Korpusová lingvistika

Koncem 19. století byly v reakci na dosavadní zkoumání jazyka z ryze historického hlediska položeny základy strukturalistické lingvistiky. Otcem tohoto hnutí byl Ženevan Ferdinand de Saussure (1857– 1913). Lingvistiku vracející se k Saussurovi lze charakterizovat s ohledem na její vědecko-teoretická východiska jakožto empirickou vědu. Jejím cílem je synchronní popis jazyka opírající se o analýzy empiricky uchopitelného materiálu jednotlivých jazyků. Z tohoto přístupu vzešla ve strukturalistické lingvistice myšlenka korpusu. Korpusy zahrnují jazykové jevy v podobě "masových dat", která lze uchovávat a zpracovávat pomocí počítačů. Tak dochází k úzkému propojení korpusově orientovaného výzkumu jazyka a počítačové lingvistiky. Z komputační lingvistiky se začíná postupně vydělovat korpusová lingvistika, lišící se od ní orientací na masové zpracování korpusových dat a na aplikace z něj plynoucí. Prvním moderním elektronickým korpusem byl *The Brown Corpus of Standard American English* většinou uváděný pod názvem *Brown Corpus*. Vytvořili jej W. Nelson Francis a český rodák Henry Kučera. Ačkoliv z dnešního hlediska jde o korpus velmi malý (1 milion slovních tvarů), jednalo se o první korpus v dnešním slova smyslu (elektronický, složený ze vzorků vybraných z široké škály textů tak, aby byl dodržen požadavek reprezentativnosti korpusu). V roce 1967 shrnuli Francis a Kučera výsledky analýz vycházejících z *Brown Copusu* v nyní již klasickém díle *Computational Analysis of Present-Day American English* (srv. W. Nelson Francis, Henry Kučera, 1967). Přestože byla korpusová lingvistika zpočátku vystavena značné kritice (zejména v 50. a 60. letech ze strany N. Chomského), stala se postupem času významným metodologickým proudem jazykovědy. Kromě mnoha korpusových projektů orientovaných na angličtinu se začínají budovat korpusy dalších jazyků. Velmi podrobné informace o korpusech jednotlivých jazyků a řadu dalších odkazů lze najít na http://www.athel.com/corpus.html. Od roku 1996 vychází International Journal of Corpus Linguistics (IJCL) přinášející širokou škálu názorů na roli korpusové lingvistiky ve výzkumu jazyka, počítačové lexikografii a NLP.
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 Schmiedtová, a další), kteří založili zájmové sdružení *Počítačový fond češtiny* (PFČ) (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 v r. 1994 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).
V rámci korpusové lingvistiky se dnes termínem korpus označuje rozsáhlý (vymezení rozsahu korpusu je dáno účelem, k němuž se korpus buduje) soubor počítačově čitelných (MRF – Machine Readable Form) textů složený ze vzorků vytvořených podle jistých

pravidel tak, aby reprezentovaly pokud možno jazyk jako celek v celé jeho pestrosti (reprezentativnost korpusu), který obsahuje standardní reference (anotace). Tyto reference zahrnují nejrůznější metatextové informace (vzhled textu, členění na kapitoly, odstavce, typografie, ale také údaje o typu textu (informace o autorovi, dataci, žánrovém zařazení atd.) a interpretace jednotek, z nichž je text složen (jazykové značky – tagy – anotace slovnědruhové, morfologické, syntaktické, sémantické, prozodické aj.). Iniciativu při budování standardního způsobu anotací převzala iniciativa TEI (Text Encoding Initiative). Jedná se o aktivitu sponzorovanou hlavními vědecky orientovanými asociacemi zabývajícími se využitím komputerů v humanitních vědách: ACL (*Association for Computional Linguistics*), ALLC (*the Association for Literary and Linguistic Computing*), ACH (*the Association for Computers and Humanities*). Cílem TEI je vytvoření standardní implementace pro operace s počítačově čitelnými texty. TEI za tímto účelem používá již existující formu značkovacího jazyka SGML (Standard Generalised Markup Language), popř. XML. Značkovací jazyk je jakýkoli jazyk, který vkládá do textu značky vysvětlující význam nebo vzhled jednotlivých jeho částí. Vůbec první obecný značkovací jazyk byl Generalized Markup Language (GML). Na jeho základě byl vytvořen jazyk SGML, který se v minulosti stal jedním z nejrozšířenějších značkovacích jazyků. V dnešní době se nahrazuje jazykem XML (EXtensible Markup Language) – což je rozšiřitelný značkovací jazyk, jenž je zjednodušenou verzí jazyka SGML, vzniklou odstraněním chyb SGML a jeho modernizací. Vlastním příspěvkem TEI je detailní návod k použití příslušných standardů. Text jako celek se v TEI popisuje pomocí DTD (**D**ocument **T**ype **D**escription). Celá řada korpusových projektů přijala TEI za své. TEI vydává mnoho návodů pro kódování korpusových textů. EU založilo dozorčí skupinu EAGLES (*Expert Advisory Groups on Language Engineering Standards*), která má za úkol sledovat různé evropské iniciativy a pomáhat jim. Jde o to vytvořit systém anotací, v němž by na jedné straně byla brána v úvahu specifika všech evropských jazyků a na straně druhé byla zachována jednota systému. (Více k tomuto tématu srv. Blatná, Čermák, 1995 a http://www.tei-c.org/).

V souvislosti s korpusovou lingvistikou se do středu zájmu dostávají aplikace NLP zaměřené na automatickou anotaci (značkování) velkých korpusů. Jsou vytvářeny počítačové programy pro automatické značkování – vkládání lingvistických informací do textu – velkých jazykových korpusů. Aplikace automatické morfologické analýzy se zaměřují na automatickou lemmatizaci (přiřazení základního tvaru tzv. lemmatu textovému tvaru slova), slovnědruhové značkování (POS – part of speech tagging) a přiřazování gramatických významů některých gramatických kategorií. Souhrnně se tento typ označuje jako gramatické značkování (tagování, popř. anotace, ale i lemmatizace). Vzhledem k tomu, že korpusy mají být zdrojem informací pro co nejširší okruh uživatelů, se teoretikové korpusové lingvistiky zamýšlejí nad tím, jak by měla lingvistická informace vkládaná do textu vypadat a čemu by měla sloužit. (Některé zásady pro vytváření anotačních schémat formuloval Geoffrey Leech, (viz Leech, 1993). Ruku v ruce s vývojem konkrétních anotačních nástrojů jdou tedy úvahy o mezích a možnostech, smyslu a účelu různých teoretických koncepcí vtělovaných do konkrétních interpretací jazykových jednotek, jimiž jsou tzv. značky (tagy) v případě slovnědruhových a gramatických značek, stromové struktury interpretující syntaktické vztahy s ohledem na sémantické popřípadě pragmatické aspekty vyšších jednotek jazyka, značky postihující prozodické vlastnosti úseků textů, jeho sociolingvistické aj. charakteristiky atp. Klíčovou roli v tvorbě automatických korpusových analyzátorů hraje řešení problému disambiguace (zjednoznačnění homonymních jazykových jednotek na všech úrovních jazyka). Konkrétně o problémech spojených s morfologickým anotováním a následnou disambiguací českých korpusů srv. např. Petkevič, 2001.