Slovenský národný korpus (2002 – 2012): východiská, ciele a výsledky pre výskum a prax

Mária Šimková, Radovan Garabík

Jazykovedný ústav Ľ. Štúra SAV, Bratislava
korpus@korpus.sk

Rozvoj jazykových technológií a zdrojov na Slovensku a vo svete
Bratislava, 7. 6. 2012 – 8. 6. 2012

Co-funded by the 7th Framework Programme of the European Commission through the contract T4ME, grant agreement no.: 249119.

Co-funded by the ICT PSP Programme of the European Commission through the contract CESAR, grant agreement no.: 271022.

Východiská

- **Vonkajšie faktory**
  - **technologizácia a informatizácia**
  - **rozvoj korpusov a korpusovej lingvistiky**
  - **prístupové konania do Európskej únie**
- **Vnútorné faktory**
  - **tvorba nového výkladového slovníka**
  - **koncepcia starostlivosti o štátny jazyk**

Východiská

- interný korpus textov slovenského jazyka v JÚĽŠ SAV
  - základné počítačové vybavenie
  - voľne dostupné programy
  - texty bez licencie – cca 25 mil. jednotiek
  - 3 prac. s čiastkovou kapacitou

- snahy o podporu v grantovej agentúre VEGA
- zapájanie sa do medzinárodných projektov

Východiská

- **2000 – 2001 vypracovanie a podanie projektov:**
  - ***Budovanie Národného korpusu slovenského jazyka***
  - ***Elektronizácia jazykovedného výskumu***
- **MK SR, MŠ SR, SAV**
- **uznesenie vlády SR č. 137 z 13. 2. 2002**

Ciele

- čo najviac čo najkvalitnejších dát
- potreby používateľov a možnosti oddelenia
- vývoj korpusov a korpusovej lingvistiky

- primárne vedecko-výskumné a učebné využitie (lingvistické/lexikografické)

Ciele

- všeobecný jednojazyčný korpus
- databáza lexikografických diel
- terminologická databáza
- korpus diachrónnych textov
- korpus nárečových textov
- paralelné korpusy
- korpus hovorenej slovenčiny

Jazykové zdroje

- prim-5.0 (720 mil.) + podkorpusy + skweb
- Slovenská terminologická databáza
- databáza lingvistických zdrojov
- http://slovniky.korpus.sk/
- http://www.juls.savba.sk/ediela/

Využitie

- národné výskumné projekty
- 500 registrovaných používateľov ročne, 40 000 dopytov denne
- slovníky (výkladový, kolokačný, prekladové), monografie, štúdie, postupové práce

Medzinárodné projekty

- **Mondilex (FP7)**
  - **Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources**
- **Slovak Online (Lifelong Learning)**
- **EuroMatrixPlus (FP7)**
  - **Bringing Machine Translation for European Languages to the User**
- **CESAR (ICT PSP)**
  - **CEntral and South-east europeAn Resources**
- **MAD: SAV + БАН**
  - **Elektronický korpus – konfrontačná štúdia so zameraním na návrh bulharsko-slovenských elektronických jazykových zdrojov**

Pôvodné východiská a ciele sa vzhľadom na súvislosti s týmto mierne posunuli (vzhľadom na podobné zahraničné projekty, zdroje a nástroje)

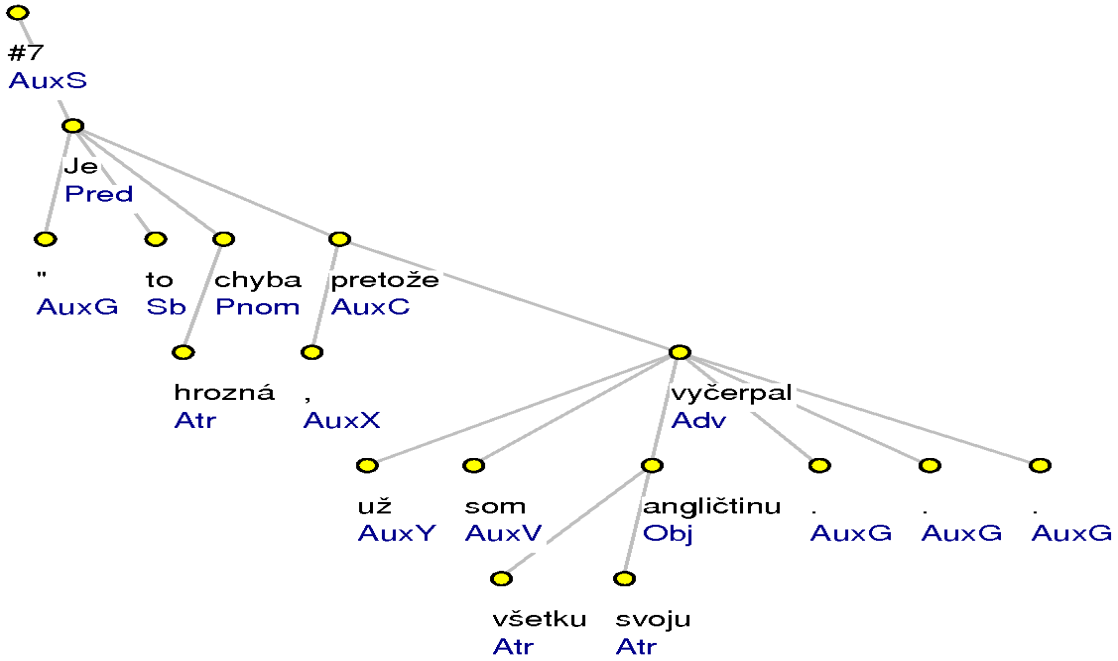
Jazykové zdroje

- **Morfologická databáza** META-SHARE
  - **pokrýva základnú slovnú zásobu jazyka (96000 lem, 1.15 ·10⁶ tvarov, 3.25·10⁶ záznamov)**
- **Slovenský morfologický (morfosyntaktický) tagset**

- **MULTEXT East (Mondilex)**
  - **MULTEXT East tagset**
  - **1984**

Jazykové zdroje

- **Syntakticky anotovaný korpus (EuroMatrixPlus)**
  - **Kompatibilný s PDT1.0 (analytická rovina)**
  - **50 tisíc viet, 2× nezávisle anotované**



Jazykové zdroje

- Paralelný slovensko-český korpus (EuroMatrix+)
- Paralelný slovensko-anglický korpus (EuroMatrix+)
- Paralelný slovensko-ruský korpus
- Paralelný slovensko-bulharský korpus (MAD: JÚĽŠ+ИМИ БАН)
- Paralelný slovensko-francúzsky korpus
- Paralelný slovensko-latinský korpus

Jazykové zdroje

4457	Obraz v zrkadle sa strácal . „ Somár ! Čo si urobil s tou topánkou ? “ skríkol otec , ktorý sa strácal v dyme , stúpajúcom z podrážky .	Rozplýval se i obraz v zrcadle . „ **Blbe** ! Cos udělal s tou botou ? “ křikl tatínek , který zmizel v kouři , stoupajícím z podrážky .
1682	Nikto tomu nerozumel , ani tučný básnik ; čítal moje básne prasačími očkami a kričal , ty sviňa prekliata , kde si toto nabral ? A potom sa šiel ožrať na oslavu poézie a plakal : pozrite na toho ťulpasa , to je básnik ! Taký potmehúd , a čo vie napísať !	Nikdo tomu nerozuměl , ani tlustý básník ne ; četl mé básně prasečíma očkama a křičel , ty svině zatracená , kdes tohle vzal ? A pak se šel ožrat na oslavu poezie a plakal : koukejte se na toho **blba** , to je básník ! Takový tichošlápek , a co dovede napsat !
8193	Fakt je jeden , ak tu chce niekto ukázať , že je schopný niečo urobiť , šancí je neúrekom . Bohužiaľ , nemožno sa vyhovárať , že šéf je hlúpy a šéfuje len preto , lebo je v partaji . Hádam preto je všade ako - tak a doma najlepšie . “	Fakt je ten , pokud tu chce někdo dokázat , že je schopen něčeho dosáhnout , šancí je nad hlavu . Bohužel , nelze se vymlouvat , že šéf je **blb** a šéfuje jen proto , že je v partaji . Myslím , že proto je všude dobře , a doma nejlíp . “
9189	ZASIAHNUTIE DOBRÉHO VOJAKA ŠVEJKA DO SVETOVEJ VOJNY „ Zabili nám teda Ferdinanda , “ povedala posluhovačka pánu Švejkovi , ktorý pred rokmi , keď ho vojenská lekárska komisia definitívne vyhlásila za hlupáka , opustil vojenskú službu a živil	ZASÁHNUTÍ DOBRÉHO VOJÁKA ŠVEJKA DO SVĚTOVÉ VÁLKY " Tak nám zabili Ferdinanda , " řekla posluhovačka panu Švejkovi , který opustiv před léty vojenskou službu , když byl definitivně prohlášen vojenskou lékařskou komisí za **blba** , živil se

Jazykové zdroje

- WordNet (Slovak Online)
- Web korpus
- Korpus právnych textov
- Slovenský hovorený korpus

Perspektívy

- rozširovanie a skvalitňovanie zdrojov
- nárečový korpus
- historický korpus
- špecializované slovníky

Ďakujem(e) za pozornosť