日本語・スロヴェニア語ウェブ辞書の開発

Tomaž Erjavec (Jožef Stefan Institute, Dept. of Knowledge Technologies) tomaz.erjavec@ijs.si
Kristina Hmeljak Sangawa (University of Ljubljana, Faculty of Arts) kristina.hmeljak@guest.arnes.si
Irena Srdanović (University of Ljubljana, Faculty of Arts) irena_srdanovic@hotmail.com
Anton ml. Vahčič (Univ. of Ljubljana, Faculty of Computer Science and Informatics) vahcica@hotmail.com

１．はじめに

辞書は自律的な外国語学習、特に語彙学習において不可欠な道具である。しかし１９９５年に初めて本格的な日本語教育が始まったスロヴェニアでは、日本語・スロヴェニア語の辞書はまだ存在せず、学習者数が少ないため出版される見込みもない。２０００年からリュブリャーナ大学文学部アジア・アフリカ研究学科日本研究講座では学生の協力を得ながら、スロヴェニア人日本語学習者のために日本語・スロヴェニア語の辞書編纂を行っている。学習者のニーズに対応できる辞書の完成までには相当な年数を必要とするが、日本語・スロヴェニア語の辞書が存在しない現状では、未完成な資料でも学習者のためになると考えられる。そこで、作成中の辞書は、辞書利用者である学生が提供するデータも随時加えながら、XML形式に編集しインターネットで公開している。[http://nl.ijs.si/jaslo/] この公開方法には学習者のフィードバックが即時に得られ、編纂方針を随時改善できるという利点もある。本稿では本辞書の内容構造、学生からのデータ収集方法、TEI準拠XML形式への変換、辞書の公開形式を紹介し、今後の展望を述べる。

２．辞書のモデルと編集の流れ

辞書の編集を始めた当初、編集はリュブリャーナ大学の日本語教師が行う企画であったが、数人の教師では、日本語の学生が学習期間に遭遇する可能性のある語彙を全て網羅することは非現実的だと判断した。そこで、編集作業を二つの流れに分けることにした。一つ目は、教師が日本語学習のための詳しい辞書項目を記述し、基礎語彙の学習辞典を構築すること、二つ目にこれと平行して上級学習者が日本語読解の際、調べた語彙（すなわち基礎語彙以外の語彙）を対応語彙表にまとめるということである。この二つの流れに沿ってデータを編集・収集した後、一つの辞書にまとめて公開することにした。

日本語学習のための詳しい記述については、上級学習者は日本で出版されている国語辞典や専門用語辞典などを利用できることから、最初の編集段階では初級・中級学習者に必要な約１万語の語彙項目を選定し掲載することにした。１万語とは、先行研究でしばしば中級日本語学習者に必要な語彙数として挙げられる数である。今回の辞書のためには、リュブリャーナ大学で使われている１年生の教科書の出現語彙をはじめ、「日本語能力試験１級出題基準語彙」、『日本語教育のための基本語彙調査』（国立国語研究所編）、『１万語語彙分類集』（専門教育出版編）を基に語彙項目を選定し掲載する企画である。

以下で述べるのは、当大学の１年生が利用している教科書に出現する約2000の語彙項目についてのモデルである。矢印の項目は、「その項目から関連項目への参照を挿入する」という意味である。

初級2000を除いた約8000語については、より簡単な記述を行う。アクセント型、表記（かなでの読みと漢字仮名交じり表記）、品詞、スロヴェニア語訳のほか、必要に応じて待遇上の制約などを掲載する。

学生がまとめたデータは現時点では教師を通してサーバー管理者に提出されるが、今後インターネット上で直接入力できるインターフェースも計画している。

（１）初級2000語彙の掲載項目	（２）初級2000語彙の掲載項目例
・アクセント ・表記：かな表記、漢字仮名交じり表記 ・文法：品詞 (スロヴェニア語訳と違う場合は注記) 動詞・形容詞の活用形と結合価 使用制限（文末制限など） ・意味：スロヴェニア語訳 類語とその使い分け（→） 対義語、上位語、下位語（→） 共起語（→） 慣用句・ことわざ ・位相：語種（和語・漢語）（→） 待遇上の制約（謙譲、丁寧、尊敬）（→） 媒体の制限（話し言葉・書き言葉） 話し手・使用場面の制限（専門用語など） 語感（雅語、俗語など） ・例文とそのスロヴェニア語訳	さむい【寒い】(Ai) 《⇔ あつい【暑い】》 1. mrzel, hladen (kraj, vreme) ‖ 去年の冬は寒かった。Lanska zima je bila mrzla. ｜フィンランドは寒い国だ。Finska je hladna dežela. 2. zebsti (v jpn. prid.!, povedniška raba samo v 1. os.) ‖ 今日は寒い。Danes zebe. ｜私は寒い。Mene zebe. ｜（*彼は寒い。）→彼は寒そうだ／寒がっている。Njega zebe. ぼく【僕】(N) [neform. moško za → わたし【私】] 《⇔きみ【君】ti》 1. jaz ‖ これは僕のだ。To je moje. ｜僕が手伝ってあげる。Ti bom jaz pomagal.

３．ＸＭＬ形式への変換

最初の辞書データ（約2000語彙項目）は表計算ソフトで編集できるような表形式になっていた。２００３年にその表データからマクロを利用してインターネットブラウザーで表示できるようなHTML形式に変換し、当学科のホームページ [http://www.ff.uni-lj.si/AzAfr] において公開したところ、このような形式に若干の問題があることが分かった。それは、辞書データを編集・拡大したり、見出し項目の構造を充実させたりするのに不便なこと、そしてデータ形式の確認が困難なことである。そこで、データの論理構造を記述することのできるXML（eXtensible Markup Language - 拡張可能なマーク付け言語）形式に変換することにした。XML形式への変換の利点は、データ形式の確認、充実、そして他の言語資源との交換が可能になることである。

XMLの要素（タグ）は自由に定義することができる。そこで、辞書データをXML形式に変換する際、まずどのXML要素を利用するか、つまりそのデータをどのように分類し記述するかをDTD（Document Type Definition - 文書型定義）において定義しなければならない。DTDを定義する際には、（１）独自のDTDを設計する、（２）既存の枠組みの中で既製品のDTDを利用する、（３）既存のタグセットを組み合わせたり、新しいタグを加えたりするという３つのやりかたがある。この辞書企画では、幅広く利用されており、かつ標準化され、検査済みで使用説明が充実しているTEIのDTDを採用することにした。

３．１．ＴＥＩ

TEI（Text Encoding Initiative）はテキスト資料を電子形式で表すための共通の方式を定めるために1987年に始まった国際的な共同研究活動である。TEIの目標は、世界中の電子化テキストの形式を統一させることによって、テキストのコンピュータ処理と交換流通を促すことにある。2002年に公開されたTEI-P4ガイドラインはウェブ上でも一般公開されている [http://www.tei-c.org]。

TEIガイドラインに対応したDTDを作成する際、TEIによって指定されているタグセットを利用する。当辞書プロジェクトでは、ＴＥＩ・Ｐ４のＸＭＬバージョンを利用し、そこで定められた辞書用のタグセットの他に、リンク用と分析用のタグセットも利用した。

３．２．ＴＥＩ規定形式への変換

元のデータは、１２のフィールド（表記・品詞・スロヴェニア語訳・例文１・例文１のスロヴェニア語訳、例文２・例文２のスロヴェニア語訳・教科書における新出課・活用形・注）からなる表になっていた。これをまず、スロヴェニア語の特殊文字が同時に処理できるようにShift-JISコードからUTF-8コードに変換し、続いて Perl フィルターを用いて自動的にTEIのXMLコードに変換した。その際、TEIタグを付加すると同時に、データの標準化（余白や句読点の排除）、検証（不正空文字列の洗い出し）と文字列パターンによる自動分類をほどこした。自動分類においては、潜在的に含まれた情報を明示的に記述することができた。例えば、「注」というフィールドの中には、位相、丁寧さ、外来語の語源などの情報が含まれており、その中で「iz...」で始まる文字列、例えば「iz angl. curtain」（「英語curtainから」の意味）を、外来語の語源を指す物として洗い出し、「<etym> <lang> angl. </lang> <gloss> curtain </gloss> </etym>」というふうに明白に記述した。

データをインターネット上公開する際、XSLTを用いてXML形式をHTML形式に変換して表示する。

（３）XML形式の見出し項目

```
<entry id="j.43">
    <form type="hw">
        <orth type="kana">あきらめる</orth>
        <orth type="kanji">諦める </orth>
    </form>
    <gramGrp>
        <pos>V1</pos>
        <subc>tr.</subc.>
    </gramGrp>
    <form type="infl">
        <orth>あきらめます<orth>
        <orth>あきらめて<orth>
    </form>
    <sense>
        <trans>
            <tr>vdati se (v usodo)</tr>
            <tr>obupati</tr>
        </trans>
        <eg>
            <q> 1 度の失敗ぐらいで
                あきらめるのはまだ早い。</q>
            <tr>Obupati komaj po enem spodrsljaju je
                &#353;e prekmalu. </tr>
        </eg>
    </sense>
    <sense>
        <trans>
            <tr>odpovedati</tr>
            <tr>opustiti (na&#269;rte, zelje)</tr>
        </trans>
        <eg>
            <q>天気が悪かったのでその計画
                をあきらめた</q>
            <tr>Ker je bilo vreme slabo, smo se
                odpovedali temu na&#269;rtu. </tr>
        </eg>
    </sense>
<entry>
```

４．辞書の公開 - http://nl.ijs.si/jaslo/

辞書のデータを書籍として出版する予定もあるが、当段階ではインターネット上で公開することにしている。その理由は、次の利点にある。まず、出版コストが低いこと、それから随時再編・加筆でき、簡便なことである。さそしてさらに重要なのは、利用者の検索語彙を記録し、その利用法の分析を基に辞書を改善していくことが可能であることである。特に日本語とスロヴェニア語のように、参考にできる辞書の伝統がない言語間では、De Schryver・Prinsloo (2000)が提唱するように利用者からの随時フィードバックを考慮した編集方法が有益だと考えられる。

電子媒体で提供する利点はもう一つある。利用者が辞書を引く場合は、必要な情報をすぐ入手でき、それ以外の情報には妨げられないのが理想だが、実際に辞書を引いてみると、そのような情報の弁別は大変な手間がかかる作業である。そのため、できるだけその手間を省くような辞書の設計を試みて、利用者に応じて必要以上の紛らわしい情報を表示しないように設定し、利用者のレベル、利用場面に応じた情報だけを表示する設計を予定している。

現時点では、約４０００語彙項目のデータが公開され、検索可能になっている [http://nl.ijs.si/jaslo/] 。検索においては、次の設定が可能である。

1）完全一致／前方一致／後方一致

2）全文検索／見出し語（日本語）検索／訳語（スロヴェニア語）検索

3）すべての品詞／名詞／動詞／形容詞･形容動詞／文句

検索画面の説明はデフォルトとしてスロヴェニア語になっているが、スロヴェニア人日本語学習者以外の利用者も想定し、日本語と英語の説明も作成した。

公開した時点から、検索のログを蓄積しており、これを分析しながら編集作業を進める予定である。

5．これからの展望

近年、本来書籍として出版された日本語の辞書の多くはIC 辞書や CD-ROM の形で出版・発売されているが、出版社のホームページ内、あるいは検索エンジンの一環としてインターネット上に提供されている辞書も多い。一方、インターネットの普及に伴い、高等教育機関を拠点に、またはインターネット上での自由な集まりとして、出版社に所属しない複数の人がインターネットを通して協同に作り上げていく辞書のプロジェクトが実現された。

日本語においては、モナシュ大学のJMDICT [http://www.csse.monash.edu.au/~jwb/wwwjdic.html] を始め、Linuxの協同の原理を辞書編纂に応用したパピヨン辞書プロジェクト[http://www.papillon-dictionary.org/]、和独辞典[http://bunmei7.hus.osaka-u.ac.jp/]、仏教用語の他言語辞書 [http://www.acmuller.net/]、英和の翻訳者が蓄積した単語・用例のデータベース英辞郎 [http://www.alc.co.jp/ 及び http://www.eijiro.jp/] など、多くのプロジェクトが挙げられる。

スロヴェニア語においては、インターネット上で自由に検索できる辞書は http://www.sigov.si/slovar.html にまとめられており、とりわけスロヴェニア標準語辞典 [http://bos.zrc-sazu.si/sskj.html] 、EU法律のスロヴェニア語訳のために作成された英語・スロヴェニア語・ドイツ語・フランス語用語辞典 [http://www.sigov.si/evroterm/] 、ドイツ語・スロヴェニア語辞典 [http://www.rrz.uni-hamburg.de/slowenisch/] などがあげられるが、日本語・スロヴェニア語という組み合わせにおいては本プロジェクトが初めての試みである。

本辞書プロジェクトのデータは「リーディング・チュウ太読解学習支援システム」（川村・北村 2001）に組み込む予定である。このシステムには現在３種類の辞書ツール（日日・日英・日独辞書ツール）が用意され、すでに当学科の学生に利用されているが、同じ形態素解析、辞書引き作業のためのツールを利用し自分の母語の訳語が得られれば、一層学習が進むであろう。これからは、利用者の希望とニーズを考慮に入れた辞書編纂、項目数の増加と記述の充実を図ると同時に、XML形式データの交換可能性と柔軟性を利用し、以上述べたような電子媒体の共同編集・増殖型辞書プロジェクトへの協力も試みたい。

参考文献

川村よし子、北村達也 （2001）「インターネットを活用した読解教材バンクの構築」『世界の日本語教育』国際交流基金、６号、pp.241-255. http://language.tiu.ac.jp/

国際交流基金 （2002）『日本語能力試験出題基準』凡人者

専門教育出版編 （1998）『品詞別・A～Dレベル別　１万語語彙分類表』専門教育出版

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. Universitaet Stuttgart.

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