Slovak National Corpus tools and resources

Radovan Garabík

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

Abstract. The article presents current state of affairs in several projects conducted by the Slovak National Corpus department of the L'. Štúr Institute of Linguistics, Slovak Academy of Sciences. We describe the Slovak National Corpus, Corpus of Spoken Slovak, tools used for linguistics analysis and an ongoing effort to create Slovak WordNet.

1 Slovak National Corpus

The Slovak National Corpus is a huge, representative corpus of modern written Slovak (since the 1953 orthography reform). Currently, the whole corpus contains over 700 million tokens. There are several specialised subcorpora (fiction, professional texts, journalistic texts, original Slovak fiction, balanced subcorpus, texts written until 1989). The corpus is automatically lemmatised and morphologically annotated and is indexed using the *Manatee* software [Ryc00]. To query the corpus, there are two possibilities – first, the users can use multiplatform (Tcl/Tk) *Bonito* client to access the *Manatee* server, using its own protocol. This approach provides the users with complete access to all the advanced querying, sorting and statistical features of the server, however requires installation of a specialized software. The other possibility is to use web based access, where only basic features are present. In both cases, the search interface provides CQL compatible query syntax.

However, in the last few years the ability of an average user to install arbitrary software (and use anything that is not web-based) declined considerably, and new corpus users often face an insurmountable obstacle in downloading, unpacking and running the *Bonito* client. Because of this, we are considering transfer of the corpus to *Manatee-2*, which provides complete web-based interface as a replacement of the Tcl/Tk client.

A separate corpus (although part of the whole Slovak National Corpus project) is a manually morphologically annotated corpus, whose main purpose is to be a source of train data for Slovak language tagger (and, to a lesser extent, for morphology annotation tools).

The size of the Slovak National Corpus source archives is 46 GB, however, a substantial percentage of this are original scan images (when converted into raw XML text, the size is about 6 GB uncompressed).

<u>M</u> anažér Korpus <u>H</u> ľadanie <u>K</u> onkordancia	Zobrazenie Výber	r <u>P</u> omocník			
Nové hľadanie 🥏	▪ meno:	rim-4.0-snk-all -			
účasné trendy rozvoja informatiky Plander I., vodňových situáciách Hluchý L., Petrovič P., diskrétnych výrobných systémov Kogel J., in Distributed Systems DUC THAI NGUYEN, počítačových systémov Kotmanová D., vied mal vraj dva platy NKÚ tvrdí, že šéf Ladislavom Hluchým, CSc., riaditeľom ckej a aplikovanej mechaniky. Dnes je z neho riaditeľom Ing. Ladislavom Hluchým, CSc. Baltazár Frankovič, DrSc., Drh., c. z stribuovaného spracovania informácií. Šéfom na úroveň miestnosti. V tom istom projekte aktivitou. Na čo je zameraná v súčasnosť ?- Defence, ZTS Elektronika, Corinex Group a necistalný tv VRUP Jozef Mikulec. Riaditeľ únii, ' hovorí Kavala. Podľa riaditeľ únii, v tovorí Kavala. Podľa riaditeľ únii, v tovorí Kavala. Podľa riaditeľ únii d v VURUP Jozef Mikulec. Riaditeľ platilo zamestnancom odmeny neoprávnene.	Ústav informatiky Ústav informatiky	y SAV , Bratislava Účasť Slovenska v 5. a y SAV , Bratislava n Paralelná sekcia : Vzdi y SAV , Bratislava Genetic code application y SAV , Bratislava Genetic code application y SAV , Bratislava Exchange 5. 5 alebo Ex- y SAV Automatizácia , kybernetika , umelá vy SAV Automatizácia , kybernetika , umelá y SAV Automatizácia , kybernetika , umelá y SAV Automatizácia , kybernetika , umelá y SAV a za vedecko - výskumnú orientáciu n: ky SAV za vedecko - výskumnú orientáciu n: ky SAV za vedecko - výskumnú orientáciu n: y SAV a báze gridov simuloval predpovec y SAV , ktorý sa teraz podieľa na piatich gri y SAV Ladislav Hluchý . Pracoviská musia i ky SAV Ladislav Hluchý . Pracoviská musia i ky SAV (2004) - riadi lokanžite orienti y SAV (2004) - riadi lokanžite orienti			
N					
doc.cong Hospodárske noviny 04/05					
doc.date 2004-05-06					
doc.orgd 2004-05-06					
doc.name Slovensko bude participovať na prieskumnom systéme 🥢					
Zobrazené: všetko/19 Riadok: 14 Vybrané:	1				

Fig. 1. Screenshot of Bonito client

2 Corpus of Spoken Slovak

Corpus of Spoken Slovak is a project to record reasonable amount of sound samples of contemporary Slovak, together with their manual phonemic transcription, automatic lemmatisation and morphosyntactic analysis. At the time of writing, the corpus contains about 160 hours of sound recordings, corresponding to 1.2 million tokens. Since the transcription is done manually (no reasonably accurate transcription software exists), the remaining task of morphosyntactic analysis is exactly the same as with the Slovak National Corpus texts.

The archive is kept in FLAC format, and we convert the whole recordings into Ogg/Vorbis and Ogg/Speex formats (for easier handling and transcription) and for the final linking through the corpus web interface we split the files into small chunks corresponding to dialogue turns. The source archive size is currently over 200 GB.

One of our primary goals was to make this corpus unencumbered by usual copyright and privacy concerns that plague similar projects. We have to take care not only of copyright law, but also the law on protection of personal data [Ná05]. We do this by removing any sensitive information (e.g. personal names) before including the recordings in the archive, and by including only those recordings where we have explicite expression of consent by all the relevant participants to include the recordings in our archive.

For transcription, we are using the *transcriber* software [BGWL01], with a detailed set of tags to annotate both internal speech features and external sound events influencing the recorded discourse.

Access to the corpus can be performed in two (almost independent) ways. One of them uses standard *Bonito* client, in the same way as the preferred access to the main Slovak National Corpus. Each token provides following attributes: *pron, lemma, tag, dcount. pron* is the transcribed pronunciation, *lemma* and *tag* come from the standard automatic morphosyntactic annotation, *dcount* is the possible number of lemma-tag pairs.

The other way to access the corpus is to use specialized web interface, offering addional visual representation of transcription and annotation, as well as links to sound recordings themselves.

🔳 🛃 John Bzbyku: " a poslal 🔳 ho späť a v leopoldove zomrel . " 🔳 🕹 Luosem Bzbyku: áno . 🔳 🗜 aj som to jeho pan"i na myjave ra:z povedav , 🔳 ked" tam bola n"ejaká | schô:dza . 🔳 🔳 🕹 lebo tam bol nejaká 🔳 pan"i , n"eviemm* 🔳 jako sa 🔳 🔳 menuvala . Bzbyku+02 ⊥ tá mala príkaz z ústredného III alebo III jakého výboru okresného , či jakého III polievat" ■ kyselinu ■ mohylu . ■ ■ 1 tak išli ■ a v@:* nedeľu/n"ed"el"u vypátral"i t"i 🔳 🔳 l"udia: 🔳 a p 🔳 dopravili ju 🔳 až z čiech tam . 🔳 🧎 a tam plakala , 🔳 abych to n"ejak 🔳 nazval , 🔳 jako stará ščetka , no . 🔳 🧎 že 🔳 ona za to n"emo:že , 🔳 tak 🔳 sme asi za to mohli my , no , 🔳 čo 🔳 sme boli v krimináli . " 🔳 👗 John Bzbyku: " a:* ona kde bývala ? 🔳 🚶 na myjave bývala , alebo že som 🔳 🧎 ra:z sa ma dva partizáni 🔳 pýtali , 🔳 z myjavy , 🔳 🖉 🖉 🖉 🖉 🖉 🖉 , a ten tvoj svoker , už som 🔳 boll* o 🔳 ženatý , 🔳 n"emajú dvestopadesátpet"ku ? 🔳 🔳 🛔 a čo ja 🔳 reku viem čo ? 🔳 🛓 načo by im to bolo ? 🔳 🔳 🚛 🏌 no n"ie , ale 🔳 to im patrí , lebo , 🔳 🔳 ten martin hovori , ja: som bol ran″ený 🔳 a ležal som u ich bratranca a on″i mi nosili jedlo . 🔳 🛓 lebo 🔳 tí chod"ili do sveta , 🔳 aj pan"i , 🔳 aj starý pán 🔳 🔳 a oni 🔳 Bzbyku+04 ma tam chovali. ■ ■ 1 d ■ reku , ■ tak ■ n°e:kedy prícte tam ■ k nám a , ■ v nedeln/n°ed'el"u , ked" budú doma , ■ ■ no , tak že dobre , aj ti ■ povieme . ■ 1 tak raz mi napísal"i , že prídu v n"ed"el"u 🔳 🔳 poobed"e . 🏽 🗮 🕹 no , a 🖄 svoker chudák , 🔳 no , pýtajú sa ho , pán 🔳 , mát"e 🔳 dvestopad"esátpet"ku ? 🔳 🔳 🛔 n"emám* . 🔳 🔳 🛔 to majú druhí . 🛃 a 🛛 ten@ 🔲 on"i skrátka 💭 @hovoria , 👘 a hovorí , prosim vás pekn"e , čo keď sa to n"e:kedy buď e muset" vracať ? 🗉] [: 🖷 § 🕹 Luosem Bzbyku: "je:žišmaria , 🔳 ja: som ostal jak tá scena bledy . 🔳 🔳 🔳 上 že , reku 🔳 tak 🔳 som ešte 🔳 na: tom mal cenu , reku tak vidiš , 🔳 no , 🔳 som t"i

Fig. 2. Screenshot of Corpus of Spoken Slovak web interface

```
<Event desc="poz" type="noise" extent="instantaneous"/>
niekedy/n"e:kedy prídťe/prícte tam
<Event desc="pi" type="pronounce" extent="instantaneous"/>
k nám a,
<Event desc="poz" type="noise" extent="instantaneous"/>
v nedeľu/n"ed"el"u, keď/ked" budú doma,
<Event desc="poz" type="noise" extent="instantaneous"/>
<Event desc="mm" type="noise" extent="instantaneous"/>
no, tak že dobre, aj ti
<Event desc="pi" type="pronounce" extent="instantaneous"/>
povieme.
<Sync time="755.906"/>
```

Table 1. Example of annotation of Corpus of Spoken Slovak trancriptions

3 Linguistic analysis

The foundation of all subsequent analysis is assignment of unique lemma and tag combination to all the words in the analysed text (e.g. in our corpus). This is realised as a two stage process, first stage is morphosyntactic analysis, i.e assignment of all the possible lemma-tag pairs to a given token. Second stage is disambiguation – selection of one (correct) lemma-tag pair for a given word. We collected semi-automatically complete paradigms for 74 000 lemmata[Gar06] and stored manually verified and into a wiki-based database[Gar08]. The database contains complete paradigms, with an exception for third person plural of L-participle, where we keep only tag for general gender (*všeobecný rod*, tag 'h'), since the forms of all the other genders are identical, and the paradigm is then automatically expanded to cover all the existing genders with corresponding tags. The morphological analysis then consists from looking up all the possible tags and lemmata for a given word form, and from guessing possible lemmata and tags for words not present in the database.

== Lema ==		
mať		
== Paradigma		
VIe+: mať		
VKesa+: mám		
VKesb+: máš		
VKesc+: má		
VKepa+: máme		
VKepb+: máte		
VKepc+: majú		
VMesb+: maj		
VMepa+: majm	e	
VMepb+: majt	e	
VHe+: majúc		
VLesam+: mal		
VLesaf+: mal:	a	
VLesan+: mal	o	
VLepah+: mal:	1	
== Homonymia	==	
[[mať]]		
KategóriaVerb	á	

Table 2. Paradigm of the verb mat'

3.1 Guessing

Quite an important part of the analysis is assigning a lemma-tag pair to words that are not present in the morphological database. While a reliable determining of lemma, part of speech and morphological tag when given an unknown word is impossible, it is nevertheless desirable to obtain at least some information about those words. E.g. even if we guess lemma incorrectly, getting at least correct part of speech will help in eventual subsequent syntactic annotation. Our guessing is based on suffix similarity – first, during the training phase, we build an array of suffices of existing wordforms. We use fixed length of 3 characters (determined empirically). During the guessing phase,

if the unknown word starts with a capital letter and is not situated at the beginning of a sentence, it is assumed to be a noun or a adjective (most common parts of speech for proper names), otherwise it could be also a verb, participle, adverb or a numeral. Special provision is implemented for potential adjectives beginning with the prefix *naj*-and verbs beginning with the prefix *ne*- (for superlatives and negated verbs).

3.2 Disambiguation

The second step is disambiguation, where each word is assigned a unique lemma and a morphosyntactic tag out of the possibilities assigned in the first step. For disambiguation, we use *morče*, an averaged perceptron model originally used for the Czech language tagging [SHRS09], re-trained on the Slovak manually annotated corpus.

<s></s>			
Po	ро	Eu6	04
chvíli	chvíľa	SSfs6	02
ste	byť	VKepb+	02
zistili	zistiť	VLdpbh+	07
,	,	Z	01
že	že	0	02
to	to	PFns1	05
nejde	nejsť	VKesc-	01
	•	Z	01

 Table 3. Example of an automatically morphosyntactically tagged sentence from the Slovak

 National Corpus

4 WordNet

There is currently an ongoing effort in collaboration with Technical University of Košice in building a basic Slovak WordNet database. We plan to use the database as a skeleton of a basic English-Slovak-German-Polish-Lithuanian dictionary¹. The building process consists of mapping automatically generated Slovak synsets to English synsets from WordNet v.3.0. The synset generation has been described in [Gen09]; the synsets are manually corrected before being added to the database. We use special annotation to mark synsets that do not have clear English equivalent. Our goal is to build synsets containing ten thousand most frequent words from the Slovak National Corpus (nouns, adjectives, verbs and adverbs), together with a complete set of their hypernyms (i.e. each Slovak synset will have a hypernym, unless mapped to those few English synsets that do not have a hypernym).

¹ As part of the Slovak Online (Lifelong Learning Programme DG EAC/31/08) project.

POS	synsets	%
noun	4669	51.6
verb ^a	1895	21.0
adjective	2265	25.0
adverbs	214	2.4

^a Negated verbs are not in the database.

Table 4. POS Composition of Slovak Wordnet Database

References

- [BGWL01] Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. Transcriber: Development and use of a tool for assisting speech corpora production. Speech Communication, 33(1–2):5–22, 2001.
 - [Gar06] Radovan Garabík. Slovak morphology analyzer based on Levenshtein edit operations. In Michal Laclavík, Ivana Budinská, and Ladislav Hluchý, editors, 1st Workshop on Intelligent and Knowledge oriented Technologies, pages 2 – 5, Bratislava, 2006. Institute of Informatics, Slovak Academy of Sciences.
 - [Gar08] Radovan Garabík. Storing morphology information in a wiki. In Olga Shemanayeva, editor, *Lexicographic tools and techniques*, pages 55 59, Moscow, 2008. IITP RAS.
 - [Gen09] Ján Genči. Synset Building Based on Online Resources. In Jana Levická and Radovan Garabík, editors, *NLP, Corpus Linguistics, Corpus Based Grammar Research*, Brno, 2009. Tribun.
 - [Ná05] Národná rada Slovenskej republiky. Zákon č. 428/2002 Z. z. o ochrane osobných údajov Z. z. v znení zákona č. 602/2003 Z. z., zákona č. 576/2004
 Z. z. a zákona č. 90/2005 Z. z. Zbierka zákonov Slovenskej republiky, Bratislava, Slovakia, 2002, 2004, 2005.
 - [Ryc00] Pavel Rychlý. *Korpusové manažery a jejich efektivní implementace*. PhD thesis, Faculty of Informatics, Masaryk University, Brno, 2000.
 - [SHRS09] Drahomíra Spoustová, Jan Hajič, Jan Raab, and Miroslav Spousta. Semisupervised training for the averaged perceptron POS tagger. In EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pages 763–771, Morristown, NJ, USA, 2009. Association for Computational Linguistics.