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.

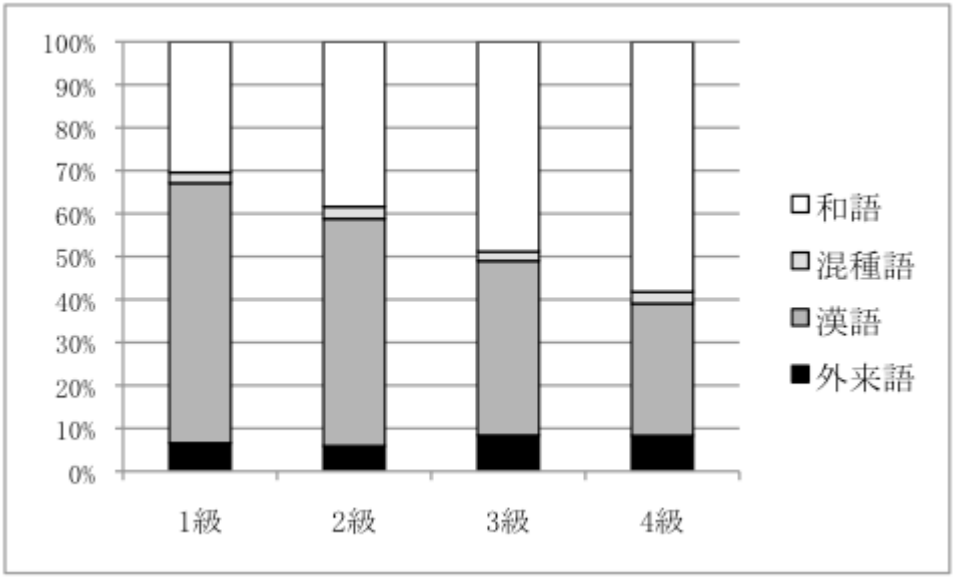

Figure 1: Vocabulary distribution in the "former test vocabulary list", □ originally Japanese words, □ words of mixed origins, □ words of Chinese origin, ■ other borrowings

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.

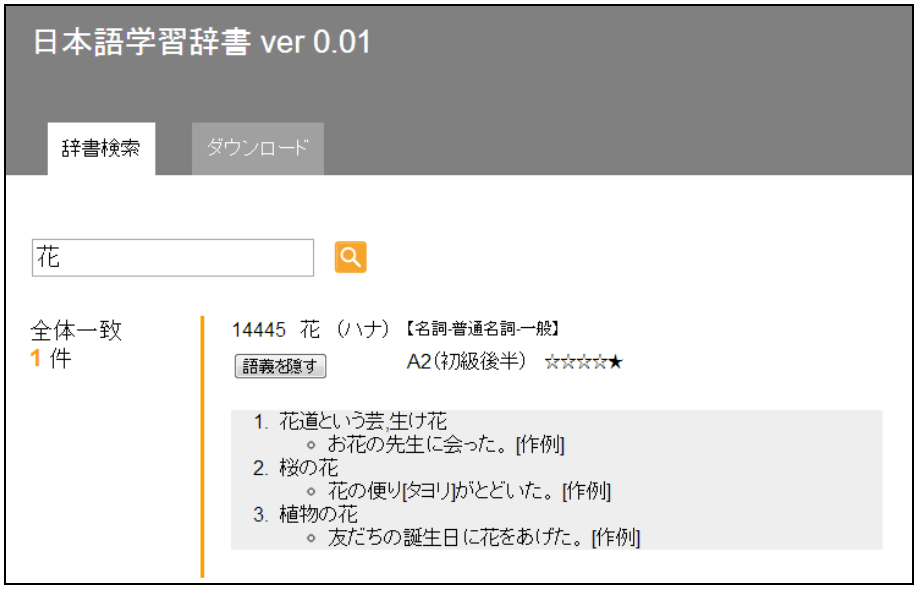

Figure 3: Display of word sense and examples

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▯ ータヘ▯ 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.