Slovak paremiography database*

Peter Ďurčo[1] and Radovan Garabík[2]

[1] Univerzita sv. Cyrila a Metoda v Trnave, Trnava
[2] Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava

Abstract. The article describes the effort to design a wiki-based paremiography database and the process used to extract data from a published dictionary of proverbs. The database is build using MoinMoin engine, offering (standard) possibilities of full text search, categorisation, on-line editing and access control lists. The process used in parsing and correcting the OCRed source is described in detail, and most common sources of errors are discussed.

1 Introduction

A paremiography dictionary (or a database) spreads our lexicographic description of a language into a broader realm of commonly used expressions, and as such, it extends and complements the (better researched and described) dictionaries of idioms.

Such a dictionary is paralleled by a collocation dictionary[2] – proverbs can be seen as a subset of collocations, however, while a proverb is a self-contained, independently functioning language unit, a collocation can be nothing more than almost a random, high frequency sequence of words.

Concerning Slovak language, so far unsurpassed paremiography collection is a compilation by Adolf P. Záturecký [6], first published in 1896. It contains over 10 000 different proverbs (not counting variants). The influence of this work on any subsequent paremiography compilations was immense, since no other collection came even close to the volume of this work, and there was virtually no need to engage in additional field research – following compilations just upgraded and refined selected subsets of Záturecký's collection. The collection itself has been reprinted several times (with the orthography and language progressively converted to ever increasingly modern Slovak, acquiring additional notes and comments), the most recent edition was published as late as in 2006 [7].

The core of the collection is made up of proverbs, sayings and locutions. However, there are also some more indefinite units (pieces of weather-lore, rhymes etc.) as well as other types of phraseologisms (similes, figurative expressions). Although the collection does not record phraseology in its entire extent but concentrates on one type of idioms – proverbs and sayings, i. e. stable sentences. Záturecký divided the entire material into 20 thematic groups (man, one's age, sex, family and home, human body, its needs, disease and death, social circumstances, social classes, status, descent and employment, possession and nourishment, food, clothes, cleanliness and dance, human intellect, general rules of wisdom and carefulness etc.). The collection includes immensely valuable material which is however only insufficiently exploited and explored from the point of view of linguistic theory and interdisciplinary research. Záturecký tried to solve the problem of variability of proverbs. His correspondence with other scholars gives also evidence of his interest in the semantics and etymology of proverbs. Záturecký, together with Dobšinský dealt also with paremiological terminology and they attempted to elaborate optimal taxonomy of thematic concepts. Záturecký combined an alphabetical order of statements within the thematic groups. He also applied the formal criterion of division within particular groups and elaborated the index of key words.

The goal of our effort is to put existing proverbs (not just from Záturecký's collection) into an easily searchable electronic database, to aid further research and to have a ready source of information. Since the Záturecký's collection was available only in printed form, this article deals

* 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.

mostly with specific issues connected with converting its scanned version into a machine readable form, while keeping most of the available information.

2 Implementation

The database has been implemented as a straight, unmodified MoinMoin installation[3]. Since the database is expected to be pre-filled with the data, it will be used mostly in passive mode (searching the data) and the editing will be limited to occasional fixing of typos and OCR errors, we do not need to concern ourselves with designing an additional user-friendly data visualization and/or editing. The database micro- and macrostructure is implemented only in a set of guidelines for the users, concerning article structure and components, while keeping standard MoinMoin syntax (in fact, only a tiny subset of it, to facilitate further automatized article parsing).

The wiki engine is centered on the concept of 'pages' – each page keeps a separate, contained information, is uniquely identified by its name and can optionally belong to one or more categories. Our database maps one (semantic) locution into one wiki page. The page starts with locution variants, separated by an empty lines (visualised as separate paragraphs), followed by an optional comment (currently used to note the locution number in Záturecký's collection, if applicable), followed by a list of categories the locution belongs to (see Tab. 1).

```
<entry> ::= <locution> {<p> <locution>} \n ---- \n [ <comment> ]
            \n ---- \n <category> { <category> }
<locution> ::= ? characters ?
<comment> ::= ? characters ?
<p> ::= \n \n {\n}
<category> ::= Category ? characters ?
```

Table 1. Formal description of an entry syntax

Initially, the core of the database consisted of proverbs from [4], extended by selected proverbs from [3, 5], giving first 2828 entries, then we added Chapter 3 of Záturecký's collection.

3 Deriving a page name

In the first version of the database, we kept the name of the wiki page to be the complete locution, including proper capitalization and punctuation. This has one undisputed advantage – when using the built-in MoinMoin search, searching for a given word will return all the pages that contain the word in their title (Fig. 1).

However, we soon hit several problems:

- There are often several variants of a given proverb. It is desirable to keep all the variants clumped together, and by listing the variants inside one page we would loose the search ability. The situation however could be remedied by choosing one of the variants as the 'main' one and keep the others as redirect pages.
- There is a technical limit for maximum page name length, arising from underlying filesystem limitation – MoinMoin stores the pages as files, with file names being an ASCII quoted version of page name (each unsafe character is stored as its UTF-8 representation in hexadecimal, enclosed in parenthesis). No matter what the actual filesystem limit, traditional Unix file

[3] `http://moinmo.in`

system (used in BSD variants) and the Linux VFS place a hard limit of 255 bytes for the file name length – many proverbs are simply longer than that.[4]
- We need to put the proverb inside the page content as well – repeating the same information in the page name and page content is redundant and prone to errors, and makes automatized manipulation with the data cumbersome (we have to watch two instances of the same information).

In the second iteration of the database, we decided to use different, unique page names. Out of several choices available (numbers, random strings, various transformations of locutions) we decided to choose a 'semantic hash', trying to reduce the locution down to as little words as possible, while keeping a hint of the meaning in the resulting name.

At a first glance, the most obvious thing to do with the locution is to lemmatise its constituent words, to get their basic forms, without ballast of additional grammar information. There are, however, two main problems arising from this approach. First, Slovak exhibits relatively high level of homonymy, so the texts should undergo also morphology disambiguation. That is, unfortunately, inherently imperfect process, with accuracy rarely exceeding 95 %. The problem is exasperated by the fact that morphology disambiguation is usually tuned to 'normal' texts (and therefore is even less precise for our locutions, with their specific kind of language) and the embarrassing nature of a bad lemmatisation – which would completely ruin the meaning of the page name[5]. The second problem is that proverbs are 'semi-frozen' expressions, with the grammar categories often strictly given for a given locution, and lematising would blur the traditional wordforms used and diminish the usefulness of semantic hints the page name could provide. At the end, we decided not to try to lemmatise the page names.

The second most obvious process is to eliminate 'unimportant' words. We keep not only lexical words (such as nouns, verbs, adverbs, adjectives), but also prepositions and two words *sa* and *si*. The (somewhat surprising) presence of preposition is necessitated by not lemmatising the nouns – the case is often governed by prepositions and excluding the preposition would lead to markedly ungrammatical sentences. *Sa* and *si* form (among other possibilities) a part of reflexive verbs, and leaving out an obligatory reflexive marker would again emphasise ungrammaticality.

To keep the page names short, we include at most two words that are either noun or verb (with the exception of forms of verbs *mať*, *byť* and *jesť*[6]). If there is a locution to be added to the database and the derived page name already exists, we keep adding another (lexical) words until the page name is unambiguous. This algorithm has an advantageous side effect: it quite reliably detects duplicates that differ mostly in function words.

4 Structure of the original text

Záturecký's collection is divided into 20 chapters (19 in the edition [7]), each concerning certain aspect of society or language. Locutions in each chapter are numbered, starting with number 1, with every 5$^{\text{th}}$ entry marked at the left text margin. Different typeface (and a smaller font size) is used for literature and references – this style is recognised by the OCR software as italics. Locution variants are (somewhat unfortunately) separated by a dash surrounded by spaces: ' – '.

Each chapter is accompanied by a comments section, containing further explanation of the locutions, often including Hungarian, German, Polish or Latin equivalents, or explanation of dated terms and expressions, otherwise incomprehensible for a contemporary reader (these comments were often not written by Záturecký himself, but by later editors). The comments are numbered by the number of the locution they refer to.

[4] Note that the new storage backed system being prepared for MoinMoin v. 2.0 is going to lift these limitations, since the pages will no longer be necessarily stored as corresponding single files in the filesystem.

[5] Given especially the two meaning of the word *mať*, 'mother' and 'to have', words that thanks to their nature occur very frequently in proverbs.

[6] 'to eat', 3$^{\text{rd}}$ person singular *je* is homonymous with the same categories of the verb *byť*, 'to be'

Fig. 1. Searching for a word 'slovo' in all the page titles, in the old version of the database. Later, we have abandoned the idea of using complete locutions as page names.

Fig. 2. Fulltext search for a word 'slovo'.

5 Scanning and conversion

Záturecký's collection has been scanned and submitted to OCR using the ABBYY Finereader software[7]. As a result we obtained document in Microsoft Word (97–2003) format, including distinct typographic styles, as present in the original scanned text. As the text style differences encode useful information, it is desirable to keep this information (it can also aid in correct parsing), therefore we needed to convert the document into some kind of a structured, easily readable format. An example of such a format is the ubiquitous XML, and since the OpenOffice uses natively OpenDocument format (XML based) [1], an obvious approach would be to open the Microsoft Word file in the OpenOffice Writer and save it as the OpenDocument text, which can be further parsed using standard XML parsing tools and libraries. Unfortunately, the conversion does not deal well with character styles – the resulting document contains 235 common styles, majority of them used to encode the same basic paragraph style. In order to parse the document, we would have to infer the original text style for each of the OpenOffice character styles, which would be a rather time consuming process.

A better approach was to convert the file into HTML, which marks the text styles into either different `FONT` tags, distinguished by the `size` attribute, or `B` and `I` tags for boldface and italics, respectively. It is worth pointing out, that due to OCR errors and imperfections, the styles are not always recognised correctly, and it has to be taken into account during parsing. To parse the HTML file, we have chosen the BeautifulSoup framework[8], which is a Python HTML/XML parser with easy API and emphasis on parsing damaged or invalid HTML/XML files. Although the HTML file we obtained is perfectly valid, in the course of processing we split the file at some specific points, sometimes obtaining HTML chunks with unpaired tags, where we conveniently use BeautifulSoup's abilities to deal with invalid HTML.

Since it is highly desirable to link comments to the proper locutions, we needed to keep correct numbering of the entries. We start by splitting the HTML file at positively identified locution index numbers – numerals dividable by 5, in italics and using smaller font size. We take into account only monotonically increasing numbers, since we have to realise that due to OCR errors, there are going to be gaps and other errors in the sequence (OCR is good in finding out sequences of numeric characters, therefore errors caused by replacing the digit '1' by lowercase letter 'l' occur only with standalone characters, rarely in multi-digit numerals, however errors produced by replacing '1' with '7', '8' with '6' or '9' and vice versa are common), and we disregard numbers that differ from previous index by more than 40 (since that is probably an OCR error).

Once we have these HTML chunks with the range of entry numbers for each of them, we split them by additional present entry index numbers (if their span is more than 5 indices), this time regardless of their visual style (which is often incorrect). With these smaller chunks, we already used up all the information that there was about the entry numbers, and we had to apply some heuristics to split the remaining locutions correctly. We conveniently used the ability of BeautifulSoup to parse invalid HTML and fed the chunks back into the parser. This time we looped through the `P` tags and separated texts from the paragraphs (which correspond to locutions). Since the OCR does not find ends of paragraphs very reliably, in the next step we joined back the sentences where the paragraph started with a lower case letter. Then we tried to find out sentence boundaries and split the text on them (fixing the cases where the OCR ignored paragraph break). The heuristics is simple – a sentence boundary is where the previous character is one of full stop, exclamation mark or question mark, and the next character is an uppercase letter. This fails in some cases (e.g. single expression *Koho bili? Petra. A kto sa bil? Peter.* will be incorrectly separated into 4 locutions), but overall significantly improves the segmentation. Then a plain text file with numbered locutions had been produces. We then apply Procrustean bed method to keep the locutions properly numbered: if the number of automatically segmented locutions is smaller than expected according to index numbers, we put dummy sentences consisting of a single character (we have chosen '@') into the file; if the number is bigger, we join excessive ones, again using character '@' as a separator.

[7] `http://www.abbyy.ru/finereader/`
[8] `http://www.crummy.com/software/BeautifulSoup/`

After the segmentation, entry numbers were manually corrected. The process consisted basically of finding all the occurrences of the '@' character and splitting or joining the lines as required. In Chapter 3, out of 1405 locutions, there were 227 incorrectly numbered ones (each one of the incorrectly parsed locutions is counted twice, first at its original number, second time the for the locution number it replaced). One of then was caused by genuinely dropped number in the printed source, 168 occurrences were caused by transposed pages in the OCRed text – that gives only 4.2 % genuine error rate, with a very quick manual correction (we have to stress here that we were not proofreading the text and fixing OCR typos, just fixing the numbering of entries).

Fixed text is then parsed again and converted into internal MoinMoin structure. For the comments, we conveniently used the subpages MoinMoin mechanism – if present for a given locution, each comment is been put into a subpage named `/poznámka`[9], with a link from the parent (locution) page. Since many of the comments were written by subsequent editors of Záturecký's collection, their copyright protection has not expired yet, and we cannot make them freely available. We used the MoinMoin's possibility to use ACL to block public access to these subpages. In the Chapter 3, there were 253 comments present.

The locutions are categorised – there is a category for the Chapter 3 locutions[10], a category for the 'core' proverbs[11] and a category for comments to Záturecký's locutions[12]. Given page can belong to more than one category, as a matter of fact, many of the core proverbs also belong to `CategoryZátureckýPomerySpoločenské`.

6 Conclusion

Presented database is intended to serve as an easily reachable source of paremiography data. To test the concept, Chapter 3 of Záturecký's collection [6, 7] has been scanned, OCRed and converted to the database, together with a few thousand other selected proverbs. The conversion process is mostly automatic, with minimal (though still substantial) human intervention, and will be used to convert remaining chapters of the collection.

[9] i.e. comment
[10] `CategoryZátureckýPomerySpoločenské`
[11] `CategoryCore`
[12] `CategoryZátureckéhoPoznámky`

References

[1] ISO/IEC 26300:2006 (2006). *Information technology – Open Document Format for Office Applications (OpenDocument) v1.0.* Geneva: International Organization for Standardization.
[2] Majchráková, D. & Ďurčo, P. (2009). Compiling the First Electronic Dictionary of Slovak Collocations. To be published.
[3] Miko, F. et al. (1989). *Frazeológia v škole.* Bratislava: Slovenské pedagogické nakladateľstvo.
[4] Mlacek, J. & Profantová, Z. (1996). *Slovenské príslovia a porekadlá, zv. 1–2. Výber zo zbierky A. P. Zátureckého.* Bratislava: Nestor.
[5] Smiešková, E. (1988). *Malý frazeologický slovník.* Bratislava: Slovenské pedagogické nakladateľstvo.
[6] Záturecký, A. P. (1896). *Slovenská přísloví, pořekadla a úsloví.* Praha: Česká akademie věd.
[7] Záturecký, A. P. (2006). *Slovenské príslovia, porekadlá, úslovia a hádanky.* Bratislava: Slovenský Tatran.