JAZYKOVÝ KORPUS: Prostředek a zdroj poznání v lingvistice

1. Pojem jazykového korpusu

Jazykovým korpusem lze rozumět vnitřně strukturovaný, unifikovaný a obvykle i oindexovaný a ucelený rozsáhlý úhrn elektronicky uložených a zpracovávaných jazykových dat většinou v textové podobě, organizovaný se zřetelem k využití pro určitý cíl, vůči němuž pak je také považován za reprezentativní. Existence a smysl tvorby korpusu vycházejí především ze dvou základních teoretických předpokladů a jazykových faktů zároveň:

(1) data nejrůznějšího druhu se v korpusu nacházejí ve své přirozené kontextové podobě a užití, které umožňuje jejich všestranné studium a indukci závěrů;
(2) velký rozsah plánovitě vybudovaného korpusu minimalizuje to, že čistou náhodou - k níž u malých rozsahů dat dochází - převládnou zvláštní a okrajová užití jazykových jednotek nad základními a typickými.

Vedle primárně sledované reprezentativnosti (viz 4.) korpusu v různém smyslu a míře (včetně škály typické/zvláštní/výjimečné) se u tvorby korpusu též obvykle zdůrazňuje i nutnost v něm zachytit variabilitu (viz 4.,5.) jazyka v různém smyslu.

Někdy se zjednodušeně a synonymně také mluví o komputerovém korpusu, pojímaném prostě jako velký soubor počítačově čitelných textů, ale to může zavádět. Žádný komputerový korpus není jen prostým a tedy třeba i náhodným souborem textů, a vždy tedy jde o korpus jazykový ve smyslu uvedeném výše (srov. mj. Aijmer et al., Johansson et al., Sinclair 1991, Souter et al., Svartvik 1992).

Možnost správy hromadných jazykových dat a práce s nimi na počítači vedou ve svých důsledcích nejen k nesmírnému zrychlení a usnadnění lingvistovy práce, ale i k jejímu dotud nevídanému zkvalitnění. Data takto získaná, která lze průběžně snadno modernizovat a doplňovat, tak představují dnes absolutně nejbohatší a nejrealističtější zdroj poznání jazyka vůbec, který vysoko předčí všechny lingvistovy pracně budované kartotéky a archívy minulosti; proto také jazykový korpus je předpokladem ke skutečné revoluci v práci s jazykem a i proto se zcela právem o posledním desetiletí tohoto století mluví také jako o dekádě korpusové lingvistiky.

Mluví-li se v přírodních a dalších vědách zcela samozřejmě o základním výzkumu, na který navazuje veškerý další výzkum a aplikace, pak v jazykovědě se právě takovým zdrojem a předpokladem základního výzkumu stává elektronický korpus. Docenění jeho prvotního významu vede pak i k pochopení toho, že jeho budoucí studium a široké využívání znamená skutečně novou epochu v lingvistice jak co do kvality a povahy dosahovaných výsledků, tak ovšem i podoby a povahy metodologie práce s ním; už na samotném začátku tu těsně spolupracují lingvisté s matematiky a odborníky v komputerové vědě a nové problémy a cíle, které se vynořují v průběhu práce, si vyžadují a budou vyžadovat zcela nové způsoby řešení a exaktnější metodologie, užité techniky a nástroje. Rostoucí význam tu nabývá lingvistické modelování a statistické metody, propojované do probabilitních modelů, ale i

fuzzy logika apod. Lingvistika se právě až v této fázi stává i prakticky plně interdisciplinární a není pochyb o tom, že k dosavadním disciplinám, které tu spolupracují, brzy přistoupí i další, jako je psycholingvistika, sociolingvistika a neurolingvistika.

První významné zužitkování neocenitelného a bezprecedentního bohatství informací uložených v korpusu se už promítlo do řady kvalitativně zcela nových slovníků některých jazyků, není však zdaleka jediné a do budoucnosti lze při využívání korpusových dat předpokládat závažnou a plodnou spolupráci mezi jazykovědou a všemi obory, které tak či onak s jazykem pracují (a to jsou téměř všechny), protože jazyk je nástrojem komunikace všech a jeho poznání a její zkvalitnění je také v zájmu všech; není v této perspektivě proto žádnou náhodou, že např. národní britský korpus sponzoruje britské ministerstvo průmyslu. Jazyk je však také odrazem kolektivního vědomí a kultury národa, resp. komunity jeho nositelů a v tomto ohledu jeho hlubší poznání může v lecčem přispět i oborům uměleckým, zvláště však literatuře.

Na krátké historii korpusu je dobře vidět, jak může být těžké být doma prorokem a jak i takoví lingvisté jako N. Chomsky, dnes zabředlí do hlubin svého materiálově nezakotveného a bezbřehého mentalismu, nemusejí potřebu budoucnosti dobře odhadnout. Už v r. 1962 se s despektem vyjadřuje o jakékoli možnosti přínosu korpusu rozvoji jazykovědy a poznání jazyka (mluví o jeho pokřivenosti), který mohl být dán snad tehdy slabým stupněm rozvoje počítačů, který však zřejmě neopustil ani dnes; aspoň se sám dosud nikdy nepokusil o studium jazyka na dnes už nepřeberné materiálové nabídce dat. Ve stejné době výrazného nástupu generativní gramatiky se naopak jiní spíše potichu a skromně pouštějí do prvních pokusů o korpusový přístup ke studiu jazyka, protože si uvědomují, nakolik je dosavadní poznání a teoretizování mezerovité a mnohdy i podložené nedostatečnými daty; za skutečné pionýry tu lze považovat kolem r. 1960 R. Quirka v Londýně s jeho Survey of English Usage (z něhož mj. vyrostla dodnes zřejmě nejobjektivnější i největší mluvnice angličtiny) a Čecha H. Kučeru spolu s Američanem N. Francisem, kteří vytvářejí první elektronický korpus angličtiny (Brown Corpus), dodnes ceněný a široce známý, který se považuje v dané oblasti za klasický počin.

Na rozdíl od dílčích nebo příležitostných užití korpusu jako zdroje dat pro určitý účel, rešerši či např. slovníkovou aplikaci, kdy se ke korpusu mohou obracet jak nejrůznější části lingvistiky vlastní, tak vědy a obory další, je však korpus i centrálním a trvalým objektem celé zvláštní disciplíny. Korpusová lingvistika je ta část lingvistiky, která systematicky pracuje s korpusem a jeho nástroji, resp. studuje zásady a praxi práce s ním s cílem lepšího poznání funkce a struktury jazyka, jaké až dosud nebylo možné. Je dnes nesporně hlavní složkou lingvistiky komputační, v jiném pohledu se však s ní značně překrývá. Rozdíl mezi obojí lingvistikou vyvstává hlavně při zdůraznění metod (na rozdíl od zdrojů) a nástrojů: vedle aplikovaných výstupů (jako je

strojový
překlad) se komputační lingvistika může zaměřovat jen na teoretické řešení otázek prostřednictvím komputerových programů a technik, avšak dříve nebo později k jejich ověřování a uplatnění na korpusu stejně přistupuje.

2. Korpusová data

Je třeba lišit mezi povahou jazykových dat (data vnější a hrubá) ještě před jejich vstupem do korpusu a po jejich vstupu do něj (data vnitřní a strojově čitelná, resp. zpracovatelná).

Zdrojem korpusových dat (vnějších) jsou obě manifestace jazyka, psaná i mluvená, resp. psané i mluvené texty, ne však zatím ve stejné míře, protože záznam mluveného jazyka a jeho převod do počítačově čitelné podoby (vlastní magnetofonová nahrávka a následný přepis) je dosud velmi nákladný (Crowdy); s ohledem na zlepšující se možnosti počítačového rozpoznávání mluvy a jejího přímého záznamu komputerem se však situace může dramaticky změnit.

Dosavadní zdroje dat se podle dostupných prostředků člení v zásadě na tři druhy. Nejlevnější a nejsnadněji využitelná jsou data v podobě elektronické sazby textů, kterou dnes užívá už většina centrálních novin a časopisů a některá nakladatelství. Druhou možností, různě úspěšnou v souvislosti s mírou typografické náročnosti textu, je načítání textů, resp. jejich skenování do počítače pomocí scannerů; na rozdíl od snímání obrázků je snímání písma v jeho různorodosti a různé velikosti (jen typografických sad jsou stovky), a tedy i optické rozpoznávání znaků (OCR, Optical Character Recognition) spíše složitější a vyžaduje velmi výkonný a specializovaný software. Zvlášť složité texty, např. slovníkové, však přesto při tom vykazují mnoho chyb a vynucují si mnoho oprav (osvědčily se tu do jisté míry pouze klasický přístroj od firmy Kurzweiler, popř. i výkonný novější ProLector).

Třetí možností je konečně manuální přepisování potřebných textů do počítače písařkou (prostřednictvím některého z běžných editorů). žádný z těchto způsobů tedy nevede k potřebným datům přímo a snadno, vždy je zapotřebí kontroly a oprav, většinou bohužel i odborných (nejjednodušší je paradoxně způsob třetí, který jazykového odborníka průběžně nepotřebuje), každý z nich však navíc vyžaduje řadu větších či menších programátorských a odborných zásahů v podobě preeditace či posteditace, různých konverzí, sjednocení aj. (viz 4, korpusová data interní), často v podobě zvláštních dalších programů. Ať už je elektronický text pořízen tím či oním způsobem (převádí se napřed většinou do neutrálního mezinárodního ASCII formátu), má pak v zásadě trvalou, neomezenou platnost a lze ho opakovaně využít při různých dalších příležitostech a to ho dělá zvlášť cenným.

Doprovodnými faktory bývají i některé aspekty právní. Závažnou součástí sběru dat je jeho uvedení do souladu s copyrightem, autorskými právy patřícími vydavateli či vlastnímu autorovi. Pokud je pro účely korpusu, a tedy zpravidla jen pro účely nekomerční a výzkumné poskytnou, pak obvykle na základě právní smlouvy či úmluvy; ta může např. připouštět jen omezené přímé citace jednotlivých autorů apod. Případná pozdější komerčně využitelná aplikace pak musí jejich dodatečnou využitelnost, jakkoliv obvykle jen nepřímou, řešit však

právně zvlášť. V případě mluveného korpusu je často potřeba respektovat případné přání mluvčích zachovat jejich anonymitu apod.

3. Typy korpusů a standardizace

Každý soubor textů v počítači však ještě korpusem není. Od vlastního korpusu (srov. vymezení v 1.) je třeba především lišit volné kolekce textů, popř. elektronickou knihovnu ("textotéku") a souhrnný elektronický archiv, jakým je např. známý Text Archive Oxfordské university. Takovýto archiv je, podobně jako tradiční knihovna, prostou rozsáhlou sbírkou různých, většinou však literárních elektronických textů v různých formátech (Oxfordský archiv, který je dostupný a určený ke studiu, má asi tisícovku textů literárních děl v 25 jazycích v různých formátech a je tříděn hlavně podle vnější dostupnosti po síti a rozsahu textů; označkované jsou však jen některé z nich). Jiná velká podobná centra jsou např. na universitách v Torontu a Brigham Young.

Vlastní elektronické korpusy vykazují značnou různorodost, zčásti dnes už považovanou za nedostatek. Liší se mnoha parametry, zvl. však podle jazyka, typu textů, zaznamenané formy jazyka (proti textům mluveným je textů psaných většina), cíle a typu využití, způsobu uchovávání, formy uchovávání, popř. i doprovodného vybavení aj. Z hlediska pokrytých jazyků jde počet korpusů do desítek, v řadě z nich je však korpusů více zároveň. V Evropě je dnes už velmi málo jazyků, které nějaký korpus nemají (srov. mj. Taylor et al.); r. 1990) eviduje A. Zampolli rozsah pokrytých jazyků mj. takto: např. pro francouzštinu existovalo dohromady 190 milionů slov (Frantext), 27,5 mil. pro němčinu, 60 mil. pro holandštinu, 30 mil. pro italštinu, 12 mil. pro srbochorvatštinu aj. V Evropě a USA se ovšem zpracovávají i některé korpusy z dalších, popř. i mimoevropských jazyků (např. arménština v Leidenu, klasická řečtina v Irvinu aj.). Zdaleka

největší
pestrost i bohatství však představuje se svými více než 20 různými korpusy angličtina (viz např. Aijmer-Altenberg, 315n.); pro účely koordinace i přehledu tu vznikla mj. i organizace ICAME (International Computer Archive of Modern English) sídlící v norském Bergenu, která některé korpusy i distribuuje (informace po síti lze získat na adrese FILESERV@HD.UIB.NO). Zabývá se však především šířením obecných informací o zpracování korpusů vůbec (srov. její elektronický časopis CORPORA) a vedle zkušeností nabízí i některé nástroje. K hlavním korpusům angličtiny patří Brown Corpus (W.N. Francis a H. Kučera, americká angl., 1 mil.), LOB Corpus (=Lancaster-Oslo-Bergen, G. Leech, S. Johansson, K. Hofland, britská angl., 1 mil.), London-Lund Corpus (mluvená angl., J. Svartvik, 0,5 mil.), Helsinki Corpus (diachronní, M. Rissanen, O. Ihalainen, M. Kytö), Cobuild Corpus (J.M. Sinclair, dnes Databank of English, 160 mil.), British National Corpus (ve spolupráci oxfordské a lancasterské
university, nakladatelství Longman a Oxford a British Library, 100 mil.),
International Corpus of English (S. Greenbaum, 10 národních skupin a variet angličtiny z celého světa), Longman/Lancaster English Language Corpus (R. Quirk a G. Leech, 30 mil.), Survey of English Usage Corpus (napůl psaný a napůl mluvený, R. Quirk, S. Greenbaum, 1 mil.), Susanne Corpus (G. Sampson, 128 000 z Brown C. s relativně plně označkovanou i syntaxí) aj. Existuje i několik korpusů dvoujazyčných, popř. i vícejazyčných, např. mezi italštinou, resp. francouzštinou či dánštinou a angličtinou aj. Pro češtinu vznikl r. 1994 z inciativy interdisciplinární
skupiny Počítačového fondu češtiny a spojením sil více univerzitních pracovišť a ÚJČ AV ČR Ústav českého národního korpusu, jehož cílem je vybudovat rozsáhlý a víceúčelový korpus češtiny obecné povahy na půdě Filosofické fakultě University Karlovy.

Z hlediska typu textů se korpusy dělí především na obecné, resp. nespecifické a specializované. Existující pestrost zaměření specializovaných korpusů naznačuje jak pestrou paletu obecných možností jejich využití, tak individuální orientaci jednotlivých korpusů, které jsou zacílené např. na skotské drama, americkou povídku, dialekty, právní smlouvy a předpisy, naftařské texty, dětský jazyk, staré a první texty, jazyk novin, jazyk jednoho autora (např. korpus Thomase Manna) aj. I velké obecné korpusy mohou být složeny z více složek, subkorpusů, např. jazyka psaného-mluveného, synchronního-diachronního,
nespecifického-specifického (např. terminologického), obecného-nářečního apod., a to navíc ve více podobách (viz dál 4.).

Z hlediska typu uložení se korpusy dělí na ty, které existují v prosté podobě (t.j. ASCII formátu), nebo navíc v různém stupni i podobě označkované, popř. řídce i syntakticky analyzované; často existují i paralelně, obv. však jen zčásti, i v podobě doprovodných frekvenčních slovníků a konkordancí. Vlastní formou uchování je hard disk na komputerech různého typu, často zároveň i ve verzi na magnetických páscích, disketách či optických discích.

Protože počet korpusů i jejich rozsah rychle roste, je dnes už zřejmá jak potřeba standardizace sběru a označování textů, tak možnost jejich vícenásobného, sdíleného použití (reusability - znovupoužitelnost). Tomu prvnímu je věnována mezinárodní iniciativa TEI (Text Encoding Initiative), sponzorovaná mj. Evropskými

společenstvími a americkou vládou; TEI v několika dokumentech (zvl. Sperberg-McQueen et al. 1990, 1993, Hockey) doporučuje společný výměnný formát textů, zásad kódování nových a způsoby převodu mezi formáty existujícími. Její různé subkomise už mj. specifikovaly a doporučily i vhodné znakové sady, zásady textové analýzy v návaznosti na různé obory i kódovací metajazyk. Za ten byl pro deskriptivní rámec syntaktické analýzy zvolen SGML (Standard Generalized Markup Language, Bryan, Burnard), uznávaný od r. 1986 jako mezinárodní standard (ISO 8879). Opakovaná, obecná znovupoužitelnost textů (Hockey-Walker, Heid et al.), aktuální zvl. ve světle nákladů na pořízení a přípravu elektronických textů i jejich mezinárodní výměny, vyžaduje ke své realizaci vyřešení především otázek polyfunkčnosti korpusu, jeho polyteoretičnosti (tj. nepoplatnosti jedné úzké teorii), dostupnosti, intelektuálních vlastnických práv, reprezentativnosti, standardizace aj. Hlavním centrem mezinárodní inventarizace elektronických
netechnických textů je od r. 1991 CETH (Center for Electronic Texts in the Humanities), situovaný na universitách Rutgers a Princeton; novým evropským střediskem se však v tomto smyslu stává i Edinburgh. Otázkám a zkušenostem práce s korpusem se věnují především dva časopisy: Literary and Linguistic Computing a Computer and the Humanities,
srov. však i elektronický časopis CORPORA (viz výše).

4. Výstavba korpusu

Hlavní fáze výstavby korpusu tvoří (A) specifikace jeho projektu, tj. jeho typu především podle cíle a použitelnosti (viz zvl. zde 3.), (B) zajištění potřebného hardwaru a softwaru, (C) sběr dat a jejich označkování (viz zde 2. a dál), (D) zpracování korpusu a integrace jeho částí (textová a relační databáze, konkordance, frekvenční seznamy, lemmatizace aj.) a (E) zajištění jeho dalšího růstu a zpětných vazeb (srov. Atkins-Clear-Ostler).

Podle povahy korpusu jsou data v zásadě standardními vzorky nebo plnými texty, a to tak, aby se co nejvěrněji zachytila jak variabilita textů z hlediska jejich typů (variabilita lingvistická), tak rozsahu a složení jejich distribuce (variabilita situační). Jejich vlastní výběr se řídí podle pojetí, vymezení statistické populace, a to především kritérii (A) recepce či (B) percepce, a tedy podle toho, jak ho lidé užívají (recepce, tj. jazyka ve skutečnosti jen několika málo spisovatelů, novinářů apod. pro velmi široké publikum různých médií) či toho, jak ho píší a mluví (produkce, tj. jazyka velmi širokého vzorku aktivních uživatelů pro stejně široké spektrum příjemců). Protože jednostranná orientace na recepci (A) by znamenala zaměnit jazyk jen několika profesionálů, jakkoliv vlivný, za skutečně reprezentativní obraz celého spektra uživatelů a naopak orientace na produkci (B) by sice zachytila pestrost typů textů, ale za cenu záznamu i velmi řídkých, ne-li ezoterických případů, je třeba volit pro výběr textů percepci i recepci v určité proporci. Především tímto ohledem je dána zásadní otázka řešení nezbytné reprezentativnosti korpusu a jeho dat. Druhým ohledem při stanovení povahy reprezentativnosti
je to, zda zapojená kritéria určující typy sledovaných textů, tj. (C) textů jakožto produktů, jsou externí či interní (srov. též Biber 1993). Interní kritéria jsou kritéria lingvistická (ne/formálnost textu, lexikon/syntax aj.), kritéria externí jsou naopak nelingvistická,

nejazyková (týkají se typologie textů, tj. jejich původu, ne/připravenosti,
žánru, situace, odbornosti, času aj.); žádná přímá či jednoduchá souvislost mezi oběma typy není. Avšak základní orientace primárně či výlučně jen na toto kritérium (C) by sice zachytila textové typy a registry, ne však tolik typické vzorce úzu různých sociálních skupin.

Zjednodušený avšak vyčerpávající obraz populace, který lze užít jako rámec pro strategii projektu korpusu, nabízí D. Biber (Biber 1993); uvažuje o osmi hierarchizovaných situačních parametrech, použitelných hlavně pro stanovení povahy vzorku (z nichž hlavní jsou první tři):

1- primární kanál (jazyk psaný/mluvený/transkribovaný)
2- formát (ne/publikovaný, uvnitř dál dělený)
3- scéna (institucionální/jiná veřejná/soukromá či osobní)
4- adresát (a-pluralita: ne/vyčíslený/plurálový/ individuální/já sám;
b-přítomnost, tj. čas a místo: ne/přítomný;
c-interaktivnost: čádná/malá/rozsáhlá;
d-sdílená znalost: obecná/specializovaná/osobní)
5- adresor (a-demografická variace: pohlaví/věk/zaměstnání aj.;
b-uznání poplatnosti/díký: obecné/specializované/osobní)
6- faktuálnost (faktuální/informační/střední/ neurčitá/imaginární)
7- účel (přesvědčit/bavit/pozvnést/informovat/instruovat/vysvětlit/
vyprávět/popisovat/zazanamenat/přiznat

se/vyjádřit postoj, názor či emoci/posílit osobní vztah aj.)

Z hlediska zastoupení různých jazykových prvků, jevů a forem podle D. Bibera platí tyto souvislosti:

1- Běžné lineární jazykové jevy mají velmi stálou distribuci a lze je získat spolehlivě i z relativně krátkých segmentů textu (často už o 1000 slovech).
2- řídké lingvistické jevy mají velkou distribuční variabilitu a vyžadují delší vzorky.
3- Jevy s distribucí probability po křivce, tj. různé typy jevů (např. kumulativnost slovních druhů) jsou relativně stálé v různých segmentech textů, ale výskyt nových typů postupně klesá. Naopak frekvence nových typů je ve vzorcích z různých textů vyšší než v textu jediném (což je důvod pro stratifikované vzorky, tj. z různých vrstev definované populace).

Skutečné řešení reprezentativnosti velkých současných korpusů je ovšem různé, často z nedostupnosti některých objektivních kritérií. Takto např. Britský národní korpus (Summers) ji chápe poměrně široce a zdůrazňuje zaměření na typické a centrální jevy. Svou metodologii opírá o základní dělení textů na informativní a imaginativní (1-8 a 9-10 dole, v poměru 60 : 40 %), které dále dělí podle tématu na těchto deset superoblastí:

1-přírodní a čisté vědy (6 %),
2-aplikované vědy (4,3 %),
3-společenské vědy (14,1 %),
4-světové záležitosti (10,1 %),
5-obchod a finance (4,4 %),
6-umění (7,9 %),
7-víra a myšlení (4,7 %, tj. jak např. náboženství tak filozofie),
8-volný čas (5,7 %),
9-umělecká próza (40 %) a
10-poezie-drama-humor (2,3 %).

Naproti dánský korpus (Norling-Christensen) vychází z kombinací tří základních kritérií či parametrů textů (jejichž zdrojem byly z celé třetiny jen noviny a z druhé knihy): obecný-odborný (91 : 9 %), psaný-mluvený (84 : 16 %) a recepce-produkce (88,8 : 11,2 %), a to v těchto proporcích:

1-obecný-psaný-recepce (71%),
2-obecný-psaný-produkce (3,9 %)
3-obecný-mluvený-recepce (9,8 %)
4-obecný-mluvený-produkce (6,8 %)
5-odborný-psaný-recepce (7,7 %)
6-odborný-psaný-produkce (0,15 %)
7-odborný-mluvený-recepce (0,32 %)
8-odborný-mluvený-produkce (0,14 %)

Korpusová data (vnitřní), získaná z vnějších (viz 2.), musejí ještě před tím, než je lze použít, projít aspoň dvěma přípravnými fázemi, (1) čištěním a (2) standardizací, resp. unifikací (ne nutně v tomto pořadí). V první fázi jsou zbavena speciálních znaků editorů či jiných programů (zvl. sázecích), v nichž vznikla či kterými naposledy prošla; mohou však být podle potřeby zbavována i textových obrázků a grafiky, překlepů, tiskových chyb apod. Ve druhé se převádějí do zvoleného jednotného formátu (zvl. mezinárodního ASCII). Často však musejí data projít ještě třetí přípravnou fází spočívající,

podle záměru a potřeby, v jejich případném scelování do větších celků apod.

Nedílnou součástí této přípravné fáze zpracovaní jazykových dat v komputeru je jejich doprovodná archivní anotace, a to jak vnější, v podobě písemného záznamu do seznamu textů, tak vnitřní. Vnitřní anotace zachycuje, obecně řečeno, demografické aspekty textu a řídí se dnes při tom zpravidla mezinárodními standardy TEI (viz výše, srov. obecně Atkins-Clear-Ostler a Čermák).

Takto pak přístupná a strojově čitelná vnitřní data v samotném počítači jsou takového druhu a povahy, jakou jim tvůrci korpusu v závislosti na zamýšleném cíli tvorby a využití korpusu dodají. Jakkoliv je to taky možné, prakticky žádný korpus dnes nedává k dispozici jen data v podobě prostých lineárních textových řetězců; jejich využití by bylo omezené jen na studium poměrně zdlouhavé vyhledávaných jednotlivých tvarů slov a jejich sousedství. V souladu s potřebou poznat skrze textové výskyty obecnější vlastnosti jazykového systému se textovým datům tudíž dodávají na škále delinearizace různě složité a často i korelované indexy, které ji různě silně ruší. Touto delinearizací, resp. zachycením a značkováním, taggováním (angl. tagging) zvolených aspektů tohoto procesu, lze tedy obecně rozumět zpětný převod lineární konkrétnější, resp. individuální syntagmatické stránky a manifestace jazyka do obecné výchozí a v různém stupni abstraktní paradigmatické stránky a podoby. Stupňů takového značkování tedy může obecně být tolik, kolik je potřeba a kolik lze komputerově (programově) úspěšně zavést a uplatnit; zdaleka však nepokrývá všechno, co by lingvista rád měl k dispozici.

Nejsilněji syntagmatickou povahu, relativně nejbližší prosté textové podobě nepřipraveného textu, mají konkordance, t.j. obv. různě velké dílčí seznamy slovních forem v jejich přirozeném (co do rozsahu volitelném) kontextu, často s dodatečnou informací o místě výskytu v původním textu, frekvencí výskytu apod. Hlavní výhodou konkordance, užívané obv. v běžném standardu KWIC (Key Word in Context), je možnost studia slova (popř. jen jeho části), resp. všech jeho forem vedle sebe (obvykle se řadí abecedně), jejich kontextů a tím také různě pevných a habituálních kolokací, obecněji pak kolokability (spojitelnosti) slova, resp. jeho formy, a valence. Je to primární nástroj např. pro lexikografa, kterému nahrazuje kartotékovou dokumentaci výskytů, protože ten svou práci musí vždy začínat, především kvůli studiu významu slova a jeho odstínů, od úhrnu kontextů studované jednotky. Protože však jde u konkordance (podle zadání) o mechanické seřazení vždy stejně velkých úseků textu s daným výskytem každého slova (lze ovšem zadat i jejich určitá omezení a vynechávky), je jednak konkordance mnohonásobně větší než původní text a jednak nemusí vedle sebe uvádět všechny flektivní tvary slova, které k sobě patří, protože mezi nimi abecedně mohou figurovat slova jiná, která sem čistě abecedně taky patří (srovnej ukázku v příloze).

Pro běžnou lingvistickou práci se tu nejvíce a hojně osvědčil v prostředí DOSu komerčně šířený oxfordský program Micro-OCP obsahující vedle vlastní konkordance řadu dalších nástrojů, popř. WordCruncher z Brigham Young University nebo KAYE od G. Kaye vyvinutý pro firmu IBM, v prostředí MacIntoshe pak zvláště úspěšně Conc, který je volně šiřitelný (mj. prostřednictvím Consortium for Lexical Research v Novém Mexiku). Pro velké počítače (mainframe, ale i Unixovské pracovní stanice) byl vyvinut OCP (Oxford Concordance Programme),

jímž byl zpracován velký oxfordský slovník a z něhož byl pro potřeby PC pak vytvořen i zmíněný Micro-OCP. Pro svůj velký rozsah může být do formy konkordance převedena též jen určitá část jazykových dat korpusu, a to v zásadě buď účelově (se zadáním omezeného a dobře vymezeného cíle, popř. i možné selekce, např. při tvorbě slovníků), nebo obecně jako omezené referenční jádro korpusu pro základní (stručné) ověřování hlavních dat a jejich rysů.

Full-textová databáze (též jen textová databáze) leží v podstatě kdesi na půl cesty mezi syntagmatickou lineární podobou řetězců textových výskytů a jejich paradigmatickou podobou v jazykovém systému. Tuto její přechodnou povahu zabezpečuje kombinace lineární textové podoby korpusu a přidaného značkování (viz i výše), které k lineárním tvarům přiřazuje jejich kategorie a tedy i paradigmatické třídy; míra, podoba a forma těchto značek může být ovšem velmi různá. Práce s takovouto podobou korpusu, která je dnes obvykle i podobou pro celý korpus základní a nejúplnější, umožňuje díky speciálnímu uložení pomocí zvláštních indexů rychlé vyhledávání a vyvolávání (angl. data retrieval) zadaných potřebných dat v celém korpusu; lze tu výhodně vyhledávat i kombinace slovních tvarů oddělené i větším počtem jiných slov, jejich souhrný výpis, statistiky apod. Pro prostředí DOSu je nejznámější zmíněný WordCruncher, popř. různé další komerční nelingvistické full-textové programy; v prostředí UNIXu je zdaleka nejrozšířenější program PAT (srov. např. Salminen et al.).

Plnou korelaci se systémem a jeho kategoriemi tento typ programu v žádné podobě pochopitelně neumožňuje; záleží to na vneseném značkování a to zase na kvalitě výchozí teorie, kterou odráží. Hlavní potřebou lingvisty je totiž obvykle převést textové flektivní tvary a varianty pod neutrální slovníkové reprezentace, t.j. jejich lemmatizace. Lemmatizátor je tudíž takový program, který sám nebo v propojení s jiným programem (např. full-textovou databází) dokáže všechny tvary lexému svést dohromady pod společné lemma, např. nominativ či infinitiv (u českých sloves může jít o desítky až stovky tvarů k jedinému slovesu); vyvíjený český lemmatizátor může navíc k danému lexikálnímu základu, resp. kořenu dodávat i pravidelné deriváty z oblasti tvoření slov, tedy slovní čeledi. Žádný lemmatizátor však dosud není schopen lemmatizovat víceslovné tvary a jednotky, vždy se jeho možnosti omezují na diskrétní hranice tvaru jediného; zde tedy zůstává celé významné pole otevřené a dosud neřešené.

Třetí formou správy a zpracování korpusových dat je databáze, obv. relačního typu, kde se buď užívají individuálně konstruované databázové programy, které však nedošly většího rozšíření, anebo komerčně šířené úspěšné programy, jako je pro DOS Oracle či Fox-Pro apod. Databáze tohoto typu je strukturována a vytvořena podle potřeby, t.j. skutečných jednotek obvykle slovníkového typu, jejich částí, hierarchie a vnitřní souvztažnosti, které jsou všechny vzájemně propojeny a lze v nich hledat podobně jako ve slovníku, ale ovšem také podle jednotlivých polí, tedy např. všechna slova/lemmata spojitelná s akuzativem, či mající ve výkladu svého významu slovo nástroj či způsob nebo barva apod. Tato nejvýrazněji paradigmatická a nejabstraktnější forma korpusu bývá jeho integrální složkou zvláště ve dvou případech: když je součástí korpusu i (obvykle velký) slovník daného jazyka, který pak může sloužit např. jako filtr pro ověřování a kontrolu dat, anebo když je takový slovník naopak jedním z cílů, o jehož dosažení výstavba korpusu usiluje. V tomto druhém případě

je jako dodatečný nástroj nezbytný i lemmatizátor (o lexikální databázi srov. mj. Calzolari 1990, 1993).

V optimální podobě jsou všechny tři formy korpusu navzájem propojeny tak, aby se data z jedné části dala spojovat s daty z části jiné, zvl. za účelem cíleného výstupu či aplikace, např. při rešerši určitého typu či tvorbě slovníku, kdy je třeba spojovat data dřívější (např. z naskenovaného slovníku, uloženého v relační databázi) s novými (zvl. v podobě konkordance). Všechny tři formy či mody existence korpusu tudíž mj. závisejí na dobrém a rychlém vyhledávacím programu; většinou je přímo součástí základních databázových programů obou typů (viz výše), popř. i programu konkordančního.

Je pochopitelné, že ať už v podobě textové databáze či databáze relační, jsou v korpusu přístupné ty aspekty a aspekty jeho jednotek, do kterých se v podobě příslušného značkování dokázala uspokojivě promítnout ověřená a fungující lingvistická teorie, resp. její model. V tomto smyslu lze v korpusu vyznačovat relativně nejspolehlivěji jasné diskrétní jednotky formy (viz však neřešený problém víceslovnosti, ať třeba některých slovesných tvarů nebo frazémů), a tedy jevy v zásadě morfologické v užším i širším smyslu a z hlediska formálního tedy i jednoslovné jevy lexikální. Vedle nejběžnějšího značkování morfologického
(zahrnujícího určení slovních druhů a různého počtu jejich kategorií) je však na kvalitě předchozí teorie neméně závislý návrh struktury databázového hesla, analogický v tomto smyslu značkování morfologickému. Vzhledem k nejednoznačné povaze řady aspektů je žádoucí, aby obojí značkování na sebe komplementárně navazovalo; doporučuje se dokonce, aby při značkování často existovalo i řešení paralelní, dvojí (McNaught).

Zachycení syntaktických aspektů, vztahů a útvarů záleží na úspěšnosti učitého parseru (syntaktického analyzátoru) a kritérií a množství značkování (pozoruhodný je např. přístup uplatněný v korpusu Susanne, srov. Garside). Zpravidla však je tu dosud mnoho nevyřešené nejednoznačnosti,
ani se tu nepřekročují přitom hranice věty. Samozřejmou možností je ovšem i analýza fonologická (srov. Leech); ta a analýza fonetická, popř. prozodická se ovšem týká korpusu mluveného jazyka. Pro jazyky s přirozeně se vyvíjející, kodifikačně neochromenou formou a tudíž i přirozenou variabilitou se ovšem nabízí i analýza ortografická. Každá další analýza, zvl. analýza významu a většiny oblasti funkce včetně aspektů pragmatických zůstává, přes nejrůznější pokusy o její částečné uchopení (Patten), mimo dosavadní možnosti; výjimkou je nabízející se možnost sémantické analýzy založené na metajazyku slovníku uloženého v databázi (Alshawi; srov. však i projekt automatické obsahové analýzy, Wilson-Rayson).

Vedle lemmatizátoru, parseru a dalších nástrojů je pro práci s korpusem, především v jeho základní podobě textové databáze zapotřebí mít k dispozici i vhodné softwarové nástroje (jako TACT, LEXA, PAT, Corpus-Bench aj.), které budou schopné splňovat aspoň tyto požadavky: rychlé a interaktivní ovládání, spolehlivé vyhledávání zjišťovaných forem i v různě modifikovatelných kombinacích, jejich různé statistické vyhodnocování včetně zjišťování frekvence, které v pozdější fázi umožní i statistické odlišování různých významů a jejich odstínů aj. (srov. Gale et al., Čermák).

První zkušenosti s korpusy v různých jazycích přinesly už i

některé zásadní zkušenosti metodologické povahy. Jednak je zřejmé, že analýza většiny sémantické stránky jazyka (srov. mj. Atkins 1987, Pustejovski, Introduction, Kay), která je na rozdíl od diskrétní formy (tu lze opřít programově o binární volbu typu "ano-ne") spíše většinou kontinuální a škálové, resp. splývavé povahy (a tedy v závislosti
především na volbě typu "spíše toto než to, popř. ono"), bude mít jinou povahu (M.A.K. Halliday: jazykový systém je inherentně probabilistický, jeho kontinuu s komplementárními perspektivami gramatiky a lexikonu lépe vyhovuje koncepce lexikogramatiky). Významný přístupem, založeným na předpokladu různé statistické pravděpodobnosti výskytu různých jevů formy, je vyhodnocování těchto aspektů na základě probabilistických odhadů, měření a různých indexů) (o vztahu kvantitativních a kvalitativních aspektů viz mj. Itkonen). Vždy však ke studiu této stránky bude možné přejít pouze skrze zřetelně a spolehlivě okódovanou stránku formální; jedno tu tudíž předpokládá druhé. I ve formální stránce jazykových dat lze však pozorovat, resp. předpokládat nejednoduchost a nejednoznačnost, především ve smyslu časté variabilnosti formy. Na druhé straně se jako odraz určité skepse v sílu jednotlivých jazykových teorií také doporučuje (srov. Leech 1993), aby značkování bylo spíše jednodušší, široké a konsensuální (viz dál) a nevycházelo z jedné konkrétní teorie, protože se později nemusí osvědčit a označkovaná data by nebyla jinak použitelná i v přístupech dalších. Nejen v této souvislosti pak nabývají na významu stále častěji produkované frekvenční seznamy, resp. slovníky jak tvarů tak lemmat, dílčí či obecnější, doprovázející vznik a rozvoj korpusů. Slouží mj. především jako neocenitelný referenční zdroj ve všech otázkách, kde selhává formální gramatika i intuice.

G. Leech (1993) shrnuje zkušenosti své i mnohých jiných s anotací a značkováním do sedmi zásad. Podle něj anotace má být

(1) postradatelná, t.j. vždy musí být možné se vrátit k původnímu syrovému korpusu,
(2) extrahovatelná (zvl. z textové databáze) a uložitelná zvlášť,
(3) opřená o zásady přístupné, srozumitelné koncovému uživateli, a ne pouze lingvistovi,
(4) autorsky vysledovatelná ke svému tvůrci (anotátorovi),
(5) jen pohodlnou pomůckou ("device of convenience"), a nemá se tedy vydávat za zjevenou pravdu; uživatel má být varován a poučen, že je na něm, zda ji přijme či ne,
(6) založena na konsensu většiny teorií a teoreticky tedy co nejneutrálnější,
(7) ne autoritativní, ale být slučitelná se standardy jinými.

Toto je třeba chápat zvl. v kontextu situace, kdy většina existujících korpusů má jen jedinou značku (tag, srov. Johansson 1991), což je spíše výhoda než nevýhoda.

Zabezpečení dalšího růstu korpusu přihlíží především k potřebě udržet korpus vyrovnaný a reprezentativní, zvl. metodami postupných, cyklických aproximací založených stejně na potřebách jako na kladech a záporech zjištěných z analýzy materiálu už dostupného. V dalších krocích může tudíž jít jak o jeho kontrolované obohacování tak případné vypouštění některých dat. Jednou z běžných zkušeností (srov. např. Summers), která k takové korekci vede, je např. to, že v korpusu záhy převažují slova užívaná "tvůrčím" způsobem, a

tedy jen okrajové důležitosti a chybějí naopak ta nejobyčejnější. Důležitost má taková zpětná vazba k jeho uživatelům, která umožní správci korpusu reagovat v kontaktu s nimi při další výstavbě korpusu na jejich zkušenosti, poznámky, varování apod. Nemalým korektivem i stimulem je ovšem i rostoucí mezinárodní spolupráce, standardizace a integrace jednotlivých národních korpusů do propojené sítě umožňující mj. i užitečnou výměnu a srovnávání dat.

5. Práce s korpusem a jeho využití

Mluví-li jeden z dokumentů Evropských společenství v r. 1991 (Commission... 1991, 20) o tom, že "Technologie mluvy a jazyka vyžadují rozsáhlé databázové korpusy... pro výzkum a rozvoj, účely testování a k podpoře spisovatelů a překladatelů" a odpovídá tak na otázku Proč korpus?, pak o dva roky později člen téže komise ES DG XIII, J. Soler (Soler 1993) si už tuto otázku vůbec neklade a uvažuje spíše o způsobech jeho využití:"...rozvoj standardizovaných korpusů a metod a nástrojů jejich správy i aplikace je dlouhodobý podnik přesahující možnosti projektu individuálního. Nesmírnost úkolu, jehož má být dosaženo, i jeho náklady naznačují, že standardizované korpusy mají být budovány spoluprací skrze evropskou koordinaci národních snah, která je otevřená mezinárodním výměnám, i to, že výsledné korpusy mají být veřejné a orientované tak, aby uspokojovaly různé potřeby uživatelů."

Základní hrubé lišení uživatelů korpusu je pochopitelně na (A) lingvisty a (B) nelingvisty, a rozpadá se dál do řady orientací a oblastí. Proti množství potřeb a specifických softwarových nástrojů, umožňující v případě první skupiny (A) elicitovat vzorce, struktury, schémata, kombinace a jejich typy, stojí v druhém případě (B) výsledky a poznatky zpravidla jen statisticky a probabilisticky zjišťované. Přes různost korpusů lze i z hlediska cílů jejich využití rozlišit především dva hlavní (Atkins-Clear-Ostler): (a) jako extenzivního zdroje dat, z něhož je možné vybírat, co je potřebné a (b) jako prostředí pro testování, trénování a vylepšování automatizovaných (lingvistických) nástrojů různého druhu.

Lingvistické využití korpusů pochopitelně závisí na tom, s jakým cílem byly vybudovány; proti specificky orientovaným stojí korpus relativně obecný a vícefunkční, který však sám může být složen z různých homogenních vrstev, resp. subjazyků (McNaught) vhodných pro specifičtější cíle. Podle stupně označkování a analýzy korpusu, u které kvůli splývavé povaze dat zkušenosti (McNaught, Leech 1993 aj.) stále více mluví jen pro obecnou skeletonovou podobu, lze základní práci s textovým korpusem vidět v pěti fázích:

1-identifikace tvarů v textu,
2-zjištění distribuce tvarů a jejich kombinací s cílem odhalit syntaktické a sémantické třídy a jejich kombinace, včetně kombinací pevných,
3-zjištění, jak tyto sémantické třídy a jejich kombinace tvoří vyšší sémantické celky a struktury,
4-zjištění, jak se tyto vyšší struktury kombinují v základní textové jednotky,
5-zjištění, jak se získané výsledky promítají/mapují do struktur jiného jazyka.

Je zřejmé, že jen skrze korpus půjde
(1) - vzhledem k tomu, že tradiční popisy leccos vynechávají - poprvé v historii o možnost relativně úplného popisu jazyka,
(2) o precizaci, resp. redistribuci hranic a podstaty mnohých tradičních jazykových kategorií a jevů (popř. testování dosavadních gramatik),
(3) o první popis jevů, pro které dostatečná data dosud nebyla k dispozici a v neposlední řadě
(4) i o reálnou šanci objevu jevů a souvislostí zcela nových.

Obecně bude pozornost věnovaná většině těchto oblastí i specifických jevů znamenat i specifickou renesanci zájmu o statistické aspekty jazyka (Baayen, Köhlert et al.), a to nejen v jevech paradigmatických ale i syntagmatických, zvl. v oblasti kolokability jazykových forem (lexémů, Church et al.). Jednou z hlavních metodologických otázek, kterou bude třeba tak či onak odpovědět (srov. Leech 1991), je to, zda bude možné budoucí analýzu jazyka na korpusu opřít už o indukované a automatizované procedury objevování (discovery procedures), či zda lingvista bude i nadále muset zůstat u své intuice a lingvistické distinkce do textu vnášet. Menší míra úspěšnosti dosavadních kognitivních přístupů ukazuje spíše na nutnou interakci komputeru a člověka, která je založená na přístupech s nižší mírou výlučnosti (zvl. na gramatikách konečných stavů), doplňovaných probabilitami pro další měření přechodů mezi stavy, kategoriemi aj. Přirozeným důsledkem takového přístupu je sebeorganizující metodologie příslužných programů, které takto učí a zdokonalují samy sebe a jedním z hlavních požadavků, na ně kladených, je pak i schopnost indukovat datové struktury v textech do značné míry samostatně. V jistém protikladu,
ukazujícím na pestrost přístupů i možností práce s korpusem, stojí naopak pokusy o generování textů na jeho základě (Bateman).

Přestože korpus je pro lingvisty všeho druhu obecným a základním zdrojem, popř. testovacím prostředím (viz výše), lze s ohledem na jejich primární orientaci na něj rozlišit především následující typy:

- lexikografové/lexikologové (zdroj informací o skutečném úzu obecně či specifických slov apod., srov.např. Atkins 1991, 1992, Atkins-Zampolli, Boguraev et al., Čermák, Fillmore et al., Karlsson, Kiefer et al., Meijs 1992),
- komputační lingvisté (zdroj zjišťovaných statistických pravděpodobností jako klíč k analýze, prostředí k aplikaci teorií a modelů jazyka),
- teoretičtí lingvisté (zdroj vzorků jazykových jevů i prostředí ověřování svých domněnek apod.),
- úzcí specialisté (zdroj specifických informací, paralelních řešení i úzu ap. pro překladatele, terminology, dialektology aj., srov. Lewis, Meijs),
- aplikovaní lingvisté (autoritativní a typický zdroj dat, zvl. pro výuku, tvorbu jazykových pomůcek, srov. např. Last, Pennington).

Mimolingvistické využití korpusu se nabízí vlastně všem oblastem a disciplínám, které pracují s jazykem, a to hlavně jako referenční zdroj informací o distribuci zjišťovaného jevu. Především tu jde však o specialisty různých oborů zaměřené na obsah textů (historikové, literární kritikové, tvůrčí autoři, sociologové, psychologové, srov. Bateman-Hovy, Burrows, Williams) či jejich formu (specialisté na média včetně např. reklamy, stejně jako právníci apod.). V řadě oblastí se však oba aspekty překrývají (právní normy), v jiných

se doceňují i souvislosti, které nejsou zřejmé na první pohled (studiu korpusu z hlediska komunikace věnují pozornost např. i projekty americké armády). Silně se rozvíjejícími oblastmi specifických aplikací pro různé obory jsou nyní systémy získávání informací (information retrieval systems) a expertní systémy, strojový překlad (založený na statistických systémech a paralelních korpusech dvou/více jazyků) a zpracování mluvy včetně její syntézy. O dalších nezřejmých možnostech využití korpusů svědčí cíle některých menších korpusů (srov. zvl. Taylor-Leech-Fligelstone), pro které byly vytvořeny: výzkum školní četby, řešení otázek psycholingvistiky či zjišťování sporného autorství.

Za specifický cíl studia jazyka na korpusu se však často považuje studium jazykové variace. Obecné možnosti z hlediska typu jazykové formy nastiňuje následující přehled; naznačeny jsou v něm i hrubé kvantitativní poměry, resp. množství dat, kterých se variace týká.

FORMY:		Kvantitativně
A invariabilní		všechny
B variabilní		
a-diachronně	(dublety..)	výjimky?
b-pozičně synchronní	(var. prep)	málo
C variabilní synchronní	(morfologie)	většina
1 nominativně	(synonyma)	hlavně autosémantika,
zvl. S/A a-substituce		hlavně autosémantika,

	zčásti pron	
	b-víceslovnost	idiomy a frazémy
	c-smíšená	(a:b)
2	gramaticky	
	a-částečná obměna	
		S
		A
		V
		ADV zčásti
		pron
		num
	b-víceslovnost	V (čas/modus/reflexivita)
		S/A reflexivní
	c-kontrakce	V víceslovná (2b)
	d-smíšení aj.	?

Vedle studia jazykové variace formy se však stále více dostává do popředí potřeba zmapovat především hlavní oblasti a aspekty sémantiky jazyka, jejich distribuci, vzájemnou souvztažnost apod. Elementární situaci cílů tu lze zachytit např. takto:

VÝZNAMY/SÉMY aj.:	A-systém	a-sémy/sémantické komponenty
		b-hyponymie/taxonomie/tezaurus
	B-text	a-rámce/scénáře/témata (typická)
		b-pragmatické funkce

Bibliografie

-Aarts J., Meijs W., eds., 1990, Theory and Practice in Corpus Linguistics.
Rodopi Amsterdam

-Aijmer K., Altenberg B., eds., 1991, English Corpus Linguistics.
Studies in Honour of Jan Svartvik, Longman London

-Alshawi H., 1989, Analysing the Dictionary Definitions.
In Boguraev et al. 153- 170

-Atkins B.T.S., 1987, Semantic ID-tags: corpus evidence for dictionary senses.
The Uses of Large Text Databases: Proceedings of 3rd Annual Conference of the UW Centre for the New Oxford English Dictionary. University of Waterloo Waterloo

-Atkins S.T.S, 1991, Corpus lexicography: The Bilingual Dimension.
In Computational Lexicology and Lexicography. Vol. I. Guardini Pisa, 43-64

-Atkins Sue, Clear J., Ostler N., 1992, Corpus Design Criteria. LLC, Vol. 7, No. 1, 1-16

-Atkins B.T.S., 1992, Tools for computer-aided corpus lexicography:
the Hector Project. In Kiefer et al., 1-59

-Atkins B.T.S, Zampolli A., eds., 1994, Computational Approaches to the Lexicon. Clarendon Press Oxford (=5. Pisa International Summer School on Computational Lexicology and Lexicography)

-Baayen H., 1992, Statistical Models for Word Frequency Distributions:
A Linguistic Evaluation. Computers and the Humanities 26, 347-363

-Baker M., Francis G., Tognini-Bonelli E., eds., 1993, Text and Technology.

In Honour of John Sinclair. J. Benjamins Amsterdam
-Bateman J.A., E.H.Hovy, 1992, Computers and Text generation: Principles and Uses. In Butler, 53-74
-Biber D., 1989, A Typology of English Texts. Linguistics 27
-Biber D., 1993, Representativeness in Corpus Design. LLC O:4, 243-257
-Boguraev B., Briscoe T., 1989, Computational Lexicography for Natural Language Processing. Longman London New York
-Briscoe T., 1991, Lexical Issues in Natural Language Processing. In Klein E., F. Veltman, eds., 1991, Natural Language and Speech, Springer-Verlag Berlin, 39-68
-British National Corpus. Written Corpus Design Specification, 1991 (informační materiál)
-Brunet É., ed., 1986, Méthodes quantitatives et informatiques dans l'étude des textes (hommage a Charles Mueller). Colloque international de CNRS. Université de Nice. Slatkine-Champion Paris
-Bryan M, 1988, SGML: An Author's Guide to the Standard Generalized Markup Language. Addison-Wesley, Wokingham (England), Reading (Mass., USA)
-Burnard L., 1991, What is SGML and How Does it Help? TEI DOcument TEI ED W25. TEI fileserver tei-l@uicvm
-Burrows J.F., 1992, Computers and the Study of Literature, in Butler, 167-204
-Butler C.S., ed., 1992, Computers and Written Texts. B. Blackwell Oxford
-Calzolari N., 1990b, Structure and Access in an Automated Lexicon and Related Issues. In Calzolari 1993a, 139-161
-Calzolari N., 1990, Lexical Databases and Textual Corpora: Perspectives of Integration for a Lexical Knowledge-Base. In Zernik U., ed., Lexical Acquisition: Using On-line Resources to Build a Lexicon. Lawrence Erlbaum Hillsdale New Jersey
-Calzolari N., 1993a, ed., Fifth European Summer School in Logic, Language and Information Course. Computational Lexicons. Reader. Faculdade de lettras universidade de Lisboa Portugal.
-Calzolari N., 1993, Detecting Patterns in a Lexical Database. In Calzolari 1993, 170-173
-Calzolari N., T. Briscoe, 1992, ACQUILEX-I and -II. Acquisition of Lexical Knowledge from Machine-Readable Dictionaries and Text Corpora. In Calzolari 1992a, 1-17
-Church K.W., Hanks P., 1990, Word Association Norms, Mutual Information and Lexicography. Computational Linguistics 16/1
-Commission of the European Communities, 1991, Language and Technology: Preliminary Consultations with Industry and User Organisations, Vol. 1 DGXIII-B, CEC, Luxembourg
-Corpusgebaseerde Woordanalyse. Jaarboek 1986-1992. Vrije Universiteit Faculteit der Letteren. Vakgroep Taalkunde Amsterdam
-Crowdy S., 1991, Spoken Corpus Design and Transcription. 1991 (dokument) Longman Dictionaries
-Crowdy S., 1993, Spoken Corpus Design, LLC 8:4, 259-265
-Čermák F., 1994?, Komputační lexikografie. In Manuál lexikografie, eds. F. Čermák, R. Blatná. H+H Praha
-Čermák F., Králík J., Pala K., 1992, Počítačová lexikografie a čeština. SaS 53, 41-48
-Evens W., 1988, Relational Models of the Lexicon. Cambridge U.P. Cambridge
-Fillmore C.J., B.T.S Atkins, 1994, Starting where the dictionaries stop: the challenge of corpus lexicography. In Atkins B.T.S., Zampolli A., eds., Computational Approaches to the Lexicon
-Gale W.A., K.W. Church, D. Yarowsky, 1992, A Method for Disambiguating

Word Senses in a Large Corpus. Computers and the Humanities 26, 415-439
-Garside R.G., G. Leech, G. Sampson, 1987, A Computational Analysis of English. Longman London
-Garside R., 1993, The Large-Scale Production of Syntactically Analyzed Corpora, LLC 8:4, 39-45
-Gunton T., 1992, The Penguin Dictionary of Information Technology and Computer Science. Penguin Books Harmondsworth
-Halliday M.A.K., 1991, Corpus studies and probabilistic grammar. In Aijmer et al. 30-43
-Heid U., M. Heyn, O. Christ, 1992, Extracting Linguistic Information from Machine-Readable Versions of Traditional Dictionaries: a Metalexicographic Method and Some Tools. In Kiefer et al., 161-174
-Hockey S., 1991, The ACH-ACL-ALLC Text Encoding Initiative: An Overview. TEI Document TEI J16. TEI fileserver tei-l@uicvm
-Hockey S., D. Walker, 1993, Developing Effective Resources for Research on Texts: Collecting Texts, Tagging Texts, Cataloguing Texts, Using Texts, and Putting Texts in Context. LLC 8:4, 235-242
-ICAME Collection of English Language Corpora (CD-ROM), 1991 (materiál)
-Ide N., 1992, Introduction: Common Methodologies in Humanities, Computing and Computational Linguistics. Computers and the Humanities 26, 327-330
-Illingworth V., ed., 1991, Dictionary of Computing. Oxford Oxford U.P., 3.ed.
-Introduction to the Cambridge Language Survey Semantic Coding Project, 1994 (dokument)
-Itkonen E., 1980, Qualitative vs quantitative analysis in linguistics. In Perry T., ed., Evidence and Argumentation in Linguistics. de Gruyter Berlin
-Johansson S., Atwell E., Garside R., Leech G., 1986, The Tagged LOB Corpus. Users' Manual. Norwegian Computing Centre for the Humanities. Bergen
-Johansson S., K. Hofland, 1989, Frequency Analysis of English Vocabulary and Grammar 1-2. Clarendon P. Oxford
-Johansson S., 1991, Times change, and so do corpora. In Aijmer et al. 305-314
-Johansson S., Stenström A.-B., 1991, English Computer Corpora: Selected Papers and Research Guide. Mouton de Gruyter Berlin
-Karlsson F., 1992, Lexicography and Corpus Linguistics. Opening Address at 5th Congress of Euralex. Tampere
-Kay C.J., T.J.P. Chase, 1987, Conastructing a Thesaurus Database, LLC 2, 161-163
-Kaye G., 1989, KAYE. The KWIC Analyser. IBM UK Scientific Centre Winchester
-Kiefer F., G. Kiss, J. Pajzs, eds., 1992, Papers in Computational Lexicography COMPLEX '92. Linguistics Institute, Hungarian Academy of Sciences Budapest
-Köhler R., Rieger B.B., eds., 1993, Contributions to Quantitative Linguistics. Proceedings of the First International Conference on Quantitative Linguistics. Kluwer Dordrecht
-Kučera H., W.N. Francis, 1967, Computational Analysis of Present-Day English. Brown U. P. Providence, Rhode Island
-Last R., 1992, Computers and Language Learning: Past, Present - and Future? In Butler 227-247
-Leech G., 1991, The State of the Art in Corpus Linguistics. In Aijmer-Altenberg, 8-29
-Leech G., S. Fligelstone, 1992, Computers and Corpus Analysis, in Butler, 115-140

-Leech G., 1993, Corpus Annotation Schemes. LLC 8:4, 275-281
-Lewis D., 1992, Computers and Translation, in Butler 1992, 75-114
-McNaught J., 1993, User Needs for Textual Corpora in Natural Language Processing. LLC 8:4, 227-234
-Meijs W., ed., 1987, Corpus Linguistics and Beyond. Rodopi Amsterdam
-Meijs W., 1992, Computers and Dictionaries, in Butler, 141-166
-Micro-OCP. User Manual, 1988, Oxford University Computing Service. Oxford, University Press Oxford
-Norling-Christensen O., 1992, Preparing a Text Corpus. Computational Tools and Methods for Standardizing, Tagging and Structuring Text Data.
In Kiefer et al., 251-259
-Patten T., 1992, Computers and natural Language Parsing. In Butler, 29-52
-Pennington M., Stevens V., eds., (in press), Computers in Applied Linguistics: an International Perspective. Multilingual Matters, Clevedon, Avon
-Procter P., The Cambridge Language Survey (nedatovaný materiál)
-Pustejovski J., 1993, Semantics and the Lexicon. Kluwer Dordrecht
-Rissanen M., 1989, Three problems connected with the use of diachronic corpora. Journal of ICAME 13, 16-19
-Salminen A., F. W.M. Tompa, 1992, PAT expressions: an algebra for text search. In Kiefer et al., 309-331
-Sampson G., 1993, The Need for Grammatical Stocktaking, LLC 8:4, 267-273
-Sinclair J.M., 1987, ed., Looking Up: An Account of the COBUILD Project in Lexical Computing. Collins Glasgow
-Sinclair J.M., 1991, Corpus Concordance Collocation. Oxford U.P. Oxford
-Smith M.W.A., 1987, Hapax Legomena in Prescribed Positions: An Investigation of Recent Proposals to Resolve Problems of Authorship, LLC 2:3, 145-152
-Soler J., 1993, Text Corpora: Meeting the Challenge of Information Excess, LLC 8:4, 1
-Souter C., Atwell E., eds., 1993, Corpus-Based Computational Linguistics.
Rodopi Amsterdam
-Sperberg-McQueen C.M., L. Burnard, eds., 1990, Guidelines for the Encoding and Interchange of Machine-Readable Texts, draft ver. 1.0, Association for Computational Linguistics/Association for Computers and the Humanities/Association for Literary and Linguistic Computing, Chicago and Oxford
-Sperberg-McQueen C.M., L. Burnard, eds., 1993, Guidelines for the Encoding and Interchange of Machine-Readable Texts, draft ver. 3, TEI DOcument P3, ACH-ACL- ALLC. Chicago, Illinois and Oxford
-Summers D., 1991, Longman/Lancaster English Language Corpus. Criteria and Design (dokument)
-Svartvik J., 1990, The London-Lund Corpus of Spoken English: Description and Research. Lund Studies in English 82. Lund Lund University Press
-Svartvik J., 1992, Lexis in English Language Corpora. In Euralex '92 Proceedings I, 17-31
-Svartvik J., ed., 1992, Directions in Corpus Linguistics. Proceedings of the Nobel Symposium 82, Stockholm 4-8 August 1991, Mouton De Gruyter The Hague Berlin
-Taylor L., Leech G., Fligelstone S., 1989, Lancaster Preliminary Survey of Machine-Readable Language Corpora. (materiál)
-Thomson N., 1989, How to Read Articles which Depend on Statistics, LLC 4:1, 6-11
-Walker D., A. Zampolli, eds., 1994, Automating the Lexicon. Research

and Practice in a Multicultural Environment. Clarendon Press Oxford
-Warwick S., J. Hajič, G. Russell, 1990, Searching on Tagged Corpora: Linguistically Motivated Concordance Analysis. In Electronic Text Research. Proceedings of the Sixth Annual Conference of the Centre for the New OED. University of Waterloo, Waterloo, 10-18
-Williams N., 1992, Computers and Writing, in Butler, 247-265
-Wilson A., Rayson P., 1993, The Automatic Content Analysis of Spoken Discourse: A Report on Work in Progress. In C. Souter, E. Atwell, eds., Corpus-Based Computational Linguistics. Rodopi Amsterdam, 215-226
-WordCruncher (IndexETC, ViewETC). Text Indexing and Retrieval Software, 1987, Electronic Text Corporation. Brigham Young University Provo
-Zampolli A., 1990, A Survey of European Corpus Resources. In SALT. Proceedings of a Workshop on Corpus Resources. LOndon DTI/Speech and Language technology Club, 64-84

-ICAME Journal. Bergen
-Journal of Literary and Linguistic Computing (LLC)
-Computational Linguistics (CL)
-La Banque des mots (zvl. numéros speciaux 1988, 1989, 1990, 1991). CNRS-INaLF, Conseil international de la langue francaise

PŘÍLOHA

Ukázka malé konkordance slov DNES, JAK a MOC z jednoho týdne novin (LN, květen 1991), vytvořené pomocí Micro-OCP. Celý text obsahuje 20 964 textových slov (tvarů) a 8957 lemmat (slovníkových hesel) ilustrující jejich úzus.

```
                          dnes 22
      jejichž dozvuky ještě dnes prolínají čas od času na stránky
                 Když jsem dnes otevřel Lidové noviny z 21.května
    JEDNÁNÍ O LUSTRACÍCH AŽ DNES o FIS: PŘERUŠENO
 a pořad pléna má přijít až dnes
          Cena, kterou jsem dnes poctíván, je udělována spíš
dém, ale poklidném venkově, dnes jich tam zbývá
stolků pár metrů od radnice. Dnes soukromá
                  jak se má dnes. Na ztracené vartě
 potrefená husa, tehdy jako dnes, na tuto
            do Bruselu, kde dnes podepíše Dohodu o půjčce mezi Evropským
kého rockového podzemí řeší dnes úplně jiné problémy
Lobkowicz, 35letý rodák ze Švýcarska, dnes
      BŘECLAV ZH Slyšíte-li dnes o záplavách v lužním lese
     1280        Bohužel se dnes lužní les zavlažuje jen na malém úseku
  regulaci Moravy a Dyje a dnes za tento hřích pyká. Lužní les
  Rozptylové podmínky budou dnes dobré, v severočeské pánvi
                          Dnes se prezident se svým doprovodem vrací do
                       na dnes večer 19.00 do Janáčkovy síně v Praze 1,
```

rican film, jenž bude mít dnes prémiéru
KVAČKOVÁ TANKOVÝ PRAPOR: DNES SVĚTOVÁ, ZÍTRA ČS. PREMIÉRA
Půjde-li dnes a ono půjde, neboť režii nemá stát
anglistická veřejnost, ho dnes mohou poznat i

jak 39
Jako Čecha mě přirozeně zajímá, jak se s podobnými
chybnými čísly. Jak však sám uznává, přesné údaje o
škodě
investice v Československu, jak pan
abídnout jiným novinám. Tak jak
rozcházejí. A jak je to v demokracii vlastně možné, že
zpravodajských prostředků. Jak zástupkyně předkladatelů
postavení FIS je nezbytná. Jak konstatoval federální ministr
fenomén moci, jak jsem ho zatím tak říkajíc zevnitř
yto tři druhy důvodů se vždycky, jak jsem si všiml
a dokonce jako svého druhu objev. Jak tak ale
ám. Je velmi zajímavé pozorovat, jak
enka či máslo, jak se vaří káva, jak se řídí
auto a jak se telefonuje. Ocitám se tedy
vrcholů. Naopak, jak působivá byla tichá ševelení, jemné
smiling. Jak ostatní, nevím
úvodu přednášeli o tom, jak se v Anglii, kde Eduard
bezvýchodné situace, k níž došlo jak pod vlivem hudební
jak už to v dnešní hudbě bývá zvykem,
jak bychom program mohli nazvat,
y Liberálně demokratické strany. Jak nám sdělil na
Jak řekl německý ministr obrany
liberalizace. Je však třeba, jak pravili přítomní znalci
praktické ukázce, jak taková záměna obsahu vypadá
lovely, i kdyby ševci padali. A jak jsou
metropoli relativní klid. Jak ČTK telefonicky sdělil zástupce
Na otázku LN, jak se cítí v roli ekonoma, do níž je
využít lukrativnějším způsobem. Jak si pomoci tady
tart mi vnukl spíše představu, jak Šimon a Matouš cestou
bojkotovala jedání o tom, jak vyplnit nynější mocenské
vé akce a s tím, co konkrétně a jak budeme dělat
Příště si probereme, jak budou navazovat jednotlivá
Ano, poprvé to bylo před více jak dvaceti lety. Svaz
z vás optimismus. Jak lze tyto dva postoje spojit
e v Praze pouze několik dní, jak na Vás
Je z ní patrno, jak obrovský obchod, s ostatními
pomněl dvţ různé cesty republik, jak se
Na otázku, jak chtějí republiky přijít k penězům,
ituaci, kdy mnohé naznačuje, že moc politického útlaku je
existenci moci
váze, neţ jaký nabízí politická moc? Vždyť ze samé své
o moc jako takovou, ale pouze o určité obecné hodnoty, a ţe
touha po výhodách, které moc přináší, anebo prostě jen
dát u těch nás, kteří žádnou moc nikdy neměli a vždycky
: na jedné straně dává politická moc člověku
ţlého rodáka parafrázovat a říct moc k smrti
si své slávy vskutku moc
Pet Shop Boys. Moc živé muziky jsme při něm neslyšeli,
ne, jestliže KSČ u nás převezme moc. Jakoby mi

ovšem si to nebudou moci dovolit, ve vlastní zemi budou
fenomén moci, jak jsem ho zatím tak říkajíc
zevnitř
tně lidé touží po politické moci a proč se této
moci když ji mají tak neradi vzdávají
politické moci a proč se jí tak nerado vzdává, je
pestrá
ţit z výhod, které z politické moci
důvodů touhy po politické moci, o níž jsem hovořil, totiž
ţábelské je pokušení moci právě v této sféře. Nejlépe to lze
jsme se náhle sami ocitli u moci
tupu od sebe sama, aby člověk u moci, byť to
Tedy znovu: jsa u moci, jsem si permanentně podezřelý
svůj zápas s pokušeními moci zvolna začínají prohrávat a
moci a ve všem, co k ní logicky patří,
budou moci v historickém centru Prahy, ale
výkonné moci
Básník v prostředí moci
příští konkurence nebude moci upřít
hladký nástup KSČ k totalitní moci v únoru 1948
X...tak se budeme moci oslovovat již za tři čtvrtě
Až začne škola, budeme si moci koupit kupónové knížky

pouze nahrazována mocí ekonomické nerovnosti. Zdůraznil
e svou touhu být mocní a svou mocí a jejím dosahem
V pokušení mocí je cosi velmi zákeřného, šálivého a
□