Korpusová lingvistika

Literatura:

Blatná, R. - Čermák, F. (eds.): **Jak využívat Český národní korpus**. Nakladatelství Lidové noviny, Praha 2005.
Čermák F. - Klímová J. - Petkevič V. (eds.): **Studie z korpusové lingvistiky**. Karolinum, Praha 2000.
Čermák, F. - Křen, M. (eds.): **Frekvenční slovník češtiny**. Nakladatelství Lidové noviny, Praha 2004.
Čermák, F. – Blatná, R.: **Korpusová lingvistika: Stav a modelové přístupy**. NLN, Praha 2006.

Základní pojmy:

korpus

korpusová lingvistika

vlastnosti korpusu: označkovanost, reprezentativnost

typy korpusů: psané, mluvené, synchronní, diachronní, historické, paralelní

korpusy národních jazyků: BROWN, BANK OF ENGLISH,
ČESKÝ NÁRODNÍ KORPUS apod.

Korpus je soubor počítačově uložených textů, který slouží k jazykovému výzkumu. Vyhledáváme v něm pomocí vyhledávacího programu např. slova a slovní spojení v kontextu, frekvenci slov, tvary slov, textové zdroje.

Korpusová lingvistika - disciplína, která zkoumá jazyk pomocí elektronických jazykových korpusů. Zabývá se také výstavbou těchto korpusů, jejich zpracováním a metodologií. Jako vědecký obor se začala korpusová lingvistika rozvíjet v posledních dvou desetiletích 20. století v souvislosti s rozvojem výpočetní techniky, i když některé malé korpusy (asi 1 milion slov) existovaly už dříve, např. *Brown Corpus* (1964), na jehož tvorbě se podílel i lingvista českého původu Henry Kučera.

Přínos korpusového přístupu k jazyku

korpusová gramatika oproti gramatice nekorpusové (příkladové):

- poskytuje podstatně lépe podložená jazyková data,
- poskytuje frekvenční a statistické charakteristiky jazykových dat, z jejichž analýzy můžeme určit jevy typické (centrální) a jevy okrajové (periferní),
- upřesňuje nebo opravuje některá tvrzení v gramatikách.

Vlastnosti korpusu:

1. Označkovanost – lemmatizace a tagování (morfologické značky)

Grafická podoba jednoho konkordančního řádku korpusu SYN2000:

kód identifikující jednoznačně text *(J. Durych, Kouzelný kočár, Torst, 1995)*
rok vydání
typ textu *(próza, román)*
typ korpusu *(SYN2000)*

<doc S|SCI|1999|ikaros99>*Počátek roku 1998 byl v Ústřední knihovně věnován všem činnostem a* <**pracem**/práce/N.FP3> , *které souvisely*

vyhledávaný výraz *(KWIC)*
lemma *(zákl. tvar)*
tag *(gramatické značky)*

2. Reprezentativnost korpusu

Reprezentativnost je často diskutovaná vlastnost korpusu. Můžeme ji chápat tak, že korpus obsahuje všechny centrální a většinu periferních gramatických jevů, které se vyskytují v textech a promluvách dané řeči. Vybudovat korpus naprosto všestranný je vyloučené.

Korpusy národních jazyků

Elektronické textové korpusy se postupně budovaly od 60. let 20. stol. v USA a v Evropě. K nejstarším korpusům 60. let patří korpus **BROWN**, vytvořený v USA.

První velké korpusy v Evropě vznikly ve Velké Británii. Dnes patří k největším britským korpusům: **Bank of English** (více než 500 milionů slovních tvarů) nebo **British National Corpus** (asi 100 milionů slovních tvarů, obsahuje i složku mluvenou), který se stal základním korpusem pro studium angličtiny. K významným korpusům jiných jazyků patří dva korpusy němčiny (v Mannheimu – různorodý - obsahuje literaturu uměleckou, odbornou, texty publicistické, protokoly z jednání spolkového sněmu i texty drobnějšího rozsahu, např. návodové, ve Stuttgartu). Pro Evropu je typické, že je obtížné najít jazyk, který by korpus neměl nebo pro který by se nebudoval.

Český národní korpus

Český národní korpus (dále ČNK) vzniká od roku 1992, v roce 1994 byl založen ÚČNK, který práci koordinuje. Na jeho budování mají podíl skupiny odborníků z pracovišť FF UK, MFF UK, FF MU, Fakulty informatiky MU, Fakulty elektrotechniky ČVUT, Ústavu pro jazyk český AV ČR a Ústavu pro českou literaturu AV ČR

Korpusy:

SYN2005 – odlišné texty od SYN2000

Pražský mluvený korpus – zejm. obecná čeština z Prahy a okolí

Brněnský mluvený korpus - subkorpus mluvené češtiny mluvčích narozených v Brně, je budován na FF MU pod vedením dr. Hladké. Skládá ze záznamů předem nepřipravených mluvených projevů – řízených i neřízených dialogů. Všichni mluvčí jsou obyvatelé Brna ve věku od 20 let výše. V mluvě rodilých Brňanů jsou obsaženy prvky spisovné i obecné češtiny, středomoravského dialektu, brněnských slangů a ve slovní zásobě je ve zbytcích znát také někdejší sepjetí brněnského mluvy s německým jazykem. Ze spolupráce pracovišť FF MU a FI MU vzešel značkovaný synchronní subkorpus **DESAM** obsahující zhruba 1 milion slovních tvarů psané češtiny.

DIAKORP – čeština 7 století od 13. stol do současnosti (pub. do r. 1989, uměl. texty do r. 1944)

Český národní korpus – http://ucnk.ff.cuni.cz/

Struktura ČNK

Český národní korpus

Synchronní část		Diachronní část
psaný	mluvený	psaný
SYN2000	Pražský mluvený korpus	DIAKORP
SYN2005	Brněnský mluvený korpus	
SYN2006PUB	ORAL2006	
PUBLIC		
SYNEK+ LITERA		

Složení korpusu SYN2000 a SYN2005

	SYN2005	**SYN2000**
beletrie	40 %	15 %
odborná literatura	27 %	25 %
publicistika	33 %	60 %

Rozsah korpusů

SYN2000 – 100 milionů slovních tvarů
SYN2005 – 100 milionů slovních tvarů
SYN2006PUB – 300 milionů slovních tvarů
PUBLIC – 20 milionů slovních tvarů
SYNEK – 10 milionů slovních tvarů
Pražský mluvený korpus – 700 tisíc
Brněnský mluvený korpus – 400 tisíc slovních tvarů
DIAKORP – 1 milion slovních tvarů

Co lze například najít v korpusu SYN2000:

1. informace o **frekvenci** tvarů slov nebo spojení slov

2. **kontext** slov

3. informace o **typech textu**

4. rozsah **užití nekodifikovaných prostředků** v současných psaných textech

5. rozsah užití **přejatých slov**, jejich pravopisná podoba

6. Jazykové **varianty**, např. **pravopisné** (*realismus/realizmus)* **morfologické** (*kope /kopá)*, **lexikální** (příslovce *alespoň*/*aspoň*).

Příklady:

1. Tvar slova
dotaz: *lemma=nabít, lemma=nabýt*
vyjmenovaná slova – nepravá homonyma – *nabít x nabýt*
kolokace: *nabít nos, zbraň, sál emocemi, pomocí kuponu, akumulátor, bateriemi, mohl si nabít…*
nabýt sil, rozměrů, objemu, vědomost, významu, nesmrtelnosti, dojmu, platnosti, rysů, přesvědčení, platnosti, intenzity, …

2. Slova začínající na *vodo-, dis-/dys-* apod.
dotaz: *vodo.*, dis.**
vodovod, vodou, vodopád, vodorovný, vodoodpudivý, …(urči složeniny)

3. Slova končící na *-ička*
dotaz: *.*ička*
lahvička, babička, sklenička, krabička, matematička, trička, cestička, alkoholička, …(vyber zdrobněliny)

4. Jazykové varianty a jejich frekvence
dotaz: *word=gymnázim*
pravopisné: *gymnasium* (32) x *gymnázium* (772)
penicilin (84) x *penicilín* (31)
tvaroslovné: *nemocem* (229) *x nemocím* (26)
pravomocech (18) *x pravomocích* (51)