KORPUS A KORPUSOVÁ LINGVISTIKA

FRANTIŠEK **ČERMÁK**

KAROLINUM

Korpus a korpusová lingvistika

František Čermák

———

Recenzenti:
prof. PhDr. Eva Hajičová, DrSc.
PhDr. Jitka Šonková, CSc.
PhDr. Věra Schmiedtová

Vydala Univerzita Karlova
Nakladatelství Karolinum
Redakce Lenka Ščerbaničová
Grafická úprava Jan Šerých
Sazba DTP Nakladatelství Karolinum
Vydání první

© Univerzita Karlova, 2017
© František Čermák, 2017

ISBN 978-80-246-3710-5
ISBN 978-80-246-3776-1 (online : pdf)

Ukázka elektronické knihy

Univerzita Karlova
Nakladatelství Karolinum 2017

www.karolinum.cz
ebooks@karolinum.cz

Slovo úvodem

Text knihy zachycuje vznik a rozvoj korpusů a korpusové lingvistiky nutně k určitému datu, vývoj jde však rychle dál. Vyjadřovalo se k němu více lidí, můj hlavní dík však patří jejím recenzentům, kterými byli prof. PhDr. Eva Hajičová, DrSc. (MFF UK), PhDr. Jitka Šonková, CSc. (University of Iowa) a PhDr. Věra Schmiedtová (FF UK).

František Čermák

Ukázka elektronické knihy

OBSAH

Ukázka elektronické knihy

I. ČÁST: TEORIE

1 ÚVOD: KORPUS A KORPUSOVÁ LINGVISTIKA

Tato kniha je stručným shrnutím a přehledem základních pojmů, zásad a metod ***korpusové lingvistiky***, tj. nauky o korpusu a práci s ním, i nového stejnojmenného oboru. Už podle svého jména se tu staví nutně na **korpusu**, tj. velkém elektronickém, systematicky budovaném a organizovaném úhrnu textů, který je conditio sine qua non pro jeho disciplínu, tj. korpusovou lingvistiku. Kniha zároveň přináší i malou ilustraci možností výzkumu s modelovými výsledky.

Korpus se objevil v lingvistice, spolu s nástupem počítačů, poměrně nedávno a výzkum jazyka v ní v důsledku existence korpusu zcela převrátil poměry a etablované postupy, především z hlediska rozsahu informací a možností, které data i počítače nabízejí. Dalo by se tu, jakkoliv se to běžně nedělá, mluvit o korpusové revoluci v lingvistice, která jde mj. ruku v ruce s rozvojem počítačů, aspoň na svém počátku.

Jestliže lingvista tradičně donedávna strádal zásadním nedostatkem dat, a tedy, metaforicky řečeno, jistou informační podvýživou, pak se s nástupem korpusů situace radikálně změnila. Tam, kde v celé dlouhé minulosti oboru lingvista pracně, dlouho a nedokonale shromažďoval svá manuální excerpta v podobě lístečků a výpisků, než je mohl začít třídit a dávat do své kartotéky (a později i většího archivu), kterými se obv. pak sekvenčně probíral, tam nastoupil počítač a jeho možnosti. Na místě někdejších lístečků a výpisků dnes korpus poskytuje doslova záplavu informací a dat, ve kterých je sice často obtížné se vyznat přímo a snadno, ve kterých se ale dá už hledat více způsoby, vždy pomocí počítače. To vedlo a vede mimo jiné k nutnosti nalezení (a dalšímu hledání) potřebných technik a metodologie, jak k tomuto množství dat smysluplně přistupovat.

Korpusové poznání přináší v mnoha ohledech překvapivá zjištění. Tato kniha výběrově přibližuje možnosti mnohostranného poznání všeho základního, co se událo za cca dvacet a více let v této nové disciplíně, **korpusové lingvistice**, jejíž

výsledky a aplikace začínají i u nás už otřásat tradiční důvěrou ve spolehlivost stávajících lingvistických prací, starých slovníků, mluvnic a dalších příruček. Mimo jiné se korpusová lingvistika postupně stala už i novým lingvistickým oborem studovaným na univerzitách.

Kniha postupuje od širšího rámce, potřeb moderní lingvistiky, povahy korpusu a jeho možností až k některým základním a konkrétním autorovým výsledkům z jeho studia a výzkumu v daném oboru. Svým systematickým důrazem na vyčerpávající přístup k datům a jejich popisu, který v lingvistice přináší a důsledně uplatňuje poprvé korpusová lingvistika, se tato kniha snaží korpusová data a přístupy k nim aspoň základním způsobem charakterizovat a narýsovat tak zde zároveň i půdorys této rychle se rozvíjející disciplíny. Stejně tak se snaží ukázat na průnik i návaznosti na lingvistické disciplíny tradiční a ovšem přitom i naznačit otevřené otázky a problémy. Jakkoliv se hlavní zkušenost opírá o práci s budováním a analýzou ***Českého národního korpusu*** (viz blíže na adrese *korpus.cz*), v pozadí stojí i zkušenost s většinou korpusů větších a řadou menších, především evropských, a kontakty s týmy, které je budují, která se prezentuje výběrově také.

Tato stručná oborová příručka a malý sborník zároveň se pokoušejí komplexně zachytit tedy jak povahu korpusových dat, tak způsoby jejich zpracování, možnosti práce s nimi i všechny hlavní souvislosti a návazné aspekty tak, aby podala stručný komplexní a ucelený, avšak zároveň i základní obraz ***oboru*** korpusová lingvistika (***I. ČÁST: TEORIE***).

Doprovázejí ji a jen volně na ni navazují reprinty vybraných autorových studií a analýz představující některé výsledky analýzy korpusu (***II. ČÁST: SONDY, STUDIE, ANALÝZY***), které na konci doplňuje ilustrativním způsobem několik konkrétně orientovaných dodatků (***III. ČÁST: PŘÍLOHY***).

Svým pojetím je kniha určena jak studentům a lingvistům či odborníkům z jiných oborů, především humanitních, tak i širší zainteresované veřejnosti. Kniha je vybavena dvojí bibliografií, obecnou, vztahující se k celé problematice disciplíny, a specifickou, vztaženou přímo k tématům v podobě stručných odkazů za jednotlivými kapitolami. Bibliografie tak do značné míry vyvažuje i specifikuje stručnost formulací a popisů zde obsažených a orientuje uživatele na další možnosti studia. Příspěvky v části II mají původní bibliografii vlastní.

ZÁKLADNÍ LITERATURA

Čermák 1995a, 1998, 2001b; McEnery – Wilson 1996; McEnery – Hardie 2012; Wynne 2005.

Speciální a další literaturu *viz v* **Korpusové bibliografii**

2 DATA A INFORMACE

Informace se chápe různě a ve více smyslech, intuitivně však především jako nějaký údaj, zjištění o něčem, resp. už i přímo zobecněné poznání či vědění. Informace se odborně někdy vymezuje jako to (tedy zjištění, poznatek), co relevantního lze získat z podkladových dat (údajů) pro analýzu jevu či problému, a to obvykle v nějakém rámci, pro lingvistiku pak z dat textových, a to zvláště v rámci a kontextu sociálním. Ten se v případě jazyka chápe především jako rámec sémiotický a je řízený pragmatickými pravidly.

Současný informační věk a jeho povahu obecně i specificky reflektuje i korpus: obojí, tj. náš věk i korpus sám, nabízí nebývalou záplavu informace. Její množství, které je zásadně vítané, však i hrozí zahlcením uživatele, a představuje proto také relativně nový problém. Pravda je, že náš informační věk se už dobře neobejde bez počítačů a globálního webu, na tyto nástroje a ***hromadné*** zdroje informace spoléháme dnes zcela samozřejmě. Této novodobé pravdě ovšem předcházela a předchází pravda starší a stále zcela základní, totiž že ***valná většina informace*** (významová, formální i pragmatická) ***je kódovaná a přenášená (přirozeným) jazykem*** a je jen málo oborů, které se při výměně informací bez jazyka obejdou (srov. část matematiky, popř. i malbu apod.); na to se často zapomíná.

Tyto informace, často po určitém zpracování a formalizaci, se obvykle předávají dál, dalším uživatelům, zájemcům. Pak se role toho, kdo informaci objevil, získal (zvl. odborník, lingvista), mění a stává se z něj ten, kdo ji předává, sděluje v nějaké podobě jiným lidem. Tyto informace, tj. poznatky zvláště o aktuální povaze a stavu toho, co nebo kdo nás zajímá, se dnes výrazně sdělují elektronicky, a to často ***hromadně***.

Stále se však dosud a hlavně sdělují i po staru také ***individuálně***, mj. ústním a soukromým kontaktem mezi lidmi, osobním pozorováním, zkoumáním, poznáním něčeho, popř. i zčásti zobecněnou a zprostředkovanou školní zkušeností, a tedy vlastní deduktivní či induktivní činností. Taková informace, např. článek, kde lingvista zformuluje své poznání a zjištění, či ho odpřednáší studentům, se tak dostává dál. Je však třeba lišit, zda takové poznání autor považuje, po ověření jeho platnosti, tj. zda skutečně odpovídá datům a obvyklým přijímaným normám, popř. i stupni poznání, za zobecnitelné, za obecně přijatelné a považuje ho tedy za poznání a informaci **objektivní**. Proti poznání objektivnímu jde však často i o poznání, zkušenost i informaci *subjektivní*, vázanou na jednotlivce apod., kde míra zobecnitelnosti nebývá jasná.

Oba zdroje poznání a informace z něj odvozené, zdroj individuální i hromadný, jsou ovšem komplementární a jeden nevylučuje druhý. Nicméně individuální a často subjektivní zkušenost a informace nabývaná jedincem obvykle nemívá

možnost ***potvrzení*** své správnosti ani obecnosti, a tedy zjištěné **platnosti** při opakovaném výskytu, tj. ověřované opakovaně, a použitelné tedy i ostatními. Taková povaha informace se naopak očekává, ba vyžaduje u informací veřejných, obecných, a zvláště vědeckých.

Data, informační jednotky a základní stavební kameny poznání, jsou podle oboru a druhu vnímání a zapojeného smyslu pochopitelně různá, (téměř) všechna jsou však převoditelná do slov, a tedy jazyka, psaného a mluveného, který nám zprostředkuje zrak a sluch. A až tady se zúročuje možnost, zprostředkovaná komputerově, dostávat se k zobecněné, popř. zobecnitelné, a tedy i zpravidla ***objektivnější a opakovatelné***, resp. opakované informaci, tj. takové, kterou lze nalézt až ve větších informačních jazykových souborech, resp. textech, tj. kterou v nich *lze ověřit* a kterou individuální a jednorázová zkušenost a poznání z vlastního výzkumu v zásadě nenabízí.

Toto ***zobecnění*** jednotlivého *faktu*, tj. poznání, ať je induktivní či deduktivní, a *jeho platnosti* je ovšem možné jen v ***kontextu*** a na základě či proti pozadí široké znalosti dané vzděláním obecným či oborovým; izolovaný fakt nám naproti tomu neřekne (téměř) nic.

Data a informace v nich však ještě nevedou k tomu, co je zvláště v případě seriózního studia na jeho konci to nejdůležitější, to jest k ***vědění***, popř. znalosti, jak informace používat a využívat. To je však už věcí integrace poznatků do platného kontextu a jejich nazírání v něm, ale i studia a učení, opřeného o opakování a postaveného kvůli souvislostem do vhodného širšího rámce, zvláště celého jazyka i disciplíny.

2.1 ELEKTRONICKÝ TEXT

2.1.1 OBECNÁ POVAHA, VLASTNOSTI A ZPRACOVÁNÍ TEXTŮ

Základní a nejčastější podoba informace je obsažená v *textu*, dnes většinou už elektronickém, tj. v textu vytvořeném pomocí počítače (a jím i dále zpracovávaném). Každý text je smysluplný řetězec znaků, pro lingvistiku a uživatele jazyka řetězec hlavně slov, který má lineární povahu. Psaný text má ale i povahu lineárně prostorovou, kde platí dimenze *vlevo-vpravo* (podle druhu písma i jiná), nebo lineárně časovou, kde platí v podstatě dimenze *napřed-potom* (u mluveného jazyka se jeho povaha lineárně časová přepisem mění na lineárně prostorovou taky). Paralelní jiná informace (vedle základní psané), zvláště intonace u mluvené komunikace (suprasegmentální rysy), se v písmu nezachycuje, a jen mluvený text je tudíž vlastně vícelineární, skládá se z více souběžných informačních linií (víc viz 2.3); mluvený text je však také obecně primární a historicky starší.

Elektronický text je pro potřeby korpusové lingvistiky text, který je počítačově čitelný a dál zpracovávatelný. V lingvistice se za text považuje libovolný souvislý text, získaný obvykle jen z autentických zdrojů, v praxi konkrétně (smysluplný) text psaný nebo mluvený. Míra a způsob zpracování a přípravy elektronického textu závisí na cíli, technických podmínkách a potřebách. Zpracování textu psaného a mluveného se liší. Podoby psaného textu se však liší také.

Obecně je při zpracování textů třeba lišit tři základní pojmy, jejichž míra obecnosti se postupně zužuje, a to nepřímo úměrně k nárůstu jejich specifičnosti: typ dat, kód dat a formát (obv. souboru). ***Typ dat*** je pojem nejširší a patří sem jen okrajově. Tvoří ho proměnlivý *způsob ukládání dat různého druhu*, dat vizuálních, obrázků (např. *jpeg, gif*) či akustických (např. *wav, wma* či kompresní *mp3*). Pro korpusy jsou však zásadní druhé dva pojmy, kód (kódování) dat a jejich formát (obv. formát souboru).

Texty se zachycují, resp. znázorňují v různých ***kódech***, které v elektronické oblasti označují *způsoby jejich převodu (a informace v nich) do způsobu jiného;* patří sem např. morseovka (písmena převáděná na tečky a čárky), (dříve) dírky a jejich uspořádání na děrném štítku, a dnes konečně a hlavně čísla (na která se převádějí písmena). V této poslední oblasti (užívající číselného kódu) jsou dnes běžné dva typy kódů a kódování, ASCII a Unicode. *ASCII*, které prodělalo více proměn a postupné rozšiřování, dnes čítá jen pro češtinu až pět různých verzí, a protože obecně stále není plně použitelný pro všechny jazyky, přechází se proto, resp. přešlo se do univerzálního *Unicode* (viz víc v 2.5.1 a 2.1.1).

Naproti tomu ***formát*** je *způsob technického záznamu a zpracování informace* v určitém kódu technických dat, který je značně proměnlivý. Formátů může být pro uložení téže informace víc než jeden (vedle netextových formátů, viz výše např. *mpeg),* specificky mj. *doc, T602, ODF pro Linux, html*, či obecný *XML*, i podle druhů operačního systému, pro Windows, Unix či Mac OS). V tomto smyslu tedy formát (resp. způsob kódování znaků, viz dál) je *takové uspořádání dat, které slouží k jejich znázorňování, ukládání a znovuužívání.* Známý je mj. formát *pdf* (portable document format) užívaný pro různé typy textů jako snadno převoditelný, určený zvláště pro tisk (s informací o charakteristikách takového textu, ale i obrázky), který je známý z běžné práce s elektronickými texty.

Elektronické texty, které dnes vstupují jako jeho stavební kameny do korpusu, se tedy vyskytují v různých podobách, *formátech,* a pro potřeby počítače se musejí konvertovat do formátu jednotného, tj. sjednotit. Vedle formátů spojených s operačním systémem (*Windows, Unix/Linux*, dříve *DOS*) se texty vyskytují ve formátu, který jim dávají různé textové editory a procesory (jako *Word, Writer* v *LibreOffice* či *TextEdit* v *Mac OS X* aj.), existují ve formátech vznikajících i ve webovém prostředí (zvl. *html, Hypertext Markup Language*) a ty je třeba podrobovat *konverzi* (viz 2.5.1).

Ukázka elektronické knihy

V praxi se po konverzi a převodu ze specifického formátu výsledný formát označuje obvykle jako *txt* (prostý text). Zachycuje ho rozšířený kód *ASCII* (pův. *American Standard Code for Information Interchange*), obvykle už v univerzálním kódování *Unicode* (viz 2.5.1), specificky v UTF-8. Tak se tvoří základ pro další korpusové zpracování a budování vlastního korpusu (srov. dál ještě následnou úpravu textu jazykem XML, zahrnujícím i další zapracování, 2.5.1).

Vybraný korpusový text, získaný různými konverzemi textů z textových procesorů užívajících různých formátů, je ve výsledku zásadně text *autentický*, a má tu podobu, kterou mu dal autor, resp. vydavatel. Je tedy nepředstavitelné, že korpusový text, který by měl ideálně zachovávat záměr autora, by se měl opravovat, dál editovat, pospisovňovat či jinak pravopisně, nebo dokonce cenzorně, a tedy krajně sporně „vylepšovat" (což je věcí konkrétní a vždy problematické jazykové politiky a inženýrství). Ochranu autora a autentičnost jeho textu je třeba dodržet za každou cenu (jakkoliv i autorská formální úprava textu může být nevyrovnaná a naznačovat problémy tvůrce textu s jazykem). Předpokládá se tedy, že korpusový text zachovává i různé individuální překlepy a omyly autora, což pro masové korpusové hledání nevadí. Takové, v tradičním pohledu „spisovníků" a tradicionalistů, nenáležitosti bývají totiž řídké a neovlivňují zásadně celkový výsledek hledání v korpusu (viz 2.5.1). Korpusy mají tedy mj. i „konzervační", archivní roli tím, že uchovávají (relativně) trvale texty v neměnné podobě (jakkoliv ale přitom z hlediska typografického už není jejich původní podoba zpětně rekonstruovatelná). Při přípravě textů pro korpus se však vyskytují i (často hromadné) chyby textu vzniklé technickým zpracováním či sazbou textu (a vzniklé tedy bez, anebo proti přání autora), a ty je naopak třeba zjistit a řadou procedur opravit, jakkoliv rozlišení obou typů nebývá snadné.

Protože zásadní podobou korpusových dat je stále text psaný, spadají sem nutně i *elektronické přepisy* mluvených textů (viz 3.3.2), které mají také textovou podobu. Pro zachování jednotnosti s texty psanými mají být s nimi ideálně srovnatelné, avšak hledání jejich optimální formy (v zásadě formy transkripce) při zachování jejich věrné původní podoby je zatím otevřená otázka (zvl. s ohledem např. na nespisovné či individuální podoby slov). Takové texty se můžou dodatečně ještě vybavovat i fonetickým či prozodickým přepisem (zachycujícím jejich suprasegmentální rysy).

Proti dřívější koncepci budovat korpus z homogenních (náhodných) ***vzorků*** textů (ty vyhovují zvl. statistickým potřebám někdy lépe) se dnes dává, mj. i vzhledem k dostupnosti elektronických textů, přednost tomu, do korpusu začleňovat ***texty celé***, které umožňují optimální zkoumání kontextu, stejně tak ale i povahu odlišných částí téhož textu (jako je specifický začátek apod., při textové analýze). Problém, kdy se může přístup založený na vzorcích textů jevit výhodnější, se např. projevuje tehdy, když jde o malou textovou kategorii či (pod)obor, jehož naplnění