Slovene Terminology Web Portal

Vojko Gorjanc
University of Ljubljana

Simon Krek
Jozef Stefan Institute

Špela Vintar
University of Ljubljana

Work in the field of terminology is extensively supported worldwide as it enhances the transfer of science and technology. In Slovenia, there is a series of terminology-related activities running, and a significant number of terminology dictionaries and terminological data exist, but they are methodologically heterogeneous and often unavailable for public use. Traditionally, terminology work in Slovenia is closely connected with other activities in the filed of lexicology and lexicography, especially regarding the methodological approach to the compilation of dictionaries of specialised languages. Therefore, terminology work is mostly regarded only as a process involving the compilation of dictionaries. The paper presents the Slovene Terminology Web Portal project. The main objective of the project is to develop the Slovenian terminology portal to offer basic information on the principles of terminological work and to present a terminological database in a unified format. In the core of the presentation, there is a process of conversion of different types of existing terminology data from different sources into XML format with a simple DTD/schema and from there to unified TBX database. Simultaneously, the feasibility of linking textual resources and the extraction of term candidates with the terminological database is also shortly presented.

1. Introduction

Work in the field of terminology is extensively supported worldwide as it enhances the transfer of science and technology. In Slovenia, numerous terminology-related activities are underway, yielding term glossaries and databases from various domains, however they are methodologically heterogeneous and often unavailable for public use. Traditionally, terminology work in Slovenia was closely connected to other activities in the field of lexicology and lexicography, especially regarding the methodological approach towards the compilation of specialized language dictionaries. Therefore, terminology work was mostly understood in terms of traditional dictionary compilation. However, we are now aware that with efficient terminology management we not only can encourage faster transfer of science and technological innovations within and beyond the Slovene language community, but also present Slovene technological achievements in foreign languages. By ensuring fast and efficient exchange of new information, the terminology portal also offers all kinds of actors operating in this field an opportunity to achieve better performance. The skilled transfer of specialized information from one language to another is also an important factor in economic development.

2. Language resources for Slovene and terminology data

Until now, projects developing language resources for the Slovene language have been aimed primarily at building text corpora, which was a reasonable choice as corpora represent the foundation of language infrastructure. In Slovenia, corpus linguistics successfully completed its first and indispensable stage with the compilation of large reference corpora, which represent the basis for further development in the field. Due to the inevitable cooperation between experts from different fields, corpus-building projects provided a solid platform for further evolution of language

resources for Slovene. Furthermore, the available Slovene corpora generated a number of comprehensive corpus studies, both monolingual and contrastive, and they have become—the FIDA and FidaPLUS corpora in particular *http://www.fidaplus.net* (Arhar et al. 2007)—an indispensable part of linguistic research in general, especially in lexical or semantic studies, and regrettably less so in terminological ones (Gorjanc and Krek 2001).

To ensure the integrity and comprehensiveness of language resources for Slovene, the existing resources have to be supplemented with new ones, among them also terminological. It would be advisable for these resources, however, to be designed so as to foster cooperation and exchange between broader professional audiences. As terminological work requires a considerable amount of time and resources, it is important to implement efficient methods of work, procedures, and tools.

The goals of the Slovene Terminology Web Portal project are, first, to compile an overview of Slovene terminology resources and to measure the interest of different professional fields in cooperation. The main objective of the project is to develop the Slovene Terminology Web Portal, which would offer basic information on the principles of terminological work and above all make available software tools for presenting terminology in a standardized format. 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 Slovene corpora with (semi)automated procedures, which can later be used as the basis for terminology database creation.

3. From dispersed terminology resources to a unified termbase

Numerous and various terminology resources are already available for Slovene. On the one hand, there are large-scale printed terminological dictionaries laboriously produced by the Commission for Terminology of the Slovene Academy of Sciences and Arts. Over a dozen such works have been compiled within the last decade, and since electronic versions exist for the majority of them, the Slovene Terminology Portal will provide for conversion into a standardized format. Secondly, terminology data is being collected in academic settings, where specialized dictionaries of different scope and for various professional fields and subfields are compiled within regular student work on research papers, and bachelor's, master's, and doctoral theses. In addition, a number of initiatives promoting specialized dictionary design exist on the web, both individual and institutional, for example on the web pages of financial and insurance institutes, mobile communication providers, etc. This indicates that there is a strong awareness of the economic importance of terminology as a basis for efficient professional and inter-professional communication, but this has thus far been left to private initiative or individual institutions. Upon analysis of existing terminology data, we can conclude that the main problems are the following:

- the dispersion of terminology resources,
- incoherent methodologies and principles, as well as formats that lead to incompatibility,
- a lack of public availability, or—if available—multiple access points with many types of data presentation.

In the preliminary phase of our project, the terminology data collection is organized according to the established availability and legal status of each terminological resource. In the first step, resources are collected regardless of their form (printed, electronic) or particular format of the electronic resources. In the second step, a priority list is formulated for digitalization if necessary and for subsequent standardization of the electronic format. This is largely dependent both on the copyright status and on the type or form of the resource, and also on the particular domain, as we wish to present samples from different terminological resources for various domains.

3.1. *Following best practices for standardized terminology databases*

The EuroTermBank project (Gorzkoś and Borkowski 2006, Rirdance 2007) was chosen as a best practice methodology for multilingual terminology management based on international standards (*http://www.eurotermbank.com*). This approach towards the treatment of copyright and application of standards was to a great extent adopted for the purpose of creating the Slovene Terminology Portal. Since it is a large-scale project and with limited financial and human resources, we also examined other terminology portals, the *German Terminology Portal* (Deutsches Terminologie-Portal) *http://www.iim.fh-koeln.de/dtp*, the Canadian *Termium http://www.termiumplus.gc.ca*, and the *Slovak Terminological Database* (Slovenská terminologická databáza) *https://data.juls.savba.sk/std/*. In terms of available resources and the structure of the database, this last database is most similar to the project of the Slovene Terminology Portal.

For the rendering of the terminology database as a concept-oriented database (Wright and Budin 1997), TBX as an open XML-based standard format for terminological data was chosen (*http://www.lisa.org/standards/tbx/*). It is also important that the conversion from the simplified basic format into the standardized format used by more sophisticated software tools such as SDL Trados Multiterm or similar is enabled. The TBX standard is perfectly suited for this task, since TBX files can be imported into and exported from most software packages that use terminological databases. On the basis of TBX, a limited set of data categories was created, following also TBX-Basic, which is a scaled-down version of TBX, particularly suited to small or medium sized terminology projects (*http://www.lisa.org/Terminology-Special.102.0.html*). Both TBX and TBX-Basic are in accordance with ISO standards. In fact, a complete inventory of types of data categories for work with terminology data has been defined in ISO 12620:1999 (Computer applications in terminology—Data categories), and TBX is in the process of being published as ISO/DIS 30042.2. But for a great part of terminology recording, a much smaller set of data categories is required than are defined in the ISO standard; as demonstrated by the studies conducted by the Localization Industry Standards Association in order to determine what types of data the language industry actually needs for managing terminology (*http://www.lisa.org/*). Since terminology data in the database will hopefully be used for many purposes, in the process of creating the final database structure, several other international standards were also consulted, especially ISO 12200:1999 (Computer applications in terminology—Machine-readable terminology interchange) and ISO 16642:2003 (Computer applications in terminology—Terminological markup framework) (Bonnono 2000, Khayari et al. 2006).

3.1.1. *Conversion to a standardized XML format*

Given the imperative that the terminology web portal should be able to integrate various existing terminology resources with heterogeneous contents, the first step was the creation of an XML DTD/schema which can seamlessly integrate different formal structures of existing data and produce a standardized terminology database structure to be integrated in software tools which recognize the standard terminological mark-up. Initially, the DTD/schema was based upon the findings of two projects which dealt with integrating different dictionary-like contents into a standardized mark-up structure: the CONCEDE project (*http://www.itri.brighton.ac.uk/projects/concede*)—Consortium for Central European Dictionary Encoding—and the TMF project (*http://www.loria.fr/projets/TMF*)—Terminological Markup Framework (Kilgarriff 1999; Erjavec et al. 2000, Ide et al. 2000). 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 creation of the database. The final DTD/schema included sixteen elements and ten attributes and was tested on twenty different dictionaries in various formats: printed (unstructured, OCR-ed), Microsoft Office with styles, lightly annotated in quasi-XML, HTML, and XML with an extensive hierarchy:

ELEMENTS		
NAME	SLOVENE	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
ATTRIBUTES		
Ime	Ime	Name
Tip	Tip	Type
Id	Identifikacija	Identification
Jezik	Jezik	Language
Oblika	Oblika	Format
Vred	Vrednost	Value

With conversion from different sources, an XML database can be created with entries as follows:

```
<geslo>
        <struk tip="p">
                <zapis tip="small">English</zapis>
                <zapis tip="br"/>
                <struk tip="big">
                        <struk tip="b">
                                <prevod jezik="ang">computational linguistics</prevod>
                        </struk>
                </struk>
                <zapis tip="br"/>
```

```
            <razlaga>Computational linguistics is a field concerned with the processing of natural language by computers. The term is more often used in an academic context. It is closely related to Natural Language Processing and Language Engineering.</razlaga>
        </struk>
        <struk tip="p">
            <zapis tip="small">Slovenian</zapis>
            <zapis tip="br"/>
            <struk tip="big">
                <struk tip="b">
                    <izt jezik="slo">
                        <zapis ime="stp_odkrit">računalniško jezikoslovje</zapis>
                    </izt>
                </struk>
            </struk>
            <zapis tip="br"/>
            <razlaga>Računalniško jezikoslovje se ukvarja z računalniško obdelovo naravnega jezika. Termin se pogosteje uporablja v akademskem okolju. Tesno je povezan z obdelavo naravnega jezika in jezikovnim inženiringom.</razlaga>
        </struk>
        <zapis tip="hr"/>
</geslo>
```

3.1.2. *From simplified XML to TBX*

In the second step the simplified format is converted into the standard TBX format. The primary aim is to identify the two elements considered to be the most important for the purposes of the web portal—the headword and the translation, and to isolate them in a formal structure that could be applied to any existing terminology database or dictionary. In comparison to this goal, identifying other content elements in the dictionary was considered of secondary importance and depends on the identifiable features of the original and the effort needed to convert the particular database. The TBX entry converted from the simplified format looks as follows:

```
<termEntry id="15">
    <langSet xml:lang="sl">
      <descripGrp>
        <descrip type="explanation">odklon svetlobnih znakov</descrip>
      </descripGrp>
      <ntig id="15-sl-1">
        <termGrp>
          <term>aberacija</term>
        </termGrp>
        <transacGrp>
          <transac type="terminologyManagementTransactions">origination</transac>
          <transacNote type="responsibility">JJJ</transacNote>
          <date>20071123T13:18:53Z</date>
        </transacGrp>
```

```
        </ntig>
    </langSet>
    <langSet xml:lang="en">

      <ntig id="15-en-1">
        <termGrp>
          <term>aberation</term>
        </termGrp>
        <transacGrp>
          <transac type="terminologyManagementTransactions">origination</transac>
          <transacNote type="responsibility">JJJ</transacNote>
          <date>20071123T13:18:53Z</date>
        </transacGrp>
        <transacGrp>
      </ntig>
    </langSet>
</termEntry>
```

3.2. *Final design of the Slovene terminology portal*

Within the project, a web site will be designed to make terminological resources publicly available along with information on terminology, procedures, and principles. Free tools will be available together with instructions and recommendations on how to use them. During the project, discussion will be encouraged among the project partners and between the partners and the interested professional public regarding the different levels of control over the data published on the terminology portal. Another key functionality we envisage is the possibility for users to upload their domain-specific text collections and have them processed with our tools, including a term extractor.

3.2.1. *Interactive term extraction functionality*

The compilation of domain-specific text collections is no longer considered a technological challenge. Since corpus-based terminography provides methods that facilitate and, in part, automate the process of termbase creation, we shall provide a tool that will allow users to upload their corpus, whereupon the tool will produce lists of term candidates for the user to evaluate.

The technology behind this feature is well known in NLP circles and is based on a combination of statistical processing steps and linguistic patterns considered terminologically relevant (Vintar 2004). Thus, the word list from a domain-specific corpus is compared with the word list of a reference corpus, which yields the keywords of the domain. These are then used to extract terminological phrases, whereby users will be given the possibility to select specific part-of-speech patterns of interest to them.

3.2.2. *Term annotation*

Another useful extension of the portal will be the term *annotation functionality*, a feature whereby the user will upload a text and the tool will automatically identify the terms in the text, using an index created from all lexical resources contained within the portal. This function is extremely useful in determining the level of specialization in a text, finding the domains a text belongs to, or for simply checking how well our termbases cover the terminology contained in a certain text (Vintar 2001).

4. Conclusions

The information society is a major challenge which has sparked lasting interest in the formulation of principles and methods for tackling the problem of transferring knowledge across languages. The recognition of the right to free communication, which can be realized only within one's mother tongue, has lead to the widely accepted principle of ensuring the creative freedom of each individual in their own language, together with the possibility of exchanging information with other languages. This is true for all domains of human activity and creativity; in a knowledge-based society, however, scientific and professional communication is particularly emphasized. Creating a unified terminology database for the Slovene language is therefore one of the steps towards effective communication in terms of specialized knowledge transfer in Slovenia that fulfils the fundamental needs of the speakers of Slovene regarding specialized communication.

References

Arhar, Š.; Gorjanc, V.; Krek, S. (2007). "FidaPLUS Corpus of Slovenian. The New Generation of the Slovenian Reference Corpus: Its Design and Tools". In *Proceedings of the Corpus Linguistics Conference* [online]. University of Birmingham. *http://www.corpus.bham.ac.uk/corplingproceedings07/* [Access date: 26 March 2008].

Bonnono, R. (2000). "Terminology for Translators – an Implementation of ISO 12620". *Meta* XLV (4). 646-669.

Erjavec, T. et al. (2000). "The CONCEDE Model for Lexical Databases". In *Proceedings of the Second International Language Resources and Evaluation Conference*. Paris: European Language Resources Association.

Gorjanc, V.; Krek, S. (2001). "A Corpus-Based Dictionary Database as the Source for Compiling Slovene-X Dictionaries". In *COMPLEX 2001, 6th Conference on Computational Lexicography and Corpus Research Computational Lexicography and New EU Languages* Birmingham: Centre for Corpus Linguistics, Department of English, University of Birmingham. 41-47.

Gorzkoś, E.; Borkowski, T. (2006). EuroTermBank – a Product of Multilateral Cooperation of Terminological Institutions Entailing Sophisticated Technology Transfer [online]. *http://www.opi.org.pl/repository/7b6bedc5afcc6e87961986055ec717b4YvkaAy.pdf* [Access date: 25 March 2008].

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

ISO 12200:1999 - Computer applications in terminology – Machine-readable terminology interchange

ISO 12616:2002 - Translation-oriented terminography

ISO 12620:1999 - Computer applications in terminology – Data categories

ISO 16642:2003 - Computer applications in terminology – Terminological markup framework

ISO/DIS 30042.2 - Computer applications in terminology – Term-based exchange format specification

Khayari M. et al. (2006). "Unification of Multi-Lingual Scientific Terminological Resources Using the ISO 16642 Standard, The TermSciences Initiative". In *Acquiring and Representing Multilingual, Specialized Lexicons: the Case of Biomedicine (LREC 2006)*. Genua, Italy.

Kilgarriff, A. (1999). Generic Encoding Principles, CONCEDE Project Deliverable 2.1. University of Brighton.

Rirdance, S. (2007). Eurotermbank: Multilingual Terminology Sharing Portal [online]. *http://www.lisa.org/Terminology-Special.102.0.html* [Access date: 25 March 2008].

TBX-Basic: Data Categories and Usage Guidelines [online]. *http://www.lisa.org/Terminology-Special.102.0.html* [Access date: 25 March 2008].

TermBase eXchange Link (TBX Link) 1.0 Specification, Initial Draft 0.1.1, 1 March 2007 [on line]. *http://www.lisa.org/standards/tbxlink/tbxlink.html* [Acces date: 25 March 2008].

Vintar, Š.; Kipp, M. (2001). "Multi-Track Annotation of Terminology Using ANVIL". In Ide, N. (ed.). *Workshop on Multi-layer Corpus-based Analysis, Eurolan 2001*. Iasi, Romania. 1-13.

Vintar, Š. (2004). "Comparative Evaluation of C-Value in the Treatment of Nested Terms". In *Memura 2004 - Methodologies and Evaluation of Multiword Units in Real-World Applications* (LREC 2004). 54-57.

Wright, S. E.; Budin, G. (1997). *Handbook of Terminology Management*. Amsterdam: John Benjamins.