

## Korpusová lingvistika u nás

V souvislosti s technickým vývojem počítačů dochází od počátku 90. let XX. stol. k prudkému rozvoji korpusové lingvistiky (srv. níže a rovněž sam. kapitola Korpusová lingvistika).

Od počátku 90. let 20. století se také u nás dynamicky rozvíjí korpusová lingvistika. Na jaře roku 1992 se sešla v Praze skupina badatelů (František Čermák, Jan Hajič, Eva Hajičová, Jan Králík, Karel Pala, Klára Osolsobě, Věra Schmidtová, a další), kteří založili sdružení Počítačový fond češtiny (PFC) (srv. podrobněji Čermák, Králík, Pala, 1992). Cílem tohoto sdružení bylo koordinovat úsilí a zajišťovat komunikaci a spolupráci odborníků, kteří mají zájem o počítačové zpracování českého jazyka. Posléze jejich snahy nabyly institucionalizované podoby. Prvním krokem byla grantová podpora (vůbec první grant nesl název "[Počítačový korpus českých psaných textů](#)", od roku 1993 koordinoval spolupráci odborníků univerzity Karlovy v Praze, Masarykovy univerzity v Brně a Ústavu pro jazyk český, byl úspěšně ukončen v roce 1995). Klíčový význam pak mělo založení samostatného pracoviště **Ústavu Českého národního korpusu** <http://ucnk.ff.cuni.cz/> v čele s Františkem Čermákem (srv. sam. kapitola Korpusová lingvistika). Na další rozvoj bohemistiky má velký vliv budování rozsáhlých jazykových korpusů a korpusových nástrojů na straně jedné a „vytěžování“ (mining) korpusů (využívání jazykových korpusů jako zdrojů informací o jazyce) na straně druhé. Nepochybný a netrpělivě očekávaný bude význam využívání korpusů pro počítačovou lexikografii (srv. Čermák, 1999, Čermák, Klímová, Pala, Petkevič 2001). Prvním slovníkem založeným na korpusových datech v moderním slova smyslu je *Frekvenční slovník češtiny* (Čermák, Křen, 2004).

**Ústav Českého národního korpusu** FF UK zajišťuje od svého založení v roce 1994 budování Českého národního korpusu (ČNK), koordinaci jednotlivých korpusově orientovaných projektů a pěstování oborů korpusová lingvistika a korpusová lexikografie (srv. více sam. kap. Korpusová lingvistika a <http://ucnk.ff.cuni.cz/>). Členové pracovního kolektivu (Renata Blatná, František Čermák, Karel Kučera Věra Schmidtová, Michal Šulc a další) publikují nejen v odborném tisku doma i ve světě, ale populárně orientovanými knižními díly i časopiseckými a novinovými články přispívají k seznámení širší veřejnosti s děním v oboru. (srv. např. Blatná, Čermák, 2005).

fi mu  
bonito  
PDT  
ff mu

## Korpusové nástroje

Problematika korpusových nástrojů je rozsáhlá a představuje pole, na kterém se setkávají požadavky uživatelů (hlavně lingvistů a lexikografů) s přístupy programátorů. Výsledkem je konkrétní programové vybavení umožňující získávat z korpusů "poklady", které jsou v nich skryty. Aby si čtenář mohl udělat představu, co je v současnosti uživatelům k dispozici, uvedeme příklady konkrétních programů (systémů) používaných ve třech hlavních centrech korpusového bádání ve Velké Británii, konkrétně v Oxford University Press (OUP) a na univerzitách v Lancasteru (UCREL) a Birminghamu (COBUILD). Poznamenejme, že jednotlivá centra si ve skutečnosti vyvíjejí svůj vlastní software a mají jej jen pro vlastní potřebu, ovšem, jak lze vidět níže, jde do značné míry o podobné programy.

Základem jsou obvykle *konkordanční programy*, které třídí a počítají objekty nalezené v korpusu, což jsou v *syrovém* korpusu slovní tvary, interpunkce, případně další znaky (vyznačující třeba hranice vět, odstavců aj.) - ty jsou typicky součástí *SGML*. Pokud není do korpusu nějak zavedena další informace, konkordanční program nemůže rozlišit určité víceznačnosti (homonymie), např. v češtině mezi tvary *ženu* (ak.sg.substantiva *žena*) a *ženu* (1.os.sg.prés.slovesa *hnát*), nemluvě již o tom, že tvar *hnát* může být také tvarem substantiva mužského rodu. Proto ke korpusovým nástrojům patří i programy, které představují svého druhu gramatické analyzátoři: orientují se na morfologii, syntax a v poslední době i na sémantiku. V současné terminologii se obvykle mluví o *značkování (tagging)* a o *značkovacích programech (taggers)* různé úrovně. Níže uvedené taggery obvykle pracují tak, že se snaží každému slovu v korpusu přiřadit jeho *gramatickou značku*, tj.jeho *slovní druh* včetně relevantních gramatických kategorií. Programy uvedené dále buď s těmito analyzátoři spolupracují, nebo je přímo obsahují jako svou součást, nicméně pro přehlednost se o nich dále zmiňujeme zvlášť.

- *Program* *TESS*  
Vytvořen v OUP, v jazyce C, běží pod X-Windows a poskytuje uživateli možnost:
  - vyhledat souvšskyty zadaných slov (v korpusu - rozumí se v *BNC*), např.*obvykle* a *pršet*
  - zjistit a porovnat užití slova, tj.porádit jeho konkordanční seznam
  - zjistit frekvenční údaje o slovu
  - zvolit korpus (častěji subkorpus)
  - vyhledat slova obsahující zadané řetězy znaků (regulární výrazy)
  - zjistit distribuci slov v korpusu
  - vyhledat nejčtenější slova (podle slovních druhů)
  - najít gramatické údaje pro zadaná slova
  - pro zadané kolokace (kombinace slov) jako *dále uvedený* vypočítat tzv.MI- a T-score (viz níže)
- *Program* *LOOKUP*  
Vytvořen J.Clearem v COBUILDU, je napsán v jazyce C, běží pod UNIXem, využívá X-Windows a uživateli umožňuje:
  - zjistit frekvence slov v celém korpusu (zde *Bank of English*)
  - třídít podle různých zadaných kritérií
  - sestavovat konkordance s různými filtry (formát KWIC - key word in context) stál v autobuse

distribuce slova v korpusu  
první den v měsíci

- vyhledávat n-místné (!) kolokace, vypočítávat MI- a T-score (Mutual Information, které udává poměr pozorované/očekávané pravděpodobnosti výskytu jednotlivých prvků kolokace v korpusu; je to tedy míra udávající kolokabilitu a čím je vyšší, tím je spojení idiomatičtější - to se uplatní ve spojeních typu *vyšoká škola*, *slaměný vdovec* nebo *horký kandidát*)
- vyhledávat výrazy na úrovni gramatiky, tj. podle slovních druhů a gramatických kategorií; lze pak pracovat i s taggerem a nechat si označkovat vyznačenou část korpusu
- poskytovat údaje k jednotlivým subkorpusům - podle volby uživatele
- *Program SARA*  
Vytvořen v Oxford University Computing Centre (k dispozici na třech CD asi za 240 liber), pracuje pod Unixem a DOSem ve Windows a uživateli nabízí možnost:
  - vytvářet konkordance z *BNC*
  - vyhledávat kolokace a k nim hlavní frekvenční údaje
  - vyhledávat výrazy v korpusu na základě regulárních výrazů
  - získávat dvouprvkové kombinace výrazů
  - v omezené míře získávat statistické údaje
- *Program TED*  
Editor pro vytváření a zpracování slovníkových hesel. Byl vytvořen v OUP a poskytuje možnost přístupu do rozsáhlé databáze asi 80 elektronických slovníků a příruček. Lze v něm získávat veškeré lexikografické údaje potřebné při tvorbě nového slovníku. O programech a databázích tohoto typu se nám zatím může jen zdát.
- *Program D4*  
Vytvořen na univerzitě v Lancasteru, umožňuje pracovat se značkovánými texty a také korpusovými texty, kde každé větě je přiřazen její odpovídající syntaktický strom (treebanks), a dovede tvořit konkordance jak s gramatickými značkami (tagy jako *subs(tantivum)*, *verb(um)*), tak i bez nich. Je to jeden z nejzajímavějších programů, který lze na tomto poli v UK vidět. Slouží k budování rozsáhlých a realistických počítačových gramatik a syntaktických analyzátorů (pro angličtinu). Je založen na tzv. skeletonové syntaktické analýze.
- *Morfologické analyzátoři*  
Nejznámější morfologické analyzátoři (značkovací programy - taggers pro angličtinu) zpracovávají data v korpusu tak, že každému slovnímu tvaru přiřadí jeho gramatickou značku (tag), tj. obvykle symbol slovního druhu (může jich být i víc). Obvykle se značkují vybrané části korpusu v rozsahu do 10 mil. slovních tvarů; vzniklé soubory jsou zhruba třikrát až čtyřikrát větší než původní, což znamená, že při jejich dalším zpracování vznikají časové problémy.
  - Probabilistický analyzátor *CLAWS* (autor R.Garside z Lancasteru): Má vysokou úspěšnost, dosahuje jen 1.7% chyb. Celkově je *CLAWS* hybridní (vedle stochastického přístupu obsahuje i jednoduchá syntaktická pravidla) a pracuje s anotovaným lexikonem, jehož součástí je i seznam základních anglických idiomů. Značkování se provádí v několika fázích, používá se rovněž Viterbiho algoritmu

(zpracovává pravděpodobnosti přechodu mezi větnými složkami). Probabilistický přístup je motivován tím, že je blízký psychologii člověka.

- Analyzátor vytvořený J.Clearem v birminghamském COBUILDU: Rovněž využívá pravděpodobnostního přístupu, je velmi robustní a jeho míra úspěšnosti je 95%- autor ji pokládá za dostačující.
- Helsinský analyzátor: Je založen na tzv. *constraint grammars* a je 60krát rychlejší než ostatní (předpokládá ale užití dvoustupňového morfologického analyzátoru Kimmo od Koskenniemiho) - je zatím ze všech zjevně nejúspěšnější, pokud jde o zvládnutí více jazyků (dosud dovede pracovat s 5 jazyky).
- Analyzátor D.Cuttinga et al. (je v public domain a dostupný v Internetu): Užívá skrytého Markovova modelu, je jazykově nezávislý, učí se od počátku na menších vzorcích, pracuje s vahami pravděpodobnostního výskytu, pracuje iterativně a ve fázi učení počítá s 18% předem označovaného textu.

## Značkování korpusů

Zmínili jsme se už o *gramatickém značkování* (tagging) - přiřazení (symbolů) značek slovních druhů každému výskytu slova v korpusu. Výsledkem je tedy *anotovaný* korpus, tj. ne již *čistý* (surový) korpus, ale jeho verze opatřená gramatickými informacemi jistého druhu.

Takto anotovaný korpus se stává odrazovým můstkem pro další výzkum: pomocí konkordančního programu v něm můžeme vyhledávat gramatické abstrakce, jako např. výskyty pasíva (seznamy tvarů jako *dělán, prodán, vyroben*), vidu (seznam všech dokonavých sloves s předponou *vy-*), různé posloupnosti slovních druhů aj. Anotovaný korpus poskytuje též výchozí statistická data pro pravděpodobnostní zpracování jazyka. Ke značkováným korpusům patří *Brown Corpus*, *Lancaster-Oslo-Bergen Corpus* (LOB) a *Spoken English Corpus*, který obsahuje fonetické a fonémické značkování.

Gramatické značkování na úrovni vyšší než slovnědruhové lze najít např. v *London-Lund Corpusu* (Svartvik, 1990). Vznikly již syntakticky analyzované subkorpusy známé jako stromové banky (*treebanks*), byly však vytvořeny jen z podčástí korpusů. Nedávný výzkum na LOB Corpusu však vedl k technice zjednodušené syntaktické analýzy známé jako *skeletonová analýza*, kterou lidští operátoři mohou provádět velmi rychle (Leech and Garside, 1991).

Anotování korpusů nekončí u syntaktické analýzy. Dalšími předpokládanými fázemi jsou sémantická a textová (promluvová) analýza. Byla již provedena anotace *London-Lund Corpusu* týkající se promluvových (textových) ukazatelů (Stentström, 1990) a dalším příkladem je anaforická stromová banka, která se vytváří u LOB Corpusu a zahrnuje nejen skeletonovou analýzu, ale i vyznačení anaforických vztahů v textu - vztahy typu *Nachystám ti tam ty diskety. Vezmeš si je tam zítra.*

## Situace v češtině

Závěrem uvedme základní informace o tom, jak vypadá situace pro češtinu. Na podzim roku 1994 byl na FF UK založen Ústav českého národního korpusu, v němž se nyní buduje Český národní korpus. Během roku 1995 byl vytvořen jeho základ, v němž je uloženo cca 20000000 slovních tvarů, a na konci r.1996 by již český korpus měl obsahovat téměř 100 mil.českých slovních tvarů. Vedle ÚČNK se na této práci podílejí další pracoviště UK, jako Ústav teoretické a počítačové lingvistiky FF UK, Ústav formální a aplikované lingvistiky MFF UK, dále Ústav pro jazyk český AV ČR a v neposlední řadě Ústav českého jazyka FF MU i Fakulta informatiky MU.

Struktura textů ukládaných do korpusu se vyznačuje analyzátozem SGML. Pro gramatické značkování se připravuje analyzátor (tagger) LEMMA vytvořený v Brně skupinou Ševeček, Osolsobě, Pala, který je dnes schopen pracovat se 164000 českých kmenů a dovede každému rozpoznanému slovnímu tvaru přiřadit jeho slovní druh(y) a odpovídající gramatické významy. Na rozdíl od pravděpodobnostně orientovaných analyzátorů pro angličtinu je LEMMA založena na úplné morfologické analýze češtiny, proti které je podobná analýza angličtiny spíše dětskou hračkou. Ze stejné dílny pocházejí i podobné lemmatizující programy pro slovenštinu a ruštinu a dále pro angličtinu, němčinu a francouzštinu.

Vedle již uvedených důvodů korpusy potřebujeme i s ohledem na náš budoucí vstup do EU: i když jedním jazykem je zde do značné míry angličtina, překládání mezi jazyky uvnitř EU je nevyhnutelné. Vznikají proto *paralelní korpusy* využívané při budování systémů strojového překladu a tvorbě vícejazyčných a dnes už primárně elektronických slovníků. Není tajemstvím, že EU počítá s Polskem, Maďarskem a Českou republikou jako prvními východoevropskými členy EU - odráží se to i v existenci společného slovníkového projektu CEGLEX (Central European Generic Lexicon) zahrnujícího primárně polštinu, maďarštinu a češtinu.