. It is an additional difficulty.

A fourth challenge is to monitor variations in genre over time. This would require all the entries to be checked on a regular basis (every three or four years) and to create a team of lexicographers each time. The research and evaluation process described above would then be repeated.

As is understandable, it is not realistic to manually describe the genre of a dictionary's entries using qualitative criteria in a scientific approach. The task obviously requires an automated procedure. To this end, our idea is to base the study on a statistical analysis of corpora. By using selected corpora to represent different genres, the genre(s) of a word will be correlated to the word's frequencies in those corpora. In addition to being easy to implement, such an automated procedure based on quantitative criteria is easy to replicate and suitable for monitoring variations over time.

This process has been applied to automatically classify the genre of 16,000 entries in a Japanese-French dictionary (Blin, 2012a; abr. "DFJC"), on the basis of frequencies obtrained from a purposely-built corpus consisting of several subcorpora. In the following chapters, we will describe the process in more detail

In section 2, we will provide a brief overview of the DFJC to show what kind of data was included. In section 3, we will discuss the genres that were studied and the corresponding corpora. In section 4, we will describe the software used to count the number of occurrences, the problems encountered and their solutions (if any).

2. The dictionary

2.1 The dictionary entries

The dictionary contains more than 16,000 Japanese common nouns and qualifying nouns (e.g.: *kinkyu*, 緊急, "urgent", *ihou*, 違法, "illegal") in Japanese, each described as the example entry in Table 1.

Table 1: Example of an entry in the Japanese-French dictionary DFJC

		White papers	Newspapers	Legal texts	QA govn.	misc. QA	Chats	Total occ.
36	*asita* 明日 (h:41); demain	11	57		5	41	215	150
<1>	<2> < 3> (<4>); <5>	<6>	<7>	<8>	<9>	<10>	<11>	<12>

Each entry includes <1> the entry number, the usual <2> reading and <3> writing of the word, <4> the list of its homographs (see section 4.2), <5> the French translation, <6-11> its frequencies in six sub-corpora/genres (see section 3.2) and <12> the total number of occurrences.

All the data used to calculate the frequencies are provided in a database distributed freely (JaLexBD[1]).

Furthermore, the dictionary provides a summary description of the sub-corpora/genres: size, list of the most frequent words, comparison of the frequencies of some morphological structures, etc.

2.2 The distributions

For each word, the dictionary provides the frequencies of the word alone and with affixes, in each of the subcorpora. Due to lack of space, the frequencies for each distribution of each word are not detailed in the paper version, but all the numbers of occurrences are provided in the JaLexBD database.

[1] http://rkappa.fr/lexic/JaLexBD/index.JaLexBD.php

We chose constructions with affixes as the most linguistically interesting ones, especially those which are less morphologically and syntactically ambiguous (see section 4.2). For example, we avoided counting strings of concatenated nouns (without particles), since this can be extremely difficult for the automatic analysis of a text. The constructions were counted as follows:

2.2.1 Word alone

The word appears without any affix.

2.2.2 Words with the adjectivising suffix *-teki* (的)

Any noun with the *-teki* (的) suffix is transformed into a *-na* adjective. N+*teki* means roughly "which has the property of N". For example, when *-teki* (的) is attached to the noun *gainen* (概念, "concept"), it forms the word *gainenteki* (概念的), which means "conceptual".

2.2.3 Words with the nominalising suffix *-sei* (性)

The suffix *-sei* (性) can be placed after a noun or an adjective. N/adj+*sei* is a (grammatical) noun and means roughly "the property of being N/adj". For example, *ensyoo* (炎症) means "inflammation", while *ensyoo* + *sei* (炎症性) means "of an inflammatory nature, type or origin". Some of the constructions suffixed by *sei* may be lexicalized, such as 経済性 (*keizai* + *sei*, "economic efficiency, economic performance, economy").

2.2.4 Words with the plural suffixes *-ra* (等／ら) or *-tati* (達／たち)

The suffixes *-ra* (等／ら) and *-tati* (達／たち) express the plural. For example, the common noun *gakusei* (学生 "student") is indefinite. Depending on the context, it can be interpreted as being either singular or plural. The construction *gakusei+ra* (学生等／学生ら), however, can only be interpreted as being plural.

In contemporary Japanese, the two suffixes are both used for human nouns, like "*gakusei*" (学生). They can also be used with a humanised entity, such as *neko+tati* (猫達 "the cats"). However, this construction is generally limited to children's language.

The difference between the two suffixes may lie in the register of the language. For example, *-ra* is known to be more formal than *-tati*. Thus, comparing the frequency of plural suffixes may provide an interesting indication of text genres.

The process of counting plural suffixes is based on the hiragana transcription, i.e. ら (*ra*) and たち (*tati*). This restriction is justified by the fact that the Chinese character (等), which can be used to write the suffix *-ra*, is ambiguous, since 等 can also be used to write *nado* ("etc."). In order to prevent possible errors, we did not count the occurrences of 等, but only the transcriptions in hiragana. In order to unify the counting procedure, we also restricted the counting of *-tati* to the hiragana

transcription. This restriction certainly comes at the expense of *-tati*, since *-tati* is written more frequently in Chinese characters. Thus, the results in the DFJC certainly minimise the number of occurrences of *-tati*, although we cannot say to what extent it is minimised.

3. Corpora and genres

The corpus is divided into seven sub-corpora. Each sub-corpus has specific characteristic(s) that we will refer to as "genre". These characteristics mainly stem from their source. For example, "journalistic genre" (i.e. "journalistic corpus") will refer to the collection of texts retrieved from newspaper websites. The details of the other sub-corpora are outlined below.

We consider that such a "genre" definition is explicit enough and does not require the laborious evaluation process described in the introduction.

The description may be supplemented with other characteristics, but they will not have been used to build the corpora and define the genres.

We can distinguish between two types of text: monologues and dialogues. A "dialogue" refers to a text which provides an answer to a question, or which is constructed to be answered by someone other than the author. A "monologue" refers to a text which is not constructed to be followed by an answer, and which does not provide an answer itself.

We also distinguish between reviewed and non-reviewed texts. The assumption is that the variety of morpho-syntactical structures and vocabulary is wider in non-reviewed texts. When reviewing is part of the production process, it is expected that the author and reviewer will agree to some (at least implicit) conventions about acceptable (or authorized) language. The author must produce a text corresponding to this agreement. If not, the reviewer will correct the text according to these constraints. In principle, there are no such limitations in non-reviewed texts.

If the set of authors is limited, the variety of structures and vocabulary is limited to the skills of those authors. Indeed, the variety is expected to be poorer in comparison with corpora produced by an unlimited set of authors. This is why we distinguished between textual corpora produced by a limited and an unlimited set of authors.

We should also take the writing time into account. We assume that the variety of vocabulary and morpho-syntactic structure is greater when writers have unlimited time to write.

Table 2 below presents a summary of those characteristics for all the corpora. In addition, it shows the size of the sub-corpora and their frequencies.

Table 2: Characteristics of the corpora used

(Conventions: "+" yes for all the texts of the corpus; "–" no for all the texts of the corpus; "+/–" depending on the text of the corpus)

		No. of sentences	frequency of updates	reviewing	only one theme for the corpus	limited set of authors	limited set of readers
monologue	White papers	105 520	one year	+	–	+	+
	Daijirin Dictionary	176 809	partial, long-term	+	–	+	–
	Newspapers	167 819	1 day	+	–	+	–
	Legal texts	34 688	1 partial year	+	+	+	+
dialogue	Q&A govn.	54 901	1 full year	+	–	+	+
	Q&A misc.	136 946	1 partial day	+/–	–	–	–
	Chats	23 547	1 full day	+/–	–	–	–

3.1 Selection criteria for the corpora

We applied the criteria below to select the sub-corpora/genres.

3.1.1 Representativeness of a subcorpus with respect to its genre

In order to obtain a better representativeness, we built the sub-corpora as follows. Our strategy differs from the well-known corpus BCCWJ (Maruyama, 2009) in many respects.

Whenever possible, we used the complete collection of texts from a source rather than using a sample. For example, we collected all the White Papers of 2009, 2010 and 2011, whereas the BCCWJ contains only samples of collections.

For the same reason and also unlike the BCCWJ, all the texts of the sub-corpora are complete. We did not use any samples.

The corpus is strictly limited to written language. Transcriptions of spoken language, such as the Minutes of the Diet (included in the BCCWJ) are excluded.

3.1.2 Development of the corpus

The DFJC, published in 2012, is the first step of the project to observe the development of genres over time. As such, it was necessary to choose genres/corpora that would change over time.

To date, statistical studies in Japan about the Japanese language have been designed for static corpora. Those corpora were created once and for all and no updates were planned. Ever since the first statistical study on a large corpus of written Japanese was conducted by the National Institute for Japanese Language and Literature in the 50's (Yamazaki, 2006), no corpus has been updated or compared to previous versions to observe its development over time.

To break away from this method, all the corpora used for the DFJC are textual collections that will be updated within a few years (or within a few months in some cases). The corpora of newspapers, chats, questions to the government, miscellaneous Q&A and White Papers will be entirely renewed in one year. Some of the sub-corpora, such as commercial dictionaries and fundamental legal texts (Constitution, etc.) should not change for a long time, but as long as they are distributed without changing (and excepting cases where they are distributed explicitly as historical texts), they should be understandable. Thus, even if such texts have not been renewed for a while, the language used represents the current language during the period of their distribution.

The sub-corpora for the DFJC consist of texts produced mostly between 2008 and 2011, which represents a span of 4 years. We had first planned to re-compile the texts annually, but the size of the corpora was not sufficient to obtain a good representativeness. Consequently, we estimate that a span of 3 to 5 years would be a good compromise.

3.1.3 Accessibility

In order to build the corpora, we also had to make do with limited financial and human resources. Scanning or manually retyping texts, as has been done for the BCCWJ, was out of the question. The solution was therefore to use the Internet. To this end, the corpora/genres selected were taken from collections of texts accessible on the Internet. However, we did not have enough technical (and financial) resources to build a corpus as large as the one described by Kawahara and Kurohashi (2006) which contains 470 million sentences. Even if the texts are easy to access on the Web, we had to limit the selection to a relatively small set of genres/corpora (710,000 sentences)

3.2 Detailed presentation of the sub-corpora

3.2.1 White papers

This is a collection of White Papers published in 2009, 2010 and 2011. Due to their thematic variety, this corpus cannot be considered as representative of one

discipline. However, we assume that the production conditions are homogenous: the texts are written by a restricted set of authors (specialists). We also assume that these texts are reviewed.

3.2.2 *Daijirin* dictionary

To our knowledge, the language genre of dictionary texts has not yet been studied, despite the fact that it is certainly original and subject to many editorial conventions.

This corpus has not yet been completed. In the online *Daijirin* dictionary (Matsumura, 2006; http://dic.yahoo.co.jp), we only used the pages corresponding to the lemmas of the DFJC. As such, this corpus only contains 16,000 entries even though *Daijirin* contains 230,000 entries in total. Furthermore, *Daijirin* cannot be considered as being representative of all Japanese dictionaries. Therefore, this corpus is a poor representation and has been used as a test corpus only. In the future, the totality of the *Daijirin* and other online dictionaries will be used.

3.2.3 Newspapers

This corpus contains the online versions of three newspapers, covering the editions from April through December 2011. The newspapers are Asahi (www.asahi.com), Nippon Keizai (www.nikkei.com) and Nikkan Kougyou (www.nikkan.co.jp). The first two newspapers are widely distributed and have a significant place among newspapers. Furthermore, newspapers are very important in the everyday lives of Japanese people (almost all Japanese households are subscribed to a newspaper). We plan to incorporate more newspapers in the future.

3.2.4 Legal texts

The legal corpus is divided into two parts. The first part is the compilation of all official legal texts produced in 2008, 2009 and 2010 (law 法律, *hooritsu*; Cabinet Office Ordinance, 内閣府令, *naikakuhurei*; decree 命令, *meirei*). The second part is the compilation of six legal codes (Constitution, 憲法, *kenpoo*; civil law, 民法, *minpoo*; commercial law, 商法, *shoohoo*; criminal law 刑法, *keihoo*; civil procedure, 民事訴訟法, *minzi soshoohoo* and criminal procedure 刑事訴訟法, *keizi soshoohoo*).The second part will not be updated, whereas the first one is renewed every year.

3.2.5 Written questions submitted to the government

A compilation of all the written questions (from the Diet) submitted to the government between 2008 and 2010.

3.2.6 Miscellaneous questions and answers

A compilation of websites taken from oshiete.goo.ne.jp. This site is equivalent to the website Chiebukuro used for the BCCWJ. Each page contains an open question and

possibly one or more answers. In some cases, there is no answer. Questions can address any subject matter.

3.2.7 Chats

This corpus is made up of pages from different chat websites. Due to financial concerns, this corpus is small. The procedure for collecting the pages included in this corpus will be changed in order to obtain more information.

It must be pointed out that the fundamental difference between this corpus and the corpus of miscellaneous questions and answers is the temporal constraint in terms of production. In miscellaneous question-answer dialogues, there is no (implicit) constraint on the time interval between the question and the answer. On the other hand, the response is usually immediate where chats are concerned.

4. Tools and analytical method

This section presents the software used, the problems encountered, their solutions and their impact on results.

4.1 Software

The number of occurrences of the words in the corpora was counted using the free software SAGACE version 4.2.0 [2] (Blin, 2012b). This tool is designed for searching patterns, but it does not analyse entire sentences. Patterns are defined as strings of words (or characters). A word can be a single word or any word of a pre-defined category. In the latter case, the category must be listed in the lexicon which is associated with SAGACE. SAGACE is only executed from the command line. To launch the query, the required pattern and the search parameters are described in a request form (a simple text file) interpreted by SAGACE.

We have chosen this software mainly on account of its ease of use. Unlike symbolical parsers, it is not necessary to develop a grammar , which requires time to carry out maintenance operations. It is sufficient to create a lexicon listing the words of the various categories. This is important since the maintenance and modification of a rule-based grammar can be a complex operation. Furthermore, there is currently no free and open grammar for Japanese. SAGACE also differs from statistical parsers (such as Mecab 3) on account of the fact that it does not require training and manual evaluation. In order to obtain good results with such tools, it is necessary to perform training and manual evaluation for each genre of text. Such a procedure is very costly.

[2] http://crlao.ehess.fr/japonais-coreen/corpus/sagace/sagace.html

Furthermore, SAGACE is autonomous and does not require any other software. It performs all the necessary functions for the analysis: a request interface, search engine and results interface. In addition, untagged plain text is sufficient. Thus, no pre-analysis is required.

As mentioned above, the advantage of SAGACE is that it is easy to use. The drawback, however, is that errors of analysis are (perhaps) more frequent than with other parsers. In order to limit the risk of errors, we selected less ambiguous patterns to be searched. As a result, not all the occurrences were counted. The frequencies indicated in the dictionary are thus slightly lower than the real frequencies. We are unable to assess the difference.

4.2 Difficulties of analysis and solutions

The automatic analysis of the Japanese language is subject to a few difficulties, which are well known in the field of Natural Language Processing. In this section, we will discuss how they have been solved (or not) using SAGACE, and what impact this solution had on the results.

4.2.1 Lemmatisation

Words are not separated graphically in written Japanese. Even if the parser analyses the entire sentence, morpho-syntactic errors may occur.

Only a semantic and pragmatic parser can prevent errors, but no such tools currently exist.

To limit the risk of errors with SAGACE, we first applied the traditional “longest match method”. Secondly, we restricted the number of searched patterns to the ones with a low risk of ambiguity (even if it is not zero). Overall, the searched pattern includes the contiguous words before and after the target structure. For example, when searching occurrences of a noun suffixed by 性 (*sei*), we used a pattern including a particle or punctuation mark on the left, and a particle, punctuation mark or copula on the right:

$$\left\{ \begin{array}{c} \text{particle} \\ \text{punctuation} \end{array} \right\} \text{ NOUN 性 } \left\{ \begin{array}{c} \text{particle} \\ \text{punctuation} \\ \text{copula} \end{array} \right\}$$

As an example, this is the description of the pattern in the request form:

>0 cat:particle | punctuation | XX	// 1
=0 cat:LEXEME /-affich:trait:lemme /-count	// 2
=0 性	// 3
=0 cat:particle | punctuation | copula	// 4

These lines are interpreted as follows:

(1) the first element of the pattern is anywhere (“>0”) in the sentence. It is either a particle, a punctuation mark, or a mark to indicate the beginning of sentence.

Formally, the description of the elements of the patterns are formulas written in a language close to propositional language. The interpretation is very similar to the interpretation of propositional logic. For example, the description of the first element can be formally interpreted as follows: the element is any word belonging to the category (“cat:”) defined as the union (“|”; disjunction) of three basic categories (propositional constant): category of particles (“particle”), category of punctuation (“punctuation”) and category “XX” (which is a singleton containing the mark indicating the beginning of a sentence). All the basic categories are listed in the lexicon associated with SAGACE. Using this description language, it is possible to “create” new categories by merely combining basic categories listed in the lexicon associated with SAGACE and without modifying this lexicon.

(2) the second element is contiguous (“=0”) with the precedent one. It is the word to be counted (“/-count”); it is any word of the category named LEXEME in the lexicon used by SAGACE.

(3) the third element is contiguous (“=0”) with the previous one. It is “ 性 ”.

(4) the third element is contiguous (“=0”) with the previous one. It is either a particle, a punctuation mark or the copula.

A more detailed description of the pattern syntax is available in the manual and online tutorials of SAGACE. All requests are provided in the DFJC.

4.2.2 Homography

Some words have the same graphic form but a different reading, and perhaps a slightly different meaning. For example, two homographic words transcribed as 魚 are almost synonymous, but have a different reading, *uo* and *sakana*. In the corpus, in order to know which reading is being referred to, a semantic (including pragmatics) analysis must be performed, but such an analysis tool is not currently available. A statistical analyser may solve the problem, but the results are not absolutely certain and the tool requires training.

For a great number of regular common nouns, there is a homographic proper noun. For example, *mori* (森, “forest”) and *hayasi* (林, “wood”) are also used as a last name. These last names are very common. Fortunately, in the corpora that have been used, they frequently appear with specific affixes, such as honorific suffixes (san, “miss, mister”) for human proper nouns. As such constructions don’t agree with the pattern we use, most of them have been excluded from the counting operation. Despite these precautions, it is possible that some occurrences may have been counted as proper nouns. Thus, the frequencies of the entries which can also be used as common nouns

may be slightly higher than the real frequencies. We assume that this is a minor problem with no significant impact on the frequencies.

In the DFJC, we tagged all the entries which can also occur as a proper noun. To this end, we used a list of 320,000 proper nouns, extracted from the mecab-naist-dic and some other resources. The list contains most Japanese personal names, place names and company names. It does not include Chinese proper nouns. This list is not very long, but it should suffice to fulfil the purpose.

4.2.3 Multiple transcriptions

All Japanese words can be transcribed in various ways, by combining three character sets: hiragana, katakana and Chinese characters. Standard dictionaries provide the standard transcription. In fact, there is no "official" or "academic" standard. The so-called standard transcription could be defined as "the one which most closely resembles government prescriptions, among the most used transcriptions". For example, the word *môsikomi* ("application") is lexicalized as 申し込み or 申(し)込み, depending on the dictionary. The parentheses indicate that the characters can be omitted (but are still pronounced). Some dictionaries used for Natural Language Processing provide all the most common transcriptions and consider them as entries. For example, mecab-naist-jdic 4 provides five transcriptions/entries for the word *môsikomi* ("application"): もうしこみ, モウシコミ, 申しこみ, 申込み and 申し込み. Such lexicalisation has many flaws: it is very redundant and not exhaustive.

The multiplicity of transcriptions is not a problem per se, since the author's choice of one transcription among many can help to characterise a written style. It rather represents an editorial problem when publishing a paper dictionary: listing all the transcriptions takes up a lot of space, even though many of them have such a low frequency that they are insignificant. The DFJC provides only the most common transcriptions. For some words, the frequency is the sum of the frequencies of two transcriptions. For example, when a noun contains the so-called honorific prefix *o*, we do not separate the transcription in kana from the transcription in Chinese characters. For example, the number of occurrences of the entry *otearai* (お手洗, "toilet") is the addition of the number of occurrences of the two transcriptions お手洗 and 御手洗.

Variations of transcriptions are not only obtained by combining different systems. Some words have two or more transcriptions in Chinese characters. In most of these cases, two transcriptions exist: an "academic" transcription and a "popular" transcription. For example, the "academic" transcription of *tamago* ("egg") is 卵. The popular transcription is 玉子. In the DFJC, the two (or more) transcriptions are clearly separated and constitute independent entries.

4.2.4 Homography *and* homophony

For some words, there are other words that are both homophonic and homographic. This is more common with monosyllabic (one kana) words. It can also

occur with the kana transcription of a word. For example, “tooth” and “blade” are homophonic: *ha*. They are usually written in Chinese characters (resp. 歯 and 刃). However, when they are written in kana in a corpus, it is necessary to perform a semantic analysis to determine what word is being referred to. For the DJFC, we chose to count only dictionary lemmas, which are mainly in Chinese characters. We assume that existing entries in hiragana do not have such homophonic and homographic equivalents.

5. Conclusion and outlook

In this paper, we explained the process we have implemented to automatically characterise the genre(s) of 16,000 words in a Japanese-French dictionary. We plan to repeat this work regularly, about every three or four years, using the same process. There is room for improvement, and we wish to improve at least two points. Firstly, the number of entries will be increased. In particular, we will add verbal nouns. Secondly, as explained above, some corpora must be changed: the dictionary will be supplemented and we need a more reliable source for chats.

We also plan to conduct the same study on inflected words, such as verbs and adjectives. Despite the fact that SAGACE is not well designed for manipulating inflected words, a large-scale test (Blin, 2012c) showed, however, that the same method can be applied with the same tool. A more detailed (and manual) assessment of the results is required.

If the results are good enough, we will apply the process to locutions, including locutions with inflected words.

References

Blin, R. (2012a). Dictionnaire de fréquence du japonais contemporain - 16 000 noms (Youfeng.). Paris.

Blin, R. (2012b). *SAGACE v4.2.0*. CNRS. Retrieved from http://crlao.ehess.fr/japonais-coreen/corpus/sagace/manuel/Manuel.pdf

Blin, R. (2012c). Fréquences des verbes japonais dans un corpus de grande taille. Blin. Retrieved from http://rkappa.fr/sagace/tutoriel/sagace4-2/data/ListeDesFrequencesDesVerbesJaponais.pdf

Kawahara, D., & Kurohashi, S. (2006). *Case frame compilation from the web using high-performance computing*. Presented at the 5th International Conference on Language Resources and Evaluation.

Maruyama T. (2009). “Gendai nihongo kakikotoba keikin koopasu” monitaa kaihatu deeta (2009nendoban) sanpuringu houhou ni tuite [About the method of sampling in the “Balanced Corpus of Contemporary Written Japanese” (v.2009)]]. National Institute for Japanese Language and Linguistics

Matsumura, A. (2006). *Daijirin Second Edition*. Tokyo: Sanseido.

Yamazaki, M. (2006). Kokuritu kokugo kenkyuuzyo no goi tyousa no rekisi to kadai [Thematics and history of the lexical surveys of the National Institute of Japanese Language]. *12th Workshop "Thematics and history of the lexical surveys of the National Institute of Japanese Language"* (pp. 168–186). Tokyo University.

THE CONSTRUCTION OF A DATABASE TO SUPPORT THE COMPILATION OF JAPANESE LEARNERS' DICTIONARIES

Yuriko SUNAKAWA
University of Tsukuba
sunakawa@sakura.cc.tsukuba.ac.jp

Jae-ho LEE
University of Tsukuba
jhlee.n@gmail.com

Mari TAKAHARA
University of Tsukuba
takahara.mari.ge@u.tsukuba.ac.jp

Abstract

The number of Japanese language learners outside Japan, especially of advanced level learners, is increasing yearly. From the intermediate level onwards, they could profit from bilingual Japanese learners' dictionaries in their native language, but in most linguistic areas of the world only very simple dictionaries for beginners and for tourists are available. Our project therefore aims at supporting the compilation of Japanese language learners' dictionaries for intermediate and advanced learners by building a database of contents needed when editing a Japanese language learners' dictionary, and offering it online. This 4 year project is going to be running from 2011 to 2014. Two surveys were conducted: a survey of the vocabulary used in textbooks of Japanese as a foreign language and a quantitative survey on the targeted area of the Japanese language in a large-scale corpus, in order to select the list of words to be included in the database, and a general list of basic vocabulary for Japanese language instruction was created. At present, usage examples are being compiled on the basis of this vocabulary list, and a database system is being developed. A prototype of a database search interface and download system has been completed. The database is going to include various types of information which are considered to be useful for learners, such as grammar, phonetics, synonyms, collocations, stylistics, learners' errors etc. These are presently being studied in detail to be made public in 2014.

Keywords

Japanese language learners' dictionary, lexicography, dictionary editing support, bilingual dictionary, database, basic vocabulary for Japanese language instruction

Izvleček

Število učencev in študentov japonskega jezika zunaj Japonske, posebej na višjih nivojih, narašča iz leta v leto. Od srednjega nivoja dalje so za učenje koristni dvojezični učni slovarji, ki vključujejo uporabnikov materni jezik, a za večino jezikov na svetu obstajajo le zelo preprosti slovarji za začetnike ali za turiste. Zato je cilj tega projekta sestaviti bazo podatkov, ki so

potrebni v učnem slovarju japonščine, in jo ponuditi na spletu, zato da bi s tem podprli urejanje japonskih učnih slovarjev za srednjo in nadaljevalno stopnjo. Projekt bo trajal 4 leta, od leta 2011 do leta 2014. Doslej sta bili izvedeni dve raziskavi, ki sta služili kot osnova za izbor besedišča v bazi podatkov: analiza besedišča učbenikov japonščine kot tujega jezika ter kvantitativna raziskava ciljnega jezikovnega področja v obsežnem korpusu. Na osnovi tega je bil izoblikovan seznam osnovnega besedišča japonščine za splošno rabo. Trenutno sta v teku urejanje primerov rabe teh besed ter razvoj sistema za urejanje in objavljanje podatkov, izdelan pa je prototip spletnega vmesnika za iskanje po bazi podatkov in prenašanje podatkov iz baze. Načrtuje se vključitev informacij, za katere se predvideva, da bodo koristne učencem, kot so informacije o slovnici, glasoslovju, sinonimih, kolokacijah, slogu in kulturi. Delo poteka s ciljem, da se baza javno objavi leta 2014.

Ključne besede

učni slovar japonskega jezika, slovaropisje, slovaropisna podpora, dvojezični slovar, baza podatkov, osnovno besedišče za učenje japonščine

1. Introduction

In 2009 there were more than 3,650,000 Japanese language learners outside Japan: a 28,7-fold increase in 30 years.[1] The number of learners taking the Japanese-Language Proficiency Test is also increasing, especially at the advanced levels. In 2009, the number of test takers at the advanced levels (levels 1 and 2) had increased by 6.4 times since 1999, and its ratio to the total number of test takers increased from 55 % to 76 % in 10 years.[2]

A useful tool for Japanese language learning is a language learners' bilingual dictionary including the learners' mother tongue and developed on the basis of the characteristics of their mother tongue. Particularly from the intermediate level onwards, students have more opportunities to read and write on their own, and therefore need a learners' dictionary which satisfies both the needs of receptive and productive tasks. However, the majority of learners around the world are provided only with simple dictionaries for beginners or for tourists, except for countries like China and Korea, where there are many learners of Japanese.

The development of dictionaries requires enormous financial and human resources. For the production of a Japanese language dictionary for native speakers in which one of the present authors was involved, for example, a strong team of experienced dictionary writers and editors together with the editorial board of a

1 http://www.jpf.go.jp/j/japanese/survey/result/index.html (July 21st, 2012) Kaigai no nihongo kyōiku no genjō: nihongo kyōiku kikan chōsa 2009-nen gaiyō ("The present situation of Japanese language education abroad: Research on institutions with Japanese language education; 2009 summary")

2 Numbers are calculated by authors based on the statistical data obtained on the site 'Changes in the number of candidates for the Japanese Language Proficiency Test' http://www.jlpt.jp/statistics/index.html (July 21st, 2012)

publishing company spent nearly 10 years of trial and error before completion. In the field of Japanese language learning around the world, which is a very poor market compared to that of the Japanese language dictionary market for native speakers, the financing and manpower needed for compiling a dictionary from scratch are simply not available.

However, the appearance of a strong medium, the Internet, has greatly changed the scene. Publishing a dictionary in paper form through a publisher involves a considerable financial and temporal investment, and its distribution in different countries may face problems due to differing publishing and marketing conditions. If it is published on the web, on the other hand, almost no extra cost is needed and only two problems need to be solved: the creation of the contents needed for the dictionary, and the development of a system that can be used by learners. The problem of distributing learners' dictionaries has largely been solved by internet use, and conditions are becoming ripe to offer a dictionary free of charge anywhere in the world.

The problems that remain to be solved are the creation of dictionary contents and the development of a system for making the contents available to users. The present project aims at building an electronic database with the contents necessary for a Japanese learners' dictionary, and offering this database to all areas of the world over the internet. Dictionary editors of individual areas may make use of any information in this database for further processing, or add new information particular to their area and eventually make their own web dictionary to be published free of charge or at a low price.

One existing web dictionary should be mentioned here: the multilingual Reading Tutor Web Dictionary (http://chuta.jp/ Kawamura et al., 2012). This dictionary was developed as a dictionary tool for Reading Tutor, a reading support system for Japanese language learners. Presently it includes 20 languages and this number is expected to increase. The Reading Tutor Web Dictionary has been an ambitious try to broaden the possibility of a bilingual dictionary in many different languages. However, since it is based on a preset monolingual Japanese dictionary, it is difficult for editors in different linguistic areas to freely reshape it and edit their own bilingual dictionary. In order to develop a dictionary which is useful for intermediate and advanced learners, the editors should be able to work on a unique dictionary for learners of their own linguistic area, taking into consideration contrastive research on Japanese and the learners' native language. The main novelty of our approach lies in the fact that the "database for Japanese learners' dictionary editing support" is not aimed at producing a dictionary, but rather at offering the general information on word usage, with appropriate usage examples, which is considered to be necessary to foreign learners of any language background. In this sense, this project is a wholly new attempt at creating the necessary environment for bilingual dictionary compilation for learners of any mother tongue.

Our project team, based on the conditions described above, is set to build a database with all necessary information for editing Japanese learners' dictionaries, and

support editors of bilingual Japanese learners' dictionaries around the world. The project is supported by a Japanese government grant-in-aid for scientific research ("Basic research A") and is running from April 2011 for 4 years up to 2014, under the name "Research for the formulation of basic grounds for the construction of a general database for the development of Japanese language learners' dictionaries". The following sections present a general description of the project.

2. Organisation of the project team

The present project has two teams, a construction team which builds the database to support the editing of Japanese learners' dictionaries, and a research cooperation team which supports the activities of the first team.

The database construction team has 30 members. Besides the leader, Yuriko Sunakawa, there are 11 research members, 18 affiliated researchers, and one part-time researcher. Members are divided into groups, including a Japanese language research group, a corpora research group etc., and investigate methods for including word-usage information into the data base, or for the use of corpus studies in dictionary description, while also being involved in the construction of the database itself. Within the Japanese language research group, there are sections for research on (1) collocation, (2) synonyms, (3) grammar information, (4) cultural information and (5) phonetic information. The corpus research group includes sections for (1) corpus information, (2) basic vocabulary, (3) learner corpora and (4) language processing. Each section is engaged in research in its own area.

The team of collaborating researchers counts 47 members, including many who reside outside of Japan. Collaborating researchers in Japan are involved in English lexicography, corpus linguistics, Japanese language research, research on foreign languages such as French, English or German, Japanese language teaching research etc. All of them are engaged in research which can contribute to Japanese learners' lexicography from different points of view, and share their research findings with all other members through oral and written presentations.

Collaborating researchers outside Japan are involved in research on Japanese lexicography, corpus compilation, Japanese language and language education research, and while sharing the results of their research with other members of the project like domestic cooperating researchers, they also conduct surveys and investigations needed for the construction of the database, such as surveys on the needs of Japanese language learners outside Japan, contribute to the compilation of learners' corpora, investigate learners' errors, etc.

3. Data base to support editorial work of Japanese language dictionary

The development of dictionaries requires a detailed description of Japanese language use based on actual research results of contrastive studies and linguistic research. Since the present project aims at supporting lexicographic work aimed at intermediate and advanced learners of Japanese, we are building a database containing the following information:

a) headword usage information (information on meaning, grammar, phonetics, synonymy, collocation, style, culture, corpus-based frequency etc.);

b) example sentences based on typical usage examples for each subsense, edited at an appropriate level for intermediate and advanced learners;

c) information on frequent errors by Japanese language learners.

This information is going to be published with a Creative Commons license, thus enabling dictionary editors anywhere in the world to freely access our database, be it for a profit or nonprofit undertaking, to process the information according to their own area's needs and eventually develop bilingual learners' dictionaries for speakers of their own native language.

In order to build the above-mentioned database, our work plan within the research period is the following:

a) selection of basic vocabulary needed by Japanese language learners;

b) research aimed at including word usage information on basic Japanese vocabulary into the database, making use of existing Japanese language corpora;

c) research in error analysis in order to include error information into the database, making use of existing learners' corpora;

d) editing of usage examples which are appropriate for intermediate and advanced learners, on the basis of typical usage examples extracted from existing Japanese corpora, for each subsense of each headword;

e) development of a system for organising word usage information, and of a corpus search tool aimed at editing word usage information;

f) development of a system to make the database public, and suitable tools for users.

4. Making the vocabulary list

As a first step for creating the database, we constructed a list of lexemes to be included and described in the database. In the field of teaching Japanese as a foreign

language, the vocabulary list of the old version of the *Japanese Language Proficiency Test: Test Content Specifications* (hereinafter "old JLPT list") is well known and is still being widely used as a basic source of data for educational yardsticks, teaching material development, vocabulary research etc. However, in the present project we developed our own basic Japanese instructional vocabulary list instead of using the above mentioned JLPT list, due to the following reasons.

1. The "old JLPT list" was created more than 30 years ago and does not reflect recent vocabulary changes.
2. Out of concern for learners abroad, it does not include culturally-bound terms.
3. Its scale of difficulty was set up for test compilation and not for language education.

First of all, concerning point 1. above, the "old JLPT list" was compiled manually in 1980s and, although twice revised, it has not changed much from the 80s and does not correspond to the new changes in Japanese language vocabulary (cf. Oshio et al., 2007). Specifically, loanwords are poorly represented and vocabulary which can enliven expression, such as onomatopoeia, is largely missing.

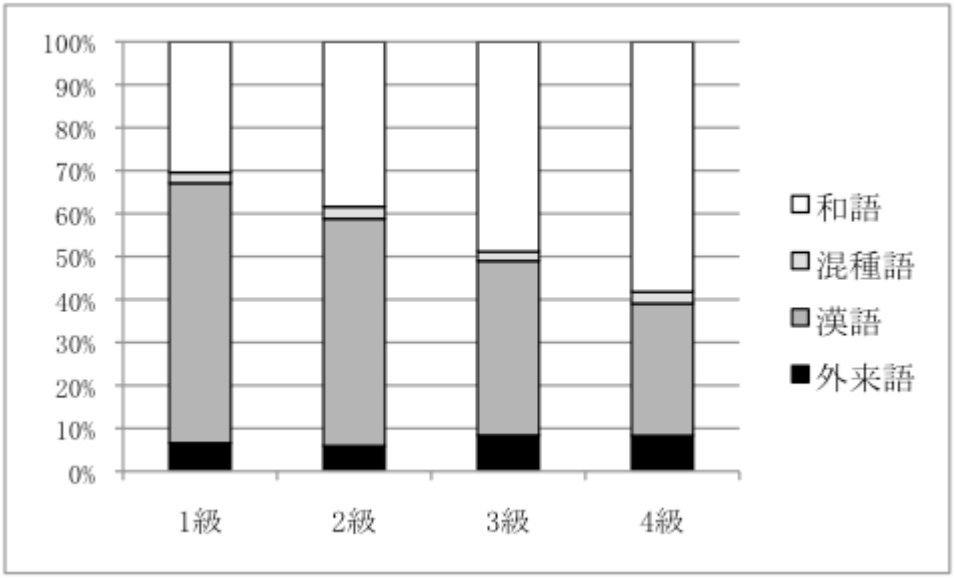


Figure 1: Vocabulary distribution in the "former test vocabulary list", □ originally Japanese words, □ words of mixed origins, □ words of Chinese origin, ■ other borrowings

Figure 1 shows the distribution of words in the "old JLPT list". Vocabulary for level 4 includes as much as 50% native Japanese words, but as the level gets increases, so does the ratio of words of Chinese origin. The most problematic is the ratio of loanwords. As can be seen in Figure 1, loanwords make up less than 10 % of each level. Such a small number of loanwords does not correspond to actual language use in contemporary Japanese society and needs to be revised. Onomatopoetic words are also poorly represented in the "old JLPT list", which includes only a few, such as *nikoniko, pikapika, furafura* and *wakuwaku*.

Concerning point 2., the intentional exclusion of culturally-bound terms, including names of food, animals and plants, out of concern for test-takers abroad is problematic. This choice is based on the understanding that the Japanese Language Proficiency Test is meant to test language ability and not cultural knowledge. Considering the list was compiled for the purposes of this kind of test, the policy of the "old JLPT list" is in itself very reasonable, but decidedly removed from the reality of Japanese language education in which Japanese society and cultural matters are part of the curriculum.

Lastly, with regard to point 3., the "old JLPT list" is aimed at the evaluation of Japanese language ability, nothing more and nothing less. The "old JLPT list" is not intended for the development of teaching materials and dictionaries, and problems will inevitably occur if it is used for these purposes. The test-making perspective diverges from the perspective of language education in many respects, particularly in with regard to the setting of a difficulty scale (levels of vocabulary items). The difficulty scale in the test is set from the perspective of "levels of Japanese which may be assumed to be known to students" and not the perspective of educational goals, as "levels of Japanese one would like students of a certain level to know".

Taking into account the three problems described above, we conducted a survey of the vocabulary of Japanese language textbooks, and quantitative research on the target language area in a large-scale corpus. On the basis of this research, we compiled a general-purpose list of basic vocabulary for Japanese language education (hereinafter "instructional vocabulary list").

The main aims of the "instructional vocabulary list" are: (1) to make a vocabulary list for Japanese language education including authentic vocabulary items; (2) to label vocabulary items according to their various characteristics so that the vocabulary list will be useful for dictionary development as well as various needs in classroom situations; (3) to create a vocabulary list which various users in and outside Japan may share through the web. To accomplish these aims, we have conducted the following:

In order to realise (1), we conducted a vocabulary survey making use of corpus data and natural language processing technology.

In order to realise (2), we added information about the degree of difficulty of each vocabulary item, based on the subjective judgement of Japanese language teachers, and decided to add semantic information according to the "categorised vocabulary list" (*Bunrui goi hyou*).

In order to realise (3), we decided to format electronic data in CSV format, which can be used with proprietary spreadsheet software (such as Microsoft's Excel®), as well as with plain text editors.

4.1 Compilation procedure

The "Instructional vocabulary list" was compiled in the following 4 steps.

1. Vocabulary extraction: vocabulary was extracted from morphologically analysed corpus data.
2. Manual editing: noise and boiler-plate was manually removed.
3. Subjective assessment: the difficulty level of the extracted vocabulary was subjectively assessed by five teachers of Japanese.
4. Index construction: each vocabulary item was tagged with semantic information and frequency data obtained from the corpus.

The following section presents each step in detail.

4.1.1 Vocabulary extraction

As a first step towards the compilation of the "Instructional vocabulary list", we extracted content words (excluding particles and auxiliary verbs) from the "Japanese textbook corpus" and from the "Yahoo!Chiebukuro" and "Books" part of the 2009 edition public data of the "Balanced Corpus of Contemporary Written Japanese" (http://www.tokuteicorpus.jp/), after having morphologically analysed all texts. We then calculated the frequency of all content words and compiled a list of all words appearing more than 5 times.

The "Japanese textbook corpus" mentioned above is a corpus of texts extracted from 100 Japanese language textbooks. It was compiled for research purposes by the present authors and is not publicly available. It includes major Japanese language textbooks used in Japan and abroad, in a balanced proportion of textbooks from beginning to advanced level. The "Balanced Corpus of Contemporary Written Japanese" is a balanced corpus of the Japanese written language developed by the National Institute for Japanese Language and Linguistics, but since at the time our project started the complete corpus was not yet publicly available, we used the monitor data version published in 2009. The "Books" section of this data amounted to 40,000,000 words, and was considered sufficient for the compilation of our vocabulary list.

Morphological analysis was conducted using MeCab (Kudo, 2011) and UniDic (Den et al., 2007). When extracting vocabulary, we used not only the short

morphological unit *tan-tan'i* (短単位), but also morpheme N-grams[3], combining multiple morphemes into longer units, as exemplified below.

1. Examples of 2-grams: *aien-ka* (愛煙家 "habitual smoker"), *aisu-koohii* (アイスコーヒー "iced coffee"), *ai-tsugu* (相次く "come in succession"), *aite-kata* (相手方 "other party"), *ao-shingou* (青信号 "green traffic light")
2. Examples of 3-grams: *ami-no-me* (網の目 "net mesh"), *iku-tsu-ka* (幾つか "a few"), *i-kko-date* (一戸建て "detached house"), *ichi-do-ni* (一度に "all at once"), *ichi-nin-mae* (一人前 "a portion for one person; a grown-up"), *itsu-de-mo* (何時で も "anytime"), *itsu-made-mo* (いつまで も "forever"), *ima-ni-mo* (今にも "at any moment"), *ima-hito-tsu* (今一つ "not quite"), *unten-menkyo-shou* (運転免許証 "driver's license"), *o-kyaku-san* (お客さん "guest"), *o-jii-san* (おじ いさん "grandfather"), *o-jii-chan* (おじ いちゃん "grandpa")

Examples in (1) are 2-grams, i.e. sequences of two morphemes. For example, *aien-ka* (愛煙家 "habitual smoker") is a word composed of *aien* (愛煙 "love of smoking") and *-ka* (-家 "person"), *aite-kata* (相手方 "other party") is a word composed of *aite* (相手 "partner") and *kata* (方 "person"), etc. Examples in (2) are 3-grams, are sequences of three morphemes, such as *ami-no-me* (網の目 "net mesh"), which is composed of *ami* (網 "net"), *no* (の "of") and *me* (目 "mesh, grain").

The extracted N-grams were manually checked and cleaned of noise, resulting in a list of 18,010 lexical units.

4.1.2 Subjective assessment

If the list is to be used in the context of Japanese language teaching, a difficulty scale needs to be designed, and lexical units must be labelled according to this scale as words to be learned at a certain level. However, vocabulary cannot be categorised only mechanically; subjective labels by teachers of Japanese, based on their experience and intuition, must also be included. However, subjective judgement is not necessarily based on scientific evidence and it is therefore difficult to handle such an index when building a database which must be consistent and systematic. In our project we therefore asked five teachers of Japanese with ten or more years of teaching experience to judge - each one by his or herself - the difficulty of the words, collected all responses, processed them statistically and labelled all lexical elements by degree of difficulty.

Raters were asked to classify the list of 18,010 words which was obtained as described in 4.1.1., dividing it into six categories: beginning - 1st part, beginning - 2nd

3 N-grams are a model of language proposed in the field of natural language processing, consisting of strings of N elements, which can be characters or morphemes: a morpheme 3-gram is composed of three consecutive morphemes, a 4-gram of four, etc.

part, intermediate - 1st part, intermediate - 2nd part, advanced - 1st part and advanced - 2nd part. The raters were instructed to judge the level of word difficulty from the perspective of classroom instruction, as the level at which words should be introduced during classroom learning.

The average rating for each word was computed, and the word list divided into six levels. The final decision of word level was taken in two rounds. During the first round, we first computed the average level score of all five raters, and then also the k-value agreement of each rater's score with the average score. When the agreement between the rater and the average score was less than 0.5, we excluded that rater's score and computed again the average of the remaining raters' scores, taking that as the final score. We were thus able to exclude those scores which were markedly different from the rest. The final results of this procedure are presented in Table 1.

Table 1: Results of subjective assessment

Vocabulary level	Number of vocabulary items	Examples
1. Beginning - 1st part	426	*oyasumi* お休み "good night", *tonari* 隣 "neighbour", *petto* ペット "pet", *onegaishimasu* お願いします "please", *ohayougozaimasu* おはようございます "good morning", *watashi* 私 "I, me", *warui* 悪い "bad", *otearai* お手洗い "toilet", *otousan* お父さん "father"
2. Beginning - 2nd part	800	*ryouri* 料理 "food", *ryokou* 旅行 "travel", *reizouko* 冷蔵庫 "refrigerator", *resutoran* レストラン "restaurant", *remon* レモン "lemon", *wakai* 若い "young", *wasureru* 忘れる "forget", *gokurousama* 御苦労様 "thank you for your work", *irasshaimase* いらっしゃいませ "welcome", *annai* 案内 "introduction, guidance"
3. Intermediate - 1st part	2,323	*ikebana* 生け花 "ikebana", *iken* 意見 "opinion", *ikou* 以降 "from ... onwards", *ikooru* イコール "equal to", *iremono* 入れ物 "container", *ironna* 色んな "various", *iwa* 岩 "rock", *iwau* 祝う "celebrate, congratulate", *ugokasu* 動かす "move", *usotsuki* うそつき "liar", *uchuujin* 宇宙人 "creature from outer space"
4. Intermediate - 2nd part	6,482	*iryou* 医療 "health care", *iryou* 衣料 "clothing", *irui* 衣類 "clothing", *irogami* 色紙 "colored paper", *iwaigoto* 祝い事 "celebration", *iwakan* 違和感 "sense of incongruity", *insutorakutaa* インストラクター "instructor", *ushinau* 失う "loose", *ushirosugata* 後ろ姿 "view from behind", *uttae* 訴え "lawsuit", *kakudo* 角度 "angle"

Vocabulary level	Number of vocabulary items	Examples
5. Advanced - 1st part	6,401	*kakudan* 格段 “remarkable”, *kakuchou* 拡張 “extension”, *kakutei* 確定 “decision”, *kakutou* 格闘 “fight”, *gattai* 合体 “union”, *gatchiri* がっちり “solidly”, *kabuseru* かぶせる “cover”, *kafusoku* 過不足 “too much or too little”, *kabunushi* 株主 “shareholder, stockholder”, *kabegami* 壁紙 “wallpaper”, *kahogo* 過保護 “overprotective”, *kankakuki* 感覚器 “sensory organ”
6. Advanced - 2nd part	1,578	*kanten* 寒天 “agar-agar”, *kannushi* 神主 “Shinto priest”, *kampa* カンパ “fund-rising campaign, contribution”, *kampan* 甲板 “deck”, *gyouten* 仰天 “astonishment”, *kyokushou* 極小 “infinitesimal”, *kirifuki* 霧吹き “sprayer”, *guzuru* 愚図る “grumble”, *kusemono* くせ者 “cunning person; fishy thing”, *kuchidutae* 口伝え “oral tradition”, *kuppuku* 屈伏 “surrender”, *kumikyoku* 組曲 “suite”
Total	**18,010**	

The results of a comparison between the vocabulary included in our “instructional vocabulary list” and the vocabulary list of the “old JLPT list” are presented in Table 2.

Table 2: Old JLPT and “Instructional vocabulary list” comparison

		Levels of the old JLPT vocabulary list					
		Level 1	**Level 2**	**Level 3**	**Level 4**	**Not included**	**Total**
Levels of the Instructional Vocabulary List	1. Beginning - 1st part	0	4	7	375	40	426
	2. Beginning - 2nd part	6	79	208	341	166	800
	3. Intermediate - 1st part	94	921	410	105	793	2,323
	4. Intermediate - 2nd part	884	1,944	93	37	3,524	6,482
	5. Advanced - 1st part	1,290	449	13	0	4,649	6,401
	6. Advanced - 2nd part	118	32	0	0	1,428	1,578
	Total	**2,392**	**3,429**	**731**	**858**	**10,600**	**18,010**

The lexical units marked as “Not included” in Table 2 are words which are part of the “instructional vocabulary list”, but not included in the “old JLPT list”, and amount to 10,600 lexical units. When comparing the “instructional vocabulary list” with the “old JLPT list”, loanwords appear to be a particularly problematic area. For example, words such as *jazu* (ジャズ “jazz”), *kameraman* (カメラマン “cameraman”), *tisshu* (ティッシュ “tissue”) are categorised as Level 1 (the most difficult) in the “old JLPT list”, while in our “instructional vocabulary list” they are set in level Beginning - 2. On the other hand, words such as *rekoodo* (レコード“ (audio) record”), *firumu* (フィルム “film”), *haadodisuku* (ハードディスク “hard disc”), which are categorised as Level 4 (the easiest) in the “old JLPT test”, are set in level Intermediate - 2 in our “instructional vocabulary list”. These differences are likely to reflect the changes in word usage which have occurred since the 1980s, when the “old JLPT list” was compiled.

4.1.3 Index construction

The “instructional vocabulary list” is now being turned into a database by adding the following indexes to each lexical item.

1. Vocabulary ID
2. Standard written form
3. Readings
4. Vocabulary difficulty level
5. Part of speech
6. Type of word by origin
7. Old Japanese Language Proficiency Test Level
8. Meaning classification
9. Accent information

1 is a unique number for the lexical item. 2 was prepared in accordance with the dictionary *Gendai kokugo hyouki jiten* (“Dictionary of modern language written forms”). 3 is the reading of the standard written form, 4 is one of the six difficulty levels determined by subjective measurement as described above. 5 complies with the part-of-speech divisions of UniDic. 6 is also based on UniDic’s labels, and indicates whether the word is a native word, loan from Chinese, loan from other languages, a word of mixed origin, or a fixed expression. 7 is the level in the “old JLPT list”, 8 is a semantic label which complies with the categorisation of NINJAL’s *Bunrui goihyou* (“Table of vocabulary by semantic categories”), while 9 indicates the accent pattern of the word. Table 3 shows a concrete example of a few indexed lexical items.

Table 3: Sample of the Instructional Vocabulary List

語彙ID	標準的表記	読み	語彙難易度	品詞	語種	旧試験語彙級	意味分類	アクセント情報
10	アート	アート	中級前半	名詞–普通名詞–一般	外来語		体–活動–芸術・美術	1
40	アイスコーヒー	アイスコーヒー	初級前半	名詞–普通名詞–一般 and 名詞–普通名詞–一般	外来語		体–生産物–食料–飲料・たばこ	6
109	明かり	アカリ	中級前半	名詞–普通名詞–一般	和語	2級	体–生産物–機械–灯火 | 体–自然一自然一光	0
222	足掛かり	アシガカリ	上級後半	名詞–普通名詞–一般	和語		体–関係–空間–点	3
294	厚かましい	アツカマシイ	中級後半	形容詞–一般	和語	2級	相–活動–心–自信・誇り・恥・反省	5
262	温まる	アタタマル	中級前半	動詞–一般	和語	2級	用–自然–物質–熱	4

The next step, based on these indexes, is going to be the compilation of definitions aimed at dictionary compiling, and the writing of usage examples.

5. Current progress

Currently, we are creating a database and developing a system aimed at dictionary compilation, on the basis of the “instructional vocabulary list” in Table 2. In particular, we are now in the process of compiling and editing usage examples on the basis of sense definitions. Definitions are compiled with reference to the database of basic words with familiarity indexes by word sense (Amano & Kobayashi, 2008) and the data being developed by Kawamura Yoshiko et al. within the system Reading Tutor. In particular, we are using the data in *The Reading Tutor Web Dictionary* (Kawamura et al. 2012), including 8000 lemmas with word ID, example ID, headword, reading, note, part of speech, sense, and 27,000 examples for particular subsenses. Conversely, the word usage data and usage examples being developed within our project are going to be included in *The Reading Tutor Web Dictionary*, and work in both projects is being carried out in close cooperation.

Usage examples are being edited by external collaborators, who were asked to write three original examples and select three corpus examples for each word sense. Original examples are to be written using only vocabulary not beyond the difficulty

level of the headword, and we are developing special software to support example compilation.

The system development group has developed a prototype system to search the database online and download data, as shown in Figure 2.

Figure 2: Prototype of a dictionary search system

When a user inputs a headword in the search box and launches the search, items which completely or partially match the search string are shown on the interface. Some lexical items are linked to pictures. By clicking on the button marked *Gogi o hyouji* (語義を表示 "Show meaning"), the user can see definitions and examples for the headword *hana*, as shown in Figure 3.

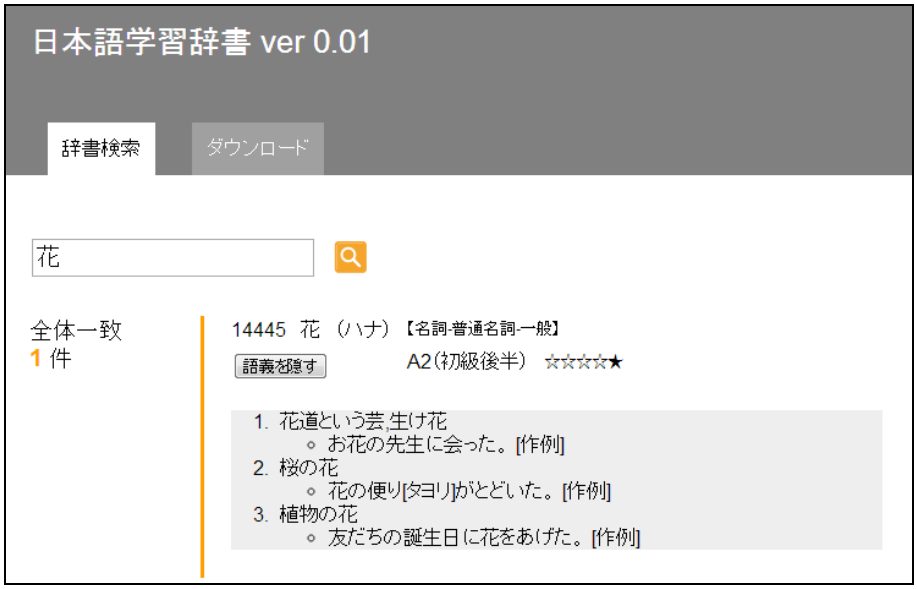


Figure 3: Display of word sense and examples

The list of partial match results includes words such as *hanabi* (花火 “fireworks”), *kaki* (花器 “flower vase”), *hanazakari* (花盛り “full bloom”), *hanataba* (花束 “flower bouquet”), *kadan* (花壇 “flower bed”), *hanabatake* (花畑 “flower field”), *kabin* (花瓶 “flower vase”), which begin with the character 花 (*hana* or *ka*, “flower”), and words such as *kaika* (開花 “blossoming”), *nanohana* (菜の花 “rape blossoms”), *ikebana* (生花 “ikebana”), *senkouhanabi* (線香花火 “toy fireworks, sparkler”), *kusabana* (草花 “flowering plant”), which include this character. As can be seen from these examples, partial match results are headwords which contain the characters of the search string, not the word *hana*.

Scrolling down the page, one can see lexical items which are semantically related to the headword, in the section labelled *kanrengo* (関連語 “related words”). For example, a search for the word *ringo* (りんご “apple”) produces the results shown in Figure 3.

Figure 4: Words related to *ringo* (“apple”)

Words listed as “related words” under the headword *ringo* (りんご “apple”) in this figure include other words for fruits and plants, such as *aamondo* (アーモンド “almond”), *appuru* (アップル “apple”), *abokado* (アボカド “avocado”), *ichou* (いちょう “ginkgo”), *ume* (梅 “Japanese apricot”), etc. The system is based on the *Bunrui goihyou* (“Table of vocabulary by semantic categories”) mentioned above. The word *ringo*, for example, is categorised in *Bunrui goihyou* as “noun > nature > flora > trees”, and all words pertaining to the same category are extracted by the system and displayed as related words.

In recently developed search systems, the user can perform complex searches and choose between complete match (searching for words which match the search string in its entirety) or partial match (searching also words which only partially match the search string), and between initial partial match (for words beginning with the search string) or final partial match (for words ending with the search string), or search by pronunciation, or by written form, etc. Non-expert users, however, may be confused by too detailed search possibilities. We therefore decided to offer a simple system where

the user only inserts a search keyword and clicks once, and the system then displays both complete and partial matches. As for the written form of the search string, in order to search for words of Chinese origin, the search string must be input in Chinese characters, while loanwords from other languages are displayed only if searched for in their standard written form, in katakana, but words of native origin are displayed both when the search string is input in Chinese characters and when it is in hiragana. Native Japanese homophones or words which are written with different Chinese characters depending on the sense in which they are used, can thus be obtained by inserting one single search string in hiragana. For example, if the search string *kiru* (きる) is searched for, the system will display information for both *kiru* (着る "wear") and *kiru* (切る"cut").

Partial match searches are useful for examining compound nouns and verbs, since by inserting a verb or part of it, one can search for all compound verbs containing it. For example, a search for the hiragana string *kakeru* (かける) produces a complete display of all compound verbs containing it, such as headwords *oikakeru* (追い掛ける "chase, pursue"), *oshikakeru* (押しかける "throng to, crash in, barge in"), *koshikakeru* (腰掛ける "to sit"), *shikakeru* (仕掛ける "start; prepare; challenge") etc. The user can check the meaning of unknown words by clicking on the button *Gogi o hyouji* (語義を表示 "Show meaning"), as explained above, obtaining sense definitions and examples as shown in Figure 3 and 5.

Figure 5: Display of compound verb senses and examples

The search function “Related words”, on the other hand, is useful for investigating synonyms and antonyms. A search for the word *sawayaka* (さわやか “fresh, pleasant”), for example, yields a result list including synonyms such as *kokoroyoi* (快い “pleasant”), *sugasugashii* (すがすがしい “fresh”), *soukai* (爽快 “fresh, refreshing”), *kokochiyoi* (心地よい “kokochiyoi”), etc., and antonyms such as *uttoushii* (うっとうしい “gloomy, disagreeable”), *fukai* (不快 “unpleasant, disagreeable”) etc.

6. Further stages and development plan

As mentioned above, the database is going to include not only semantic information and usage examples, but also other pieces of information that are useful for users, i.e. information on phonetics, synonymy, collocation, stylistics, culture and errors. At present, each team is working on how to describe these items and in what form to upload them on the database. The progress and results of these teams will be shared by all members of the project by holding research meetings.

The time plan for the coming 3 years is the following:

Year 2012

- work on the basic design of the database
- start with description of basic word usage
- set up the environment for data processing
- public release of a part of the data (vocabulary list for Japanese language learning)
- start publishing information on the project’s homepage
- compilation of usage examples

Year 2013

- construction of a corpus retrieval system
- partial release of the data (the system and the corpus tools)

Year 2014

- completion of the corpus
- release of the final set of data with usage examples
- workshops to popularise the database and its use

By the end of 2012, we will start advertising on our homepage and make the prototype of our database public. These will be improved in 2013 by adopting users’ feedback. By the end of 2014, the last year of the project, the database will be completed. After completion, results of our project will be made public through workshops, targeting particularly users outside Japan in order to encourage practical use of the database as a resource for developing dictionaries for learners of Japanese. We plan to continue with our project according to the time line as described above.

(This study is subsidised by the Japan Society for the Promotion of Science, Grants-in-aid No. 23242026.)

References

Amano, S. [天野成昭], Kobayashi, T. [小林哲生] (ed.) (2008). *Kihongo deetabeesu - gogibetsu tango shinmitsudo* [基本語テ▯ ータへ▯ ース-語義別単語親密度] (*"Database of basic words - Word familiarity index for single subsenses"*). To□kyo□: Gakken [学研].

Den, Y. [伝康晴], Ogiso, T. [小木曽智信], Ogura, H. [小椋秀樹], Yamada, A. [山田篤], Minematsu, N. [峯松信明], Uchimoto, K. [内元清貴] & Koiso, H. [小磯花絵] (2007). Koopasu nihongogaku no tame no gengo shigen: keitaisokaisekiyou denshika jisho no kaihatsu to sono ouyou [コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用] ("The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics"), Nihongokagaku [日本語科学] ("Japanese linguistics") 22: 101-122.

Kawamura, Y. et al. (2012). *The Reading Tutor Web Dictionary - Chuta no web jisho* [チュウ太の web 辞書]. Retrieved from: http://chuta.jp/

Kudo, T. (2011). MeCab: yet another part-of-speech and morphological analyzer. Retrieved from http://mecab.sourceforge.net/

Lee, J.[李在鎬] (2011). Nihongo nōryoku shiken no chōsen: Atarashii nihongo nōryoku shiken wo rei ni (Kokusai kōryū kikin jigyō repōto 14) [日本語能力試験の挑戦～新しい日本語能力試験を例に(国際交流基金事業レポート 14)] (“The challenge of the Japanese Language Proficiency Test: The case of the new Japanese Language proficiency test” [Japan Foundation Project Report 14]), *Nihongogaku* [日本語学] *(“Japanese Language”)* 30(1), 95-107.

National Institute for Japanese Language and Linguistics [国立国語研究所] (ed.) (2004). *Bunrui goihyō zouhokaiteiban* [分類語彙表増補改訂版] *("Table of vocabulary by semantic categories").* Tokyo: Dainihontosho [大日本図書].

Oshio, K. [押尾和美], Akimoto, M.[秋元美晴], Takeda, A.[武田明子], Abe, Y.[阿部洋子], Takanashi, M.[高梨美穂], Yanagisawa, Y.[柳澤好昭], Iwamoto, R.[岩元隆一], & Ishige, J.[石毛順子] (2008). Atarashii nihongo nōryoku shiken no tame no goi-hyō sakusei ni mukete [新しい日本語能力試験のための語彙表作成にむけて] (“Towards a new vocabulary list for the new Japanese Language Proficiency Test”), *Kokusai kōryū kikin nihongokyōiku kiyō* [国際交流基金日本語教育紀要] *("Japan Foundation Japanese Language Journal"),* 4,71-86.