Computer(ized) linguistic resources at the Ľ. Štúr Institute of Linguistics

Radovan Garabík

Jazykovedný ústav Ľ. Štúra SAV

813 64 Bratislava, Slovakia

korpus@korpus.juls.savba.sk, http://korpus.juls.savba.sk

Abstract. Ľ. Štúr Institute of Linguistics, as the main linguistic institution in the Slovak Republic, publishes a lot of information either in the form of dictionaries, scientific publications or periodicals, targeted both to scientific community and general audience. With increasing amount of information produced, information retrieval and effective searching can be greatly assisted by organizing and indexing electronic versions of the documents. As a public service, the institute offers some of the data to be accessed and searched on the Internet.

Introduction

Ľ. Štúr Institute of Linguistics is the main linguistic institution in the Slovak Republic. Among its main areas of research and activities is the research of contemporary Slovak language, Slovak dialectology, study of history of the Slovak language, and also connections to general and comparative linguistics. Recently, the institute started to work in the fields of etymology and corpus linguistics. The sociolinguistic and culturology approaches are gaining momentum across various research topics. The institute has a long tradition of producing dictionaries either for academic or general environment, linguistic atlases and specialized publications presenting different aspects of linguistic research. “Flagships” of institute's publication activities are three specialized journals:

- *Kultúra slova* is a scientific popular journal, serving mostly as an outpost for prescriptivism activities, promoting the usage of cultural, codified literary Slovak language, introducing new professional terminology into common usage and providing general discussion of language culture in both professional literature and fiction.
- *Slovenská reč* is the oldest scientific journal focused on slovakistics. It is the base scientific journal dedicated to a generic research of the Slovak language, including grammar, morphology, syntax, onomastics, theoretical description of orthography, but also dialectology and phraseology. The journal serves as a place for professional discussion for Slovak linguists, brings information about oncoming events in Slovak linguistics, informs about new important publications and publishes concise lists of bibliographies of Slovak linguists.
- *Jazykovedný časopis* is a scientific journal oriented towards general linguistics research, not limited to Slovak language, bringing also reviews of important linguistic publications and informations about current events in linguistics.

Rather important publication is a tetralogy *Slovenskí jazykovedci – Súborná personálna bibliografia slovenských slovakistov a slavistov*, containing detailed list of publications of Slovak linguists, covering the years 1925 – 2000, with fifth volume in preparation [1].

Two publications targeted towards general public audience of Slovak language users are:

- *Krátky slovník slovenského jazyka* (KSSJ) – *the* dictionary of contemporary Slovak language, containing definitions of 63818 headwords [2].
- *Pravidlá slovenského pravopisu* (PSP) – detailed description (definition) of Slovak orthography, together with dictionary-like section describing (defining) inflectional paradigms of most common Slovak words [3].

Importance of these two publications stems from the Slovak language law, which defines codified version of Slovak language as that described in above mentioned publications.

Dictionary interface

KSSJ is by far the most demanded publication, because non-linguistically oriented people see the language codification mostly through the prism of "proper" and "improper" vocabulary items. Searchable version of KSSJ is accessible on-line at the address http://kssj.juls.savba.sk/ Backend is based on homegrown version of DICT server[4] software (modified[1] serpento server, written in Python). Frontend is a simple cgi script, with a possibility to either directly find the definition of the word entered, or to search for entries beginning with given prefix, suffix or containing given substring. The maximum number of obtained results is limited to 30 (for various reasons, most important one being the fear of unauthorized copying of the dictionary data). If the query has been given without diacritics (i.e. containing only 7-bit ASCII characters), the search is performed on the index version built with diacritics removed – this is for the benefit of users who either cannot enter 8-bit characters in proper encoding (such as older or embedded browsers), or if entering proper diacritics is cumbersome (e.g. systems without Slovak keyboard layout), and also taking into account widespread custom to drop diacritics in internet context (reflecting the past era. when using non ASCII characters was often met with technical problems and encoding incompatibilities). The output – i.e. entry definition – is with full diacritics in any case. The script works fully in UTF-8 encoding, but it is possible to force the input and output encoding to be one of ISO 8859-1, ISO 8859-2,

1 In fact, this will become next version of serpento in due time

CP1250, CP1252, CP850, CP852, to facilitate the access with older or restricted browsers.
Similar interface is planned for the dictionary part of PSP, once all the necessary copyright issues are resolved.

Publishing non dictionary information

There are many different ways how to effectively publish information on the WWW. Ideally, the system should not put much demand on system and human resources, should be easy to install and use, information stored inside should be easily modified. Plain static html files provide by far the best performance, they however fail short in terms of easy modification (that is, preferably by untrained users). In addition, an ability to provide an easy feedback from the users about the information is desirable.

Relatively recent phenomenon is a wiki collaboration software, a system that allows users to edit the information presented via a WWW interface[5]. Overwhelming success of a wiki-based collaborative approach can be seen on the development of Wikipedia[6].

Our system accessible from the institute's main WWW page (http://www.juls.savba.sk) uses wiki engine written in the Python programming language, using the *Karrigell* application framework[2]. Thanks to the user access control list support, it is easy to implement the possibility to modify existing data for selected users from the institute, and have a safe read-only access from anonymous users from the Internet. While at the first glance most of the data included is frozen (coming from already published printed information), there is a need to modify the data occasionally – especially the texts obtained from scanned printed sources, with scanning error remaining even after proofreading (the number of errors approaches zero as the amount of proofreading man-hours approaches infinity, and it is often desirable to make the information available immediately after the first basic proofreading, with further corrections carried out while the data are already accessible). As the texts would be edited mostly by untrained persons (often students), the classic wikipedia syntax is deemed too difficult, with a steep learning curve. Therefore we opted for a much simpler and clearer reStructuredText syntax[7], backed up by an excellent *docutils* parsing module[8]. Effectively, from the point of view of proofreaders, the text data is stored as a plain text, with reStructuredText markers perceived only as a decoration of the text.

2 http://karrigell.sourceforge.net

MÁRIA ČIŽMÁROVÁ, rod. DZIAKOVÁ **————————————** **Nar. 22. 6. 1969, Hrabová Roztoka, okr. Snina. Študovala v r. 1978 – 1983 na Filozofickej fakulte v Prešove Univerzity P. J. Šafárika v Košiciach (ukrajinský jazyk – slovenský jazyk). 1993 PhDr., 1995 CSc. V r. 1983 – 1985 pôsobila na Základnej škole v Humennom (učiteľka), v r. 1985 – 1997 pracovníčka Katedry ukrajinského jazyka a literatúry Filozofickej fakulty v Prešove Univerzity P. J. Šafárika v Košiciach, od r. 1997 Prešovskej univerzity (odborná pracovníčka, odborná asistentka).** **Venuje sa konfrontačnému výskumu ukrajinčiny a slovenčiny, pracuje aj v oblasti dialektológie.** **1996** **——** **1. Konfrontačný výskum frazeologizmov v ukrajinčine a slovenčine. – In: Studia Philologica. 3. Acta Facultatis Paedagogicae Universitatis Šafarikanae. Red. Z. Stanislavová et al. Prešov, Univerzita P. J. Šafárika v Košiciach, Pedagogická fakulta v Prešove 1996, s. 143 – 145, nem. res. s. 145.** **2. Do vlasnoho bereha. – Dukľa, 44, 1996, č. 2, s. 104 – 105.** **1997** **——** **3. Uvaha botaničnij terminolohiji. – Dukľa, 45, 1997, č. 2, s. 110 – 111.**	**MÁRIA ČIŽMÁROVÁ, rod. DZIAKOVÁ** Nar. 22. 6. 1969, Hrabová Roztoka, okr. Snina. Študovala v r. 1978 – 1983 na Filozofickej fakulte v Prešove Univerzity P. J. Šafárika v Košiciach (ukrajinský jazyk – slovenský jazyk). 1993 PhDr., 1995 CSc. V r. 1983 – 1985 pôsobila na Základnej škole v Humennom (učiteľka), v r. 1985 – 1997 pracovníčka Katedry ukrajinského jazyka a literatúry Filozofickej fakulty v Prešove Univerzity P. J. Šafárika v Košiciach, od r. 1997 Prešovskej univerzity (odborná pracovníčka, odborná asistentka). Venuje sa konfrontačnému výskumu ukrajinčiny a slovenčiny, pracuje aj v oblasti dialektológie. **1996** 1. Konfrontačný výskum frazeologizmov v ukrajinčine a slovenčine. – In: Studia Philologica. 3. Acta Facultatis Paedagogicae Universitatis Šafarikanae. Red. Z. Stanislavová et al. Prešov, Univerzita P. J. Šafárika v Košiciach, Pedagogická fakulta v Prešove 1996, s. 143 – 145, nem. res. s. 145. 2. Do vlasnoho bereha. – Dukľa, 44, 1996, č. 2, s. 104 – 105. **1997** 3. Uvaha botaničnij terminolohiji. – Dukľa, 45, 1997, č. 2, s. 110 – 111.

Table 1: Example of reStructuredText format (left) and the final rendered output (right)

Fulltext search

In addition to providing access to "raw" information and texts, we need to effectively search in the data. Fulltext search is provided by a Xapian library (http://www.xapian.org). As already mentioned above, users often enter queries without diacritics, even if it is otherwise available. Therefore we indexed the texts with stripped diacritics, and the query interface strips it from the query before passing it over to xapian – obtaining the effect of ignoring diacritics in the query system. In the future, our plans envisage to use a simple Slovak language stemmer when indexing the texts. In this context, lemmatization is less suitable then stemming, because of the need to disambiguate the lemmatized texts, while stemming (often) does not care about preserving grammar information, thus increasing recall at the expense of precision.

What is included

Since the first volume of [1] has been typeset before the onset of computer typography, using traditional methods[3], it had to be scanned and proofread, the other three volumes were readily available as digital texts. The text was split into small units, each unit corresponding to one author mentioned in the volume, and then converted into reStructuredText format and placed into the wiki. All the four volumes have the size of 7.9 MB and contain together 1.17 million words.

At the time of writing, the archives of *Kultúra slova* from 1994 and *Slovenská reč* from 1997 onwards are processed and made available (with the exception of the latest year, which is available only to the institute's employees, due to copyright reasons). The total amount of texts in these two journals is 17 MB, containing 2.28 million words.

Slovak National Corpus

Separate from the resources mentioned above, Slovak National Corpus is a project to build an annotated representative corpus of contemporary written Slovak language. At the time of writing, it contains about 256 million words, covering the language as published during 1955 – 2005 (obviously, with a huge bias towards the later date). The texts are automatically morphologically annotated and lemmatized, containing detailed bibliographic annotation. Smaller, balanced subcorpus containing 60% journalistic texts, 30% professional (scientific) texts and 30% fiction has the size of 56 million words. Another subcorpus contains manually lemmatized and annotated texts that could be used to train morphological analysers and other NLP tools. Size of this subcorpus is 322 600 words.

Access to the whole corpus and all the subcorpora (excluding texts with too prohibitive copyright licenses) is publicly available free of charge on internet[4], requiring a simple registration procedure.

Literature

1. Dvonč, L.: Slovenskí jazykovedci – Súborná personálna bibliografia slovenských slovakistov a slavistov, Veda, Bratislava, 1987, 1997, 1998, 2003.
2. Krátky slovník slovenského jazyka.Veda, Bratislava, 2003.
3. Pravidlá slovenského pravopisu. Veda, Bratislava, 2000.
4. Faith, R., Martin, B.: “A Dictionary Server Protocol”, Request for Comments 2229, Network Working Group, October 1997, ftp://ftp.isi.edu/in-notes/rfc2229.txt

3 i.e. phototypesetting

4 http://korpus.juls.savba.sk

5. Ebersbach, A, Glaser, M.: Towards Emancipatory Use of a Medium: The Wiki. In: International Journal of Information Ethics, 2 (2004).
6. Ciffolilli, A.: Phantom authority, self-selective recruitment and retention of members in virtual communities: The case of Wikipedia. In: First Monday, 12 (2003).
7. Goodger, D.: Python Enhancement Proposal 287: reStructuredText Docstring Format, http://www.python.org/peps/pep-0287.html
8. Goodger, D.: Python Enhancement Proposal 258: Docutils Design Specification, http://www.python.org/peps/pep-0258.html