Korpusová lingvistika

Literatura:

Blatná, R., Čermák, F. (eds.) *Jak využívat Český národní korpus*. Praha: Nakladatelství Lidové noviny, 2005. ISBN 80-7106-736-9.

Čermák, F., Blatná, R. (eds.) *Korpusová lingvistika: Stav a modelové přístupy*. Praha: Nakladatelství Lidové noviny, 2006. ISBN 80-7106-861-6.

Čermák F., Klímová J., Petkevič V. (eds.) *Studie z korpusové lingvistiky*. Praha: Karolinum, 2000. ISBN 80-7184-893-X.

Bartoň, T. a kol. *Statistiky češtiny*. Praha: Nakladatelství Lidové noviny, 2009. ISBN 978-80-7106-5944.

Čermák, F., Křen, M. (eds.) *Frekvenční slovník češtiny*. Praha: Nakladatelství Lidové noviny, 2004. ISBN 80-7106-676-1.

Základní pojmy:

korpus

korpusová lingvistika

vlastnosti korpusu: označkovanost, reprezentativnost

typy korpusů: psané, mluvené, synchronní, diachronní, 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 atd.

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 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 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

vyhledávaný výraz *(KWIC)*
lemma *(základní /slovníkový/ tvar)*
tag *(morfologická značka, např. označení gr. rodu, čísla, pádu apod.)*

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. K největším britským korpusům patří např. **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.

Co lze například najít v synchronních korpusech psaného jazyka

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

2. **kontext** slov

3. informace o **zdrojích textů**

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

5. jazykové **dublety**, např. **pravopisné** (*realismus/realizmus)*, **morfologické** (*kope/kopá)*, **lexikální** (příslovce *alespoň*/*aspoň*).

Příklady jednoduchého vyhledávání:

1. Všechny tvary slova
dotaz: *[lemma="nabít"], [lemma="nabýt"]*

Výsledek:
nepravá homonyma – *nabít x nabýt*
Úkol: Vyhledejte kolokace.
nabít nos, zbraň, sál emocemi, pomocí kuponu, akumulátor, bateriemi, mohl si nabít apod.
nabýt sil, rozměrů, objemu, vědomost, významu, nesmrtelnosti, dojmu, platnosti, rysů, přesvědčení, platnosti, intenzity apod.

2. Tvar slova
dotaz: *[word="nabít"]*

3. Slova začínající na *vodo-, dis-/dys-* apod.
dotaz: *vodo.*, dis.**
dotaz: *[word="vodo.*"]*

Výsledek:
vodovod, vodou, vodopád, vodorovný, vodoodpudivý, ...
Úkol: Určete složeniny.

4. Slova končící na *-ička* apod.
dotaz: *.*ička*
dotaz: *[word=".*ička"]*

Výsledek:
lahvička, babička, sklenička, krabička, matematička, trička, cestička, alkoholička, ...
Úkol: Vyberte zdrobněliny.

4. Jazykové dublety a jejich frekvence (SYN2005)
dotaz: *[word="gymnázium"], [word="gymnasium"]*

Například: pravopisné: *gymnasium* (21) x *gymnázium* (549)
morfologické: *pravomocech* (10) *x pravomocích* (26)

5. V korpusu SYN2010 vyhledejte možné tvary slov a uveďte jejich frekvenci. Kodifikované podoby tvarů ověřte ve Slovníku spisovné češtiny.

Trenér v takových bezesných (noc) _____ probírá každý detail.
Studenti byli brzy hotovi se svými (odpověď) _____.
K závěrečným (část) _____ práce nemáme žádné další připomínky.
Do společnosti nosí pánové klasický oblek, ke slavnostním plesovým (noc) _____ patří smoking.
Pampeliška pomáhá při chronických (nemoc) _____ kloubů a žlučníkových potížích.
Mnohé městské (čtvrt) _____ byly zničeny při posledním nočním útoku.
Ve (zeď) _____ stavení se drolil písek.
O tvých (lest) _____ vím své.
Takovým (past) _____ se raději vyhýbám.
Dozvěděli jsme se o mnoha (oběť) _____.
Nevěřím tvým (řeč) _____.