COGNITIVE STUDIES | ÉTUDES COGNITIVES, 10
SOW Publishing House, Warsaw 2010

**KRISTINA HMELJAK SANGAWA**[1]
**TOMAŽ ERJAVEC**[2]
[1]kristina.hmeljak@ff.uni-lj.si, University of Ljubljana
[2]tomaz.erjavec@ijs.si, Jožef Stefan Institute

# THE JAPANESE-SLOVENE DICTIONARY JASLO: ITS DEVELOPMENT, ENHANCEMENT AND USE

**Abstract.** The paper presents the on-line Japanese-Slovene dictionary jaSlo, in particular the ways in which it has been used, and how it has been extended with examples mined from a parallel corpus. The paper first describes jaSlo and the structure of its dictionary entry, its Web interface for searching, and an analysis of the access logs. The use of jaSlo in the context of the Japanese reading-support tool Reading Tutor is described next, again followed by an analysis of the access logs. Also discussed is the manner in which usage examples were added to the dictionary, and an evaluation of their usefulness. The paper concludes with directions for further work.
**Keywords**: digital dictionaries, Japanese language, Slovene language, usability studies.

## 1 Introduction

A bilingual dictionary is one of the most basic and indispensable tools when learning a foreign language, but dictionary compilation is a resource-intensive process that requires a considerable investment of time and human resources. When the first program of Japanese studies in Slovenia was established at the University of Ljubljana in 1995, the need arose for Japanese language teaching materials and dictionaries for Slovene speaking students. However, Slovene language publications have a very small market, and the market for this rare language pair is even smaller, which means that the production of a Japanese-Slovene dictionary for a few hundred possible users is not a particularly profitable project that could interest a publishing house. The teachers of the Japanese studies program at the University of Ljubljana therefore decided to compile a dictionary and publish it progressively on the web, continuously striving to fully exploit available resources for this low-budget project. The first version was published in 2002 (Hmeljak Sangawa 2003), in 2003 it was converted into XML and moved to a server at the Jožef Stefan Institute (Erjavec et al. 2004). The 3rd version released in 2006 added more information to the dictionary entries, mostly acquired via third party resources (Erjavec et al.

2006), and added more entries: it has now approximately 10,000 Japanese headwords with about 25,000 Slovene translation equivalents. The dictionary is available for searching at the address `http://nl.ijs.si/jaslo/`.

The dictionary is progressively growing, both in terms of headword numbers, and in terms of structure improvement, by inclusion of publicly available data and programs, and by automatic collection of examples from a bilingual parallel corpus and from a monolingual Japanese corpus harvested from the Web. The dictionary was also included into the Japanese reading-support tool Reading Tutor (Kawamura 2000). Both tools, Reading tutor and the Web interface to the jaSlo dictionary, keep a log of user lookups. Reading Tutor's log records include the date and time of access and the text looked up, while the dictionary jaSlo records the date and time of access, the string looked up, and the number of returned hits.

In the following sections the dictionary, its enhancement and an analysis of its use are presented. Section 2 presents the jaSlo dictionary, its compilation, structure and means of access. Section 3 presents an analysis of user logs which points out which editorial decisions were effective and which need improvement. Section 4 introduces the Reading Tutor application and, again, an analysis of user logs within this reading-support tool. Section 5 deals with current work on enhancing the dictionary with usage examples automatically extracted from a Japanese-Slovene parallel corpus, describing how the corpus was compiled and how the examples are included into the dictionary. Section 6 concludes with some directions for further work.

## 2 The jaSlo Dictionary

The jaSlo dictionary began as a set of separate small glossaries prepared by the teachers and students at the Department of Asian and African Studies of Ljubljana University. The glossaries, which were mostly in tabular and HTML format, were first converted into a common encoding, all dictionary entries from these separate files merged, and manually checked. The encoding of the dictionary took into account international standards in the field, which brings with it a number of well-known advantages, such as better documentation, the ability to validate the structure of the document, simpler processing, easier integration into software platforms, longevity and easier Web deployment. The dictionary is available on the Web, via a search interface which keeps a log of user accesses.

### 2.1 Dictionary structure

For encoding the dictionary we used the XML version of the Text Encoding Initiative Guidelines, TEI P4 (Sperberg-McQueen & Burnard 2002), in particular its module for dictionary encoding.

Figure 1 presents two typical dictionary entries in jaSlo, the first one for a verb and the second for a noun. The first element, the <form> of the headword, is given in three scripts: "roma" — the transcription of the headword into Latin script (called romaji in Japanese), followed by the Japanese phonetic script kana

```
<entry id="jaslo.6784">
  <form type="hw">
    <orth type="roma">tsunagaru</orth>
    <orth type="kana">つながる</orth>
    <orth type="kanji">繋がる</orth>
  </form>
  <gramGrp><pos>V5</pos><subc>intrans.</subc></gramGrp>
  <form type="infl">
    <orth type="v-masu">つながります</orth><orth type="v-te">つながって</orth>
  </form>
  <trans><tr>biti povezan z</tr></trans>
  <eg>
    <q>工場（こうじょう）と宿舎（しゅくしゃ）は廊下（ろうか）でつながっている。</q>
    <tr>Tovarna in nastanitveni (bivanjski) prostori so povezani s hodnikom.</tr>
  </eg>
  <eg>
    <q>国（くに）の安全（あんぜん）につながる問題（もんだい）</q>
    <tr>problem, povezan z varnostjo države</tr>
  </eg>
  <xr type="related"><lbl>prim.</lbl><ref>つなぐ</ref><lbl>tr.</lbl></xr>
  <xr type="related"><lbl>prim.</lbl><ref>つなげる</ref><lbl>tr.</lbl></xr>
  <usg type="level">2</usg>
  <note type="admin" resp="KHS">2006-09-02 Check</note>
  <note type="admin" resp="TER">2005-07-11 Add Romaji</note>
  <note type="admin" resp="TER">2005-07-10 Add levels</note>
  <note type="admin" resp="VOJ">2005-02-22 V (465)</note>
</entry>
<entry id="jaslo.7601">
  <form type="hw">
    <orth type="roma">donata</orth>
    <orth type="kana">どなた</orth>
  </form>
  <gramGrp><pos>N</pos></gramGrp>
  <trans><tr>kdo</tr></trans>
  <xr type="politeness"><lbl>spošt. za</lbl><ref>だれ</ref></xr>
  <xr type="lesson" n="L1.1"><xref>1. letnik, lekcija 1</xref></xr>
  <usg type="level">4</usg>
  <note type="admin" resp="KHS">2006-09-02 Check</note>
  <note type="admin" resp="TER">2005-07-11 Add Romaji</note>
  <note type="admin" resp="TER">2005-06-06 Merge</note>
  <note type="admin" resp="KHS">2003-03-12 L1 (945)</note>
  <note type="admin" resp="KHS">2004 N (4862)</note>
</entry>
```

**Figure 1.** Two sample dictionary entries in jaSlo

(hiragana or katakana), as is standard in Japanese learner dictionaries, and finally in its standard written form, which in the case of verbs includes Chinese characters (kanji) and syllabic script (kana). In the case of words which in standard Japanese are only written in phonetic script, the third orthographic type does not appear. The <form> segment is followed by grammatical information, which in most cases

only indicates part of speech, but in the case of verbs also transitivity information, followed by another <form> segment which includes conjugated forms (for verbs and adjectives). The next element, <trans>, is the most central one, and gives the Slovene translation equivalents of the headword, in some cases followed by examples (the <eg> element). The pronunciation of Chinese characters which appear in the examples are given in brackets, in the Japanese phonetic script hiragana.

The next element contains references to related words: synonyms, especially those of a different politeness level (second sample), verbs with a different transitivity value (first sample), antonyms etc. In the case of words from the core vocabulary which is included in the Japanese language textbooks currently used at Ljubljana University for introductory courses, a reference is also made to the number of the lesson in which the word appears for the first time. Vocabulary used in a certain lesson can therefore easily be collected through the dictionary interface, which is convenient both for teachers when preparing vocabulary exercises, as well as for students reviewing language material. The last content element which also appears to users on the webpage is the difficulty level of the entry according to the Japanese Language Proficiency Test Specifications (Japan Foundation 2002). The last part of each entry is administrative information tracing the compilation history, which is not made visible to users. In addition to the elements given in the example, a subset of the entries also includes etymology (for loan-words), and encyclopaedic descriptions of proper names and Japanese culturally bound terms.

The grammatical tagset (content of <gramGrp>) was devised on the basis of the set used by the Japanese morphological analyser Chasen (Matsumoto et al. 2007), one of the most widely used morphological analyzers for the Japanese language. 19 different labels for the main parts of speech were adopted from Chasen. The part of speech labels used in the legacy files were semi-automatically converted to this common standard, which makes it easier to use jaSlo with Chasen tagged corpora, as explained in Section 4 (use of Chasen to morphologically analyze texts for reading support) and Section 5 (extraction of examples from Chasen-tagged corpus).

### 2.2 Using the dictionary

The dictionary is deployed via a Web-based interface, available at `http://nl.ijs.si/jaslo/`, which allows full text searches by string or word on the dictionary, with optional restriction of the match to headword or translation, and filtering by word class or difficulty level. The Web interface is localised to Slovene, Japanese and English. The user's browser is assumed to offer Unicode support and have a Japanese-language font installed but, apart from that, no requirements are imposed on the client architecture. The server is implemented as a Perl CGI script, which accepts the search parameters and returns the entries that match the query, and displays them in HTML.

The dictionary can be searched from the interface with the following options (c.f. Figure 2): limit search by word class (nouns, verbs, adjectives, phrase), and search in whole text, headwords or headword translations only. While the dictionary was conceived as a Japanese-Slovene dictionary and its Slovene-Japanese counterpart is

planned for a later stage, it can be (and indeed is – see section 3) used to look up translations of Slovene words into Japanese, by entering a Slovene search string and searching through the entire dictionary text (headword translations and examples), although the information thus obtained can be confusing when there are numerous Japanese headwords with the same Slovene translation.


**Figure 2.** Search interface and display of results

## 3 Analysis of lookups in jaSlo

Each query to jaSlo is logged together with the time and number of returned entries (without client machine address, thus preserving privacy). Being able to analyze what users search for in the dictionary and how, helped us to begin tailoring the

dictionary to user needs. This section presents an analysis of the log file in order to guide our further work on the dictionary.

From September 2006 (when the current 3rd version of the dictionary was released) to April 2008, 19,579 searches were recorded, corresponding to 11,938 search string types. Most strings were single words, but quite some phrases were also found as input strings (e.g. "všeč si mi" (I like you), "ime mi je" (my name is), "vse najboljše za rojstni dan" (happy birthday), "ikaga desu ka" etc.). With the exception of some common phrases (greetings, politeness formulae etc), most of such searches do not return any hit. While there is another tool on our server – Reading Tutor – which can analyze longer stretches of texts to find translational equivalents for each word included, this tool is offered separately and cannot be recalled from within the dictionary search interface. Here a clearer explanation of the function of each of the two tools on this server (Reading tutor for longer texts, jaSlo for single words or multi-word units), would help users avoid this kind of unhelpful searches.

Less than half of the search strings were in Japanese characters, as can be seen in Table 1.

| | |
|---|---|
| hiragana | 33% |
| kanji or kanji-kana mixture | 20% |
| katakana | 5% |
| romanized Japanese words | 17% |
| Slovene words | 24% |
| proper names | 0.6% |
| English words | 0.4% |

**Table 1.** Percentage of search strings by character type

Possible explanations for this massive use of romanized Japanese are that users are not comfortable with typing Japanese (beginners or students with little typing experience) or that they access the dictionary from computers with no Japanese script support (on public computers in libraries, internet cafés etc.). Our dictionary includes romanized forms of all headwords alongside their kana and kanji forms, and the very large amount of romanized Japanese search strings confirmed our choice of including them. However, since only headwords contain Latin script forms, it might be useful to romanize the whole Japanese content of the dictionary (inflected forms and examples) in order to improve the search hit ratio.

Given the considerable number of searches for proper names (mostly Slovene personal names), it might be useful to include katakana forms of the most common Slovene names, as well as Japanese proper names which could be easily obtained from freely available vocabulary lists (e.g. Unidic). Searches for English words only return a hit when they appear in an etymological note for katakana words, since the dictionary does not include English otherwise, so users must have soon realised that this is not an English dictionary.

There were even two Hangul and two Arabic search strings, and a few meaningless character strings, but overall users seem to be using the dictionary for what

it is made for: looking up Japanese words, often also Slovene words. The top 20 search strings are given in Table 2.

| **search string** | **No. of searches** | **translation** | **search string** | **No. of searches** | **translation** |
|---|---|---|---|---|---|
| dober dan | 82 | good day | medved | 32 | bear |
| ljubezen | 67 | love | hiša | 32 | house |
| dan | 44 | day | a | 26 | |
| ljubim te | 41 | I love you | pes | 25 | dog |
| 日本 | 36 | Japan | igra | 24 | game |
| zdravo | 36 | hallo | tsuku | 23 | reach, be attached |
| hvala | 35 | thank you | mesto | 21 | city, place |
| jaz | 34 | I | san | 20 | Mr./Ms. |
| avto | 34 | car | love | 20 | |
| sonce | 32 | sun | ljubiti | 20 | to love |

**Table 2.** Most searched-for strings

The very small number of Japanese search strings among the top 20 is mostly due to the fact that each Japanese word can be and presumably was actually searched for in different forms: latin script, hiragana or kanji, which are counted as separate searches.

We were particularly interested in search strings which did not return any hits, because these are ideal candidates for inclusion in our next dictionary revision. Many words were found which were not included in the dictionary because they do not figure in Japanese vocabulary frequency lists, although they denote concepts which are used rather frequently in Slovenia (čebela – bee, šipek – rosehip, stalagmit – stalagmite etc.), or simply interesting to the users of our dictionary (prevajalec – translator, pozitivna energija – positive energy, idiot – idiot etc.).

Many zero hit logs were searches of Japanese words in latin script while the function "search Slovene translations only" was on, or searches of words in the wrong word class, e.g. searching the word カボチャ (cabocha, "pumpkin") among verbs etc. Such searches returned 0 results although words were actually in the dictionary. Here a simpler default user interface (combined with a non-default advanced interface where searches could be limited to determined categories, as in our first interface), a function which looks up words in fields other than headwords, or a fuzzy search function when no result is found in a limited part of the dictionary might help avoid such problems.

A third cause for missed hits was the use of capitals: the search mechanism of our dictionary is cap-sensitive, so that e.g. "IGRA" in capital letters returns 0 hits, while "igra" in small letters returns a few appropriate lines. Given the rather erratic use of capitals in search strings, it would be better to make searches case-insensitive.

This analysis of user logs thus confirmed some of our choices, foremost the decision to log all queries and keep a track of the way the dictionary is being used, but also pointed out possible improvements which could make the dictionary more

user-friendly: a clearer explanation of the function of each tool on the same server, a simpler user interface, less restrictive default search options, more comprehensive romanisation and the inclusion of words which do not appear in general Japanese vocabulary lists.

## 4 Reading Tutor

The "Reading Tutor" (`http://language.tiu.ac.jp/`) is a Web based on-line Japanese reading support system composed of a dictionary tool, a level detection tool, and a collection of learning materials and quizzes. The dictionary tool analyses any text input by on-line users using the Japanese morphological analyzer Chasen (Matsumoto et al. 2007), links every token in the text to one of Reading Tutor's dictionaries (Japanese definitions, Japanese-English and Japanese-German in the original version), and then presents the hyperlinked text alongside a glossary of all words it contains. Users can then read through the text and summon up readings and meanings of unknown words by simply clicking on them.

The Reading Tutor lexica are encoded in XML, according to their own document type definition. The Reading Tutor DTD is quite complex, with numerous elements, quite a few of them required. In order to include jaSlo into Reading Tutor we wrote an XSLT stylesheet that converts our TEI encoding into the schema for Reading Tutor, and the jaSlo dictionary was then added to the Reading Tutor. In 2007 a mirror server was set-up at `http://nl.ijs.si/jaslo/chuta/` and an screenshot example is given in Figure 3.

### 4.1 Analysis of lookups in Reading Tutor

As with jaSlo, we log the accesses to Reading Tutor, by recording the time of the request and the submitted text. In its first year of public access, from May 2007 to April 2008, Reading Tutor's Slovene module has recorded 592 lookups, scattered over 153 days in the whole year, with an average of slightly less than 2 accesses per day and a peak of 44 accesses on 2nd August 2007. Considering that in Slovenia there are presumably not more than 400 Slovenian speaking learners of Japanese at present, these rather modest figures are not very surprising, but they do indicate that the service needs some more publicity.

Especially in the first months of operation there were many access logs of texts which were clearly input only in order to try how the tool works: chunks of words copied from Reading Tutor's homepage itself, very basic words like ありがとう (arigatou "thank you"), こんにちは (konnichiwa "hello"), 日本語 (nihongo "Japanese language") etc.


**Figure 3.** Reading tutor interface and results

Among the longer texts recorded, about one third were mail messages or personal letters, e.g.

1) 「今、コーヒーを飲みながらメールを書いています（＾□＾）」
Ima, koohii o nominagara meeru o kaite imasu :-).
"I am now writing e-mails while drinking coffee (smiley)"
2) 「レポート受け取りました。どうもお疲れ様でした。」
Repooto uketorimashita. Doomo otsukaresama deshita.
"I received your report. Thank you"

The rest were mostly web pages, including many Wikipedia articles, Japanese articles about Slovenian companies, song lyrics, and other texts.

A surprising fact which emerged from the logs is that more than half of the "texts" which were input into Reading Tutor to be analysed were actually single words of phrases, e.g. "こんにちは" (konnichiwa "hello"), "あなた" (anata "you"), "縄文時代" (jômon jidai "Jômon period"), "首都" (shuto "capital") etc. In some cases, single words were input at first e.g. "に際して" (ni sai shite "regarding"), "追っていて" (otte ite "following"), followed by longer texts of a few sentences containing these words when the user probably realised that the tool is capable of analysing and providing translations for words in longer texts.

Surprisingly many input strings (around 30%) were not Japanese character strings: most of them romanised Japanese words or phrases (e.g. "arigatou", "kawai", "naruto" etc.), but also a few Slovene and English words ("ljubezen", "love" etc.), URL addresses, and even one Chinese text. Romanized Japanese words were possibly input by users who were not able to use Japanese fonts because of their computer

settings, or because they did not know enough Japanese to do so (as in the case of song lyrics with evident spelling mistakes, e.g. "kuchimoto no ugokoki [instead of: ugoki] ni yure ugoku" "wavering at lip movements"). Slovene search strings were probably input by users who overlooked the main function of the tool and used it as an online dictionary.

The frequent use of Reading tutor as a dictionary to look up single words might stem from the fact that the Slovene Reading tutor mirror is made available as a new tool alongside the online dictionary upon which it is built. Users might have had the impression that the tool with a slightly larger search box is actually only a variation or maybe newer version of the online dictionary which has been running on the same server for 5 years. A more explicit explanation of Reading tutor's functions and peculiarities is probably needed to help users better understand which of the tools is best suitable for which activity, i.e. that Reading Tutor is meant to help users read Japanese texts, while the dictionary is meant for looking up single words.

It was interesting to note also that even some users who clearly understood Reading tutor's function and input text rather than single words, still preferred to input several single sentences, which were clearly extracted from one longer text, at a few minutes intervals, rather than the whole text. One explanation for this is the users' reading habits or preferences – maybe they preferred not to rely too much on dictionaries, or preferred to read the text in its original formatting; another reason could be that they overlooked the possibility of clicking on any word in the analysed text to summon it up in the right-side vocabulary list, and therefore input shorter sentences in order to quickly scroll down the vocabulary list. This possible overlook could also be solved by a more thorough explanation of Reading tutor's functions.

## 5 Adding examples to jaSlo via a parallel corpus

The 3rd version of jaSlo, released in 2006, had approximately 10,000 Japanese lemmas with cca. 25,000 Slovene translation equivalents, but only 2,375 usage examples. As examples are a very useful source of information on a particular word's semantic, syntactic, collocational and pragmatic behaviour, and as some parallel texts were already available, we decided to enhance the example data-base by building and exploiting a parallel Japanese-Slovene corpus (Hmeljak Sangawa & Erjavec 2008). In this section we describe the methods and resources used to build the corpus, how examples were extracted to be included into the dictionary, and a short evaluation of the examples retrieved.

### 5.1 Corpus building

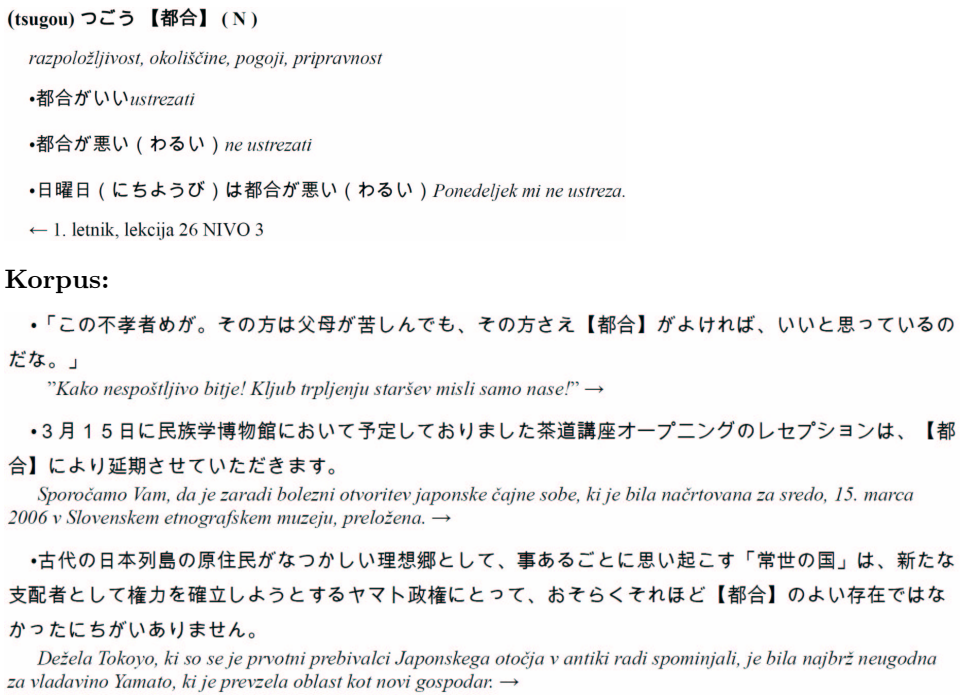
There are nowadays large amounts of parallel texts in digital form, even already aligned texts (translation memory data-bases) for combinations of major world languages, in particular for those which include English, but very few Japanese-Slovene parallel texts in digital or printed form, especially texts that have been translated directly from Japanese to Slovene or vice-versa, because there were hardly any translators for this language pair up to about 10 years ago. Out of the not very

numerous Japanese-Slovene translations we collected texts which can be divided into the following 4 categories: Slovene and Japanese internet culture-specific texts, which where then translated into the other language as part of students' course-work; handouts and course materials prepared by Japanese invited lecturers at the University of Ljubljana and translated into Slovene by department staff and students; translated fiction; and selected Web pages.

The collected texts were normalised into plain text files and aligned at sentence level, and the alignment manually validated. Japanese texts were then morphologically analysed and lemmatised using Chasen (Matsumoto et al. 2007), while the Slovene part was lemmatized using the program ToTaLe (Erjavec et al. 2005). This process yielded a parallel corpus which has 4,227 translation units (sentence pairs), 109,785 Japanese tokens (morphemes) and 83,113 Slovene tokens (words).

### 5.2 Extracting usage examples

All lemmas included in the Japanese-Slovene dictionary were searched for in the parallel corpus, and all parallel sentences containing one of the dictionary lemmas appended to the respective lemma. 4,648 headwords in the dictionary were thus augmented with new examples. In the case of very frequent words, only the shortest 6 examples were chosen.



**(tsugou) つごう【都合】(N)**

*razpoložljivost, okoliščine, pogoji, pripravnost*

•都合がいい*ustrezati*

•都合が悪い(わるい) *ne ustrezati*

•日曜日(にちようび)は都合が悪い(わるい) *Ponedeljek mi ne ustreza.*

← 1. letnik, lekcija 26 NIVO 3

**Korpus:**

•「この不孝者めが。その方は父母が苦しんでも、その方さえ【都合】がよければ、いいと思っているのだな。」

*"Kako nespoštljivo bitje! Kljub trpljenju staršev misli samo nase!"* →

•３月１５日に民族学博物館において予定しておりました茶道講座オープニングのレセプションは、【都合】により延期させていただきます。

*Sporočamo Vam, da je zaradi bolezni otvoritev japonske čajne sobe, ki je bila načrtovana za sredo, 15. marca 2006 v Slovenskem etnografskem muzeju, preložena.* →

•古代の日本列島の原住民がなつかしい理想郷として、事あるごとに思い起こす「常世の国」は、新たな支配者として権力を確立しようとするヤマト政権にとって、おそらくそれほど【都合】のよい存在ではなかったにちがいありません。

*Dežela Tokoyo, ki so se je prvotni prebivalci Japonskega otočja v antiki radi spominjali, je bila najbrž neugodna za vladavino Yamato, ki je prevzela oblast kot novi gospodar.* →

**Figure 4.** Entry with added examples extracted from the parallel corpus.

In the dictionary interface, corpus examples are graphically separated from previous constructed examples (which were already present in the dictionary), indicating to the user that they are not edited specifically for the dictionary, but rather naturally occurring examples containing the word in question, as in the sample dictionary entry given in Figure 4 (p. 213).

All corpus examples are linked to a page with information on the text where the example comes from: authors of the original text and of the translation (when known), date and place of publication or URL, source language and target language of the translation pair.

### 5.3 Evaluation of extracted examples

Usage examples play a very important role in a learners' dictionary, since they provide implicit information on a word's semantic, syntactic, pragmatic and collocational behaviour, and as such support both passive (reading) and active (writing) use of the target language. Exposure to multiple examples of usage of the same word contribute to its better retention, and in the context of data-driven learning they form the basis of learning itself. While explanations and definitions of a word's meaning can contribute to vocabulary acquisition through deduction, usage examples are the basis for inductive acquisition of vocabulary knowledge.

Examples which are automatically extracted from a corpus do not go through the usual editorial process of dictionary entries, i.e. analysis of a corpus of examples, synthesis of the dictionary entry meaning description 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. However, examples thus extracted were found to generally represent common collocational and syntactical patterns, and often contributed new translational equivalents for multi word units which had not been covered in the previous draft of the dictionary.

Thus for example the lemma **あわせる** (awaseru) had been translated only as »nastaviti« and »sešteti«, as in the examples given in the first part of Figure 5. On the other hand corpus examples for the same lemma offered other translation equivalents for the units **顔をあわせる** – *srečati – to encounter, videti se – to meet*, and **声を合わせて歌う** *peti skupaj – to sing together*, as in the second part of Figure 5 (p. 215).

Corpus examples certainly require more effort on the part of the user, who should be aware that they are not edited examples, but rather excerpts from parallel texts which have been translated in a given translational situation and may not be exact renderings of the original text, due to pragmatical or situational constraints. Indeed, some of the examples retrieved do not contain any element which could be considered as the concrete rendering of the word for which the example was extracted.

Constructed examples in the existing dictionary:

腕時計を駅の時計に合わせた。*Nastavil sem ročno uro glede na uro na železniški postaji.*
２と３を合わせると５になる。*Če seštejemo 2 in 3, dobimo 5.*

Corpus examples:

顔を合わせたくなかったから。*Nisem te hotela srečati.*
六九年の冬から七〇年の夏にかけて、彼女とは殆んど顔を合わせなかった。*Od zime do poletja sem jo komaj kdaj videl.*
私は流しをみがきながら、雄一は床をみがきながら、声を合わせて歌を続けた。*Ko sem drgnila po koritu in je Juiči brisal tla, sva pela skupaj.*

**Figure 5.** Comparison of constructed and corpus examples for the word あわせる (awaseru).

## 6 Conclusion and further work

The paper presented the Japanese-Slovene dictionary jaSlo, the Slovene module of the reading-support tool Reading Tutor, and an analysis of search logs in both tools.

Overall both tools, Reading tutor's Slovene module and the Japanese-Slovene dictionary were found to be used quite often. An analysis of frequent searches and problems brought to light a few possible improvements, the need for a new Slovene-Japanese dictionary, and the need for more publicity among the users, who are mostly presumably students of our University.

A method for the collection of a parallel corpus and extraction of examples to be used in a learners' dictionary was also presented. The corpus collected so far was found to be useful in the sense that it provided new examples to about half the entries in our dictionary, but enlarging the corpus would give a better coverage, both in terms of number of entries covered an in terms of coverage of each entry's patterns. Given a larger amount of examples for each entry, it would be useful to measure each example's level of lexical and syntactical difficulty, as proposed in (Kobayashi et al. 2007) and (Yoshihashi et al. 2007), and of its typicality, as measured by MI score of collocational patterns with reference to a large balanced Japanese corpus (Srdanović et al. 2008).

## References

Hmeljak Sangawa, K. (2003). Slovar japonskega jezika za slovenske študente japonščine. In *Konferenca Jezikovne tehnologije*, pages 102–105 , Ljubljana: IJS.

Erjavec, T., Srdanović, I., Hmeljak Sangawa, K. (2004). Suroveniajin nihongo gakushuushayou jisho no xmlka. Nihongo Kyouiku Renraku Kaigi rombunshuu, vol. 16, pages 45–52.

Erjavec, T., Hmeljak Sangawa, K., Srdanović, I. (2006). jaSlo, a Japanese-Slovene learners' dictionary: methods for dictionary enhancement. In *Proceedings XII EURALEX international congress*, pages 611–616. Torino: Edizioni dell'orso.

Kawamura, Y. (2000). EDR denshika jisho o katsuyou shita nihongo kyouikuyou jisho tsuuru no kaihatsu. Nihon kyouiku kougaku zasshi. 24(Suppl.), pages 7–12.

Sperberg-McQueen, C., Burnard, L. (Eds.). (2002) Guidelines for Electronic Text Encoding and Interchange, The XML Version. The TEI Consortium, 2002. [http://www.tei-c.org/]

Japan Foundation (2002). Japanese Language Proficiency Test: Test contents specifications. Tokyo: Bonjinsha.

Matsumoto, Y., Takaoka, K., Asahara, M. (2007). Keitaisokaiseki shisutemu ChaSen version 2.4.0 User's Manual. {http://sourceforge.jp/projects/chasen-legacy/document/chasen-2.4.0-manual-j.pdf/ja/2/chasen-2.4.0-manual-j.pdf}

Hmeljak Sangawa, K., Erjavec, T. (2008). A low cost approach to building a Japanese-Slovene parallel corpus. *Denshi Jôhô Tsûshin Gakkai gijutsu kenkyû hôkoku*, 2008, vol. 108, no. 50, pages 7–10.

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, Poznan, Poland*. pages 32–36.

Kobayashi, T., Oyama, H., Sakada, K., Taniguchi, Y., Ota, F., Evans, N., Asahara, M., Matsumoto, Y. (2007). Nihongo dokkai shien no tame no gogigoto no yourei chuushutu kinou nitsuite. In *Proceedings of the 13$^{th}$ Annual Meeting of the Association for Natural Language Processing*, Tokyo: Association for Natural Language Processing.

Yoshihashi, K., Fu, L., Nishina, K. (2007). Gakushuusha ni awaseta reibun hyouji tsuuru. In *CASTEL-J in Hawaii 2007 Proceedings*, pages 223–226. Honolulu: University of Hawaii.

Srdanović I., T. Erjavec, A. Kilgarriff. (2008). A web corpus and word sketches for Japanese. *Shizen gengo shori*, 15(2), pages 137–159.