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, 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)

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:

SYN2000, SYN2005, **SYN2010 –** synchronní korpusy současné psané češtiny

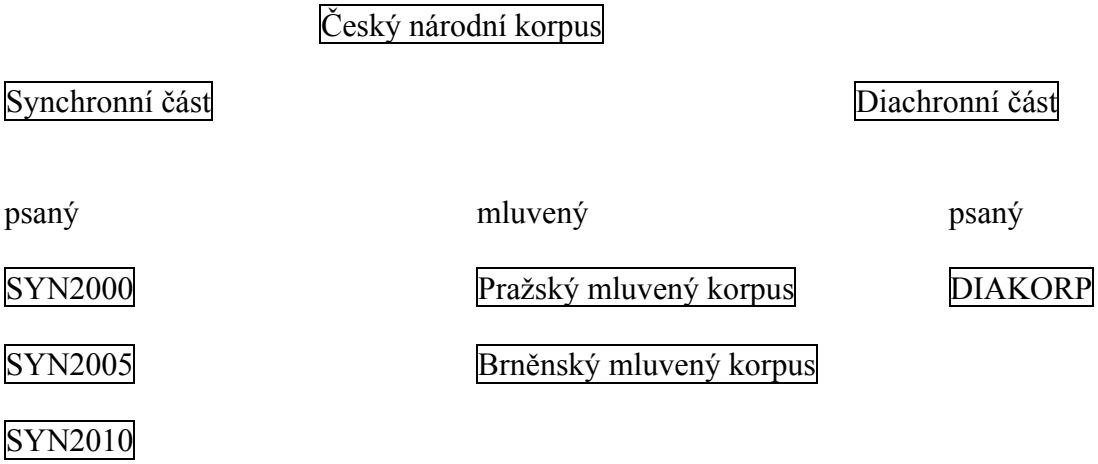
Pražský mluvený korpus – zejm. obecná čeština z Prahy a okolí

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 (publ. do r. 1989, uměl. texty do r. 1944)

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

Struktura ČNK



SYN2006PUB

ORAL2006, 2008

SYN2009PUB

ORWELL

Rozsah korpusů

SYN2000 – 100 milionů slovních tvarů
SYN2005 – 100 milionů slovních tvarů
SYN2010 – 100 milionů slovních tvarů
SYN2006PUB – 300 milionů slovních tvarů
SYN2009PUB – 700 milionů slovních tvarů
ORWEL – 80 000 slovních tvarů

Pražský mluvený korpus – 675 000 slovních tvarů
Brněnský mluvený korpus – 490 000 slovních tvarů
ORAL 2006 – 1 mil. slovních tvarů
ORAL2008 - 1 mil. slovních tvarů

DIAKORP – 1,95 milionu slovních tvarů

Co lze například najít v korpusu SYN2000:

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) _____.

Pracovní materiál - cvičení

Skloňování podstatných jmen rodu mužského neživotného

1. Doplňte spisovné tvary genitivu singuláru:

Na semináři byl od (pondělek)_____ do (pátek)_____. Skočil rovnou do (rybník)_____. V celé oblasti už nebylo jediného (rybník)_____. Pravidelně se s ním setkávali od (srpen)_____ do (listopad)_____. Dožil se jednoho (rok)_____. Bez (budík)_____se neobejdu. Z továrního (komín)_____ se valil šedý kouř. Koně vjížděli do (dvůr)_____. Mluvil ze (sen) _____. Příběh vyprávěl od (prostředek) _____.

2. Doplňte stylově rovnocennou, případně hovorovou variantu tvaru:

v hotelích	po kouscích
po okresech	o obláčcích
ve sklepech	po chodnících
po jezech	v teplákách

3. Některá podstatná jména rodu mužského mají v lokále plurálu pouze jednu koncovku:

rybník	břeh
cíl	prostředek
plech	šachy

Skloňování podstatných jmen rodu ženského

1. Doplňte správné tvary množného čísla (skloňování podle vz. *píseň, kost*).

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. Musíme zabránit častým (krádež) _____. Mnohé městské (čtvrt) _____ byly zničeny při posledním nočním útoku. Na (hráz) _____ rybníka se shromáždily stovky lidí. 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) _____ .

Skloňování podstatných jmen rodu středního

1. Slova v závorkách převeďte do správných tvarů lokálu plurálu:

Vítr se prohánět po (strniska) _____ . Už se nezmiňuj o těch (psiska) _____ . Cestovali jsme po českých (městečka) _____ . Umíš to vyjádřit v (procenta) _____ ? V obou (stanoviska) _____ byly nepatrné rozdíly. Na nedělních (korza) _____ bývalo mnoho lidí.

Časování sloves

1. Infinitivy v závorkách nahraďte přítomnými (budoucími) tvary oznamovacího způsobu:

Některé druhy hmyzu (přenášet) _____ nemoci. Oni to neradi (vidět) _____ . Lidé si někdy zbytečně (nerozumět) _____ , často se na sebe (rozzlobit) _____ pro maličkosti. Oni mi vždy (chtít) _____ pomoci. Rodiče mi ve všem (věřit) _____ . Vegetariáni (nejíst) _____ maso. Děti na podzim (pouštět) _____ draky. Oni (dělat) _____ , jako když to (nevidět a neslyšet) _____ , ale zatím o tom dobře (vědět) _____ . Na kraji lesa (křičet) _____ sojky. Mouchy (nesnášet) _____ průvan.

2. Infinitivy v závorkách nahraďte tvary minulého času:

Zubní lékař mi (vytrhnout) _____ bolavý zub. Čestní hosté (zaujmout) _____ místo na tribuně. Obránce (vykopnout) _____ míč z prázdné branky. Soud (přihlédnout) _____ k dobrému chování viníka. Auto (zahnout) _____ za roh. Policisté pachatele (dostihnout a zatknout) _____ . Bylo ticho. Nikde se nic (nepohnout) _____ . Stromky zasazené podél silnice (uschnout)_____ docela. (Usednout) _____ do křesla, aby (odpočnout si) _____ .

3. Nahraďte infinitiv správnými tvary přítomného času (uveďte všechny možnosti):

Čestní lidé nikdy (nelhat) _____. Všichni naši studenti (podepsat) _____ dohodu. Přítomní (prokázat) _____ svou totožnost. Roztoky různé nasycenosti (vřít) _____ při různých teplotách. Stroj (příst) _____tenké vlákno lépe, než to (moci) _____ dokázat lidské ruce. Jsem tak unaven, že se sotva (vléci) _____. Počkejte na mě, až se (vykoupat, obléci, učesat) _____ . (Vrzat) _____ nám dveře.