Category: Reports on Lexicographical and Lexicological Projects

JASLO, A JAPANESE-SLOVENE LEARNERS' DICTIONARY: METHODS FOR DICTIONARY ENHANCEMENT

Tomaž Erjavec[*]
Kristina Hmeljak Sangawa[†]
Irena Srdanović Erjavec[†]

[*] Department of Knowledge Technologies
Jožef Stefan Institute
Jamova 39, Ljubljana, Slovenia

[†] Department of Asian and African Studies
Faculty of Arts
University of Ljubljana
Aškerčeva 2, Ljubljana, Slovenia

ABSTRACT

The paper presents our experiences in producing a hypertext learners' Japanese-Slovene dictionary jaSlo, which currently contains over 10,000 entries. The paper discusses the conversion of the dictionary from the legacy encoding, which consisted of many separate files in a mixture of different tabular formats, into to a standardised XML format. The conversion consisted of up-translation from the legacy formats, the enrichment of the dictionary with third-party resources, merging of the data, manual verification, and the deployment of the dictionary via a Web interface. The presented methodology ensures that the resulting dictionary is of a high quality, addresses user needs, and is suitable for re-purposing and interchange. We conclude with plans for further work.

Category: Reports on Lexicographical and Lexicological Projects

JASLO, A JAPANESE-SLOVENE LEARNERS' DICTIONARY: METHODS FOR DICTIONARY ENHANCEMENT

1 Introduction

The establishment of a new Department of Asian and African studies at the University of Ljubljana and a course of Japanese studies within it in 1995 brought with it the need for Japanese language teaching materials and dictionaries for Slovene speaking students. However, due to the limited number of potential users, probably not much more than the current 180 students of Japanese at the department, the compilation of such materials and dictionaries is not a particularly profitable project that could interest a publishing house. The teachers at the department therefore decided to create it with the help of our students, the final users of the dictionary (Hmeljak Sangawa, 2002).

The compilation of a dictionary that would satisfy the needs of Japanese language students both in terms of macrostructure and of microstructure, i.e. with enough lemmas and a detailed enough description for each lemma to cover users' needs, both for passive and for active use, is going to last for many years. Meanwhile, even incomplete data can be useful to users of a language pair for which no dictionary exists at all. We therefore decided to merge a few glossaries created at our department and publish them on the web. Initially, the dictionary was conceived in a tabular format, suitable for editing in a spreadsheet program, and from which it would be possible to directly derive an HTML format. However, after a few years of this undertaking, it became apparent that this model of compilation exhibited serious drawbacks: it was difficult to enforce consistency across the dictionary additions that the students and professors produced, i.e. it was problematic to validate and exchange, as well as being unsuitable for accommodating a more complex dictionary structure.

The first stage of converting an initial dictionary (1000 entries) into XML was reported in Erjavec et al. (2004). The target encoding takes into account international standards in the field and 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, interchange and longevity, as well as easier usage of data for linguistically oriented research. This format also facilitated the Web deployment of the dictionary, which offers a full-text search facility, as well as narrowing the search to particular elements, e.g. headword or part-of-speech.

In this paper, we discuss the second stage of the project, where the dictionary, named jaSlo, was expanded to contain over 10,000 entries; for this, the various separate and partially overlapping legacy tabular dictionaries were converted into the common XML encoding and merged into the master dictionary. Furthermore, the dictionary content was normalised and enriched by various third party resources. The focus of the paper is on presenting the methodology used in producing this new dictionary, which could also benefit other similar collaborative lexicographic projects.

The rest of the paper is structured as follows: Section 2 overviews the XML structure of the dictionary; Section 3 explains the dictionary enhancement and its building process; Section 4 briefly introduces the Web interface to the dictionary; and Section 5 gives conclusions and directions for further work.

```
<entry id="jaslo.6557">
  <form type="hw">
   <orth type="kana">ちょうせつする</orth> <orth type="kanji">調節する</orth>
   <orth type="roma">chousetsusuru</orth>
  </form>
  <gramGrp><pos>Vs</pos> <subc>trans.</subc></gramGrp>
  <trans><tr>uravna(va)ti</tr></trans>
  <eg>
   <q>室内（しつない）の温度（おんど）をちょうせつする</q>
   <tr>uravnavati temperaturo v sobi</tr>
  </eg>
  <xr type="lesson" n="L1.23"><xref>1. letnik, lekcija 23</xref></xr>
  <usg type="level">0</usg>
  <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="ISE">2005-02-28 Merge</note>
  <note type="admin" resp="VOJ">2005-02-22 V (440)</note>
   <note type="admin" resp="KHS">2003-03-12 L1 (850)</note>
</entry>
```

Figure 1. A typical dictionary entry in jaSlo

2 Dictionary encoding

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 a typical dictionary entry, which includes the form of the headword given in kanji, kana (hiragana or katakana), and in Latin transcription. This is followed by grammatical information, translation into Slovene, examples, and a reference to the lesson where the word is introduced, the difficulty level of the entry, and finally the administrative information tracing the history of the compilation of the entry.

In addition to the elements given in the example, the following information is also present in a subset of the entries: cross-reference to related entries (esp. as regards levels of politeness, synonyms etc.), inflected forms of verbs, the etymology of loan-words, and encyclopaedic descriptions of proper names, esp. place names.

3 The Compilation Process

The process of compiling the present dictionary consisted of two parts: the up-translation of the partial tabular dictionaries into TEI, and the enrichment of the dictionary with additional data using third-party resources. The basic approach in producing the new version of the dictionary therefore consists of the following steps:

1. up-translation of the tabular files into the TEI format,
2. extracting relevant information from third-party resources,
3. merging of the new data with the existing dictionary,
4. manual verification.

3. 1 Up-translation

The process of up-translation to the standard TEI encoding had to cope with a plethora of file and input formats, some of them containing implicit structures. It was therefore essential to choose a flexible language with good regular expression support. We used Perl, which additionally offers easy file manipulation and linking to external programs.

For up-translation of the tabular dictionaries, the source character encoding was first converted from Shift-JiS (still the most popular encoding for Japanese, esp. for Macs) to UTF-8, and then converted to TEI. While the core of the transformations was the same for all cases, each tabular file had its own peculiarities, which had to be addressed separately. The transformations, for most fields, simply wrapped their content into the appropriate TEI tags. Additionally, however, the programs also performed some normalisation (e.g. stripping superfluous whitespace and punctuation, normalising variant spellings of labels), verification (e.g. detecting illegal empty fields and flagging suspicious elements with a question mark) and assignment of tags according to detected string patterns.

This last feature is the most interesting, as information that was implicit in the original format becomes explicitly marked. So, for example, the note column of some source files can contain remarks on usage, but also the etymology of borrowings. Where the pattern »(iz ... ...)« is found, e.g. "*(iz nemšč. Arbeit)*" ("*from German Arbeit*") this is converted to `<etym><lang>nemšč.</lang> <gloss>Arbeit</gloss></etym>`.

As was seen in Figure 1, each entry also contains administrative notes on the history of the entry – when it was created or modified, who modified it, and what the modification was – in cases where the "modification" is up-translation from a legacy dictionary, the content of the note specifies the name of the dictionary and its entry (line) number. Such a revision history significantly helps in debugging the transformations as well as giving per-entry author attribution.

Writing the up-translation programs was a long, and, to an extent, frustrating task, familiar to anyone who has attempted to automatically clean and flag 'dirty' data. It consists of a cycle where a transform is written, run over the input, the results evaluated and the transform modified, and the process repeated, all the time striving to find a balance between the precision and recall of the filter. The process typically terminated when we judged that the effort to further modify the program would exceed the effort to manually verify and correct the actual XML dictionary.

3.2 Adding external information

The dictionary entries were also automatically enriched or normalised via – mostly – third party resources, in particular, the following: the transcription of the hiragana headword into the Latin alphabet (romaji); the difficulty level of the headword; part-of-speech normalisation; the addition of the caron diacritic to Slovene characters. We present these in turn.

Japanese has a very complex writing system consisting of Chinese kanji characters and the phonetic (syllabic) scripts katakana and hiragana. For our dictionary it was initially decided that the primary headword of each entry should be in hiragana or katakana, as in traditional Japanese dictionaries, as it is much simpler to learn and search. Entries are accompanied by their kanji (or mixed kanji-kana) orthography, if the word is usually written using Chinese characters. However, analysing the log files of the usage of the first version of the dictionary showed that people often search for Japanese words using the Latin alphabet and were, of course, coming up empty-handed. We therefore decided to add this information into the dictionary and were lucky to find a freely available kana to romaji converter program (available at http://raa.ruby-lang.org/project/kana2rom/) with the help of which we could produce the Latin transcription for all entries.

The second addition was that of the difficulty level of the headword, according to the vocabulary list used by the Japanese Language Proficiency Test, which is divided into four difficulty levels, with 4 containing the basic words of the language, and 1 the ones assumed for advanced speakers. This was also a useful validation step, as it helped us identify basic words which had been previously missing from the dictionary.

A major difficulty with the legacy tabular dictionaries was their inconsistency with each other, but also internally. This was especially true for the part-of-speech of the headwords, since the tabular dictionaries were labelled according to different part-of-speech sets. We therefore spent quite some time first devising the PoS set to be used, and then semi-automatically converting the legacy PoS labels to this common standard. Our set of categories contains 19 different labels and is based on the set used in the Japanese morphological analyzer Chasen (Matsumoto et al. 2003). The PoS conversion procedure consisted of first normalising the obvious cases of mismatch with Perl (e.g. differences in capitalisation, punctuation or shortenings) and then dumping the list of remaining PoS appearing in the dictionary into a table – each of the rows (over 100) was then manually assigned the canonical PoS, as well as PoS labels from a few 'standard' PoS schemes. This mapping was then used to correct the source dictionary; as a side benefit, we included the mappings (e.g. localisations) into the jaSlo TEI header; this enables flexibility in the display of the dictionary.

Finally, there was the problem of č, š, ž, which are the only three characters (and their upper case equivalents) that Slovene uses which are not in the ASCII character set. These characters are also not part of Shift-JIS, the encoding used in most of the legacy dictionaries, so the authors usually chose to substitute them with c,s,z. Reversing this simplification is unfortunately, in the general case, impossible with automatic means. Although more sophisticated algorithms exist for automatic diacritic insertion (e.g. Tufiş and Chiţu, 1999) we chose a relatively simple method, where each Slovene word containing one of these characters was matched against a large dictionary; if it was found to correspond to an unambiguous dictionary word it was replaced; if it was not found, or was ambiguous, e.g. *resen (serious)* vs. *rešen (saved)* it was flagged for manual verification.

3.3 Merging the data

Each piece of the newly acquired data had to be merged with the evolving dictionary. This procedure consisted of first identifying whether a new entry was being added or an existing one (and which) modified, and, if the latter, how to add the new information to an existing entry. The identification of the entry is complicated by the fact that its unique key would have to be a combination of the kana headword string, the kanji, and the part-of-speech; if any of these fields is missing from the key, we can be faced with ambiguities. As we, at the outset, did not have a consistent set of PoS categories, and entries could have missing kanji information, the merge program identified ambiguous situations and flagged these for manual verification.

Merging existing entries with new ones presented special challenges. The problem was not marginal as the legacy dictionaries had a significant amount of overlap in contained entries. However, it was in general not possible to simply discard duplicate entries, as they could each contain valuable information, e.g. one examples, and the other the reference to the lecture number where the word is introduced. The merge therefore identified several possible situations. When the information in the new entry was already contained in the dictionary, the new information was simply ignored; when the information from the two sources could be unified monotonically, and the two entries matched in the complete key, the new information was straightforwardly added to the existing entry. When, however, the entries had incompatible information or they did not match in the PoS, all entries were wrapped in a <hom> (=homonym) element, with its n attribute giving the number of 'homonymous' entries, and these were then merged manually.

3.4 Manual verification

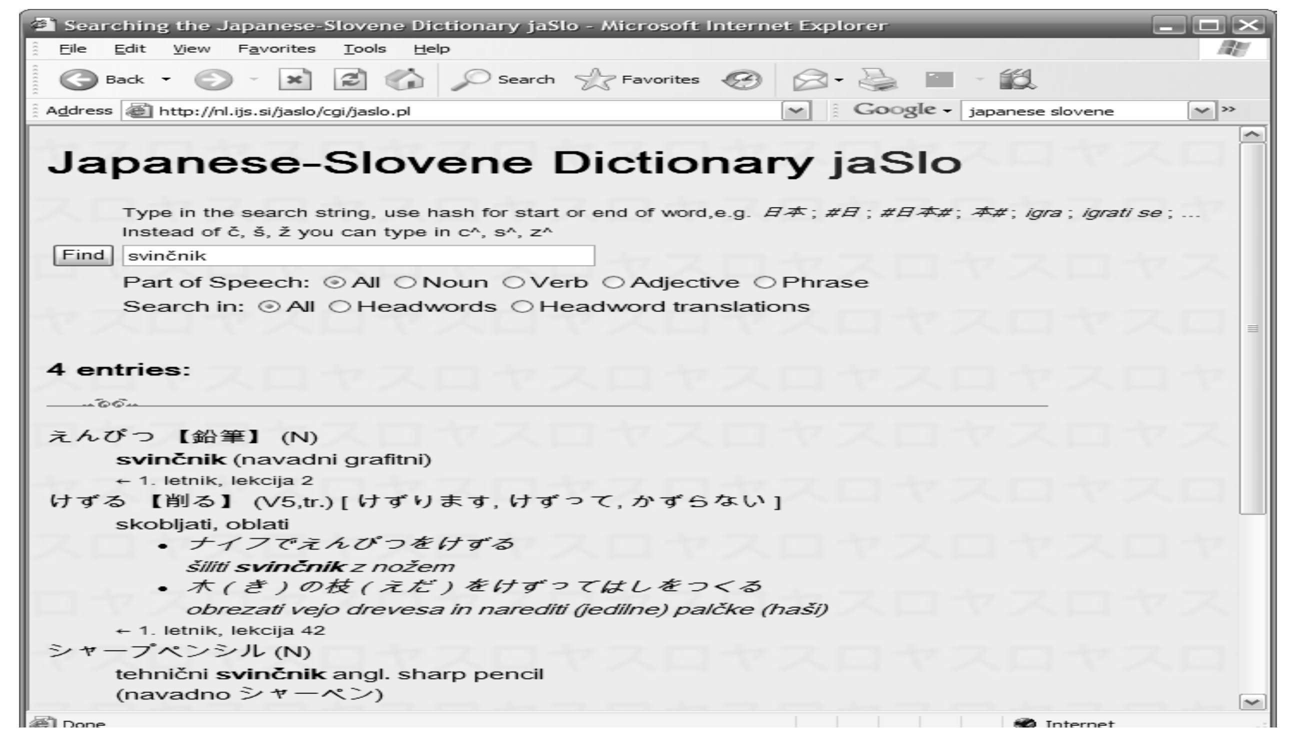
The general method of producing the dictionary involved programs that had a less than perfect precision but did strive to identify dubious cases and flag them for manual verification. This involved suffixing question marks to suspicious element content or, in the case of merging, the use of an extra tag. For easier handling and to enable simultaneous work, the automatically produced dictionary was split into 11 files and these were then manually verified with the help of an XML editor. To make editing easier, the XML is assigned an XSLT stylesheet, which converts it to nicely formatted HTML, and also produces an index of the document entries.

At the time of writing this draft, the process is still on-going (we estimate there are still about 80 hours of work to complete the clean-up), but should be finished in time for the final version of this paper.

4. Using the dictionary

The dictionary was deployed via a Web-based interface (Figure 2), 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 PoS or difficulty level. The interface is also localised to Slovene, Japanese and English. The user's browser is assumed to offer Unicode support and have installed a Japanese-language font 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 sequentially, via a SAX filter, returns the entries that match the query. While this means that for each query the complete dictionary has to be processed, this does not present problems with the current size of the dictionary and user load.


Searching the Japanese-Slovene Dictionary jaSlo - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
Address
Google
Done
Internet

http://nl.ijs.si/jaslo/cgi/jaslo.pl

japanese slovene

Japanese-Slovene Dictionary jaSlo

Type in the search string, use hash for start or end of word,e.g. *日本* ; *#日* ; *#日本#*; *本#*; *igra* ; *igrati se* ; ...
Instead of č, š, ž you can type in c^, s^, z^

Find svinčnik

Part of Speech: ● All ○ Noun ○ Verb ○ Adjective ○ Phrase

Search in: ● All ○ Headwords ○ Headword translations

4 entries:

えんぴつ 【鉛筆】 (N)
svinčnik (navadni grafitni)
← 1. letnik, lekcija 2

けずる 【削る】 (V5,tr.) [けずります, けずって, かずらない]
skobljati, oblati
- *ナイフでえんぴつをけずる*
  šiliti ***svinčnik*** *z nožem*
- *木 (き) の枝 (えだ) をけずってはしをつくる*
  obrezati vejo drevesa in narediti (jedilne) palčke (haši)

← 1. letnik, lekcija 42

シャープペンシル (N)
tehnični **svinčnik** angl. sharp pencil
(navadno シャーペン)

Figure 2. The Web interface to jaSlo

Each query together with time and number of returned entries is also logged (without client machine address, thus preserving privacy), which enabled us to begin tailoring the dictionary to user needs. Already mentioned was the discovery of the desirability of the romaji transcriptions; a further surprise was the heavy use of searching by lesson number only, i.e. the students obviously find it convenient to extract from the dictionary the complete set of entries introduced in a given lesson. It was also interesting to note that a common failure to return an entry is that the search is made by a synonym of the Slovene word, e.g. *ogledalo (mirror)* vs. *zrcalo (~looking glass)*, where the latter is not in the dictionary. The log file will also come in useful for further expansion of the dictionary, by isolating the most frequently searched-for but not found words.

After collecting the relevant TEI-encoded entries, the CGI formats them in HTML. For this a similar XSLT stylesheet to the one used by the editors in the manual verification process is used, but with less explicit rendering and ignoring certain information, e.g. the admin notes, entry ID etc.

6. Conclusions

The paper presented the encoding, production, and Web deployment of the Japanese-Slovene learners' dictionary jaSlo, meant primarily for Slovene students of Japanese at the University of Ljubljana. The home page of the dictionary is http://nl.ijs.si/jaslo/ from where the search interface to the previous version is also available. The new version of the dictionary will be made publicly available in time for EURALEX 2006; as the clean-up has not yet been finished, it is at the moment accessible only via the unadvertised URL http://nl.ijs.si/jaslo/cgi/jaslo-v3.pl

"Collaborative bottom-up editing" and open-source lexicographical projects have been criticized (Docherty 2000) for their poor quality, which is indeed often the case. However, collaborative editing can produce useful data and may be the only viable means of producing a dictionary for a non-profitable language pair. Looking back we can conclude that it would have been well worth investing time up-front to specify exact guidelines for the dictionary encoding, to use a platform that prevents syntactically ill-formed input, and to coordinate the dictionary making activity to prevent duplication. Still, for others in a similar situation, i.e. faced with varied and incosistent legacy data, we believe that our approach presents a viable method for arriving at a high-quality canonical dictionary and deploying it on the Web. The approach is predicated on the use of open platforms and tools (Linux, Apache, Perl, Saxon), standards (XML, TEI, XSLT, HTML), and on the use of supplementary resources (Slovene lexicon; kana2rom, Chasen) and consists of an up-translation, followed by a merge operation, manual post-editing, and Web deployment.

There are a number of improvements to jaSlo we are planning in our future work. For further additions as well as corrections to the dictionary we will implement a Web-based form interface, with a human editor checking the proposed updates prior to incorporation into the master dictionary. An interesting venue of further work is also to enrich the dictionary searching and display by automatically creating links between the dictionary and a kanji database and to external Web dictionaries, e.g. WWWDict (Breen 2003), one of the best-known Japanese-English Web dictionaries, or the Slovene-German-Slovene dictionary for German students of Slovene (Lönneker and Jakopin, 2003), a project similar to ours. We are also planning to add jaSlo into the on-line Japanese reading support tool “Reading tutor” (Kawamura et al. 2003).

References

Breen, J. (2003). *The Japanese-Multilingual Dictionary and the Japanese Proper Names Dictionary.* http://www.csse.monash.edu.au/~jwb/japanese.html

Docherty, V.J. (2000). Dictionaries on the Internet: an Overview. *Proceedings of the Ninth Euralex International Congress, Euralex 2000*. pp.67-74, Stuttgart, 2000.

Erjavec, T., Hmeljak Sangawa, K., Srdanović, I., Vahčič, A. ml. (2004). Making an XML-based Japanese-Slovene Learners' Dictionary. In *The 4th International Conference on Language Resources and Evaluation (LREC) proceedings*, ELRA, pp. 1059-1062. Lisbon, 2004.

Hmeljak Sangawa, K. (2002). Slovar japonskega jezika za slovenske študente japonščine (A Japanese Dictionary for Slovene Students of Japanese). In *Proceedings of the Conference on Language Technologies*. pp. 102-105, Ljubljana: Jožef Stefan Institute.

Kawamura, Y., Kitamura, T., Hobara, R. (1997-2002). *Reading Tutor*. http://language.tiu.ac.jp/

Lönneker, B., Jakopin P. (2003). Contents and evaluation of the first Slovenian-German online dictionary. *Proceedings of the 10th EACL - Conference Companion. ACL.* pp. 119-122. http://www.rrz.uni-hamburg.de/slowenisch/

Matsumoto, Y. et al.? (2003) *Morphological Analyzer Chasen*. http://chasen.aist-nara.ac.jp/

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

Tufiş, D., Chiţu, D. (1999) Automatic Diacritics Insertion in Romanian Texts. *Proceedings of COMPLEX’99 International Conference on Computational Lexicography*, Pecs.