Slovene Terminology Web Portal
and the TBX-Compatible Simplified DTD/schema

Simon Krek,* Vojko Gorjanc, Špela Arhar,*****
* Department for Knowledge Technologies, "Jožef Stefan" Institute,
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
** Department of Translation Studies, Faculty of Arts, University of Ljubljana,
Aškerčeva 2, SI-1000 Ljubljana, Slovenia
*** Amebis d.o.o., Bakovnik 3, SI-1241 Kamnik, Slovenia
E-mail: simon.krek@ijs.si, vojko.gorjanc@ff.uni-lj.si, spela.arhar@amebis.si

Abstract

The paper describes the project whose main purpose is the creation of the Slovene terminology web portal, funded by the Slovene Research Agency and the Amebis software company. It focuses on the DTD/schema used for the unification of different terminology resources in different input formats into one database available on the web. Two projects involving unification DTD/schemas were taken as the model for the resulting DTD/schema – the CONCEDE project and the TMF project. The final DTD/schema was tested on twenty different specialized dictionaries, both monolingual and bilingual, in various formats either without any existing markup or with complex XML structure. The result of the project will be an online terminology resource for Slovenian which will also include didactic material on terminology and free tools for uploading domain-specific text collections to be processed with NLP software, including a term extractor.

1. Terminology in Slovenia

Until recently, the development of Slovene language resources has been primarily focused on building text corpora which is understandable as they represent the core of language resources infrastructure. As a result of the cooperation between experts from different fields, corpus-building projects provided a solid platform for further evolution of language resources. The available Slovenian corpora also generated a number of comprehensive corpus studies, both monolingual and contrastive, and they have become – the FIDA[1] and FidaPLUS[2] corpus in particular – an indispensable part of linguistic research in general, especially in lexical or semantic studies, and regrettably less so in terminology. To ensure their integrity and comprehensiveness, the existing resources have to be supplemented with new ones, among them also terminological resources. In that area, one of the major problems in Slovenia is the lack of cooperation. There is a series of terminology-related activities running and a significant number of terminological dictionaries exist, but they are not collected in one accessible resource, they are methodologically heterogeneous and often unavailable for public use.

2. The Web Portal Project

The goals of the Slovene Terminology Web Portal research project are, first, to make an overview of Slovenian terminology resources and to measure the interest for cooperation in different professional fields. The main objective of the project is to develop the Slovenian terminology portal which would offer basic information on the principles of terminological work and above all make available software tools for presenting terminology in a unified format. It will be designed to enable the compilation of different types of terminological dictionaries for the interested parties, presented in a unified manner and publicly available on the web. Thirdly, the feasibility of linking textual resources with the terminological database will also be tested. We wish to examine to what extent terminologically relevant data can be extracted from Slovenian corpora with (semi)automated procedures, which can be later used as the basis for specialized dictionary compilation.

Given the imperative that the web portal should be able to integrate various existing terminology resources with heterogeneous contents, at the core of the Web Terminology Portal there is an XML DTD/scheme which can rather seamlessly integrate different formal structures of existing data and on the other hand produce a standardized terminology database structure to be integrated in software tools which recognize the standard terminological markup. Another important assumption which determines the design of the DTD/schema is the end-user profile. As the web portal is primarily intended for native speakers of Slovene, it is considered important that the markup should be in Slovene to enable better understanding of the database structure. It is assumed that the end-user of some of the more technical features of the portal would not be computer science experts, but typically linguists, translators, students etc. Therefore it seemed important to help them by simplifying the more technical aspects which cannot be avoided when working with the portal, for instance, downloading the terminological content in order to be used in an offline computer environment.

[1] http://www.fida.net
[2] http://www.fidaplus.net

3. Elements, Attributes

The article mainly focuses on the DTD/schema which lies at the core of the web portal, unifying heterogeneous data and providing a very simple initial markup for new terminology resources. Initially, it was based upon the findings of two projects which dealt with integrating different dictionary-like contents into a unified markup structure: the CONCEDE project[3] – Consortium for Central European Dictionary Encoding – and the TMF project[4] – Terminological Markup Framework. The latter was seen as more helpful since it involved terminological data, as opposed to the first which dealt with general monolingual dictionaries and one bilingual dictionary. TMF also has a more close relation to the current encoding standards such as ISO 16642, DXLT, TBX etc. The aim was to use a minimal set of elements enabling the integration of any existing dictionary without the loss of vital structural and linguistic information needed for the web portal. The final DTD/schema included sixteen elements and eight attributes and was tested on twenty different dictionaries:

ELEMENTS:

NAME	SLOVENIAN	ENGLISH
slovar	ROOT	ROOT
geslo	geslo	entry
izt	iztočnica	headword
struk	struktura	structure
zapis	zapis	feature
razlaga	razlaga	gloss
prevod	prevod	translation
primer	primer	example
kvalif	kvalifikator	label
izg	izgovorjava	pronunciation
besvrs	besedna vrsta	part-of-speech
kaz	kazalka	cross-reference
loc	ločilo	punctuation
obl	oblika	style
atr	atribut	attribute
mmd	multimedija	multimedia

Table 1 DTD/schema elements

ATTRIBUTES:

NAME	ENGLISH
status	status
avtor	author
datum	date
ime	name
tip	type
id	id
jezik	language
oblika	layout

Table 2 DTD/schema attributes

The original dictionaries were in various formats: printed and OCR-ed (therefore without a pre-existing easily identifiable structure), or in different electronic input formats:

- Microsoft Office with styles
- text, lightly annotated in quasi-XML
- HTML or
- proper XML with extensive hierarchy.

Different routines as well as manual tagging were used for converting the original formats into XML for the purposes of testing the minimal DTD/schema. Several principles were applied:

- existing content elements and structural elements were transformed into elements "structure" and "feature", respectively
- if original markup existed, the information about the original elements was kept in the attribute "type"
- if no markup was present and it seemed sensible to include information about the content, it was registered in the attribute "type" but marked differently to be recognized as internally coded
- elements or text recognized as belonging to the set of elements with specific content were tagged as such: translation, gloss, cross-reference, example, label, part-of-speech and pronunciation; these can only be content or mixed content elements
- headword and translation elements may contain only one headword (can be multiword) or translation; if the original element contained more, it was marked as a structural element containing only other elements.

4. Examples

We show two simple examples from two of the dictionaries: one comes from the bilingual *Dictionary of Nuclear Physics*, originally lightly annotated in near-XML format and the other is from the bilingual *Dictionary of Rubber Industry*, originally in MS Word, with different parts of the dictionary microstructure marked with different styles in MS Office:

1. Dictionary of Nuclear Physics

```
<idpo>2</idpo>
<po>Obsev, ki ga povzročajo naravni viri.</po>
<idkv>8</idkv>
<kv>varstvo pred sevanji</kv>
<idang>1714</idang><ang1/>
<ang>natural exposure</ang>
<idslo>950</idslo><slo1/>
<slo>naravni obsev</slo>
```

2. Dictionary of Rubber Industry

dinamométer -tra m
naprava za merjenje sil ob obremenjevanju vzorca: elektronski dinamometer; ugotavljanje natezne trdnosti vulkanizatov na dinamometru
AN **tensile testing machine, dynamometer**
NE **Zugprüfmaschine** ž

[3] http://www.itri.brighton.ac.uk/projects/concede
[4] http://www.loria.fr/projets/TMF

The two dictionary entries are converted into the uniform markup according to the DTD/schema:

1. Dictionary of Nuclear Physics

```
<geslo>
    <struk tip="stp_razlaga">
        <zapis tip="idpo">2</zapis>
            <razlaga tip="po">Obsev, ki ga
            povzročajo naravni viri.</razlaga>
    </struk>
    <struk tip="stp_kvalif">
        <zapis tip="idkv">8</zapis>
            <kvalif tip="kv">varstvo pred
            sevanji</kvalif>
    </struk>
    <struk tip="stp_prevod">
        <zapis tip="idang">1714</zapis>
        <zapis tip="ang1"/>
        <prevod tip="ang" jezik="ang">natural
        exposure</prevod>
    </struk>
<izt jezik="slo">
    <zapis tip="idslo">950</zapis>
    <zapis tip="slo1"/>
    <zapis ime="stp_odkrit" tip="slo">naravni
    obsev</zapis>
</izt>
</geslo>
```

2. Dictionary of Rubber Industry

```
<geslo>
<izt jezik="slo">
    <zapis tip="stp_skrit">dinamometer</zapis>
    <zapis ime="stp_odkrit">dinamométer</zapis>
</izt>
<struk tip="stp_zaglavje">
    <zapis tip="stp_sklon">-tra</zapis>
    <besvrs>m</besvrs>
</struk>
<struk tip="stp_razlaga">
    <razlaga>naprava za merjenje sil ob obremenjevanju
    vzorca</razlaga>
        <loc>:</loc>
    <primer>elektronski dinamometer</primer>
        <loc>;</loc>
    <primer>ugotavljanje natezne trdnosti vulkanizatov
    na dinamometru</primer>
</struk>
<struk tip="stp_prevod">
    <zapis tip="stp_jezik">AN</zapis>
    <prevod jezik="ang">tensile testing
    machine</prevod>
        <loc>,</loc>
    <prevod jezik="ang">dynamometer</prevod>
</struk>
<struk tip="stp_prevod">
    <zapis tip="stp_jezik">NE</zapis>
    <prevod
    jezik="nem">Zugprüfmaschine</prevod>
    <zapis tip="stp_nemspol">ž</zapis>
</struk>
</geslo>
```

The primary aim is to identify the two elements considered as the most important for the purposes of the web portal – the headword (bold and underline in the above example) and the translation (bold in the above example), and to isolate them in the formal structure which could be applied to any existing terminology database or dictionary. Compared to this goal, to identify other content elements in the dictionary is considered to be of secondary importance and mainly depends on the easily identifiable features of the original structure and the overall effort needed to convert the particular database. Because of very diverse input formats some of the identification and segmentation procedures are expected to be fully automated but most of the input material will require at least some additional processing. Therefore, the effort to uniquely identify elements of dictionary structure other than headwords and translations will largely depend on the existing information in the input format and the available funding or time constraints of the project.

The second requirement is the possibility of conversion of the data back into its original form. It is important to provide a simple procedure to record and store original markup which can be either virtually non-existent or extremely rich. The small extract from the Dictionary of Nuclear Physics shows the main principles of recording the existing structure:

```
...
<idslo>950</idslo>
<slo1/>
<slo>naravni obsev</slo>
...

<izt jezik="slo">
    <zapis tip="idslo">950</zapis>
    <zapis tip="slo1"/>
    <zapis ime="stp_odkrit" tip="slo">naravni
    obsev</zapis>
</izt>
```

The headword "naravni obsev" is identified via the <izt> element and the attribute "stp_odkrit". The names of the existing tags are recorded in the attribute "tip". The empty element <slo1/> is also recorded, exclusively for the purposes of back conversion.

5. DTD

The testing of twenty different dictionaries yielded the following DTD (only elements without attributes are presented here):

```
<!ELEMENT slovar  (geslo+) >
```

```
<!ENTITY % osnovni
   '(kvalif | razlaga | prevod | primer | izg | izt | besvrs | zapis
| kaz | loc | atr | mmd)'   >

<!ELEMENT geslo          (%osnovni; | struk)* >
<!ELEMENT struk          (%osnovni; | struk)* >
<!ELEMENT izt            (zapis , (loc | atr | mmd)*)+ >
<!ELEMENT razlaga        (#PCDATA | atr | obl | mmd)* >
<!ELEMENT zapis          (#PCDATA | atr | obl | mmd)* >
<!ELEMENT prevod         (#PCDATA | atr | obl | mmd)* >
<!ELEMENT primer         (#PCDATA | atr | obl | mmd)* >
<!ELEMENT kvalif         (#PCDATA | atr | obl | mmd)* >
<!ELEMENT izg            (#PCDATA | atr | obl | mmd)* >
<!ELEMENT besvrs         (#PCDATA | atr | obl | mmd)* >
<!ELEMENT kaz            (#PCDATA | atr | obl | mmd)* >
<!ELEMENT loc            (#PCDATA) >
<!ELEMENT atr            (#PCDATA) >
<!ELEMENT mmd            (#PCDATA) >
<!ELEMENT obl            (#PCDATA) >
```

It is foreseeable that the content of the web portal will be downloaded and used in specialized software for building individualized terminological databanks, as well as for educational purposes at the Department of Translation Studies at the Faculty of Arts, Ljubljana. It is therefore important that the conversion from the simplified basic format into the standard format used by more sophisticated software tools such as Trados Multiterm or similar is enabled. An XSLT procedure is available which converts the simplified format into the standard TBX format. However, it must be stressed that in transformation process from the simplified unification format to TBX, most of the original encoding (if existent) is lost and only selected elements such as headwords and translations (together with the corresponding grammatical information) are rendered. This is due to the nature of the original goals of the project which focuses on fully automatic processing of very different input dictionary formats to be put in an online searchable database, with segments of the database downloadable in the TBX compatible format, according to different selected criteria.

If the TXB-compatible transformation is applied to the above examples, the following result is rendered:

1. *Dictionary of Nuclear Physics*

```
<martif type="TBX" xml:lang="en">
<martifHeader>
<fileDesc>
    <sourceDesc>
        <p>from STP web portal</p>
    </sourceDesc>
</fileDesc>
</martifHeader>
<text>
<body>
<termEntry>
    <langSet xml:lang="eng">
<tig>
    <term>natural exposure</term>
</tig>
</langSet>
    <langSet xml:lang="slo">
        <tig>
            <term>naravni obsev</term>
        </tig>
</langSet>
</termEntry>
</body>
</text>
</martif>
```

2. *Dictionary of Rubber Industry*

```
<?xml version="1.0" encoding="iso-8859-1"?>
<martif type="TBX" xml:lang="en">
<martifHeader>
<fileDesc>
    <sourceDesc>
        <p>from STP web portal</p>
    </sourceDesc>
</fileDesc>
</martifHeader>
<text>
<body>
<termEntry>
<langSet xml:lang="slo">
<tig>
    <term>dinamométer</term>
</tig>
</langSet>
<langSet xml:lang="ang">
<tig>
    <term>tensile testing machine</term>
</tig>
<tig>
    <term>dynamometer</term>
</tig>
</langSet>
<langSet xml:lang="ger">
<ntig>
<termGrp>
    <term>Zugprüfmaschine</term>
        <termCompList type="grammaticalGender">
<termCompGrp>
    <termComp>Zugprüfmaschine</termComp>
        <termNote
            type="grammaticalGender">
            feminine</termNote>
</termCompGrp>
</termCompList>
</termGrp>
</ntig>
</langSet>
</termEntry>
</body>
```

```
</text>
</martif>
```

6. Conclusions

The paper describes the Slovene Terminology Web Portal applicative research project and the process of converting a set of terminological dictionaries in different input formats into one uniform structure with a core DTD/schema. The principles of conversion are examined and presented together with some of the more revealing examples. In addition, XSL transformation which transforms the converted data into TBX-compliant structure is explained and examples are given. The project is viewed as the first step towards a freely available online resource of terminology data in Slovene with a standardized procedure from the existing data in various formats to the web portal with the possibility of its conversion and loading into other terminology management systems. In addition, the tool for uploading domain-specific text collections which can be processed with NLP software, including a term extractor, is scheduled as the next step in the project.

7. References

Erjavec, T., Evans, R., Ide, N., Kilgarriff, A., (2000). The CONCEDE model for Lexical Databases. In *Proceedings of the Second International Language Resources and Evaluation Conference*, Paris: European Language Resources Association.

Ide, N., Kilgarriff, A., & Romary, L. (2000). A Formal Model of Dictionary Structure and Content. *Proceedings of Euralex 2000*, Stuttgart, pp. 113-126.

Ide, N., Romary, L., and de la Clergerie, E. (2003). International standard for a linguistic annotation framework. In *Proceedings of the HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems - Volume 8 (May 31 - 31, 2003)*. Human Language Technology Conference. Association for Computational Linguistics, Morristown, NJ, pp. 25-30.

Khayari M., Schneider S., Kramer I., Romary L. (2006), Unification of multi-lingual scientific terminological resources using the ISO 16642 standard. The TermSciences initiative. Acquiring and representing multilingual, specialized lexicons: the case of biomedicine. In *Proceedings of the Fifth International Language Resources and Evaluation Conference*, Genua, Italy.

Kilgarriff, A. (1999). Generic encoding principles. *CONCEDE Project Deliverable 2.1*, University of Brighton, UK.