Design of a Multilingual Terminology Database Prototype*

Mária Šimková[1], Radovan Garabík[1], Ludmila Dimitrova[2]

[1] Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava
[2] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia

Abstract. A prototype of a multilingual terminology database has been designed and implemented, with the intention to facilitate collaboration among MONDILEX member institutes, where either missing or incompatible Slavic languages terminology of modern aspects of linguistics can be a hindrance of mutual communication. The database is intended to contain entries with specialized corpus linguistics terms, and the prototype is filled with terms in Bulgarian and Slovak, with relevant English equivalents. The plan is to add terms in all the MONDILEX languages, and eventually release the database with the hope that its content will grow beyond the very narrow terminology of corpus linguistics.

1 Introduction

As the corpus linguistics is relatively new in Slavic languages – the development began only after the personal computer boom – there is no unified terminology of this field. The terminology started to develop uncontrollably, either by directly adopting English terms or by calquing the English expressions, or by embracing and extending existing linguistic terminology in each country. This development lead to widely varied terminology in different countries, and even to different terminology used by different institution in the same country, while sometimes the English terms are considered to be just a part of an informal slang.

The key issue is to harmonise the definitions and thus ensure consistency and clarity of information across the languages, especially when communicating with experts from various countries, where the use of bridge language is often not sufficient, or when dealing with bi- or multilingual resources, with the consequent need of multilingual documentation.

Since "the ultimate purpose of any terminological resource is to facilitate and enhance knowledge acquisition" [2], the database has been designed in a way to function as a quick reference source of terms in different languages, which has influenced its overall design.

The database, once finished, could be also used to compare the usage and acceptance of English terms in various languages.

Extensive and theoretical study on definitions and formalism is beyond the scope of this paper – we describe only the technical implementation and general features of our database.

2 Implementation

Multilingual terminology database (MLTD) is developed using the MoinMoin wiki engine as a backend. The data is kept in plain text files, with one file (MoinMoin page) corresponding to one terminology entry. The technical implementation, and to an extent a terminology entry structure has been inspired by the Slovak Terminology Database design [4, 5].

As a minimum, a terminology entry in MLTD should contain a term, its definition (explanation) and a source of the definition. Intentionally, MLTD tries to keep the minimalistic approach and therefore adds no additional data.

Compared with the simplicity of MLTD, Slovak Terminology Database entry has 13 fields, 5 of them are obligatory (*term, field, definition, biblio, acceptability*). *Field* is simulated by the page category, and *acceptability* (pragmatic term character, one of *normalised, legislative, recommended,*

* The study and preparation of these results have been partly supported by the EC's Seventh Framework Programme [FP7/2007-2013] under the grant agreement 211938 MONDILEX.

suggested, incorrect, archaic, neologism) is mostly relevant for national terminology systems dealing with terminology standardisation, and as such has no place in MLTD – it is implicitely included in the information about definition source.

This design allows the internal format of the database entry to be kept very simple, nothing more than a plain text file with a minimal layout, without any special formatting markup. By a design decision, internal page format does not use any immediately visible markup language. The motivation stems from our empirical observation regarding usability – the presence of any, even the most incopious markup distracts the editors, unless they are reasonably well trained in the markup (and discourages them to learn to use the system). Our markup is hidden in the overall text structure, using nothing more than strategically placed paragraph breaks, colons and parentheses used in a relatively (hopefully) intuitive way.

Each page consists of several entries (one for each language), separated by an empty line. Each entry starts with a term name, prefixed with an ISO 639-1 language identifier separated by a colon (:), followed by an empty line, followed by a definition, followed (immediately) by a source of the definition. Each page can belong to one or more categories – these are expressed by using the usual category mechanism (adding `Category*` link to the end of the page). For the prototype described, there is just one category used, `CategoryCorpusLinguistics`.

Terms in corpus linguistics entered Slavic languages mostly from the English language. The origin of a significant number of them (mostly purely linguistic terms), however, was known long before the differentiation of corpus linguistics as an independent branch of linguistics. These terms have originated either in Greek or Latin: for example, *corpus* and *segment* came from Latin, *lemma* and *lexeme* from Greek. It is even possible that the same term entered different target languages through different intermediaries.

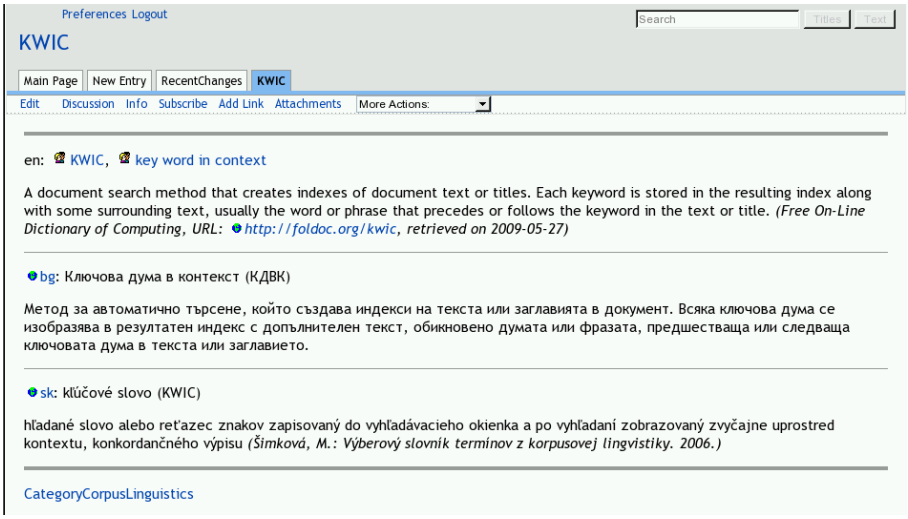

Fig. 1. Example of an entry

A special parser for MoinMoin has been written to display the entries in a distinct graphical way. Main features of the parser are:

- language entries are separated by a horizontal ruler
- ISO 639-1 language identifiers point to an external URL with more information about the language used
- English term is hyperlinked with the corresponding English Wikipedia entry
- definition source is emphasized
- URLs in definitions or sources are automatically recognized

```
en:KWIC, key word in context

A document search method that creates indexes of document text
or titles. Each keyword is stored in the resulting index along
with some surrounding text, usually the word or phrase that
precedes or follows the keyword in the text or title.
(Free On-Line Dictionary of Computing, URL:
http://foldoc.org/kwic, retrieved on 2009-05-27)

bg:Ключова дума в контекст (КДВК)

Метод за автоматично търсене, който създава индекси на текста
или заглавията в документ. Всяка ключова дума се изобразява в
резултатен индекс с допълнителен текст, обикновено думата или
фразата, предшестваща или следваща ключовата дума в текста или
заглавието.

sk:kľúčové slovo (KWIC)

hľadané slovo alebo reťazec znakov zapisovaný do vyhľadávacieho
okienka a po vyhľadaní zobrazovaný zvyčajne uprostred kontextu,
konkordančného výpisu (Šimková, M.: Výberový slovník termínov
z korpusovej lingvistiky. 2006.)


----
CategoryCorpusLinguistics
```

Fig. 2. Internal representation of an entry

```
<entry> ::= <language entry> {<p> <language entry>} \n ---- \n <category>
<language entry> ::= bg | cs | en | pl | ru | sk | sl | uk : <terms>
                     \n {\n} <definition> ( <bibliography> )
<definition> ::= ? characters ?
<bibliography> ::= ? characters ?
<terms> ::= <term> | <term> , <terms>
<term> ::= ? characters ?
<p> ::= \n \n {\n}
<category> ::= Category ? characters ?
```

Fig. 3. Formal description of an entry syntax

The points outlined are implemented in order to make the navigation around the database more efficient – they should be thought of as a visual and formatting aid to the database representation, not as a part of the database itself. In fact, the parser can be very easily modified to accommodate different visual styles and different formatting representations.

The database can use all the usual MoinMoin features concerning efficient collaborative editing. The most relevant ones, emphasised by the database design are:

- efficient indexing and searching, using the built-in Xapian search engine (even if for the database of the intended size – hundreds of entries at most, any search engine is more than sufficient)
- full Unicode support, with only some limitations concerning right-to-left scripts (irrelevant for Slavic languages)
- full editing history with backup of page revisions, allowing to see the complete history of previous entry versions
- review of differences between arbitrary page versions, using diff-like output with colourised differences
- multiuser support with full access control list – however, our database does not use complicated permission schemes, relying on the ease of reverting unwanted changes instead
- warnings to avoid editing conflicts, in case when two users intend to edit the same entry simultaneously

As a prototype, the database has been filled with corpus linguistics entries from [3], which has been compiled as a concise list of term (cf. the needs of colleagues from Czech Republic, where two different lists have been compiled: [1], [7], including data from other areas of linguistics).

We faced following problems when converting the data into MLTD:

- Homonymy:
  - Corpus linguistics is an intradisciplinary research field, where two different areas meet – computer science and linguistics, and these two areas sometimes use the same word to denote (often a little) different objects. Traditional lexicography deals with this polysemy using numbered entries for each meaning (e.g. **corpus 1.** database of digital texts. . . , **2.** collection of texts for a specific kind of research). The Slovak Terminology Database separates the meanings into different entries, with headwords marked by the numeral.
  - Often encountered problem is a dichotomy of meaning of verbal derived nouns, where a noun can mean both a process and its result (e.g *annotation* can be both the process of annotating and the resulting data). In the area of terminology, these two meanings are considered to be strictly separate.
- Traditionally, synonymy in dictionaries is reflected in a lexical entry either in the heading (as two or more equal headwords) or after a definition, while they can form a reference to a relevant entry (e.g. **anotácia – tagovanie – značkovanie**). In the Slovak Terminology Database, synonyms are stored in a separate input field (and are automatically hyperlinked). In the MLTD, different terms have to be kept separately. There is no provision in MLTD for entering synonyms.
- Terminology entries have been often described using encyclopadic style and format – under the general headword there are often specified other, narrow meanings (e.g. **korpus** — **korpus hovorených textov**: elektronická databáza hovorenej formy jazyka; – **korpus písaných textov**: elektronická databáza písanej formy jazyka; — **národný korpus**: jednojazyčný korpus textov konkrétneho národného (jazykového) spoločenstva; — **synchrónny korpus**: korpus jazyka v jeho súčasnej vývinovej fáze; — **všeobecný korpus**: nešpecifický, základný korpus zahŕňajúci široké spektrum jazykových štýlov a žánrov, vecných oblastí (domén), autorských generácií, vydavateľských úzov, regiónov a pod.). However, in the MLTD, each of the meanings has to be entered separately.
- In the Slovak Terminology Database, each term has a facultative field for storing (arbitrary) foreign language equivalents; in the MLTD, the only equivalents are those given in the other languages present.

3 Conclusion

The database is envisaged to contain entries in following languages: Bulgarian, Czech, English, Polish, Russian, Slovak, Slovene and Ukrainian. The English has been added as a semi-bridge language, unifying the entries (and taking into account that most of the terminology originates in the English language).

As a prototype, the database has been filled with corpus linguistics entries from the Slovak from the Slovak Terminology Database, together with their English equivalents (but missing English definitions), and with Bulgarian terms added later. Overall, considering the abovementioned discrepancies in database designs, 45 corpus linguistics terms were imported, out of about 150 terms present in the Slovak Terminology Database.

Bibliography

[1] Čermák, F. (1997). Slovník lingvistických termínů. In *Jazyk a jazykověda*, Prague. Pražská imaginace.

[2] Faber, P. & Sánchez, M. T. (2004). Codifying conceptual information in descriptive terminology management. *Meta*, *46*(1), 192–204.

[3] Šimková, M. (2006). Výberový slovník termínov z korpusovej lingvistiky. In M. Sokolová, M. I. (Ed.), *Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli.*, Prešov. Filozofická fakulta Prešovskej univerzity.

[4] Levická, J. (2007). Terminology and Terminological Activities in the Present-Day Slovakia. In J. Levická, R. G. (Ed.), *Computer Treatment of Slavic and East European Languages*, Brno. Tribun. Proceedings of the conference Slovko 2007.

[5] Levická, J. (2008). Analysis of "classical" and legislative definitions for the term records of the Slovak terminology database. Proceedings of the Third Conference on Translation, Interpreting and Comparative Legi-Linguistics. Poznań, Poland. In print.

[6] Popov, D. (1994). *Bulgarian Explanatory Dictionary.* Sofia: Nauka i Izkuvstvo Publishing House. In Bulgarian.

[7] Šulc, M. (1999). Výběrový slovníček pojmů z lingvistiky. In *Korpusová lingvistika*, Prague. Karolinum.