Automated collection of Japanese word usage examples from a parallel and a monolingual corpus

Kristina Hmeljak Sangawa[1], Tomaž Erjavec[2], Yoshiko Kawamura[3]
University of Ljubljana, Jožef Stefan Institute, Tokyo International University

Abstract

Examples are an important source of information on word usage for language learners, but existing reference sources for Japanese as a second language are limited. This paper describes two projects for the automated collection of word usage examples. Examples extracted from an ad-hoc Japanese-Slovene parallel corpus were included into jaSlo, a Japanese-Slovene learners' dictionary, and examples extracted from a monolingual web-harvested 400 million word corpus of Japanese were selected to be used as supplementary examples for Chuta, a multilingualized dictionary for learners of Japanese as a second language.

Keywords: example, word usage, corpus example, Japanese web corpus, Japanese-Slovene parallel corpus, readability.

1. Introduction

Examples are an excellent source of semantic, syntactic, morphological, collocational and pragmatic information for dictionary users, especially for those who are not familiar with lexicographic metalanguage and prefer inferring (or guessing) word usage from examples rather than from definitions or symbols. However, although many examples can be included into electronic dictionaries where space is not as limited as in paper dictionaries, good examples, which should be typical, natural and surrounded by typical context (Fox 1987:37, Atkins and Rundell 2008:330), are costly to produce (Rychlý *et al*. 2008:425). This is especially crucial in the case of voluntary-based or low-budget academic lexicographic projects with limited human and financial resources.

For learners and teachers of Japanese as a second language, examples of word usage can be found in existing dictionaries, textbooks and corpora, but each of these sources has some limitations. Starting with the *Dictionary of basic Japanese usage for foreigners* (Bunkachô 1971), a number of monolingual, bilingual and bilingualized dictionaries for learners of Japanese as a second/foreign language have been produced in the last three decades, and all of them contain usage examples. However,

[1] University of Ljubljana, kristina.hmeljak@ff.uni-lj.si
[2] Jožef Stefan Institute, tomaz.erjavec@ijs.si
[3] Tokyo International University, kawamura@tiu.ac.jp

dictionaries covering at least 10,000 headwords (*i.e*. the vocabulary considered to be needed by intermediate to advanced learners of Japanese, cf. Tamamura (1984), Japan Foundation (2004)) such as the *Informative Japanese dictionary* (Nihongo no kai 1995) with 11,000 headwords, or *Kodansha's Communicative English-Japanese Dictionary* (Sharpe 2006) with 22,000 entries usually do not offer more than 2-3 examples per headword. On the other hand, those containing more examples per headword such as the monolingual *Japanese dictionary: learning language the fun way* (Takano 2004) with 750 headwords, the dictionary of functional words *Nihongo bunkei jiten* (Group Jamashii 1998) with 3000 entries, the bilingual *Kodansha's Basic English-Japanese Dictionary* (Makino et al. 1999) with 4500 headwords, or the *Effective Japanese Usage Guide* (Hirose & Shoji 1994) with 708 headwords, do not cover all the vocabulary needed by intermediate to advanced learners of Japanese.

There is a very large number of monolingual and bilingual dictionaries for native speakers of Japanese, which also contain examples. However, Japanese large monolingual dictionaries for native speakers, such as *Kôjien (*Shinmura 2008) or *Daijirin* (Matsumura 2006), are generally too difficult for foreign learners, especially those based on historical principles, which contain examples of archaic language, such as *Kôjien.* The very numerous monolingual dictionaries for elementary-school children, such as *Reikai shôgaku kokugo jiten* (Tajika 2009), *Shôgaku shin kokugo jiten* (Kai 2002) or *Challenge shôgaku kokugo jiten* (Minato 2008), or for high-school native speakers, such as *Meikyô kokugo jiten* (Kitahara 2002) or *Shin meikai kokugo jiten* (Yamada 2005), which do contain easier examples with phonetic script, do not usually contain more than 2-3 examples per word. On the other hand, examples in bilingual dictionaries for Japanese native speakers are usually targeted at explaining and translating idiomatic examples in the foreign language. The Japanese translations in these dictionaries do not always exemplify the most typical usages of Japanese words, but rather collocations and phrases which are difficult to translate.

Another obvious source of examples are corpora. Some Japanese corpora have been made available in the last few years. A 39 million word demo version of the Balanced Corpus of Contemporary Written Japanese (BCCWJ) being compiled at the National Institute for Japanese Language (Maekawa 2008), was made available on the institute's portal in March 2009. Examples from the Japanese Web as Corpus (JpWaC), a 400 million word corpus of web text (Srdanović Erjavec *et al*. 2008), can be looked up via the Sketch Engine (Kilgarriff *et al*. 2004). Search engines such as Google or Yahoo can also be used as a source of usage examples, although the lists of search results given in such engines are neither linguistically representative nor sortable by linguistic criteria. However, results from such corpora searches can be overwhelming for language learners with limited linguistic ability.

We therefore decided to create an intermediate tool which would give the users (especially language learners) more examples than a conventional dictionary or textbook, but which would be less overwhelming than corpora or search engine

results. Having limited human and financial resources, we tried to make the best possible use of available resources.

In the following sections we present two projects where usage examples were automatically collected to be included in two electronic dictionaries for learners of Japanese. Both dictionaries are being compiled as academic projects with the help of volunteer editors and progressively published on the web.

2. Examples from a parallel corpus

Bilingual usage examples were collected for jaSlo (Erjavec *et al.* 2006), a Japanese-Slovene dictionary for Slovene learners of Japanese, which is being compiled at the University of Ljubljana and has been gradually published at http://nl.ijs.si/jaslo/ since 2001. The dictionary's latest edition (2006) had c. 10,000 Japanese headwords and c. 25,000 Slovene translational equivalents, but only 2,370 usage examples. We therefore decided to use an existing collection of parallel texts, augment it and structure it into a parallel corpus to use it as a source of examples.

2.1. Parallel corpus compilation

Due to a lack of competent translators for the language pair Japanese - Slovene, hardly any translation had been produced between these two languages before the establishment of the Japanese studies program at the University of Ljubljana in 1995. However, since its establishment, a small collection of parallel texts in electronic form accumulated at the department, consisting of lecture handouts (academic texts on the history, literature, geography and society of Japan, prepared in Japanese by visiting lecturers and translated into Slovene by university staff), and student coursework (texts on the Japanese society translated from Japanese into Slovene and texts on tourism translated from Slovene to Japanese, in both cases translated by students and thoroughly revised by the teacher in charge of the translation course). Given their availability in electronic form, we decided to use them as sources of examples for our Japanese-Slovene dictionary. However, since most texts were quite challenging for language learners, we decided to add some more readable texts, from which examples for intermediate students could be obtained. We therefore digitized parts of 6 contemporary Japanese novels which were translated into Slovene in the last decade and the only novel that has been translated from Slovene into Japanese up to now, by scanning them and manually correcting OCR output.

Lastly, in order to obtain a larger corpus, we searched the web for translated pages in Japanese and Slovene. We searched for pages in Japanese script within the domain .si (Slovenia), and found 150 pages, of which only 4 were relevant translations, while the others were either brief Japanese summaries of longer Slovene texts or poor quality machine translation products. We also searched for pages written in Slovene in the .jp domain, using Google's advanced search function to limit the target language, which yielded a few hundred pages, but only two of them were found to be relevant, while all

the others were wrongly identified as Slovene and actually written in some other Slavic language. We carried out searches for the words “Japanese” and “Slovene” without specifying their internet domain and found pages, mostly in English, localized into many languages including Japanese and Slovene, where the names of the two languages appeared in a menu for language selection. Such indirect translations are certainly not ideal sources of dictionary examples, but given the lack of direct translations, we decided to include them, after manually checking them and removing all segments with unreliable or missing translations.

All texts were normalized into plain text files and sentence-aligned using Wordfast PlusTools (www.wordfast.net). Alignment was manually validated, and paragraphs with missing or mistaken translations were removed. The complete corpus was lemmatized using Chasen (Matsumoto *et al*. 2007) for the Japanese part and “totale” (Erjavec *et al*. 2005) for the Slovene part. Combining all available parallel texts, we obtained a parallel corpus of 7,914 translation units (sentence pairs), corresponding to 226,220 Japanese morphemes and 171,261 Slovene words, and composed of the following subcorpora: translated lecture handouts (13.5%), revised student translations (24.5%), literary fiction (15.7%), and multilingual web pages (46.3%).

2.2. Extraction of examples from the parallel corpus

All Japanese headwords in the dictionary were searched for in the corpus, yielding examples for 4,648 lemmas, *i.e*. approximately half the dictionary entries. When more than 6 examples were found, only the shortest 6 were retained, since short sentences tend to be syntactically simpler. Lexical complexity was not taken into account at this stage, but all sentences are accompanied by a translation into the user’s native language, and therefore presumably understandable.

Examples were appended to the dictionary entries, and graphically separated from existing examples, as can be seen in Figure 1. This version of the dictionary was published at http://nl.ijs.si/jaslo/cgi/jaslo-eg.pl.

Each corpus example is followed by a link (in the form of an arrow), which leads to a page with information on the title, place and date of publication or URL of the source text and translated text, source language and target language, author’s and translator’s name when available, thus indicating in what sort of genre the word can be found.

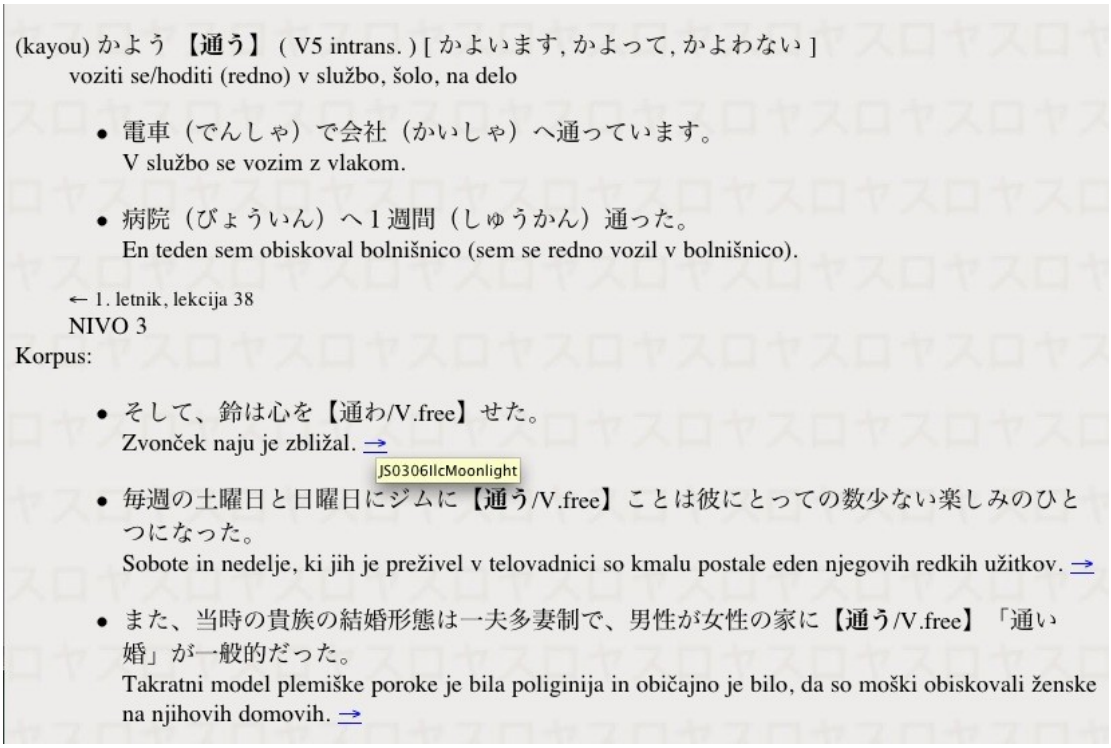

Figure 1. Example of a jaSlo dictionary entry with corpus examples

2.3. Evaluation of extracted examples

Automatically extracted corpus examples did not go through the usual editorial process of dictionary entries, i.e. analysis of a corpus of examples, synthesis of the dictionary entry and editing of appropriate examples. It cannot therefore be expected, especially given the very small size of our corpus, that automatically extracted examples should cover all senses of a word or give all its most typical syntactic and collocational patterns. We evaluated a sample of 80 lemmas of intermediate difficulty, randomly chosen from the Japanese Language Proficiency Test specifications (Japan Foundation 2004) to test coverage of word senses and usefulness.

We found that for 51% of these lemmas, all senses were covered, while for the remaining half of the lemmas some senses did not appear in any of the examples. This indicates the need for a larger corpus to achieve better coverage. For 10% of the lemmas, new translational equivalents were found which had not yet been included in the latest version of the dictionary: 2% were context dependent or unnecessarily liberal translations which were not deemed useful, but as much as 8% were useful additions to the dictionary. Moreover, corpus examples for 4% of the sampled headwords contained idiomatic expressions, collocations or multi-word units which were not present in the original dictionary, and therefore useful additions to it.

Given the fact that the dictionary was compiled by a small team of contributors with little lexicographic experience and that there are few Japanese-Slovene contrastive studies or other reference materials, it is not surprising that the dictionary still needs improvement. These corpus examples are therefore going to be useful not only for the general users, but also for the editors of the dictionary during future revisions.

Regrettably, 8% of the examples were assigned to the wrong dictionary entry because of lemmatization errors in Chasen's morphological analysis. This indicates the need for a future manual validation of example selection, while for the time being users are warned that corpus examples were extracted automatically and may contain errors.

In the future, we plan to augment the parallel corpus to achieve better coverage, and to enhance the example selection procedure to include readability criteria (including vocabulary coverage, syntactic patterns and context independence) and typicality criteria (including collocational, morphosyntactic and stylistic patterns).

3. Examples from a monolingual corpus

In a similar pilot study, a corpus of usage examples was collected to be combined with Chuta (Kawamura & Kaneniwa 2006, published at http://chuta.jp/), a multilingualized Japanese learners' dictionary in which sense divisions, definitions and usage examples are first prepared by a team of Japanese native speakers, teachers of Japanese as a second language, and subsequently translated into different languages by an international team of editors (Vietnamese, Russian, English, Turkish, Bulgarian, Korean, Chinese, Portuguese, Spanish, German, Czech, Malay, Kirghiz, Marathi, Slovak, Thai, French, Italian, Finnish, Nahuatl, Slovenian, Indonesian, Hungarian, Tagalog, Arabic and Romanian, in decreasing order of number of edited lemmas). 8,721 Japanese entries have been published at present, while bilingual entries are still being edited. To increase the number of examples for general users and also help editors of bilingual entries, examples were extracted from a web-harvested, lemmatized and PoS tagged 400 million word corpus of Japanese, JpWaC (Srdanović *et al*. 2008).

3.1. Compilation of JpWaC-L2, a monolingual corpus of example sentences

A 100 million word sample of the JpWaC corpus was extracted, starting from the beginning of the corpus until the required size was obtained. As the corpus texts are sorted according to the URL, and the start of the URL is essentially random, this does not unduly bias the corpus composition. The corpus is composed of texts, each marked by its source URL, and these, in turn, composed of sentences, each annotated by the sequential number of the sentence in the text. All words in the corpus were annotated with their difficulty level according to the Japanese Language Proficiency Test specifications, ranging from 4 (easiest words) to 1 (hardest words). Words not appearing in the JLPT list were assigned level 0. Each sentence was furthermore annotated with quantitative information for the number of tokens in the sentence, words by levels, punctuation symbols and numerals.

Single sentences were extracted from this sample corpus to create a corpus of example sentences, JpWaC-L2, according to the following criteria. We retained sentences which:

a) are not duplicate (only the first occurrence of duplicate sentences is retained);

b) are between 5 and 25 tokens in length (to exclude very short sentences, which are usually only sentence fragments, and very long sentences, which are difficult to understand);

c) contain less than 20% of punctuation marks or numerals (to retain only text rich sentences);

d) contain at most 20% of level 0 words (to exclude sentences with a high proportion of difficult or foreign words);

e) do not contain words written with non-Japanese characters (to exclude strings such as URLs, e-mail addresses or text in other languages);

f) do not contain any opening or closing quotes or parentheses (to avoid segmenting errors);

g) do not start with punctuation (to exclude improperly segmented fragments);

h) end in the kuten character, "。", the Japanese equivalent of a full stop or period (to include full sentences); i) contain at least one predicate – verb or adjective (again, to exclude sentence fragments).

The intention of the above filters, obtained by empirical testing and evaluation, is to retain only well-formed, text-rich and relatively simple sentences.

Five subcorpora of different difficulty levels where then extracted from this collection of sentences, by selecting – for each subcorpus – only sentences with at least 10% of words belonging to the subcorpus difficulty level, and no words from a more difficult level. The size of the subcorpora is shown in Figure 2. Since both corpus and example collection were automated, relatively little manual labour was required to obtain a sizeable collection of examples.

Corpus	Sentences	Words	% jpWaCS-L2
jpWaCS	3,225,572	100,001,186	
jpWaCS-L2	859,416	13,395,667	100.00
jpWaCS-L2_0	351,935	5,536,969	40.95
jpWaCS-L2_1	34,777	403,470	4.05
jpWaCS-L2_2	96,161	1,172,911	11.19
jpWaCS-L2_3	26,894	264,979	3.13
jpWaCS-L2_4	9,830	79,473	1.14

Figure 2. JpWaC-L2 corpus and subcorpora contents

The corpora are available for Web concordancing at http://nl.ijs.si/jaslo/cqp/ through the search interface shown in Figure 3.

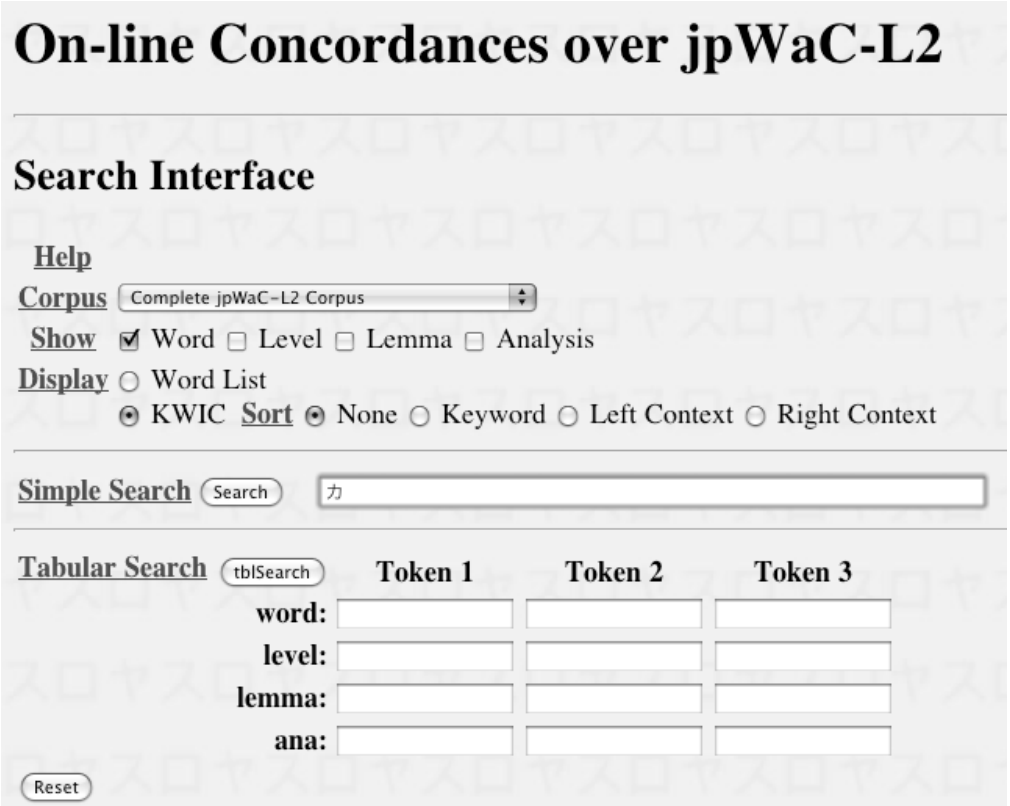

Figure 3. JpWaC-L2 corpus search interface

Users can choose to search the complete JpWaC-L2 corpus or only one subcorpus of the desired difficulty level. The "simple search" box can be used to search for any string (one or more words), while the "tabular search" section allows for combinations of searches of specific word forms, any form of a certain lemma, any word of a certain level, or any occurrence of a certain part of speech. The search result is a concordance where each line (sentence) is linked to its wider context within the original JpWaC corpus, so that users can see the paragraph containing the sentence by clicking on the word. Concordances can be sorted on the left or right context, to investigate frequent collocational patterns and – in the case of verbs and adjectives – also flectional patterns. By selecting the option "Show: Word / Level / Lemma / Analysis" the user can choose to see the difficulty level, lemma and part-of-speech tag assigned by Chasen to each word in each concordance line, as seen in Figure 4.

1	**word**	それ に は もう	慣れ た	。
	lemma	それ に は もう	慣れる た	。
	ana	N.Pron.g P.c.g P.bind Adv.g	V.free Aux	Sym.p
	level	4 4 4 4	3 4	-
2	**word**	早い 英語 に も だいぶ	慣れ た	ころ な のに 。
	lemma	早い 英語 に も だいぶ	慣れる た	ころ だ のに 。
	ana	Ai.free N.g P.c.g P.bind Adv.g	V.free Aux	N.bnd.Adv Aux P.Conj Sym.p
	level	4 4 4 4 3	3 4	4 4 3 -
3	**word**	この 生活 に も	慣れ た	と いう 事 。
	lemma	この 生活 に も	慣れる た	と いう 事 。
	ana	Adn N.Vs P.c.g P.bind	V.free Aux	P.c.r V.free N.bnd.g Sym.p
	level	4 3 4 4	3 4	4 4 3 -

Figure 4. JpWaC-L2 concordance for the verb form "nareta"

3.2. Evaluation of extracted examples

A sample of 10 lemmas for each level (4 to 0), for a total of 50 lemmas, was randomly extracted to evaluate the quantity and quality of examples in the corpus.

The average number of examples for these lemmas was 717 examples when searching through the whole JpWaC-L2 corpus, 497 examples in the level 0 subcorpus, 80 in the level 1 subcorpus, 278 in the level 2 subcorpus, 134 in the level 3 subcorpus, and 73 in the level 4 subcorpus, indicating that a sufficient amount of examples was found in each subcorpus.

Evaluating the grammaticality and acceptability of extracted sentences, it was found that less than 5% of the sentences were ungrammatical (containing garbled content or evident mistakes). A small percent of sentences were found to be assigned to the wrong lemma, due to Chasen's lemmatization error, and consequently sometimes also to the wrong difficulty level. The vast majority of the sentences, however, were well formed. The difficulty level assigned to the sentences, as measured according to the vocabulary contained, generally reflected their readability and comprehensibility. Shorter sentences were sometimes found not to be very informative without a wider context, while longer sentences, even if containing only basic vocabulary, sometimes contained challenging multi-word idiomatic expressions and syntactic structures. Although context for short sentences can be retrieved with a click, the criteria which define sentence length need further investigation.

4. Conclusion and further work

Two projects for the extraction of word usage examples from a parallel and a monolingual corpus were presented. In both cases, existing resources and automated processes were used to produce a collection of examples with relatively little manual

labor. Plans for further work include a usability study, parallel corpus enlargement, and an enhancement of the selection procedure (applying criteria proposed by Mizuno *et al*. 2008 and Nishina & Yoshihashi 2007) and the measurement of example typicality, which has not yet been addressed by previous research on Japanese dictionary example selection, both in terms of vocabulary (collocations) and in terms of structure (morphological and syntactic patterns).

References

A. Dictionaries

BUNKACHÔ. (1971). *Gaikokujin no tame no kihongo yourei jiten - Dictionary of basic Japanese usage for foreigners*. Tokyo: Oogurashô insatsukyoku – Ministry of Finance Printing Bureau.

GROUP JAMASHII (1998). Nihongo bunkei jiten. Tokyo: Kurosio.

HIROSE, M. and SHOJI, K. (1994). *Effective Japanese Usage Guide*. Tokyo: Kodansha.

KAI, M. (2002). *Shôgaku shin kokugo jiten*. Tokyo: Mitsumura kyôiku tosho.

KITAHARA, Y. (2002). *Meikyô kokugo jiten*. Tokyo: Taishukan shoten.

MAKINO, S., NAKADA, S., OHSO, M. and JACOBSON, W.M. (1999). *Kodansha's Basic English-Japanese Dictionary*. Tokyo: Kodansha.

MATSUMURA, A. (2006) *Daijirin. Dai 3 han*. Tokyo: Sanseido.

MINATO, Y. (2008) *Challenge shôgaku kokugo jiten. Dai 4 han shin dezain han*. Tama: Benesse corporation.

NIHONGO NO KAI (1995). *Informative Japanese Dictionary*. Tokyo: Shinchosha.

SHARPE, P. (2006). *Kodansha's Communicative English-Japanese Dictionary*. Tokyo: Kodansha.

SHINMURA, I. (2008). *Kôjien. Dai 6 han*. Tokyo: Iwanami Shoten.

TAJIKA, J. (2009). *Sanseidô reikai shôgaku kokugojiten. Dai 4 han*. Tokyo: Sanseido.

TAKANO, T. (2004). *Gaikokujin no tame no tanoshii nihongo jiten – Japanese dictionary, learning language the fun way*. Tokyo: Sanseidô.

YAMADA, T. (2005). *Shin meikai kokugo jiten. Dai 6 han*. Tokyo: Sanseido.

B. Other references

ATKINS, S. and RUNDELL, M. (2008). *Oxford guide to practical* lexicography. Oxford: Oxford University Press.

ERJAVEC, T., HMELJAK SANGAWA, K. and SRDANOVIĆ ERJAVEC, I. (2006). jaSlo, A Japanese-Slovene Learners' Dictionary: Methods for Dictionary Enhancement. In E. Corino *Proceedings of the 12th EURALEX International Congress*. Alessandria: Edizioni dell'Orso: 611-616.

ERJAVEC, T., IGNAT, C., POULIQUEN, B., and STEINBERGER, R. (2005). Massive multi-lingual corpus compilation: Acquis Communautaire and totale. In *Proceedings of the 2nd Language & Technology Conference, April 21-23, 2005*, Poznan: Wydawnictwo Poznańskie: 32-36.

FOX, G. (1987). The case for examples. In J. Sinclair (ed.). *Looking up. An account of the Cobuild project in lexical computing*. London: Collins: 37-49.

JAPAN FOUNDATION AND ASSOCIATION OF INTERNATIONAL EDUCATION JAPAN. (2004). *Japanese Language Proficiency Test. Test Content Specification*. Tokyo: Bonjinsha.

KAWAMURA, Y., KANENIWA, K. (2006). Kokusai kyôdô henshû ni yoru nihongo gakushûsha no tame no tagengoban web jisho no kaihatsu (Development of a multilingual web-dictionary for learners of Japanese through international collaborative editing). In Nihongo Kyôiku Gakkai (ed.) *2006nendo nihongo kyôiku gakkai shunki taikai yokôshû (Proceedings of the 2006 Japanese language teaching association spring conference)*. Tokyo: Nihongo kyôiku gakkai:61-66.

KILGARRIFF, A., RYCHLÝ, P., SMRŽ, P., and TUGWELL, D. (2004). The Sketch Engine. In Williams, G. and Vessier, S. (eds.) *Proceedings of the Eleventh EURALEX International Congress*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud: 105-116.

MAEKAWA, K. (2008). Compilation of the Balanced Corpus of Contemporary Written Japanese in the KOTONOHA Initiative. In *Proceedings of the Second International Symposium on Universal Communication – ISUC 2008*. Los Alamitos-Washington-Tokyo: IEEE: 169-172

MATSUMOTO, Y., TAKAOKA, K., ASAHARA, M. (2007). *Morphological analyzer chasen, version 2.4.0* {http://sourceforge.jp/projects/chasen-legacy/document/chasen-2.4.0-manual-j.pdf/ja/2/chasen-2.4.0-manual-j.pdf}

MIZUNO, J., OOYAMA, H., KOBAYASHI, T., SAKATA, K., EVANS, N., TANIGUSHI, M. and MATSUMOTO, Y. (2008). Nihongo dokkai shien no tame no gogigoto no yôrei chûshutsu shisutemu no kôchiku. In *Proceedings of the Workshop on Natural Language Processing for Education – The 14th Annual Meeting of the Association for Natural Language Processing*. Tokyo: Gengo shori gakkai: 31-35.

NISHINA, K. AND YOSHIHASHI, K. (2007). Japanese composition support system displaying co-occurrences and example sentences. In Furui, S. *Proceedings of the Symposium on large-scale knowledge resources (LKR2007)*, Tokyo: Tokyo Institute of Technology: 119-122. []

RYCHLÝ, P., HUSÁK, M., KILGARRIFF, A., RUNDELL, M. and MCADAM, K. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In E. Bernal and J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress*. Barcelona: Institut Universitari de Lingüística Aplicada: 425-432.

SRDANOVIĆ ERJAVEC, I., ERJAVEC, T. and KILGARRIFF, A. (2008). A web corpus and word sketches for Japanese. *Journal of Natural Language Processing* 自然言語処理 15/2: 137-159.

TAMAMURA, F. (1984) *Goi no kenkyuu to kyouiku*, Tokio: Kokuritsu Kokugo Kenkyûjô.