Handbook of Slovak Nouns

Radovan Garabík
Ľ. Štúr Institute of Linguistics
Slovak Academy of Sciences, Bratislava
garabik@kassiopeia.juls.savba.sk

1 Introduction

1.1 Dictionary interface

Dictionary interface[1] presented at the pages of the Slovak National Corpus Ľ. Štúr Institute of Linguistics (JÚĽŠ) is a very popular webpage providing access to some of the most important Slovak language dictionaries, together with several other non-dictionary databases. Since the first version available in 2003 (and serving as an interface to the Short Dictionary of the Slovak Language) it grew up to include 14 public resources (and several hidden ones, available only internally in JÚĽŠ).

Internally, the interface uses dict (RFC 2229) server[2] as a backend, with a CGI frontend for queries and formatting. There are several search strategies possible: match exact words, match prefix, match suffix, match substring, regular expression.

1.2 Handbook of Slovak Nouns

The dictionary conceived as an interface for entries from Slovak morphology database, with examples taken from the corpus – a sort of a corpus interface turned inside out.

The dictionary can be classified (by classification described in [Klo13]) by its access as and online dictionary and by its features as first published as online dictionary, (permanently) under construction, (planned) hypertextualization, with interaction with the user, dictionary with text (and rudimentary charts), with access via search option.

The Figure 1 displays a screenshot of the dictionary interface. Various user interface elements are annotated with alphabetical indices in Comic Sans MS font[3]:

a) The headword.
b) Short description of grammatical categories. The description in the picture says *masculine, animate, singular, substantive paradigm*.
c) Corpus the example comes from and the logarithm of Average Reduced Frequency [HR99], displayed as a bar consisting of '+' characters, the number of characters corresponds to the order of magnitude (i. e. logarithm) of the word form frequency. Note that the plural forms have the frequency lower by two or three orders of magnitude.
d) Case number. In Slovak, the cases are often customarily numbered: 1=Nominative, 2=Genitive etc.
e) Typical 'priming' context for the word form. This includes either prepositions, numerals (for nominative, which does not have a typical preposition to bind with), a verb *vidím* ("I see") for the accusative, 2nd person pronoun for vocative.
f) Word form from morphological database.

[1] http://slovniky.korpus.sk/
[2] http://www.dict.org/
[3] http://bancomicsans.com/main/?page_id=7

g) Context extracted from the corpus.
h) Word from from the corpus. Sometimes different from f).
i) Popup with information about relative frequency (instances per million words).

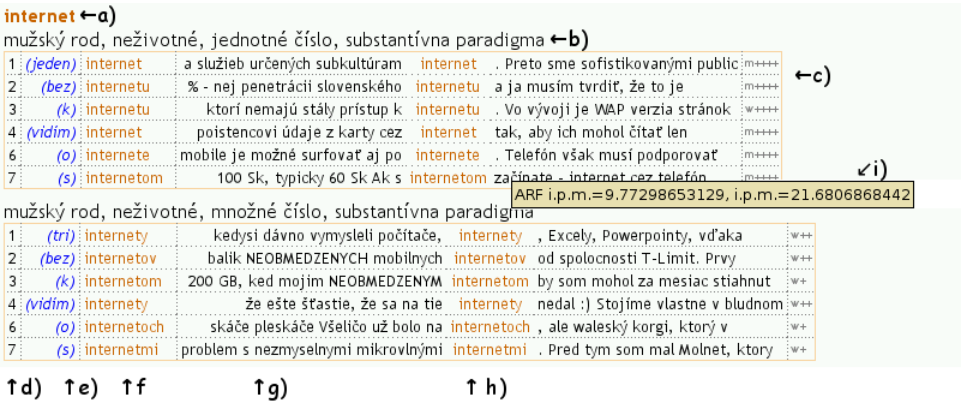

Figure 1: Screenshot of the interface, querying the word *internet*, the results for singular and plural[4].

2 Lexicographical workflow

The motive for the dictionary was to exploit existing rich data in two resources: Slovak Morphological Database and Slovak National Corpus. The whole project is somewhat an experiment in digital lexicography – not even the ultimate goals are fixed, the 'official' goals (provide interface to the morphological database) were already met and exceeded in the zeroth version. We adhere to the principle known from OpenSource community "release early, release often" and thus the database improvements are reflected at the public interface continuously, with just a minimal necessary amount of testing.

As such, the process does not fit into an orthodox lexicographical workflow at all, as the whole idea of the database is somewhat experimental. In the following, we will try to match the description by [Klo13] in order to facilitate comparing with other lexical databases/dictionaries, even if it is not really applicable to our case.

2.1 Preparation

As for the organizational plan, the database was compiled in the Slovak National Corpus department in an attempt to make available its database of morphological inflections to the users. Thus, the idea arose at the time when there were enough data and their incorporation in an on-line dictionary is well in line with Slovak National Corpus activities.

The conceptional design was however not completely set up at the beginning of the work – about the only design requirement was that the interface display the complete paradigms of nouns, with (optional) corpus examples. Before publishing the first version, we experimented with several UI concepts, but these were limited mostly to design issues, not major conceptual changes.

[4] Yes, the plural exists.

2.2 Data acquisition

The dictionary builds upon two primary sources: morphological database (where it takes the list of headwords from) and several of Slovak language corpora. This means that the data already existed well before the work on the dictionary started; on the other hand, the data is continuously being improved and upgraded (the morphological database gradually and incrementally, the corpora in discrete releases). The dictionary reflects these changes.

2.3 Computerisation & Data processing

Computerisation and data processing (these phases cannot be reliably distinguished here) is not very relevant in this case – the corpora, the data, the corpus manager had already been existing before the work began. The only issue was in testing two different approaches to extraction of corpus examples – pre-generated versus live API access. Pre-generated database had the advantage of fast access and not being dependent on corpus manager running, but took a lot of space (several gigabytes) and a long time to generate ($\sim$ one week of CPU time) – on the order of the corpus itself, for obvious reasons. It was also too rigid, necessitating regeneration in case anything in the data or structure changed. Therefore we finally decided to use live access via NoSketchEngine[5] API[6]. The examples are drawn from several corpora, in this order: 1. manually lemmatized and morphologically annotated corpus (because the cases are presumably correct here); 2. corpus of Slovak Wikipédia and Necyklopédia[7] (for copyright reasons) and 3. the corpus *omnia-2.0* [8] – a deduplicated union of all the available Slovak texts ('main' written corpora + web corpus).

2.4 Data analysis

The initial phase of data analysis was tantamount to the previous two phases. However, any new feature added to the dictionary (so far, the first one was addition of live corpus API and the second addition of relative frequencies of wordforms) required analysis (and careful testing).

2.5 Preparation for online release

Since the dictionary has been online from the beginning, this phase was parallel to the other ones.

2.6 Afterlife

The work, obviously, won't stop at nouns – this is more like a testbead for future planned versions including other significant parts of speech. However, before that happens, there are features that we plan yet to implement. These include:

- Better indication of frequency of given word form (relative to lemma). At the time of writing the relative frequency is given with respect to the corpus the example comes from, therefore it is not really comparable across different words.
- Link to corpus manager. This is subject to finding a way of allowing access to the corpus without compulsory registration and as such might not be feasible due to copyright concerns.
- Indication of (lexical and inflectional) homonymy – at least a list of homonyms with their POS tags.

[5] `http://nlp.fi.muni.cz/trac/noske`

[6] This included finding a bug in the API preventing its use; this has been reported and fixed upstream.

[7] `http://korpus.juls.savba.sk/wiki.html`; equivalents of English language Wikipedia and Uncyclopedia

[8] `http://korpus.juls.savba.sk/omnia.html`

The medium term goals include extracting a subset of the dictionary (several thousands of the most frequent nouns, together with 'problematic' less frequent ones), manually checking the examples (with emphasis on homonymy) and releasing it as a handbook for students of Slovak as a foreign language. We are also planning to add sentences not only from written corpora, but also from the Corpus of Spoken Slovak, with corresponding audio files, but first we have to evaluate possible effects on user experience.

3 Time span of the different phases

Since the 'canonical' timeline of lexicographical work is not really relevant here (the phases had been carried out either long before the project started, or are being worked upon continuously while the dictionary is being improved), we decided to display the timeline of implemented or proposed changes instead – see Figure 2.

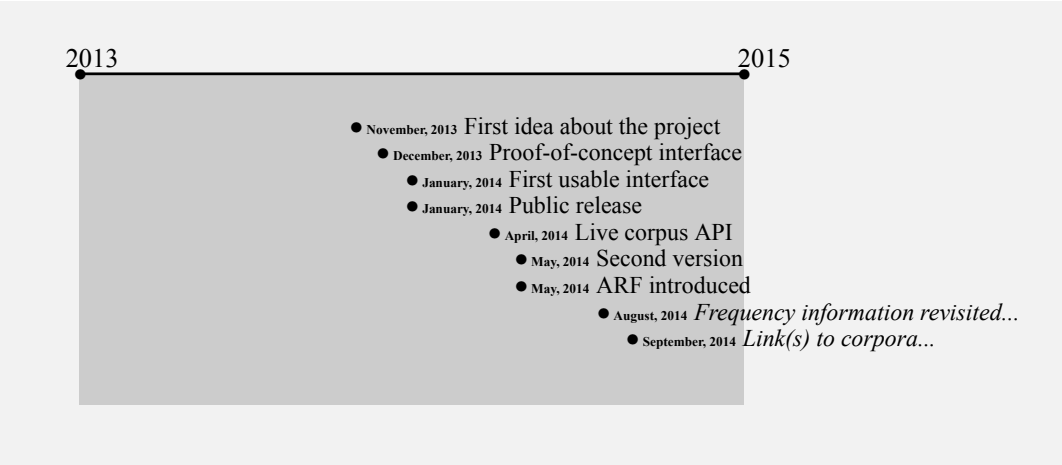

Figure 2: Timeline of the dictionary construction. Planned additions are in italics.

References

[Gv12] Radovan Garabík and Mária Šimková. Slovak Morphosyntactic Tagset. *Journal of Language Modelling*, 0(1):41–63, 2012.

[HR99] Jaroslava Hlaváčová and Pavel Rychlý. Dispersion of words in a language corpus. In *Text, Speech and Dialogue*, pages 321–324. Springer, 1999.

[Klo13] Annette Klosa. The lexicographical process (with special focus on online dictionaries). In Rufus H. Gouws, Ulrich Heid, Wolfgang Schweickard, and Herberst Ernst Wiegand, editors, *Dictionaries. An international Encyclopedia of Lexicography. Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography*, pages 517–524. de Gruyter, Berlin, Boston, 2013.