Making an XML-based Japanese-Slovene Learners' Dictionary

Tomaž Erjavec

Department of Knowledge Technologies
Jožef Stefan Institute
Jamova 31, Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Kristina Hmeljak Sangawa

Department of Asian and African Studies
Faculty of Arts
University of Ljubljana
Aškerčeva 2, Ljubljana, Slovenia
kristina.hmeljak@guest.arnes.si

Irena Srdanović

Department of Comparative and General Linguistics
Faculty of Arts
University of Ljubljana
Aškerčeva 2, Ljubljana, Slovenia
irena_srdanovic@hotmail.com

Anton ml. Vahčič

Faculty of Computer Science and Informatics
University of Ljubljana
Tržaška cesta 25, Ljubljana, Slovenia
vahcica@hotmail.com

Abstract

In this paper we present a hypertext dictionary of Japanese lexical units for Slovene students of Japanese at the Faculty of Arts of Ljubljana University. The dictionary is planned as a long-term project in which a simple dictionary is to be gradually enlarged and enhanced, taking into account the needs of the students. Initially, the dictionary was encoded in a tabular format, in a mixture of encodings, and subsequently rendered in HTML. The paper first discusses the conversion of the dictionary into XML, into an encoding that complies with the Text Encoding Initiative (TEI) Guidelines. The conversion into such an encoding validates, enriches, explicates and standardises the structure of the dictionary, thus making it more usable for further development and linguistically oriented research. We also present the current Web implementation of the dictionary, which offers full text search and a tool for practising inflected parts of speech. The paper gives an overview of related research, i.e. other XML oriented Web dictionaries of Slovene and East Asian languages and presents planned developments, i.e. the inclusion of the dictionary into the Reading Tutor program.

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 forward 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 our department, the compilation of such materials and dictionaries is not a particularly profitable project that could interest a publishing house. The teachers at our 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. However, adopting the "dictionary-making process with 'simultaneous feedback' from the target users to the compilers" which has been proposed by De Schryver and Prinsloo (2000) can help us turn the drawback of having few users into an asset: we can have direct contact and feedback from most of the users at all stages of compilation.

Initially, the dictionary was conceived in a tabular format, suitable for editing in a spreadsheet program, and from which it was possible to directly derive an HTML format. However, it became apparent that this structure exhibited various drawbacks; in particular, it was difficult to extend to accommodate a more complex dictionary structure, as well as being difficult to validate and exchange.

This paper describes the conversion of the dictionary format into XML (eXtensible Markup Language) (W3C, 2000), using a document type definition that complies with the TEI (Text Encoding Initiative) Guidelines (Sperberg-McQueen and Burnard, 2002). This approach takes into account international standards in the field and focuses on describing text properties, i.e. what a particular part of the text means. It brings a number of advantages, such as better documentation, ability to validate the structure of a document, simpler processing, better integration, interchange and longevity, as well as easier usage of data for linguistically oriented research. This format also enabled Web deployment of the dictionary, which offers a full-text search facility, as well as grouping the entries into “learning blocks”, ordered by lessons and part-of-speech.

2. The Dictionary Model

Ideally, a dictionary should contain all items its users might ever want to look up. However, striving to cover all vocabulary our students might possibly encounter during their undergraduate study would be unrealistic in our situation. We therefore decided to cover only the core vocabulary encountered up to an intermediate level of language study, and not to include the more specialized or rare vocabulary. Such a vocabulary is presumably encountered at a time when the students' knowledge of Japanese enables them to use the wealth of existing

Japanese monolingual dictionaries. The core vocabulary for learners of Japanese has been variously identified in the literature as amounting from around 5.000 to around 10.000 words (Tamamura, 1990, 1995). In a study conducted by the Japanese National Language Research Institute (1962), 10.000 words cover around 90% of the vocabulary used in present-day newspapers. This is also the number of words required for two examinations of Japanese as a foreign language on the highest level of proficiency (Japan Foundation, 2002; Senmon Kyouiku Publishing, 1998). Our selection of lemmas is primarily based on these latter lists.

The dictionary we are compiling is an electronic dictionary composed of two databases. The first is a word database, containing 10.000 words, which is the vocabulary expected from students taking the Japanese language proficiency test at the highest, i.e. 1st level. At the present moment we have a simple database containing the Japanese words, their part of speech, and one or two Slovene translation equivalents, but our aim is to compile a dictionary with the following lemma structure.

- written form of the word (in Chinese characters, hiragana and/or katakana);
- pronunciation (in kana) and accentuation;
- grammatical information: word class (with a note when this differs from the word class of its translational equivalent); inflected forms for inflecting word classes; syntactic patterning;
- meaning (denotative): definitions in Japanese; Slovene translational equivalents for each meaning; synonyms, antonyms, other related words; common collocations and multiword units;
- connotation: level of formality; written or spoken usage; male/female speech; category of origin (Japanese, Chinese or other foreign);
- usage examples with Slovene translation.

One thousand basic words, which are used in our textbook for 1st year students, have already been compiled according to this structure, except for related words and Japanese definition.

The kanji database should contain 2000 kanji and the following information for each kanji.

- written and printed form of the character;
- number and order of strokes it is composed of;
- radical;
- readings, i.e. words and morphemes written with the character;
- compounds containing the character.

All information that directly relates to other lemmas in the dictionary should be hyperlinked to the relevant lemma.

3. Converting to XML

The dictionary that is encoded as a table or in HTML format is quite usable when viewed with internet browsers, but there is a number of disadvantages of using such formats: the tabular one is too rigid for the rich structure that we find in dictionaries, while HTML is too unconstrained and is primarily oriented towards a visual representation of the data, rather than towards its semantics, i.e. what a particular part of the text actually means. These characteristics make further development, maintenance and processing of the dictionary difficult. We therefore decided to convert the dictionary into XML (the 1.000 word database, for the time being), which offers us more flexibility in encoding and in choosing different display possibilities and search mechanisms.

For the conversion, we first had to decide on what we were converting to, i.e. what XML elements (tags) we are to use in the dictionary. Formally, these can be defined using an XML Document Type Definition (DTD). Although it is, of course, possible to write one's own DTD, we decided to rather use an already established standard for text encoding, namely the TEI (Text Encoding Initiative), since it offers us a standardised, documented and tested set of elements, as well as offering supporting software and user community.

3. 1 The Text Encoding Initiative (TEI)

TEI was established in 1987 under the joint sponsorship of the Association for Computers and the Humanities, the Association for Computational Linguistics, and the Association for Literary and Linguistic Computing. The aim was to reduce the diversity of existing encoding practices, simplify processing by machine, and encourage the sharing of electronic texts. It became the only systematised attempt to develop a fully general text encoding model and set of encoding conventions based upon it, suitable for processing and analyzing any type of text, in any language, and intended to serve the increasing range of existing (and potential) applications and uses.

Currently, the newest version of TEI is TEI P4, published in 2002, providing equal support for XML and SGML applications, and being compatible with the previous version, TEI P3. Although the TEI P4 Guidelines (Sperberg-McQueen and Burnard, 2002) are published in book form, they are also freely available on the TEI Consortium web site, http://www.tei-c.org/.

To create a TEI compatible DTD that would answer the needs of a particular project, a combination of tagsets defined in TEI should be used. For our project we used the XML version of TEI P4 with the base tagset for dictionaries and additional tagsets for linking and language analysis. To validate against this DTD, either the full set of TEI P4 DTD fragments can be dowloaded, and the chosen modules selected in the XML prolog, or the »TEI Pizza Chef« Web service from the TEI Consortium can be used, which produces, once we select the modules, a one-file DTD suitable for local use. Such a DTD was used to define our document model, and validate our converted dictionary.

3. 2 The structure of the dictionary entry in TEI

The DTD defines elements that are used for encoding the dictionary. The whole dictionary is encoded as a <TEI.2> element, containing the header <teiHeader>, and the <text>, in turn containing the <body> of the document. The <body> contains one or more divisions , and these contain the dictionary entries.

The TEI header contains meta-data, i.e. gives information about the resources, about its source or sources, manner of encoding, revision history, etc.

An example of a dictionary entry from the dictionary is given in Figure 1. The meaning of the elements is largely self-explanatory, and, of course, the TEI Guidelines offer a detailed description for each element. As we see, the <entry> first contains the element <form type="hw">, which gives the headword both in the phonetic hiragana / katakana scripts, and in kanji. The headword is followed

```
<entry id="j.68">
  <form type="hw">
    <orth type="kana">あんないする</orth>
    <orth type="kanji">案内する</orth>
  </form>
  <gramGrp>
    <pos>Vs</pos>
  </gramGrp>
  <trans>
    <tr>voditi [koga po mestu]</tr>
  </trans>
  <eg>
    <q>みちを あんないします</q>
    <tr>pokazati pot </tr>
  </eg>
  <xr type="course">24</xr>
</entry>
```

Figure 1: Example entry in TEI P4

by the <gramGrp>, which groups the grammatical properties of the entry. With verbal entries, their inflected forms, grouped in <form type="infl"> come next. These elements are followed by the translation into Slovene, contained in <trans>. Next come the examples, <eg>, and finally the cross-reference to lesson 24 of the course book. Eventually other, more complex structures will also be added to the dictionary as e.g. the level of politeness, the pronunciation of the words in the examples etc. We are planning a gradual enriching of the dictionary structure as well as a linguistic analysis of the words in the examples.

3. 3 Up-converting to TEI

The source format of the dictionary was a tabular file with 12 fields per entry (not all are necessarily used for each entry): orthography, part-of-speech, translation(s) of the headword into Slovene, two examples with their translations, the number of the lesson in which the word is introduced, three inflected forms of the verb headwords, and notes. This format was automatically converted to the TEI XML encoding. We first converted the character encoding from Shift-JiS (still the most popular encoding for Japanese, esp. for Macs) to UTF-8 and then used a Perl filter, which, for most fields, simply wraps their content into the appropriate TEI tags. However, the Perl program also performs some normalisation (i.e. stripping superfluous whitespace and punctuation), verification (e.g. it complains about illegal empty fields) and assignment of tags according to string patterns.

This last feature is the most interesting, as information that was implicit in the original format becomes explicitly marked by appropriate elements. So, for example, the note column of the 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>. There are remain other such patterns waiting to be taken advantage of, but first the text will have to be further normalised.

The current dictionary in TEI obtained in this manner has the following elements, where the number in parenthesis is the tagcount in the dictionary: <eg> (237), <entry> (2681), <etym> (142), <form> (2920), <gloss> (142), <gramGrp> (2545), <lang> (142), <orth> (5349), <pos> (2545), <q> (237), <tr> (2898), <trans> (2661), <usg> (85), <xr> (2684).

4. Using the dictionary

Having arrived at the interchange XML format the next step was actually making use of the dictionary. We have so far experimented with three different usage scenarios. First, we created an XSLT style sheet to enable a straightforward HTML presentation, suitable for browsing the dictionary. We developed also two other, more substantial applications, which are described next.

4.1 The Dictionary on the Web

Rather than writing specific programs for different operating systems (Windows, Linux or MacOS), we decided to offer a Web based application, which can be used by anyone with a browser that offers Unicode support, e.g. Netscape 6 or up, Internet Explorer 5.5.

Unicode support is necessary as even though Shift-JIS is the de facto encoding standard in Japan and is quite suitable for Japanese and English text presentation, it does not support Slovene specific characters i.e. č, š, and ž.

We have developed two suites of Web applications. The first is a set of applications for dictionary editing which allows registered users to create and edit dictionary entries via the internet. This should help speeding up dictionary build-up by allowing multiple editors (students, teachers or anyone else) to work on the data, and help the coordination of teacher and student work by marking entries according to whether they are newly added items or entries checked by the dictionary main editor.

The second set of applications is meant for the use of the dictionary database: the main application is a Web dictionary search interface where users can search words or parts of words according to various criteria (part of speech, lesson number, full text search, etc.). Words that are looked up and not yet contained in the dictionary are logged for possible later inclusion into the dictionary. Further programs offer practice on adjectives for beginning students of Japanese.

These applications were developed in asp.net technology because of its excellent Unicode support and of its long tradition at the Faculty of Computer and Information Science.

4.2 Reading Tutor

The next use we are planning for our dictionary database is its insertion into the "Reading Tutor" (Kawamura et al. 2003), 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. 2003), links every token in the text to one of Reading Tutor's dictionaries (Japanese definitions, Japanese-English and Japanese-German at present), 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 also encoded in XML, according to their own document type. The Reading Tutor

DTD is quite complex, with numerous elements, quite a few of them required. We wrote a preliminary XSLT stylesheet that converts our TEI into the Reading Tutor DTD.

5. Related Work

This section gives an overview of some other projects which are centered on Web based dictionary access and XML encoding of dictionaries, esp. those that deal with Japanese or Slovene.

Two interlinked projects are the Chinese/ Japanese/Korean/Vietnamese-English Dictionary and the Digital Dictionary of Buddhism (Müller, 2003), started already in 1986. They are both XML encoded in accordance with TEI standard, although they also include quite a number of user-defined elements to cover the specifics of the dictionaries.

Another example is the Japanese Multilingual Dictionary and the Japanese Proper Names Dictionary (Breen, 2003), which both use their own XML DTD. The Japanese multilingual dictionary includes bilingual dictionaries with English, French, German and Russian translations.

The Papillon project (Mangeot and Sérasset, 2003) aims to build a multilingual lexical database created from a set of interlinked monolingual dictionaries. Presently, it covers French, Japanese, English, German, Chinese, Korean, Lao, Thai, Vietnamese and Malay. The dictionary structure is defined as an XML schema, named DML (Dictionary Markup Language), which takes some account of existing international standards, among others also TEI.

There are also examples of Web based dictionaries for Slovenian. The project by Lönneker and Jakopin (2003), similar to ours, mounted the first HTML encoded Slovenian-German dictionary on the Web. The dictionary is meant for foreign (German) students of the Slovenian language. There are other examples of digital (XML) dictionary production in Slovenia, esp. at the publishing house DZS, which also participated in the EU project "Consortium for Central European Dictionary Encoding". The project developed a framework for encoding computational lexica, based on machine readable dictionaries (Erjavec et al. 2003).

There have also been various EU projects addressing Web based dictionary servers (Popescu-Belis et al., 2002), although they are mostly meant for larger amounts of data, and for cases where protecting access to the data is of more concern than in our academic project.

6. Conclusions

The paper presented the production, encoding and Web deployment of an electronic Japanese-Slovene learners' dictionary, meant for Slovene students of Japanese at the University of Ljubljana. The dictionary is currently available from http://nl.ijs.si/nihongo/

In our on-going work we plan to automatically create links between the word database, the kanji database, and appropriate indexes for dictionary searching, and, as mentioned, to convert the word database and insert it into the on-line Japanese reading support tool "Reading tutor". We also plan to continue our work on other vocabulary-building tools and activities for beginning and intermediate students. Of course, the dictionary is also being revised and enlarged on basis of data obtained by automatic monitoring of on-line dictionary searches and by explicit elicitation of user feedback.

References

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

De Schryver, G.-M. and Prinsloo, D. (2000). Dictionary making process with "simultaneous feedback" from the target users to the compilers. In U. Heid etc. (ed.), *Proceedings of the 9th Euralex congress*. Universität Stuttgart.

Erjavec, T., Evans, R., Ide, N., Kilgarriff, A. (2003). Machine Readable Dictionaries to Lexical Databases: the Concede Experience. In *Proceedings of the 7th International Conference on Computational Lexicography*, COMPLEX'03. Budapest, Hungary.

Extensible Markup Language (XML), W3C, 2000. http://www.w3.org/XML/

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.

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

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 - Research notes and demos*. http://www.rrz.uni-hamburg.de/slowenisch/

National Language Research Institute. (1962). *Gendai zasshi kyuujuushu no yougo youji (1)(2)(3)*. Tokyo: Shuuei Shuppan, Japan Foundation.

Popescu-Belis, A., Armstrong, S. Robert, G. (2002). Electronic Dictionaries - from Publisher Data to a Distribution Server: the DicoPro, DicoEast and RERO Projects. In the Proceedings of the Third International Conference on Language Resources and Evaluation, LREC'02. ELDA, Paris.

Mangeot, M., Sérasset, G. (2003). *The Papillon project.* http://www.papillon-dictionary.org/

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

Müller, C. (2003). *The Digital Dictionary of Buddhism, The Chinese/Japanese/Korean-English Dictionary.* http://www.acmuller.net/

Senmon Kyouiku Publishing. (1998). *Ichimango goi bunruishuu.* Tokyo: Senmon Kyouiku Publishing.

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/

Tamamura, F. (1990). Jisho. In *Kouza nihongo to nihongo kyouiku 7, Nihongo no goi - imi (ge)*. Tokyo: Meiji shoin.

Tamamura, F. (1995). Gaikokujin no tame no nihongo jisho kouzou. *Gekkan gengo 6.*