Design of a New Slovak-Czech Lexical Database*

Radovan Garabík[1] and Jana Špirudová[2]

[1] Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava
[2] Institute of the Czech Language, Academy of Sciences of the Czech Republic, Prague

Abstract. We present an electronic Slovak-Czech lexical database, being compiled with the help of the MoinMoin wiki system. The lexical entry microstructure is organised into a tabular form and special plugins have been written to support easy compiling and editing of entries. Streamlined, traditional-like dictionary entries are then created of the data entered, with the aim to obtain create a printed dictionary.

1 Introduction

Czech and Slovak belong to the West Slavic languages. They have a lot of common in their morphology, phonology, lexicon and syntax. The languages are generally considered to be mutually intelligible.

After the break-up of Czechoslovakia in 1993, sociolinguistic connections between the languages started to weaken, and with the loss of perceptive bilingualism (predominantly on the side of Czech speakers), the mutual intelligibility is no longer universal, however, it is still sufficient for general communication. The situation is highly asymmetrical: while in Slovakia, Czech (both spoken and written) is ubiquitous in the TV, books and other media, in the Czech Republic, presence of Slovak language is rather rare[8]. Slovak speakers have nearly 100% understanding of all the varieties of Czech, but the Czech speakers (especially the younger ones) have sometimes troubles coping with Slovak, in particular with lexical items which are considerably different in the two languages. Consequently, a pressing need for general purpose dictionaries helping the Czech speakers in reading and understanding Slovak texts has emerged.

Ideally, we would like one single dataset to be used to construct all the possible dictionaries, and even a database to be used in all sorts of NLP (e.g. machine translation). This puts additional, often conflicting requirements on the design and building process of the lexical database, and therefore some compromises need to be made.

The primary design goals of the dictionaries to be obtained are:

- to be primarily a passive readers' dictionaries
- to be general purpose, "traditional" middle sized (cca. 20–30 thousand entries) dictionaries, with good coverage of different expressions and false friends
- to contain information on levels of usage

From this it follows that the lexical database has to meet the following requirements:

- to be a web based database with queries performed not just by lemmata, but also by varying wordforms
- to include links into various entry related information (such as morphology paradigm)
- to enable easy, online updating and editing by multiple editors

The last two points can be easily met by a wiki based software. We decided to use the MoinMoin wiki engine, because it supports custom page parsers and plugins that can be tailored to the needs of an online lexical database. On the other hand, MoinMoin full-text search is not really scalable – it is a problem especially concerning the *Category* pages, which internally use the full-text search mechanism. Therefore we refrained from using category pages in the database design.

* 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. The lexical database project has received support from the National Scholarship Programme of the Slovak Republic for the Support of Mobility of Students, PhD. Students, University Teachers and Researchers.

2 Basic structure of the database

Basic building block of the database is an entry, which we call a *page*[3]. A page is used to cover information pertaining to strictly one word meaning, information about homonyms is delegated to the overlying database structure. Each page is uniquely identified by its name, which by convention corresponds to the lemma, or, in case of homonymy, the page name consists of a lemma and a disambiguation identifier (Roman or Arabic numeral).

3 Lexical entry microstructure

Each page (database entry) is kept in a tabular form, where each item (row) has a predefined form and/or content. As an aid for the editors, fields that contain primary linguistic information have a language flag that indicates the language of that field (i.e. either *sk* or *cs*).

3.1 Paradigm (sk)

The *paradigm* field contains an identification of lemma's inflectional paradigm, as used in the morphology database[6]. Since the morphology is also stored in a MoinMoin wiki, the identifier is formatted and displayed as an interwiki link, to allow easy one-click access to the complete word morphology. Since all the word forms are available, the entries do not contain any other inflectional information (traditionally, Czech and Slovak dictionaries contain genitive singular and nominative plural suffixes for nouns, or the 3rd person singular and plural indicative forms for verbs). Similarly, since the paradigm contains a complete morphosyntactic specification including a part of speech category, we do not need to indicate the part of speech separately in the database.

3.2 Translation (cs)

The *translation* field contains direct Czech translation of the Slovak word (or of its particular meaning). We choose the best Czech equivalent. In case there are two or more equally good possibilities, we introduce them all, separated by a semicolon (;). We also take into account etymological relation between the words, and use preferably etymologically related translation[4].

In case there is no direct or indirect Czech equivalent of the Slovak word (e.g., *pahreba*), this field should contain a description of the semantic content.

3.3 Number specification (sk)

This field contains the classification of typical or prevalent number or gender characteristics of the word (for nouns). Possible values are:

- usually plural
- usually masculine or feminine
- masculine or feminine
- feminine or neuter
- feminine, usually plural
- masculine, usually plural
- neuter, usually plural
- exclusively plural
- exclusively singular

[3] Using MoinMoin terminology.

[4] For example, we translate the Slovak word *jazykoveda* by the Czech *jazykověda*, even if we can also translate it by Czech *lingvistika*, and we translate the Slovak word *lingvistika* as *lingvistika*, even if the Czech *jazykověda* would be an equally good translation.

3.4 Qualifier (sk)

This field contains a terminological and/or style qualifier(s), or a special keyword denoting a phrase. The qualifiers are taken out of a fixed set of abbreviated words. When editing this field, the lexicographer is provided with a checkbox entry for each of the qualifiers.

3.5 Gloss 1 & 2

Gloss 1 narrows down the semantics – shade of meaning of the entry word or its semantic and functional equivalent. *Gloss 2* comments on the typical usage of the word.

3.6 Exemplification

The *exemplification* is not a single field, but consists of a variable number of Slovak-Czech exemplification pairs. The Slovak exemplification is primary, the Czech exemplification should be an appropriate translation of the Slovak one. The table displays all the non-empty exemplifications, plus an empty input field for the last Slovak one (to enable the editor to add another exemplification pairs).

3.7 Note

The *note* contains assorted notes for the dictionary user, relevant to the entry. By convention, we use a magic word *viz*[5] to denote a reference to another entry (such as a close synonym, an antonym, comments on significant style characteristics of the Czech equivalents or other related word).

3.8 False friends

This field contains a list of false friends, separated by a semicolon. We do not distinguish between variants of false friends (originating in Slovak or Czech, with a similar meaning, with a completely different meaning...)

3.9 Comment

This field is intended for any other comments by the editors – as such, it will not be displayed in the final entry form.

4 Sense disambiguation mesostructure

There is (intentionally) no place in the entry microstructure to be filled in with hints concerning homonymy disambiguation. We opted to encode this information into the overlaying database nomenclature of entries instead, following to some extent the usual lexicographic classification. At the lowest level, an entry is identified by its headword (MoinMoin page name), which – as its first function – directly encodes the lexeme's lemma. If there are two or more closely related, functionally and pragmatically identical word variants (e.g. spelling variations, such as *mliekar; mliekár*), a headword can contain more variants, separated by a semicolon (;) as a convenient shortcut. This should be thought of as a shorthand for database compilers, nothing more – functionally, such an entry is equivalent to describing both (or more) variants in full.

A headword can have a trailing uppercase Roman numeral, separated by a space. This is used to mark off major homonyms (or even homographs – such as part of speech homonymy, or a completely – even etymologically – unrelated meaning).

An entry can be created as a subpage of an already existing entry, by using MoinMoin's mechanism for subpages. A subpage *XX* of a page *YY* is an ordinary page, with a special name written as *YY/XX* (i.e. the

[5] Czech for *cf.*

subpage name follows the main page, separated by a slash). Subpages of a given page are logically clumped together, in the formatted entry output they are displayed nested with the primary page. We use subpages to connect diminutives, augmentatives and phrasal units to the principal word. Although MoinMoin allows for the whole hierarchy of subpages, we use only the first level subpages in our dataabse (with the exception of sense disambiguation, as outlined the following paragraph).

A headword can have a trailing slash and an Arabic numeral. While technically a subpage, this is used as a weaker variant of a Roman numeral disambiguation in cases, where the words are related and the meaning does not diverge that much. A Roman numeral major disambiguation can be combined with an Arabic numeral minor one (e.g. *čap I/1* – a pivot, journal (mechanical device), *čap I/2* – a hinge, *čap II/1* – a splash, *čap II/2* – a catch (act of catching)).

A headword can contain parenthesized reflexive pronouns *(sa)*, *(si)*[6]. This is used with those cases which are either very frequent, or where the reflexive form diverges in its meaning from the non-reflexive one.

Also, this is used with words which do not have straight one-to-one Czech equivalent, in case the presence of the reflexive does not change the basic meaning and usage of the word (e.g. *dopukat' (sa)* – to crack (about skin)).

5 Technical implementation

The dictionary has been pre-filled with a bilingual glossary of about 60 thousand word pairs[7] and with links into the morphology analyzer wiki, in order to ease the initial editing and to enhance the usefulness of the database by offering at least the first-guess translation and morphology paradigm of the words that would not get into the "core".

A page is internally stored as a flat plain text file (see Fig. 2), with each line corresponding to one table row, with the field name followed by a colon (:), followed by a field value (which can be empty). We have written a special MoinMoin formatter plugin that displays the table in a human-friendly way, together with a final, streamlined formatted entry (Fig. 1). We have also written a MoinMoin action that is used to edit just one specific table row. The action code has hardwired fields that can contain only a fixed set of values (number specification and qualifier) and provides the editor with checkboxes for all the possible values.

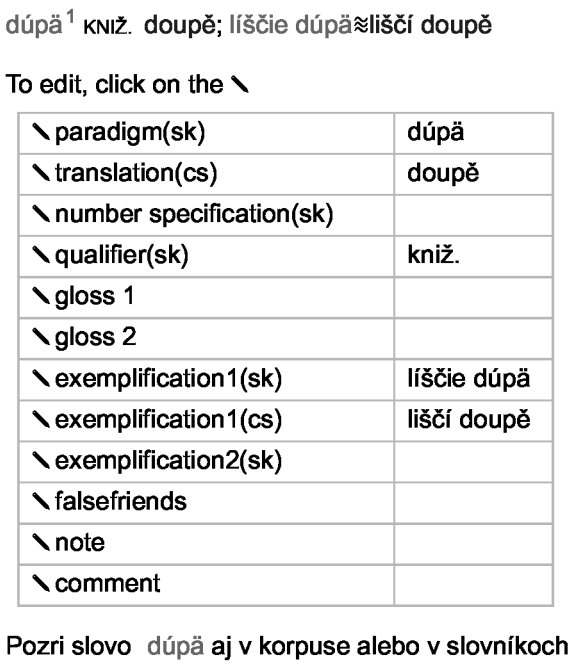
6 Formatted entry output

The tabular format of the dictionary entries displays the information in a clear and obvious way, however it is quite unsuitable for the intended published (paper) dictionary, and there is also the need to present the information in a more compact, concise form also for the internet-based version. Therefore the table is parsed and formatted into a traditionally looking entry.

7 Licensing issues

From the very beginning, we intended to publish the online dictionary entries under an open source/documentation license, in order to facilitate linguistic research and use of data in various NLP applications. The database is publicly accessible and editable under a triple license, GNU Free documentation license v. 1.2 [5] and Creative commons Contribution-Share alike (CC-BY-SA) license v. 3.0 [3] for the use in text document, and under Affero GNU Public license v. 3 [4] for use in computer programs (where by *linking* as specified in the license text we understand any use of the dictionary data by a computer program). This licensing concernes individual entries, while both our institutes keep special rights as a database compiler [1, 2] for the whole dictionary.

[6] Note that *sa* can be added to almost any transitive Slovak (and as *se* to a Czech) verb to express reflexivity, and *si* can be added to almost any verb.



dúpä [1] KNIŽ. doupě; líščie dúpä≈liščí doupě

To edit, click on the ↖

↖ paradigm(sk)	dúpä
↖ translation(cs)	doupě
↖ number specification(sk)	
↖ qualifier(sk)	knuž.
↖ gloss 1	
↖ gloss 2	
↖ exemplification1(sk)	líščie dúpä
↖ exemplification1(cs)	liščí doupě
↖ exemplification2(sk)	
↖ falsefriends	
↖ note	
↖ comment	

Pozri slovo dúpä aj v korpuse alebo v slovníkoch

Fig. 1. An example of a dictionary entry. Final, formatted output is displayed at the top.

```
paradigm (sk): dúpä
translation (cs): doupě
number specification (sk):
qualifier (sk): kniž.
gloss 1:
gloss 2:
exemplification1 (sk): líščie dúpä
exemplification1 (cs): liščí doupě
exemplification2 (sk):
exemplification2 (cs):
exemplification3 (sk):
exemplification3 (cs):
exemplification4 (sk):
exemplification4 (cs):
exemplification5 (sk):
exemplification5 (cs):
false friends:
note:
comment:
```

Fig. 2. Internal representation of a dictionary entry.

References

[1] §72 – §77, 618/2003 Z. z. *Zákon o autorskom práve a právach súvisiacich s autorským právom (autorský zákon), v znení neskorších predpisov*. Copyright Law of the Slovak Republic.

[2] §88 – §94, 121/2000 Sb. *Zákon o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů*. Copyright Law of the Czech Republic.

[3] Creative Commons (2009). Creative Commons Attribution-Share Alike 3.0 Unported. `http://creativecommons.org/licenses/by-sa/3.0/`. [Online; accessed 9 March 2009].

[4] Free Software Foundation (2007). GNU Affero General Public License. `http://www.gnu.org/licenses/agpl.html`. [Online; accessed 9 March 2009].

[5] Free Software Foundation (2008). GNU Free Documentation License. `http://www.gnu.org/licenses/gfdl.html`. [Online; accessed 9 March 2009].

[6] Garabík, R. (2008). Storing morphology information in a wiki. In *Lexicographic Tools and Techniques. MONDILEX First Open Workshhop. Proceedings*, pages 55–59, Moscow, Russia. IITP RAS.

[7] Hajič, J., Kuboň, V., and Hric, J. (2000). Machine Translation of Very Close Languages. In *6th ANLP Conference / 1st NAACL Meeting. Proceedings*, pages 7–12. Seattle, Washington.

[8] Nábělková, M. (2007). Closely Related Languages in Contact: Czech, Slovak, "Czechoslovak". *Small and Large Slavic Languages in Contact. International Journal of the Sociology of Language*, (183):53–73.