Словацкий национальный корпус

Radovan Garabík

Jazykovedný ústav Ľ. Štúra
Panská 26
813 64 Bratislava 1
Slovakia

e-mail: korpus@korpus.juls.savba.sk
www: http://korpus.juls.savba.sk

Принципы и цели корпуса

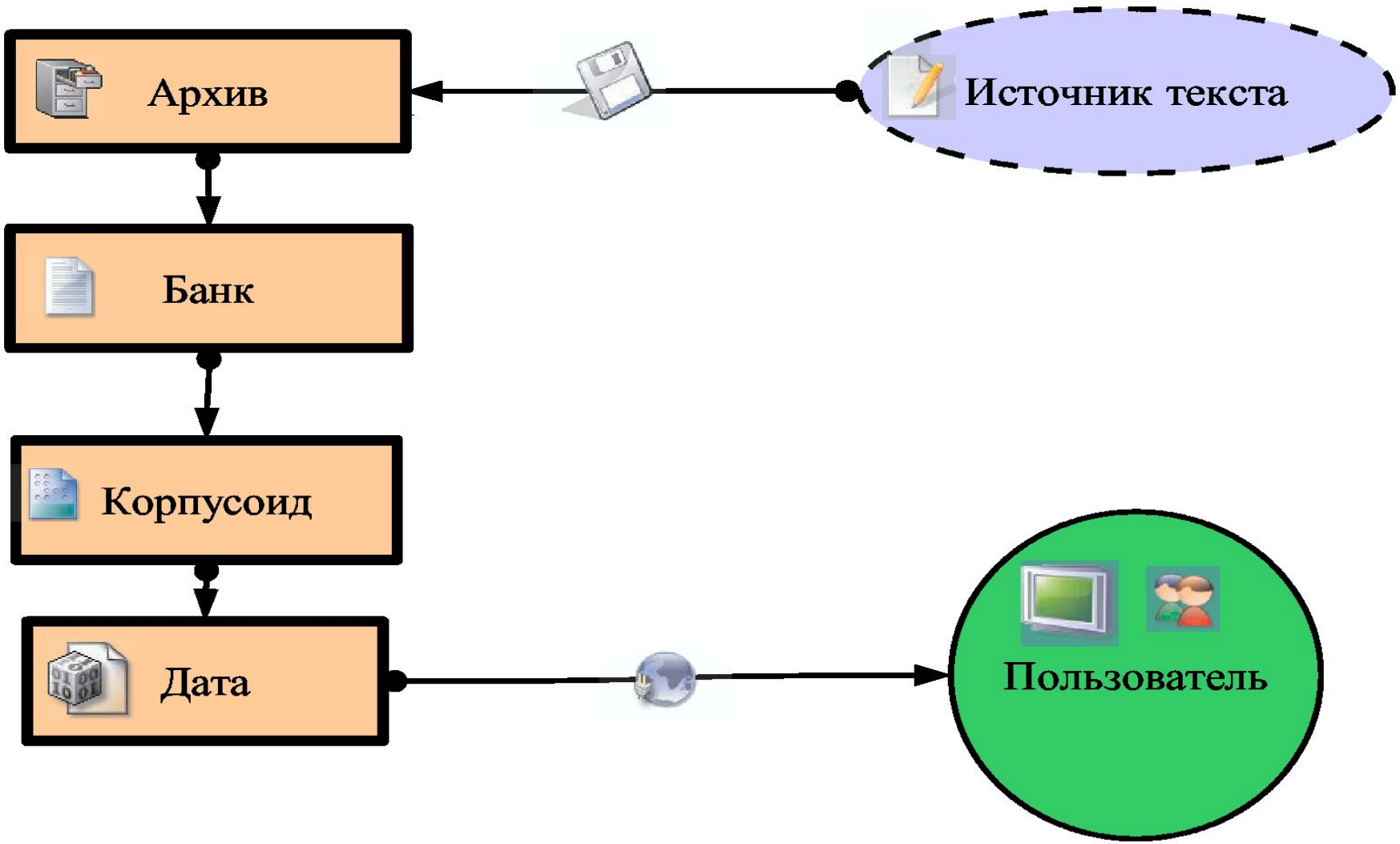
- корпус текстов современного словацкого языка (1955 – 2005)
- лингвистическая разметка (лемматизация, морфология)
- представительный корпус
- библиографическая аннотация
- доступен

- до 1989 – ничего
- 1989 – 1991 – первые разговоры о необходимости корпуса
- 1991 – Текстовой корпус словацкого языка
  - без лингвистической обработки
  - лексикография
- 2002 – проект Словацкого национального корпуса
- половина 2003 – 30 млн слов
- конец 2003 – 110 млн слов
- начало 2004 – лемматизация, морфосинтаксическая разметка
- **сегодня – 192 млн слов**
- *конец 2004 – сбалансированный корпус 30 млн слов*
- *2005 – параллельные корпусы*
- *2006 – представительный корпус 200 млн слов*

Структура данных в корпусе

- архив
  - документы в оригинальном виде
- банк
  - общий формат текстов (почти XML)
- корпусоид
  - стандарт XCES (TEI), лингвистическая разметка
- дата
  - цифровой формат для корпусного менеджера

```
<tok>
  <orth>meč</orth>
  <disamb>
    <base>meč</base>
    <ctag>SSis1</ctag>
  </disamb>
  <lex>
    <base>meč</base>
    <ctag>SSis1</ctag>
  </lex>
  <lex>
    <base>mečať</base>
    <ctag>VMesb+</ctag>
  </lex>
</tok>
```


Архив
Источник текста
Банк
Корпусоид
Дата
Пользователь

Поиск в корпусе

Система Manatee (http://www.textforge.cz)

Поиск слова, фразы

Регулярные выражения

Поиск по лемме, тэгу

Частотный анализ

Колокации

MI-score, T-score

Элементы структурной разметки

Библиографическая информация & стиль, жанр текста

Лингвистическая разметка

лемма

морфология

синтаксис?

морфосинтаксический тэг

1.автоматически

- существующие инструменты
    - Hajič & Hric (Прага)
    - Ajka (Брно)
- ~95% аккуратность

2.вручную

- небольшое количество текстов
- цель: несколько сто тысяч слов
- «ядро» корпуса

Система морфосинтаксических тэгов

строго морфологический принцип (неужели?)

один знак ⇔ одна грамматическая категория

знаки по определённой очереди

знаки (почти) не повторяются

„čas“ SSis1 – *S-существительное имя, S-существительный парадигм, i-неодушевлённый мужской род, s-единственное число, 1-именительный падеж*

1.		2.		3.		4.		5.		6.		7.	
S	**Substantives**		**Paradigm**		**Gender**		**Number**		**Case**				
		S	substantive	**m**	masc. animate	**s**	singular	**1**	nominative				
		A	adjective	**i**	masc. inanimate	**p**	plural	**2**	genitive				
		F	mixed	**f**	feminine	**o**	unknown	**3**	dative				
		U	incomplete	**n**	neutrum			**4**	accusative				
								5	vocative				
								6	locative				
								7	instrumental				
								o	unspecified				
A	**Adjectives**		**Paradigm**		**Gender**		**Number**		**Case**		**Grade**		
		A	adjective	**m**	masc. animate	**s**	singular	**1**	nominative	**x**	positive/irrelevant		
		F	mixed	**i**	masc. inanimate	**p**	plural	**2**	genitive	**y**	comparative		
		U	incomplete	**f**	feminine	**o**	unknown	**3**	dative	**z**	superlative		
				n	neutrum			**4**	accusative				
				o	unspecified			**5**	vocative				
								6	locative				
								7	instrumental				
								o	unspecified				
P	**Pronouns**		**Paradigm**		**Gender**		**Number**		**Case**		**Agglutinated**		
		S	substantive	**m**	masc. animate	**s**	singular	**1**	nominative	**g**	agglutinated		
		A	adjective	**i**	masc. inanimate	**p**	plural	**2**	genitive				
		P	pronoun	**f**	feminine	**o**	unknown	**3**	dative				
		F	mixed	**n**	neutrum			**4**	accusative				
		U	incomplete	**o**	unspecified			**5**	vocative				
		D	adverbial	**h**	general			**6**	locative				
								7	instrumental				
								o	unspecified				
N	**Numerals**		**Paradigm**		**Gender**		**Number**		**Case**				
		S	substantive	**m**	masc. animate	**s**	singular	**1**	nominative				
		A	adjective	**i**	masc. inanimate	**p**	plural	**2**	genitive				
		N	numeral	**f**	feminine	**o**	unknown	**3**	dative				
		F	mixed	**n**	neutrum			**4**	accusative				
		U	incomplete	**o**	unspecified			**5**	vocative				
		D	adverbial					**6**	locative				
		X	solitaire use					**7**	instrumental				
								o	unspecified				

1.		2.		3.		4.		5.		6.		7.	
V	**Verbs**		**Form**		**Aspect**		**Number**		**Person**		**Gender**		**Negation**
		I	infinitive	**d**	perfective	**s**	singular	**a**	first	**m**	masc. animate	+	affirmative
		K	indicative	**e**	imperfective	**p**	plural	**b**	second	**i**	masc. inanimate	–	negative
		M	imperative	**j**	ambivalent			**c**	third	**f**	feminine		
		H	transgressive							**n**	neutrum		
		L	*l*-participle							**o**	unspecified		
		B	futurum							**h**	general		
G	**Participles**		**Type**		**Gender**		**Number**		**Case**		**Grade**		
		k	active	**m**	masc. animate	**s**	singular	**1**	nominative	**x**	positive/irrelevant		
		t	passive	**i**	masc. inanimate	**p**	plural	**2**	genitive	**y**	comparative		
				f	feminine	**o**	unknown	**3**	dative	**z**	superlative		
				n	neutrum			**4**	accusative				
				o	unspecified			**5**	vocative				
								6	locative				
								7	instrumental				
								o	unspecified				
D	**Adverbs**		**Grade**										
		x	positive/irrelevant										
		y	comparative										
		z	superlative										
E	**Prepositions**		**Form**		**Binds with**								
		v	vocalised	**2**	genitive								
		u	non-vocalised	**3**	dative								
				4	accusative								
				6	locative								
				7	instrumental								
				o	unspecified								
O	**Conjunctions**	**Y**	contains conditional morpheme *by*										
T	**Particles**	**Y**	contains conditional morpheme *by*										

J	**Interjection**	**#**	**Not a word**
R	**Reflexive particle/pronoun *sa, si***	**%**	**Citation element (e.g. foreign language word)**
Y	**Morpheme *by***	**0**	**Digits**
Z	**Punctuation**	**:r**	**Proper noun**
W	**Abbreviation**	**:q**	**Incorrect spelling**
Q	**Unknown POS type**		