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

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

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

Abstract

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

Keywords

bilingual lexicography; corpus search; parallel corpus; readability level

Izvleček

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

Ključne besede

dvojezično slovaropisje; korpusno iskanje; vzporedni korpus; stopnja berljivosti

Acta Linguistica Asiatica, 2(3), 2012.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.2.3.125-140

1. Introduction - background to the project

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

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

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

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

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

1.1 Corpus-based lexicography

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

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

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

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

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

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

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

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

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

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

2. Resources for Slovene-speaking learners of Japanese

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

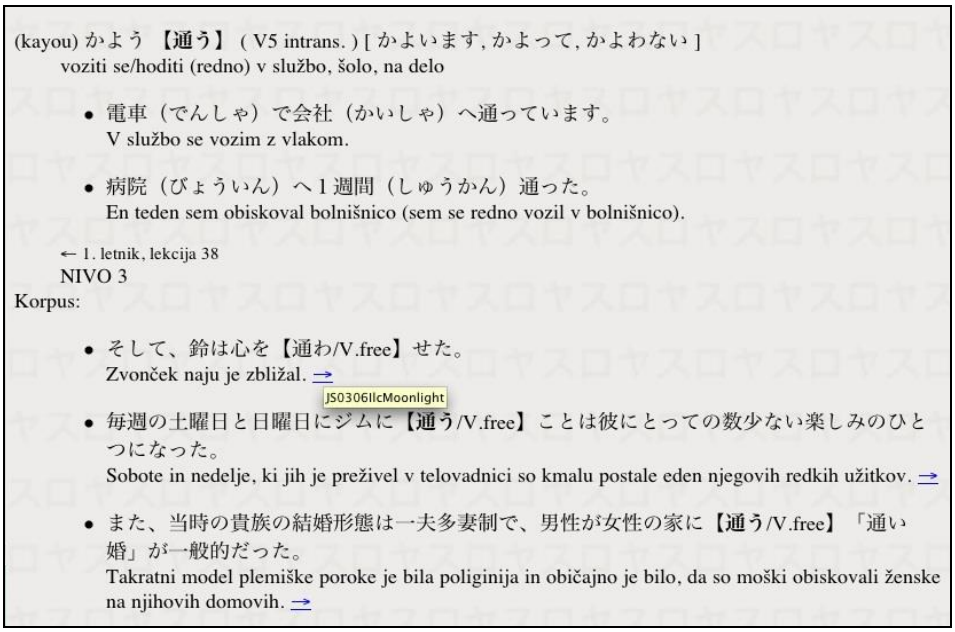
2.1 The Japanese-Slovene dictionary jaSlo

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

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

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

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


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

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

← 1. letnik, lekcija 38
NIVO 3
Korpus:

- そして、鈴は心を【通わ/V.free】せた。
  Zvonček naju je zbližal. →
  JS0306IlcMoonlight
- 毎週の土曜日と日曜日にジムに【通う/V.free】ことは彼にとっての数少ない楽しみのひとつになった。
  Sobote in nedelje, ki jih je preživel v telovadnici so kmalu postale eden njegovih redkih užitkov. →
- また、当時の貴族の結婚形態は一夫多妻制で、男性が女性の家に【通う/V.free】「通い婚」が一般的だった。
  Takratni model plemiške poroke je bila poliginija in običajno je bilo, da so moški obiskovali ženske na njihovih domovih. →

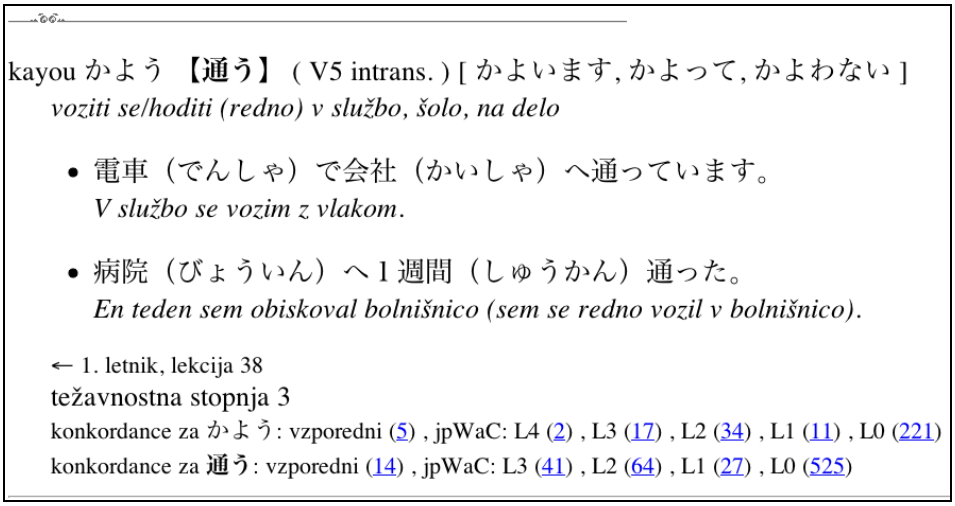
Figure 1: Example of a jaSlo dictionary entry with corpus examples in the 2009 version

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

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

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

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


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

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

← 1. letnik, lekcija 38
težavnostna stopnja 3
konkordance za かよう: vzporedni (5) , jpWaC: L4 (2) , L3 (17) , L2 (34) , L1 (11) , L0 (221)
konkordance za 通う: vzporedni (14) , jpWaC: L3 (41) , L2 (64) , L1 (27) , L0 (525)

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

2.2 The Japanese-Slovene parallel corpus jaSlo

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

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

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

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

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

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

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

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

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

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

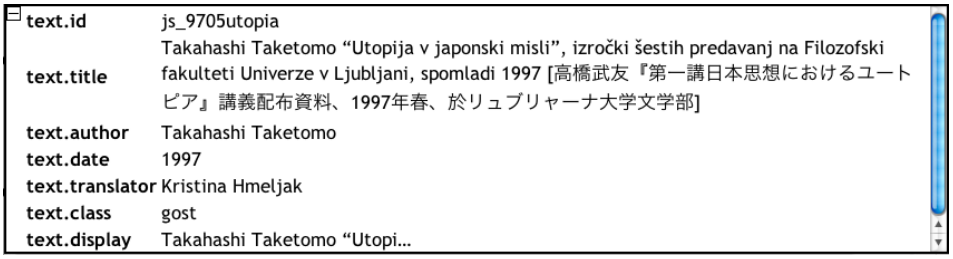


Figure 4: Display of source information for one of the documents in the corpus

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4. Conclusions and directions for further work

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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