VÁCLAV CVRČEK – DOMINIKA KOVÁŘÍKOVÁ

Možnosti a meze korpusové lingvistiky[1]

Opportunities and limitations of Corpus Linguistics

This paper addresses two most common comments on corpus linguistics: 1) a corpus is merely a card file index in electronic form and 2) corpus linguistics covers only corpora construction and linguistic marking. We argue that a corpus consists of much more complex material and it can be exploited in unprecedented ways. In response to the second question, we point out that corpus linguistics is an independent linguistic discipline with substantial contributions to linguistic theory and language description.

Key words: corpus, corpus linguistics, methodology

Klíčová slova: korpus, korpusová lingvistika, metodologie

Úvod

Využití korpusů v jazykovědném výzkumu si ve světovém měřítku už vydobylo poměrně pevnou pozici. Od doby vzniku prvního korpusu (Brown Corpus, viz Kučera – Francis, 1967) v 60. letech bylo vytvořeno mnoho nových korpusů, obecných i specializovaných, roste i jejich uplatnění v jazykových příručkách různého druhu (sestavení moderního slovníku se bez korpusové základny dnes už neobejde, korpus si nachází svoji cestu i ke gramatickému popisu, viz např. Biber et al., 2000). Za tuto dobu se dokázali teoretikové korpusové lingvistiky vyrovnat i s většinou připomínek, které v anglosaském prostředí přicházely zejména ze strany generativismu (Beaugrande, 2007). Korpusová lingvistika se tak postupně etablovala jako samostatná jazykovědná disciplína (na úrovni např. kognitivní lingvistiky, generativní lingvistiky nebo psycholingvistiky), která se zaobírá všemi běžnými rovinami jazykového popisu (Aarts, 2002). Svými metodologickými východisky navazuje na strukturalismus, přináší ovšem do jazykovědného výzkumu specifické nástroje a metody.

Zajímavý v tomto ohledu je fakt, jak korpusy a korpusovou lingvistiku přijímá česká jazykověda. Připomínky nebo odmítavá stanoviska jsou zde formulovány značně odlišně od výtek, s nimiž se korpusová lingvistika musela vyrovnávat ve světovém měřítku. Drtivá většina z nich[2] však podle našeho názoru pramení z nepřesných

[1] Tento článek vznikl v rámci výzkumného záměru MSM 0021620823.

[2] Vycházíme zde nejen z výtek publikovaných v odborných textech, ale také z těch, které jsme zaznamenali při osobní komunikaci nebo v rámci odborných setkání.

představ o korpusech a korpusové lingvistice. I po více než deseti letech, kdy je českým badatelům přístupný rozsáhlý a v kontextu evropských jazyků nadprůměrně dobře propracovaný jazykový korpus psané i mluvené češtiny (viz Ústav Českého národního korpusu, http://www.korpus.cz), se tak setkáváme s nepochopením toho, co korpus a korpusová lingvistika přináší nového a jaké možnosti se tím lingvistickému popisu otevírají. Dokládá to např. i absence korpusové studie v monotematickém číslem NŘ 5/2009, které bylo věnováno empirickému výzkumu.

Svoji roli v tom možná hraje vrozená nedůvěra humanitních vědců ke statistickým postupům a technickým prostředkům, jako je počítač. Soudíme tak podle toho, že základní kurz statistického zpracování empirických dat figuruje jenom výjimečně v rámci osnov jazykovědných magisterských oborů, a také podle úrovně počítačové gramotnosti studentů přicházejících na korpusové semináře. Zvládnutí základů statistického a komputačního zpracování empirických dat se přitom postupně stává nedílnou součástí základního vědeckého instrumentária každého oboru, lingvistiku nevyjímaje (viz např. Baayen, 2008; Gries, 2009). Takovýto exaktní přístup se postupně etabluje v některých jazykovědných disciplínách, např. ve fonetice (viz z výukového hlediska nesmírně cenná monografie Jana Volína, 2007), při zpracovávání dotazníků v sociolingvistice a psycholingvistice[3] nebo samozřejmě v kvantitativní lingvistice (např. texty G. Altmanna, L. Hřebíčka, J. Králíka nebo M. Těšitelové).

Existují dvě zkreslené představy, s kterými se jako představitelé korpusové lingvistiky mezi studenty i odborníky setkáváme a s kterými bychom rádi v tomto textu polemizovali: 1) korpus je pouze výzkumný nástroj podobný vylepšenému lístkovému katalogu a mělo by se s ním tak zacházet, 2) korpusová lingvistika se zabývá pouze (nebo většinově) sestavováním korpusů a jejich značkováním.

Vedle toho můžeme pozorovat i námitky metodologického rázu (viz část 2.4), které se ovšem podle nás zakládají na zkreslené znalosti toho, co využití korpusu v jazykovědném výzkumu může přinést. Ačkoli požadavek intenzivnějšího empirického výzkumu češtiny považujeme za oprávněný, fakt, že korpusové studie jsou v monotematickém čísle Naší řeči 5/2009 zmíněny jen v negativním kontextu (Chromý, 2009, s. 225), ukazuje na přezíravý postoj části bohemistické veřejnosti ke korpusovému zkoumání.

V tomto článku bychom se proto chtěli (bez nároku na úplnost) zamyslet nad tím, co korpusová lingvistika jazykovědcům nabízí, jaký inovativní výzkum se na korpusu u nás nebo ve světě provádí, jaké jsou další možnosti využití korpusů, ale také nad tím, co leží mimo epistemologické možnosti tohoto přístupu k jazyku.

---

[3] Zajímavé skloubení dotazníkového a korpusového výzkumu spolu se solidní statistickou analýzou získaných dat představuje např. přednáška N. Bermela *Korpusová data vs. hodnotící soudy rodilých mluvčích (pilotní studie o tvarech 2. pádu)* pronesená v Jazykovědném sdružení 13. listopadu 2008.

1. Korpus jako materiál

O výjimečnosti korpusu jako vzorku textů jazyka bylo napsáno už hodně (Čermák – Blatná, 2006; Čermák et al., 2000; Biber et al., 1998). Pro nás je podstatný zejména ten fakt, že korpus jako zdroj lingvistického materiálu (přes všechny nedostatky v realizaci konkrétních korpusů) výrazně převyšuje jakýkoli jiný dostupný zdroj dat, a to co do kvantity, kvality i vytěžitelnosti.

1.1 Korpus jako vylepšený lístkový katalog?

Jedním z omylů, s kterým se česká korpusová lingvistika setkává, je chápání korpusu jako pouhého vylepšeného lístkového katalogu[4], v kterém se sice hledá rychleji a snáz, ale mimo to žádné výhody a žádné nové informace o jazyce nepřináší. Tento názor ukazuje na zásadní podcenění možností, které korpus nabízí, a na nedostatečnou znalost možností práce s korpusovým vyhledávačem.

Problém si můžeme rozdělit do tří okruhů: 1) technické výhody korpusu oproti klasickému lístkovému katalogu, 2) koncepční výhody korpusu oproti vylepšenému (elektronickému, digitalizovanému) lístkovému katalogu a 3) metodologické výhody vyplývající z prohloubené možnosti popisu při užití korpusu (viz 2.1).

Už samy způsoby vytěžování korpusu umožněné rozvojem informačních technologií jsou přirozeně nesrovnatelně efektivnější než u klasických lístkových katalogů (ad 1). Představme si jednoduchou úlohu: chceme najít v lístkovém katalogu čítajícím několik milionů excerpčních lístků výskyty jednotlivých tvarů slovesa *nechat*. Pomiňme, že údaj o frekvenci jednotlivých tvarů nebude nikdy reflektovat skutečnou frekvenci jevu v úzu, protože lístkový katalog je z velké části budován výběrově (nevypisovaly se všechny výskyty slov ve všech excerpovaných textech). Tento úkol může zabrat až desítky hodin, zatímco v korpusovém vyhledávači to bude otázka několika vteřin, přitom se údaj bude vztahovat na mnohonásobně větší objem dat. Tato výhoda je natolik zjevná, že o srovnání korpusu s klasickým lístkovým katalogem by se v tomto případě dnes už asi nikdo ani nepokoušel; je ovšem dobré si uvědomit, že rychlost a jednoduchost vyhledávání umožňuje uživateli klást i zdánlivě nesmyslné dotazy a ověřovat si velmi nepravděpodobné hypotézy, za nimiž se může skrývat zárodek zcela nového poznání.

Podívejme se ale na složitější dotaz, který je možné korpusovému prohlížeči zadat: vyhledej všechny výskyty slovesa *nechat* v určitém tvaru (dejme tomu v imperativu)

---

[4] Což naznačuje například srovnávání korpusové lingvistiky s manuálně excerpční nebo dotazníkovou metodou: „*Když už chce být korpusová lingvistika lingvistickou disciplínou – a já myslím, že jí není (jinak bychom měli mít i lingvistiku manuálně excerpční nebo lingvistiku dotazníkovou) –, neměla by se omezovat jen na to, že bude počítat, kolik je čeho v korpusu a kolikrát je tam toho víc nebo míň než něčeho jiného.*“ (Veselý, 2008, s. 216)

a najdi slova / tvary slov, která se v okolí[5] tohoto imperativu vyskytují významně častěji než jiná. Dalo by se říct, že takovýto úkol by byl na materiálu klasického (nedigitalizovaného) lístkového katalogu v podstatě nesplnitelný, pokud bychom neměli k dispozici armádu pomvědů a neomezené časové možnosti. Výsledkem poměrně jednoduché kombinace dotazů na korpus je zjištění, že sloveso *nechat* se v imperativu vyskytuje v dobře známých frazémech: *nech někoho na pokoji*, *ne/nech si ujít*, *nech někoho bejt*, *nech se překvapit*, *nech něco/někoho plavat*, kdežto lemma *nechat* obecně (tedy sloveso *nechat* ve všech tvarech) má následující nejčastější kolokace: *nechat vychladnout*, *nechat okapat*, *ne/nechat si ujít*, *nechat přejít varem*, *nechat odležet*. Výsledek tohoto dotazu je zpracovaný během několika minut včetně interpretace; rozdíl mezi kolokacemi tvaru imperativu a celého lemmatu – tedy rozdíl v jejich použití – je viditelný na první pohled.

Existují ale i výzkumy (ad 2), které jsou v lístkových katalozích naprosto nerealizovatelné, mezi něž patří především zjišťování frekvence v textech a problémy týkající se kontextu. Lístkový katalog (i digitalizovaný) je totiž vždy výběrový a neobsahuje celé texty, ale jen jejich části (typicky věty).

Frekvence je pro zkoumání jazyka velmi podstatná zejména tím, že podává informaci o centru a periférii jazykových jevů, podle čehož by měl být strukturován i popis jazyka tak, aby nedocházelo k podrobnému popisu marginálií, zatímco pozornosti unikne celá rozsáhlá oblast jazykových jevů.[6] Výběrovost lístkového katalogu přitom znemožňuje efektivní a objektivní zjišťování celkové frekvence jevů v textech.

Dále jde o otázky vyhledávání ve vzdáleném kontextu slova, přesahujícím hranici několika vět (např. kontext celkem šedesáti slov). Existují korpusové výzkumy, které se zaměřují právě na široký kontext, kde se porovnává vliv slova a dosah tohoto vlivu na výběr lexémů v bezprostředním a vzdáleném kontextu (Cvrček, 2010a). V lístkovém katalogu není obvykle takto široký kontext vůbec k dispozici. Jevy, které by také byly pomocí lístkového katalogu zkoumatelné jen velmi omezeně (v závislosti na kvalitě excerpt), jsou například vnitrotextové odkazování (anafora, katafora) a/nebo aktuální členění věty, kde kontext jedné věty často nemusí postačovat.

Dalším z mnoha příkladů zkoumání, které není možné realizovat jinak než pomocí korpusu, je třeba variantnost lexikální kvazireduplikace typu „*hlava nehlava*“. Spojení „*X neX*“ je ve stomilionovém korpusu psané češtiny SYN2005 ve více než 150 obměnách a nejčastější jsou tato (vypisujeme pouze lemmata s frekvencí vyšší než 2): *hlava nehlava* (142), *cesta necesta* (nejčastěji v podobě *cestou necestou*) (63), *volky nevolky* (46), *čas nečas* (5), *odchod neodchod* (4), *dešť nedešť* (4), *zima nezima* (3), *zákon nezákon* (3).

---

[5] V tomto případě jde o vyhledávání kolokací v kontextu tří slov nalevo a tří slov napravo od klíčového slova.

[6] Roli takového frekvenčního korektivu na lexikální rovině plní Frekvenční slovník (Čermák – Křen, 2004), v oblasti gramatiky pak Statistiky češtiny (Bartoň et al., 2009).

Korpus tedy není v žádném případě možné přirovnávat k lístkovému katalogu, kromě efektivnosti a jednoduché zpracovatelnosti dat (kterou by mohl umožnit i digitalizovaný lístkový katalog) poskytuje i mnohem širší spektrum možností lingvistického výzkumu zejména v tom, že obsahuje celé texty, a jevy je tak možné zkoumat v prakticky neomezeném kontextu.

Korpus je navíc nedocenitelným nástrojem popisu, protože je vzorkem skutečného a realizovaného úzu, a tím dává badateli představu o jinak velmi problematické hranici mezi jazykovou realitou a potencialitou jazykového systému (viz 2.4.1).

Co do rozměrů i rozmanitosti je korpus velmi rozsáhlým a bohatým vzorkem úzu (a rozhodně tím nejlepším, který máme v lingvistickém zkoumání k dispozici, viz 1.2). Není důvod např. předpokládat, že za hranicemi stomilionového korpusu se nacházejí další desetitisíce dosud neregistrovaných lexémů (rozuměj neterminologických). Tím samozřejmě nechceme naznačovat, že by neplatila jednoduchá teze „čím větší korpus, tím lépe". Ačkoli je reprezentativnost velkým tématem řady korpusových statí a polemik (viz 2.4.3), rozsáhlost dat je věrohodným podkladem pro zobecňování empirických pozorování s poměrně velkou mírou jistoty (hypotézy o jazyce jsou ověřované na obrovských datech a dosahují proto větší míry adekvátnosti k popisované jazykové realitě).

1.2 Typy lingvistických dat

Se vznikem a využitím korpusů se znovu objevuje otázka využitelnosti a spolehlivosti různých typů lingvistických dat. Jaké druhy materiálu máme jako lingvisté v současnosti k dispozici?

1.2.1 Lingvistická introspekce

Donedávna byla hlavním zdrojem informací o jazyce data intuitivní, získaná většinou introspekcí samotného badatele. Už mnohokrát ale bylo konstatováno a prokázáno[7], že není možné se spoléhat na vlastní povědomí o jazyce, a to hned z několika důvodů. Místo pomyslného internalizovaného jazykového systému odrážejí data získaná introspekcí badatelovu představu o vlastním jazykovém chování v konkrétní situaci. Tyto představy jsou subjektivní (navíc ovlivněné znalostí předchozích popisů jazyka), a proto se nedají považovat za spolehlivé. Navíc se při opírání o badatelskou introspekci mísí dvě neslučitelné role: lingvisty, který zkoumá, a mluvčího, který výzkumu poskytuje data (Aarts, 2002). Zkoumání je tak vystaveno zvýšenému riziku tzv. observer-effectu, podle kterého pozorující chtě nechtě ovlivňuje zkoumaný objekt.

[7] Např. Beaugrande (2007), Čermák (1997), Fillmore-Atkinsová (2000, s. 393), další literatura v Aarts (2002).

Oproti intuici má korpusový materiál výhodu, že je objektivní (případná subjektivnost je potlačena množstvím dat), a to nejen ve smyslu evaluativním (např. co je ještě gramatické a co není), ale i ve smyslu mapování inventáře. Přináší totiž i poznatky, které jsou neočekávané, na které by lingvista „z hlavy" nepřišel (viz níže příklad variant frazému *Vlk se nažral a koza zůstala celá*). S tím souvisí i fakt, že korpus jako objektivní zdroj informací o jazyce nedává možnost jazykovou realitu při popisu vnímat výběrově s ohledem na jazykový vkus badatele (nekodifikovaný jazyk, vulgární výrazy, překlepy, nevhodné nebo nesrozumitelné formulace apod.).

Podívejme se například na slova, která jsou vnímaná jako synonymní. Slova *šedý* a *šedivý* jsou tradičně považována za velmi blízká synonyma, která se liší (v jednom možném pohledu) jen tzv. prázdným morfémem *-iv-*, tedy morfémem, který už podle názvu nemá jednoznačný význam, jímž by obě slova odlišoval. Když ale pomocí korpusu vyhledáme slova často se vyskytující v okolí slov *šedý* a *šedivý*, zjistíme, že se liší víc, než bychom u takto blízkých synonym čekali; velkou roli tu totiž hrají terminologie a kolokace (vyznačená slova se pojí jen s jedním ze slov):

lemma *šedivý*: *vlasy*, *oblek*, *mraky*, *obloha*, *nebe*, *vousy*, *kalhoty*, *šaty*, *knír*, *bradka*, *myš*, *mlha*, *popel*.

lemma *šedý*: *vlasy*, *oblek*, *plášť*, *kalhoty*, *obloha*, *mraky*, ale také *eminence*, *ekonomika*, *zákal*, *hmota (mozku)*, *zóna*, *kůra mozková*, *litina*.

Kombinace lexému *šedivý* a slov *eminence, ekonomika, zákal* apod. se nevyskytují.

Takovéto příklady by se introspekcí nacházely jen velmi obtížně (zvlášť pokud bychom apriorně předpokládali, že obě slova – *šedý* a *šedivý* – jsou volně zaměnitelná synonyma). Ty kontexty, v kterých je možné použít jen jedno slovo z dvojice *šedý/šedivý*, pak mohou sloužit jako základ pro analýzu rozdílů jejich významů.

To, že je korpus zdrojem objektivních dat ve smyslu inventáře jazykových jevů, můžeme dokumentovat příkladem z frazeologie. Frazémy jsou často vnímány jako neměnný celek. Díky korpusu je možné zjistit, že se s nimi obvykle zachází daleko volněji, než by se dalo čekat (Čermák, 2007, s. 584). U větného frazému *Vlk se nažral a koza zůstala celá* jsme zjišťovali, v jakých obměnách se vyskytuje v korpusu SYN2005. Následující příklad mimo jiné ukazuje, že korpus může být cenným zdrojem i pro zkoumání potenciality jazyka (viz níže). Ani po letech lingvistického tréninku není možné si představit a z hlavy odvodit, kam až může kreativní zacházení s větnými frazémy a dalšími jevy zajít (podobně je tomu i ve výše uvedeném příkladu lexikální kvazireduplikace *hlava nehlava*):

Jenže vlk se nikdy nemůže nažrat tak, aby koza zůstala celá.

… byrokratický vlk nažral a podnikatelská koza zůstala celá.

Aby se prostě vlastenecký vlk nažral a národní koza zůstala celá.

Aby se vlk federální kasy nažral a domácí špeditérská koza zůstala celá.

To odpovídá představě o otcovském vlku, který se nažere tak, aby genetická koza zůstala celá.

Dalším jevem, který je možné zkoumat jen pomocí korpusu, a nikoli introspekcí, je sémantická prozodie. Tímto termínem (alternativně se užívá i discourse prosody, tj. prozodie diskurzu, nebo sémantická preference) se tradičně myslí vztah mezi lemmatem či slovním tvarem a skupinou sémanticky vymezených slov (Baker et al., 2006). Např. přívlastek *údajný* má jednoznačně negativní sémantickou prozodii, protože se pojí se substantivy jako *pachatel*, *podvod*, *znásilnění*, *terorista*, *vrah*, *nedostatek*, *dluh*, *odposlech*, *zločin*, *vražda*, *porušení*, *korupce*. Slovo *pěkný* může mít sémantickou prozodii pozitivní: *počasí*, *vzhled*, *výhled*, *ženská*, *pozdravení*, *pláž*, *rozhled*, *podívaná*; v přeneseném významu vyjadřujícím množství nebo míru pak nejen pozitivní: *řádka*, *balík*, *hromádka*, *sumička*, ale i negativní: *blbost*, *kaše*, *průšvih*, *fuška*, *pitomost*, *potvora*, *brynda*.

To, že intuitivní data jsou nedostatečně spolehlivá, samozřejmě neznamená, že by měla být úplně zavržena. Naopak, jsou velmi důležitou součástí lingvistického výzkumu v několika jeho fázích. V první řadě lingvistickou introspekci využíváme pro zvolení vhodné a smysluplné výzkumné otázky (viz níže fáze corpus-driven výzkumu). Zároveň je introspekce cenným vodítkem při hodnocení výsledků nalezených pomocí korpusu (Tognini-Bonelli, 2001). Pokud takový výsledek nedává smysl (např. kvůli chybě v sestavení nebo značkování korpusu) nebo je nezajímavý, poznáme to právě díky introspekci. Proto zůstává cenným prostředkem lingvistického zkoumání.

1.2.2 Dotazníky

Problém mísení rolí lingvisty a mluvčího, s kterým se setkáme u introspekce, by bylo možné vyřešit obrácením se na informanty, ostatní mluvčí jazyka, nejlépe nefilology. Někteří jazykovědci totiž předpokládají, že skutečný úzus (tak jak je zachycen v korpusu) není dostatečným zrcadlem postojů mluvčích (Adam, 2009, s. 152). V této otázce je výhodné odlišit postoje deklarativní, to jsou ty, které mluvčí projevují při dotazníkové evaluaci jevů, a postoje reálné, tj. ty, které se projevují právě v úzu, např. výběrem jazykových prostředků pro stylizaci, jinými slovy jazykové chování (Cvrček, 2008, s. 150n.).

Deklarativní postoje jsou dnes nejčastěji sledovány pomocí dotazníků, příp. dalších sofistikovanějších metod měření veřejného mínění. Zastánci dotazníkové metody vycházejí z přesvědčení, že deklarativní postoje mluvčích a pisatelů je třeba zkoumat, protože vypovídají o skutečném vztahu mluvčích ke konkrétním jazykovým jevům, a to i přes námitky, že mohou být předsudečné a realitě (skutečnému jazykovému chování) odpovídat nemusejí. K takovému výzkumu, který je jinak zcela legitimní a vypovídá cosi o jazykové realitě, korpus využít nemůžeme.[8] V každém případě je

[8] Otázkou pak zůstává, jaký závěr vyvodíme ze zkoumání, v němž dotazníkové šetření ukáže, že postoje mluvčích jsou v příkrém rozporu s jejich vlastním jazykovým jednáním. Je pak věrným obrazem reality skutečný úzus, nebo představa o vlastním vyjadřování vtělená do odpovědí na anketní otázku?

i tato metoda (stejně jako introspekce) náchylná k ovlivňování zkoumaného předmětu skrze pozorování (observer-effect), ke kterému může dojít při aranžování výzkumu (formulace dotazů atp.).

Na druhou stranu, pokud budeme vycházet z toho, že postoje projevované v úzu jsou skutečným odrazem preferencí pisatelů a mluvčích (a postoje deklarativní jsou např. odrazem jazykového vzdělání), pak se bude jevit jakékoli dotazníkové šetření jako vypovídající o veřejném jazykovém mínění, ale nikoli o stavu jazyka.[9]

1.2.3 Příležitostný sběr materiálu

Dalším zdrojem jazykových dat je pro tradiční lingvistiku příležitostné sbírání materiálu, např. ve chvíli, kdy badatele něco zaujme v náhodně čteném/zaslechnutém textu. Takový přístup má tu nevýhodu, že se pozornost přirozeně soustředí především na zvláštnosti nebo neobvyklé jevy – právě takové, které badatele zaujmou. Pak ale stejně neví, jaké je místo jevu v celku jazyka, a takový přístup je vlastně subjektivně deformující. To, co je obvyklé, frekventované, pravidelné nebo centrální, většinou pozornost nepřitahuje. Výsledky, ač mnohdy obdivuhodně ucelené a zajímavé (např. Hronek, 1972), jsou touto výběrovostí nutně poznamenány.

To, že korpus je budován jako kompilát různých textů, které jsou nedělitelné, zamezuje na rozdíl od jiných lingvistických dat problematické výběrovosti. Přítomnost určitého jevu v korpusu nezáleží na pozornosti a svědomitosti badatele (k problematice výběru a složení textů, které přirozeně ovlivňují výsledky, ovšem v rámci rozsáhlých korpusů v mnohem menší míře, viz 2.4.3).[10] Takovou výběrovostí ze své podstaty nutně musí trpět např. lístkový katalog, při jehož vytváření se bránilo přílišnému narůstání materiálu používáním tzv. sít, tedy seznamů slov, které už nemají být dokumentovány (viz http://www.lexiko.ujc.cas.cz, sekce Lexikální archiv). Je tedy zřejmé, že přesný obraz o frekvenci a způsobech užívání nemohl poskytnout ani u některých běžných slov, jejichž úzus byl vyřazován z excerpce ve prospěch jevů zvláštních a raritních.

Korpus je navíc sestavován jako nespecifický zdroj informací o jazyce, tzn. že není vytvářen s ohledem na jeden konkrétní výzkumný úkol (např. zkoumání morfologie nebo tvorbu slovníku), a může tedy sloužit různým oborům, vedle těch lingvistických i didaktickým, psychologickým nebo sociologickým.

---

[9] Je to podobné, jako kdybychom zjišťovali čtenost deníků sociologickým průzkumem. V takovém případě riskujeme, že mluvčí nebudou z různých důvodů odpovídat na otázku, jaké noviny čtou, pravdivě. Namísto toho, tedy namísto deklarativních postojů, můžeme zjistit relativně přesně (i zde je třeba počítat s chybou měření) prodejnost jednotlivých periodik, tedy postoje reálné, projevené, a podle toho si udělat ucelenější obrázek.

[10] O nespolehlivosti badatelovy pozornosti svědčí i lingvistická anekdota z přípravy Příručního slovníku jazyka českého, kdy se v lístkovém katalogu nenašlo ani jednou slovo *koza*, protože si nikdo neuvědomil, že by mohl tak běžné a obyčejné slovo vypsat (F. Čermák – osobní sdělení).

1.2.4 Korpusová data

Z výše uvedených příkladů vyplývá, že korpusová data jsou pro lingvistické zkoumání cenná, a to zejména z těchto důvodů:

- data jsou vzorkem skutečného úzu, a to jak mluveného[11], tak psaného[12]
- rozsah lingvistického materiálu obsaženého v korpusu je bezprecedentní
- korpus principiálně není budován selektivně, rozhodnutí o zařazení textů není subjektivní (možná subjektivnost je potlačena množstvím textů)
- korpus je nespecifický zdroj informací o jazyce, může se využívat k výzkumu nejrůznějších jevů
- pomocí korpusu lze realizovat i zkoumání, která s jiným jazykovým materiálem nejsou možná
- data jsou pohotově k dispozici každému uživateli
- korpusový prohlížeč je schopný vyhledávat i složité dotazy během několika vteřin
- data v korpusu se neopírají o deklarativní postoje mluvčích typu „já si myslím, že mluvím takto“; jsou to skutečné promluvy/texty, neboli to, jak mluvčí jazyk skutečně používají
- velké a nespecializované korpusy v ČNK jsou statické, texty v nich neubývají ani nepřibývají; proto je možné každé vyhledávání zopakovat a ověřit, a to i po několika měsících či letech
- data lze pomocí subkorpusů jednoduše rozčlenit podle zadaných údajů o textu. V ČNK jsou to mj. tyto: registr[13], datum vzniku a pohlaví autora/autorky, u překladů zdrojový jazyk, médium textu (kniha, noviny) a bibliografické údaje, jako je ISBN.
- kontinuální budování korpusů plní i velmi žádoucí vedlejší funkci – tvoří bázi pro mapování jazykového vývoje a slouží jako archiv dobového jazykového úzu

2. Korpusová lingvistika

Sporná je i druhá představa, s kterou bychom rádi polemizovali, a totiž že korpusová lingvistika se zabývá pouze metodami výstavby korpusů a jejich (více či méně technického) zpracování a lingvistického značkování. Tento omyl může být částečně způsoben tím, že drtivá většina článků prezentujících korpus v dobách jeho začátků se věnovala právě technické stránce věci a celkovému představení konceptu korpusu (např. Čermák, 1995, a další).

---

[11] Viz korpusy mluveného jazyka řady ORAL o celkovém rozsahu 2 miliony slov (ORAL2006, ORAL2008), kde číslo odkazuje na rok vydání.

[12] Synchronní korpusy řady SYN o celkovém rozsahu 1,2 miliardy slov (SYN2000, SYN2005, SYN2006PUB, SYN2009PUB, SYN2010), kde číslo odkazuje na rok vydání.

[13] Příkladem takového rozčlenění může být korpusová anglická mluvnice rozdělující texty do čtyř kategorií/registrů: odborná literatura, beletrie, publicistika a konverzace (Biber et al., 2000).

Dnešní situace je však jiná. Korpusová lingvistika u nás zatím sice není samozřejmou součástí studijních programů filologických oborů, nicméně existuje etablované pracoviště zabývající se korpusovou lingvistikou, výstavbou korpusů, zdokonalováním nástrojů pro jeho vytěžování a výukou. Čím dál víc se hlásí o slovo ta část korpusové lingvistiky, která se primárně nezabývá pouze sestavováním korpusů, otázkami reprezentativnosti, možnostmi lemmatizace a morfologického nebo jiného tagování. Její součástí jsou způsoby vytěžování korpusů, tedy způsoby zkvalitňování a doplňování popisů jazyka, objevování nových vztahů a vynalézání nových konceptů sloužících pro popis jazykové reality. Tagování a lemmatizaci (které ideálně mohou být víceré, podle typu užité teorie, na kterých jsou založeny) chápeme spíše jako technologickou záležitost, jako nástroj, ale nikoli jako hlavní téma korpusové lingvistiky. Uvědomujeme si však, a to je třeba zdůraznit, že cesta k morfologickému značkování a lemmatizaci často podněcuje další velmi cenný jazykovědný výzkum nejen v oblasti samotné (korpusové) lingvistiky, ale také v rámci počítačového zpracování přirozeného jazyka a formálních gramatik, viz např. Petkevič, 2006. Řada problémů pramení z toho, že se značkování i lemmatizace původně zakládaly na nekompletních a nedůsledných předkorpusových popisech. Důsledný formální přístup uplatňovaný dnes upozorňuje na nedostatky minulosti a klade lingvistům mnohdy podnětné otázky.

2.1 Corpus-based a corpus-driven přístup

Výzkum v oblasti vytěžování korpusu je možné rozdělit na dva přístupy (které nemají dosud ustálené české ekvivalenty): c o r p u s - b a s e d (tedy výzkum na korpusu založený, resp. korpusem ověřovaný) a c o r p u s - d r i v e n přístup (tj. výzkum korpusem řízený či inspirovaný). Rozdíl, který podrobně popsala Elena Tognini-Bonelli (2001), v českém prostředí A. Čermáková (2009), spočívá především v míře vlivu, kterou je badatel ochoten při formulování hypotézy o jazyce přenechat jazykovým datům. Zatímco v corpus-based přístupu badatel přistupuje k jazykovým datům s předem vytvořenou, na introspekci založenou hypotézou a v korpusu hledá argumenty pro její potvrzení (resp. vyvrácení), corpus-driven přístup je charakteristický vytvářením konceptů a popisných struktur až v závislosti na zkoumání dat.[14] První přístup je v naší tradici zdá se už pevně zakořeněn (viz např. Štícha, 2006), druhý je ovšem stále ještě v plenkách. Je zjevné, že lingvista není schopen se zcela oprostit od svých předchozích znalostí (vždy je minimálně alespoň mluvčím) a přistupovat k materiálu bez jakýchkoli očekávání (není to možné a ani by to nebylo vhodné). Podstatný na corpus-driven přístupu je ale fakt, že lingvista je kdykoli ochoten „lešení“[15] introspektivně vytvořené počáteční hypotézy shodit

[14] Fakt, že korpusová lingvistika v sobě zahrnuje dva odlišné metodologické přístupy, je přitom pro některé badatele důkazem, že se jedná o samostatnou disciplínu (Aarts, 2002, s. 14).

a na základě pozorování dat stavět popis daného úseku jazykové reality úplně znovu a jinak.

Jako příklad můžeme uvést jeden ze studentských výzkumů s y n o n y m i e v Českém národním korpusu. Studentka (Jana Mataruga) se rozhodla popsat rozdíly mezi vybranými dvojicemi synonym typu *abonent – předplatitel*, *medicína – lékařství* apod. v korpusu SYN2005. Při zkoumání významu těchto lexémů přitom vycházela z jejich kolokací a kontextů po vzoru J. R. Firtha a jeho známého výroku „*you shall know a word by the company it keeps*“[16] (Palmer, 1968, s. 179). Výsledkem bylo konstatování, že dvojice synonym mají jen minimální počet společných kontextů, výjimku tvoří pouze gramatická slova jako spojky a předložky (k synonymii viz i 1.2.1). Corpus-based přístup by se spokojil se závěrem, že tato synonyma nemají (navzdory očekávání) téměř žádné společné kontexty. Corpus-driven přístup by ovšem měl ve vytěžování korpusových dat pokračovat tím, že by minimálně zpochybnil synonymii daných dvojic: nemají-li podobné kontexty, nemohou mít tedy ani shodný význam, nejsou to tedy pravděpodobně ani úplná synonyma. Případně by mohl vést až k hypotéze, která by potřebovala řádně otestovat, že úplná synonymie lexémů jako taková je v jazyce jevem okrajovým či neexistujícím a základním organizačním principem lexikonu tedy musí být něco jiného (např. antonymie/opozitnost, spíše však hypo/hyperonymie). Podstatný je zde fakt, že původní premisa o synonymii, která v počátku nebyla předmětem testování, je pod tíhou pozorování reálných jazykových dat zpochybněna a místo toho nastupuje hypotéza nová (o marginálnosti lexikální synonymie v jazyce obecně).

Jiným příkladem stejného rozdílu v přístupech by mohla být otázka rozdělení slov do s l o v n í c h d r u h ů. Přistoupíme-li k tomuto úkolu stylem corpus-based, budeme rozdělovat slova do deseti předem vymezených tradičních skupin (to je v podstatě úkol, před kterým stojí morfologické značkování korpusů), příp. můžeme zkoumat formu, funkci a význam jednotlivých slov, jejichž klasifikace je nejasná, a zjišťovat míru shody s prototypickými zástupci jednotlivých slovních druhů. Corpus-driven přístupem však můžeme dojít k závěru, že slovnědruhová klasifikace by mohla vypadat zcela jinak. Slovní druhy, jakožto lexikologická klasifikace, rozdělují lexémy do skupin podle jejich nejobecnějšího významu (substantiva jsou prototypicky substance, slovesa děje, adjektiva vlastnosti atp.). Když víme, že se význam skrývá v kontextech (viz Firth, cit. výše), je možné měřit míru podobnosti slovních druhů poměřováním podobnosti jejich kontextů. Zjistíme tak například, že na základě kontextů jsou řadové číslovky tak blízké adjektivům, že by bylo možné uvažovat o jejich sloučení do jedné významově homogenní skupiny (Cvrček, 2010b).

---

[15] Za metaforu lingvistických teorií jako lešení, které si lingvisté staví před členitou budovu jazyka, pocházející původně od V. Skaličky, děkujeme prof. P. Sgallovi (osobní sdělení).

[16] „slovo poznáš podle toho, v jaké společnosti se vyskytuje“ (překlad VC a DK).

Můžeme takto vytvořit alternativní klasifikaci lexémů na úrovni slovních druhů (jejich počet může být odlišný od tradičních deseti), jejímž kritériem se mohou stát kontextové vlastnosti slov (Cvrček, 2010a). Taková klasifikace nebude vycházet z introspekce, ale přímo z jazykových dat.

Třetím příkladem corpus-driven přístupu může být zkoumání t e r m i n o l o g i e na základě automatického vyhledávání termínů v korpusu (Šrajerová, 2009). Hledání vlastností termínů sice vychází z předpokladu, že o některých slovech můžeme s jistotou říct, že se jedná o termíny, ale v dalších fázích už je výzkum veden pouze korpusovými daty, nikoli badatelskou introspekcí. Formulování definice termínu, která by vycházela z jiných než jen sémantických vlastností, probíhá v několika fázích. Po ručním označení termínů[17] v odborných textech (za pomoci terminologických slovníků) se automaticky vyhledají slova, která sdílejí podobné formální, statistické a lingvistické vlastnosti (Šrajerová et al., 2009a). Zároveň se pomocí stejných nástrojů vyhodnotí, které z těchto rysů mají na označení slova za termín největší vliv. Konečnou hodnotu termínovosti (nebo terminologické platnosti) slova přitom určuje až kombinace těchto vlastností.

Mezi vlastnosti, které jsou pro vyhledání termínů v textech nejdůležitější, patří například:

1. frekvence slova v obecném korpusu ve srovnání s texty odbornými (pokud je frekvence slova výrazně vyšší v odborných textech než v korpusu složeném z beletrie a publicistiky, je pravděpodobnost, že jde o termín, vyšší),
2. distribuce slova v jednotlivých odborných disciplínách (v čím menším počtu disciplín se slovo vyskytuje, tím větší je pravděpodobnost, že půjde o termín),
3. struktura slova (čím je struktura slova neobvyklejší, měřeno obvyklostí grafémových bigramů, tím je pravděpodobnější, že se jedná o termín).

Na základě posouzení váhy těchto vlastností je pak možné uvažovat o podobě budoucí definice termínu (viz Šrajerová, 2009b).

2.2 Korpusová lingvistika languová nebo parolová?

Jednou ze sporných výtek vůči korpusové lingvistice, která se objevuje i v zahraničních publikacích, je to, že se zaměřuje pouze na úzus. Korpus podle této výtky dokumentuje pouze parolovou stránku jazyka (performanci) a není schopen vytvářet hypotézy o jazykovém systému (langue, kompetenci). K tomu můžeme aspoň stručně mít minimálně dvě poznámky.

1) Jazykový systém není bezprostředně zkoumatelný žádnou ze známých lingvistických metod; jediné, co máme k dispozici, jsou projevy abstraktního jazykového

[17] V současné fázi výzkumu se jedná jen o jednoslovné termíny (příp. části víceslovných termínů).

systému, tedy konkrétní texty a promluvy. Fakt, že někteří kabinetní lingvisté („armchair linguists“[18]), zakládající svoje výzkumy na introspekci a dedukci, se domnívají, že jsou schopni svými modely podchytit celý systém včetně jevů potenciálních, není víc než jen proklamací. Ověření, zda mentalisticky vytvořené modely tzv. „nepřegenerovávají“ (tzn. umožňují označovat jevy neexistující za součást jazyka) nebo tzv. „nepodgenerovávají“ (tzn. vylučují z jazyka jevy, které se v něm reálně nacházejí), je možné pouze při srovnání s reálnými promluvami (ať už v korpusu nebo jinde). Intuice badatele (nebo skupiny respondentů) je k tomuto úkolu nepoužitelná, protože je nespolehlivá[19], výběrová a navíc vždy poplatná idiolektu (viz 1.2.1). V konečném důsledku jsou takovéto teorie schopny popsat systém jazyka jen v mezích toho, co dovolí kontrolní data, a nejsou o nic víc popisem skutečného jazykového systému než induktivní popisy parole.[20]

2) Zmiňme v této souvislosti postřeh připisovaný jednomu ze zakladatelů korpusové lingvistiky J. Sinclairovi o konkordančním seznamu zobrazujícím jednotlivé výskyty hledaného slova v různých kontextech. Příkladem může být spojení „to máš“ v korpusu SYN2005 (výběr):

(1)	Debora . „ Jo ,	to máš	pravdu , “ přiznal Geoffrey
(2)	. “ „ No ,	to máš	asi pravdu , “ připustil
(3)	, my víme , že	to máš	nejradši na zaďoura a je
(4)	. „ No dobře ,	to máš	pravdu , neviděl , “
(5)	, „ znamená , že	to máš	hodit do koše . Však
(6)	jsem ji . „ Co	to máš	, Jarouši ? “ ozval
(7)	z postele . „ Kde	to máš	? Ten List paní a
(8)	. “ „ Hm –	to máš	nejspíš pravdu . „ Pat
(9)	“ „ No ano –	to máš	vlastně pravdu . Poslyš ale
(10)	stavu elektrárny . Zítra mu	to máš	odevzdat . Co tam napíšeš
(11)	? “ „ Jo ,	to máš	pravdu . Dostane to kapitán
(12)	Ale jen si dej ,	to máš	od nás gratis . “
(13)	pohodlně natažené . „ Už	to máš	? “ zeptala se otráveně
(14)	kvalitu ? Za ty prachy	to máš	vlastně zadarmo . “ Dopil
(15)	hlavu . „ Ty už	to máš	? “ zašeptala . „
(16)	ním nelíbí ! A teď	to máš	! “ Zřejmě patřil k
(17)	se nikdo nedoví , že	to máš	ode mě . . .
(18)	tebe možná vypláznul . Natrénovaný	to máš	, ne ? A kdybys

[18] Termín pocházející od Ch. J. Fillmora.

[19] Srov. „Pro potřeby jazykové analýzy je co do intuice […] lingvista jakožto informant nespolehlivý, protože má v důsledku své průpravy svou přirozenou intuici naivního mluvčího zničenou.“ (Čermák, 1997, s. 38)

[20] Nemluvě o tom, že introspekce není schopna umístit jevy na škálu. Např. existují lexémy, které jsou více či méně monokolokabilní (viz 2.3). Rozhodnutí o míře monokolokabilnosti je introspekci v podstatě nedosažitelné.

Sinclair poukazuje na fakt, že čteme-li jednotlivé konkordanční řádky horizontálně, máme představu o performanci/parole, zatímco když se na seznam díváme vertikálně, zjišťujeme možnosti kompetence/langue (Beaugrande, 2007, s. 100). Tedy čteme-li seznam po řádcích, získáváme informaci o tom, co bylo skutečně realizováno, zatímco čtením po sloupcích si můžeme udělat představu o tom, jaké možnosti doplnění za naším klíčovým slovním spojením následují a jaké mu předcházejí. V našem případě se zde na lexikologické rovině demonstruje, že význam spojení *to máš* často není odvoditelný z významů složek (v mnoha případech nejde o význam 'vlastnit' nebo 'muset'). Velká část dokladů ukazuje např. to, že *to máš* je jen počáteční částí frazému *to máš pravdu* (řádky 1, 2, 4, 8, 9, 11), příp. *to máš gratis/zadarmo* (12, 14) a nebo částí frazému *a teď to máš* (16). Další realizace svědčí o významu ‚mít hotovo' *už to máš* (13, 15) a rovněž o ustáleném spojení (v ukázce zastoupeném pouze minimálně) *mít rád* (3). Z pohledu gramatiky z ukázky můžeme např. vyčíst, že akuzativní doplnění substantivem *pravda* odsune zájmeno *to* do pozice partikule. Řádek (18) je dokladem pomocné funkce slovesa *mít*, které se podílí na perfektním významu slovesného tvaru (Cvrček et al., 2010, s. 241).

Ačkoli je možné se Sinclairovým bonmotem různě polemizovat, je nesporné, že vertikální čtení nám podává empirický obraz o kompetenci (jinak přímo nezkoumatelné), jehož míra spolehlivosti je dána velikostí a vyvážeností korpusu.

2.3 Nové koncepty popisu

Na základě zkoumání materiálu v korpusu se ale kromě zpřesňování dosavadních vědomostí vytvářejí i nové koncepty popisu. Jsou povětšinou důsledkem přesunutí pozornosti z tradičního paradigmatického aspektu směrem k syntagmatice jevů a změnou v chápání vzájemného vztahu lexikonu a gramatiky, který byl tradičně vnímán jako vztah podřízenosti prvního druhému.

Zde leží základní argument proti chápání korpusu jako pouhého výzkumného nástroje. Jedna z oblíbených metafor korpusových lingvistů hovoří o tom, že korpus by se otevřením nových perspektiv pro badatele dal přirovnat k mikroskopu. Ačkoli můžeme chápat mikroskop jako pouhý nástroj, s jeho vynálezem byly vytvořeny podmínky pro vznik zcela nových disciplín (např. mikrobiologie), které zkoumají dotud nepozorovanou a nepozorovatelnou realitu, a tím nalézají zcela svébytný předmět popisu, a tedy i raison d'être. Stejně tak korpus svým neobyčejným rozsahem předkládá badateli vhled do oblastí jazykových vztahů, které na izolovaném materiálu jedné věty nebo jednoho textu nebyly viditelné.

Příkladem korpusového konceptu, který je odrazem zvýšené pozornosti věnované syntagmatice, je otázka víceslovných jednotek. Nejedná se přitom pouze o samotný fakt existence systémových víceslovných termínů, kolokací a frazémů v popisu lexikonu (některé z těchto pojmů jsou navíc známy už relativně dlouho, např. víceslovné předložky nebo spojky). Pozoruhodné jsou zde především způsoby, jak takové

jednotky identifikovat a jaké důsledky ze zapojení těchto konceptů do popisu jazyka vyplývají.

Pro účely identifikace víceslovných jednotek byla vynalezena zcela nová metodologie a statistické postupy. Jedná se zejména o kolokační míry jako MI-score, t-score a celou řadu dalších (viz např. Stubbs, 2007; Pecina, 2010), které identifikují na základě frekvence slov ty dvojice, jejichž souvýskyt je statisticky významně častější, než bychom mohli očekávat na základě pouhé pravděpodobnosti. Výzkum v této oblasti je poměrně bouřlivý a s každou větší korpusovou konferencí se objevují nové a nové postupy identifikace.

Korpusová lingvistika promýšlí i konsekvence zapojení víceslovných jednotek do celého popisu jazyka. Mezi důsledky můžeme zařadit tyto poznatky:

1) Existují monokolokabilní jednotky, které se vyskytují pouze v rámci konkrétních slovních spojení (slovo *eminentní* se ve více než 80 % případů spojuje se slovem *zájem*, srov. další slova jako *dokořán*, *lelky*, *tratoliště* aj.). Termín monokolokabilita zde přitom není možné brát doslovně, jedná se o spojitelnost slova s jednotkami (nikoli desítkami nebo stovkami) jiných slov.
2) Přistoupíme-li na předpoklad, že existují víceslovné lexémy, musíme se s touto realitou vyrovnat nejen při popisu lexikonu, ale i gramatiky. Je-li oprávněný požadavek, aby každá lexikální jednotka měla svoji funkčněmorfologickou platnost, je zjevné, že i víceslovné jednotky musí být zařaditelné např. ke jmennému rodu nebo slovnímu druhu[21]. Tento přístup (zmiňovaný už F. Čermákem, viz např. Čermák, 2010), systematicky uplatněný v Mluvnici současné češtiny (Cvrček et al., 2010, s. 132n.), vede k závěru, že existují víceslovné ekvivalenty ke každému ze slovních druhů: podstatné jméno – *podvěsek mozkový*, *druhá světová (válka)*; přídavné jméno – *světle hnědý*, *zbrusu nový*; zájmeno – *ten samý*, *vůbec nic*; číslovka – *osmdesát pět*, *devět desetin*; sloveso – *přijít nazmar*, *být nasnadě*; příslovce – *s nepořízenou*, *všeho všudy*; předložka – *nehledě na/k*, *na rozdíl od*; spojka – *a proto*, *místo aby*; částice – *popravdě řečeno*, *vždyť přeci*; citoslovce – *to jo*, *a basta*.
3) Korpusová lingvistika zapojuje do svých popisů široké spektrum konceptů založených na syntagmatice jednotek: kolokace, koligace[22], sémantická prozodie (definice viz 1.2.1), chunks nebo lexical bundles[23]. Většina těchto konceptů vypovídá o tom, že existují sémanticky motivovaná pravidla omezující domnělou

[21] Slovní druh chápeme primárně jako klasifikaci lexikologickou (viz výše), funguje však zároveň jako základní organizační princip gramatiky.

[22] Vztah mezi slovem a skupinou slov, která je určena gramatickou kategorií (např. víceslovná předložka *s cílem* se pojí s infinitivem, adj. *četný* se pojí s plurálem apod.).

[23] Terminologicky blízkými koncepty chunks (Sinclair – Mauranen, 2006) nebo lexical bundles (Biber, 2000) se v korpusové lingvistice označují frekvenčně vymezené ustálené víceslovné jednotky

volnou kombinovatelnost slov, která stojí v počátku úvahy o služebné úloze lexikonu vůči gramatice (zejména v generativním přístupu), s níž korpusová lingvistika polemizuje.

4) Na reálných datech se demonstruje škálovitá povaha většiny popisných konstruktů (např. úplná monokolokabilnost – *křížem krážem*, většinová monokolokabilnost – *eminentní zájem*, poloviční, hraničící s pojmem sémantická prozodie – *nedozírné následky/důsledky* apod.).

Důsledkem prohloubené znalosti o syntagmatice jednotek a jejich kombinovatelnosti je přesun popisu některých oblastí ze sféry gramatiky do lexikonu (např. valence slov jako primárně lexikologická charakteristika, viz Čermáková, 2009). V první řadě musíme zmínit fakt, že koncepty, jako jsou např. kolokace, revidují dřívější představu o volné kombinovatelnosti slov. Spojitelnost slov – a potažmo i konstrukce celých vět – totiž není prioritně podřízena gramatickým vztahům, ale významové kompatibilitě. O tom svědčí i případy srozumitelných vět oproštěných od gramatiky (*Začátek říjen přijít noční mrazík.*) ve srovnání se slavnou nesrozumitelnou, ale gramatickou větou N. Chomského: *Colorless green ideas sleep furiously – Bezbarvé zelené myšlenky zuřivě spí*.

Dalším důsledkem vyrovnání vztahu mezi gramatikou a lexikonem je fakt, že slovník ve světle korpusových dat přestává být inventářem lexémů a základní jednotkou se stává tvar slova. Ukazuje se totiž, že lexikální význam není vázaný na jednotky, z nichž byla vyabstrahována gramatická informace, ale přímo na konkrétní slovní tvary s konkrétními gramatickými významy. Příkladem toho, že význam je vázán na slovní tvar spíš než na lexém nebo lemma, mohou být následující dvojice:

– lexém *(veřejná) knihovna* × tvar *(programátorské) knihovny* = ‚součást programu‘
– lexém *(citlivý) nerv* × tvar *(šlo mu na) nervy*
– lexém *mít* × tvar *mějme* = ‚matematické nechť‘
– lexém *být* (ve funkci plnovýznamové, sponové i pomocné) × tvar *seš* (pouze ve funkci sponové a plnovýznamové, srov. **kupoval seš*)

Nebylo by spravedlivé zastírat, že některé ustálené analogické případy (jako lexém *člověk* vs. tvar/citoslovce *člověče!*, lexém *bůh* vs. tvar/citoslovce *bože!*, lexém *šepot* vs. tvar/adverbium *šeptem*, příp. lexém *hledět* vs. tvar/předložka *nehledě na*) jsou už tradiční lingvistice známy.

V takovýchto případech můžeme jen obtížně uvažovat o tom, že pouze lexikální obsazení věty je dáno smyslem sdělení a gramatické kategorie jsou vyžadovány strukturou věty. Už volbou konkrétních významů si mluvčí často zvolí nejen lexém, ale i konkrétní gramatickou realizaci.

(většinou v rozmezí 3 až 5 slov) mající povahu formulí nebo frází. V mluvené češtině se jedná např. o struktury: *já si myslim, že*; *to víš, že jo*; *tak v tom případě*; *jako by se* apod.

2.4 Meze korpusového přístupu

Ani korpusová lingvistika není bez limitů. Ten základní je dán rozsahem dat (není jich nikdy dost[24]) a jejich povahou. Krom toho můžeme identifikovat několik oblastí, kam pravděpodobně korpusový přístup nikdy nedosáhne a kde bude třeba zapojit i jiné výzkumné metody.

V rámci specifikování mezí korpusového přístupu je třeba lišit mezi výhradami, které se týkají konkrétních realizací korpusů (např. ČNK), a těmi, které se vztahují k principu korpusového bádání.

2.4.1 Výzkum potenciality

Korpusové lingvistice je často vytýkáno to, že pracuje pouze se vzorkem promluv a není proto schopna postihnout jevy potenciální. Podívejme se tedy, jaké možnosti ve studiu potenciality nabízí korpus a jaké introspekce.

V rámci potenciálních jevů můžeme rozlišit ty, které v korpusu nenajdeme z toho důvodu, že v úzu nejsou realizované, a ty, které korpus nezachycuje proto, že jejich frekvence je velmi nízká a nacházejí se na periférii jazyka. Samotný fakt nedoloženosti jevu v rozsáhlém a reprezentativním korpusu je ovšem cenným výsledkem zkoumání. Oba typy potenciálních jevů je možné do určité míry zachytit introspektivním zkoumáním, nemáme-li ovšem korektiv v podobě reálných promluv, můžeme jen velmi obtížně rozhodnout, zda se jedná o jevy realizované, realizovatelné nebo nerealizovatelné. Proto ani introspekce není schopna postihnout potencialitu komplexně.

Víme například, že velice řídce existuje adjektivum *zbůhdarmý* (v korpusu SYN2005 je pouze pětkrát):

(1)	V podstatě ještě pořád takovým	zbůhdarmým	psaním pohrdal …
(2)	… dávný pocit , že po všem tom	zbůhdarmém	šlapání po světě …
(3)	… nač říkat zrovna	zbůhdarmé	, třeba je to naopak …
(4)	… více rozjímat , například nad	zbůhdarmým	těkáním turistů z místa …
(5)	Nejde ani tak o	zbůhdarmé	surfování , kterého …

Můžeme ale říct něco víc o slovotvorné potencialitě původní spřežky *zbůhdarma*? Existuje-li *zbůhdarmý*, může existovat např. i **zbůhdarmost*? Nemáme-li doklady takového tvoření, s pomocí korpusu o tom nemůžeme rozhodnout. Introspekce nás však také bez opory v datech nezavede k lepšímu poznání.

Jiným příkladem může být otázka existence přechodníků v systému slovesných tvarů. Z hlediska introspektivního zkoumání se nám může jevit přechodník jako integrální a synchronní součást systému slovesné morfologie, z korpusového hlediska je to jev spíše diachronní (v současném systému existující spíše okrajově).

[24] Nedosažitelným ideálem by bylo zachycení celého úzu, tedy všech promluv a textů daného jazyka. Zároveň je ale třeba připomenout, že rozsáhlejší než korpusová data k dispozici nejsou.

Frekvence tvarů přechodníku přítomného je méně než 0,1 %, frekvence přechodníku minulého je dokonce méně než 0,01 % všech slovesných tvarů. Navíc se tvoření těchto tvarů týká dohromady asi 5 % slovesných lemmat (cca 2 tisíce slovesných lemmat). Může tedy většina sloves potenciálně tvořit přechodník, nebo se jedná o diachronní kategorii, jejíž zbytky nacházíme tu a tam (podobně jako nacházíme zbytky supina pouze u lexému *spát*[25])?

Druhý pohled na potencialitu se týká jevů, které mohou zůstat introspekci skryty, ale které jsou realizovány v korpusu (viz výše příklad variability frazému *Vlk se nažral a koza zůstala celá.*). Zde se naopak introspekce jeví jako nástroj méně spolehlivý, protože tvořivost mluvčích a pisatelů se často vymyká lingvistově fantazii.

Korpus ve své rozsáhlosti a reprezentativnosti zobrazuje zejména jevy široce centrální v jejich prototypickém užití. Není proto možné chtít ani po korpusové lingvistice komplexní popis potenciality. Z pozorování doloženého úzu ale můžeme vyvodit pravidelnosti a tendence, jimiž se potencialita jazyka bude řídit.

2.4.2 Limity dat

Další limit korpusové lingvistiky je spojen s procesem získávání dat. Je zjevné, že promluvy z některých vypjatých situací (partnerská hádka, poslední slova umírajícího, intimní rozhovor milenců apod.) nebudou v korpusu nikdy dostatečně zastoupeny. Důvodem je, že neseženeme informanty, kteří by takové texty v dostatečné míře poskytli. Můžeme se proto asi rozloučit s myšlenkou, že bychom měli někdy relativně úplnou korpusem podloženou představu např. o inventáři českých nadávek, o jazyce rozčílených lidí nebo o způsobech komunikace v emocionálně excitovaných situacích apod. Je ovšem otázka, zda je toto hendikep pouze korpusové lingvistiky a zda se k uceleným výsledkům můžeme dopracovat s pomocí jiné metody.

2.4.3 Problém reprezentativnosti

Otevřenou otázkou z oblasti tvorby korpusů zůstává jejich reprezentativnost (viz např. Králík, 2001; Šulc, 2001), tedy vyváženost korpusu s ohledem na různé typy textů, žánry a témata. Korpus je sice rozsáhlým vzorkem, ale stále je jen částí celkové populace, kterou představují všechny texty a promluvy daného jazyka. Je jasné, že průzkum provedený na nereprezentativním vzorku může být významně zkreslený (vzhledem k celkové populaci), stejně jako by byl zkreslený průzkum stranických preferencí provedený pouze na jedné sociální skupině nebo v jednom místě.

Korpusy psané češtiny, které jsou součástí ČNK, se vydaly cestou reprezentativnosti určované recepcí (čteností) textů, což je jenom jedna z možností (vedle repre-

[25] V SYN2005 najdeme 70 dokladů supina *spat* (0,5 % všech výskytů slovesa *spát*), kterému předchází lemma *jít* nebo jeho předponová varianta.

zentativnosti založené např. na produkci textů, příp. reprezentativnosti demografické uplatňované zejména v korpusech mluveného jazyka, viz Waclawičová, 2009). Předpokládá se zde, že texty, které mají hodně čtenářů, výrazněji ovlivňují jazyk než texty v podstatě soukromé nebo neveřejné. Jsou ovšem výzkumné úkoly, u nichž je s ohledem na jejich cíl možné dospět k všeobecně přijatelnému řešení (např. pro studium jazyka K. Čapka je vhodným nástrojem korpus všech jeho publikovaných textů).

Ať už je ale koncept reprezentativnosti jakýkoli, vždy existuje možnost vytvořit si vlastní výběrový korpus (virtuální subkorpus) podle požadavků daného výzkumu.[26] Zdaleka ne každý výzkum si klade za cíl popisovat celý jazyk, např. zkoumáme-li lexikon současné publicistiky, potřebujeme beletrii maximálně pro srovnání. Z toho plyne, 1) že všechny výtky vůči reprezentativnosti jsou výtkami proti konkrétnímu sestavení daného korpusu, nikoli proti principu korpusové lingvistiky, a dále pak, 2) že je plně v moci badatele se s tímto nedostatkem vyrovnat sestavením vlastního materiálu na základě nabídky poskytovatele korpusu.

Závěr

Česká lingvistika se působením mnoha osvětových textů a také zásluhou propagační činnosti zakladatele ÚČNK prof. F. Čermáka stala komunitou akceptující korpus jako fakt nebo historickou danost, kterou nelze přehlížet. Pokusili jsme se v tomto přehledovém článku ukázat, že v případě české korpusové lingvistiky se nejedná o několik „*povrchních korpusových* [sond]“ (Chromý, 2009, s. 225) a že česká empirická bohemistika má pevnou oporu a nevyčerpatelný zdroj inspirace v korpusové lingvistice (ale samozřejmě nejen v ní). Věříme, že se nám podařilo představit možnosti, které nabízejí korpusy a korpusová lingvistika, aby se i česká jazykovědná obec stala komunitou korpus plně využívající.

LITERATURA

AARTS, J. (2002): Does corpus linguistics exist? Some old and new issues. In: L. E. Breivik – A. Hasselgren (eds.), *From the COLT's mouth… and others: Language corpora studies in honour of Anna-Brita Stenström*. Amsterdam: Rodopi, s. 1–16.

ADAM, R. (2009): Nad knihou o jazykové regulaci. *Naše řeč*, 92, s. 145–154.

BARTOŇ, T. – CVRČEK, V. – ČERMÁK, F. – JELÍNEK, T. – PETKEVIČ, V. (2009): *Statistiky češtiny*. Praha: Nakladatelství Lidové noviny.

BAAYEN, H. R. (2008): *Analyzing Linguistic Data.* Cabridge: Cambridge University Press.

[26] Některé zahraniční korpusy se téměř úplně distancují od otázky sestavení korpusů a nechávají ji čistě na badateli, který si před začátkem práce musí určit materiál, na kterém hodlá průzkum provést. Praktickým důsledkem takového přístupu je nutnost ve výstupech zkoumání přesně specifikovat, jaká data byla použita.

BAKER, P. – HARDIE, A. – MCENERY, T. (2006): *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.
BEAUGRANDE, R. de (2007): „Corporate Bridges" Twixt Text and Language. In: W. Teubert – R. Krishnamurthy (eds.), *Corpus Linguistics. Critical Concepts in Linguistics (vol. I)*. London/ New York: Routledge, s. 93–118.
BIBER, D. – CONRAD, S. – REPPEN, R. (1998): *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
BIBER, D. – CONRAD, S. – LEECH, G. – FINEGAN, E. – JOHANSSON, S. (2000): *Longman Grammar of Spoken and Written English*. England: Longmann.
CVRČEK, V. (2008): *Regulace jazyka a Koncept minimální intervence*. Praha: Nakladatelství Lidové noviny.
CVRČEK, V. – KODÝTEK, V. – KOPŘIVOVÁ, M. – KOVÁŘÍKOVÁ, D. – SGALL, P. – ŠULC, M. – TÁBORSKÝ, J. – VOLÍN, J. – WACLAWIČOVÁ, M. (2010): *Mluvnice současné češtiny*. Praha: Karolinum.
CVRČEK, V. (2010a): Contextual Approach to Parts of Speech. In: F. Čermák – A. Klégr – P. Corness (eds.), *InterCorp: Exploring a Multilingual Corpus*. Praha: Nakladatelství Lidové noviny, s. 190–204.
CVRČEK, V. (2010b): Korpusový pohled na postavení číslovek v systému slovních druhů. *Sborník z X. mezinárodní setkání mladých lingvistů*. Olomouc, s. 104–110.
ČERMÁK, F. (1995): Jazykový korpus: Prostředek a zdroj poznání. *Slovo a slovesnost*, 56, s. 119–140.
ČERMÁK, F. (1997): *Základy lingvistické metodologie*. Praha: Karolinum.
ČERMÁK, F. (2007): *Frazeologie a idiomatika česká a obecná*. Praha: Karolinum.
ČERMÁK, F. (2010): *Lexikon a sémantika*. Praha: Nakladatelství Lidové noviny.
ČERMÁK, F. – BLATNÁ, R. (eds.) (2006): *Korpusová lingvistika: Stav a modelové přístupy*. Praha: Nakladatelství Lidové noviny.
ČERMÁK, F. – KLÍMOVÁ, J. – PETKEVIČ, V. (eds.) (2000): *Studie z korpusové lingvistiky*. Praha: Karolinum.
ČERMÁK, F. – KŘEN, M. (eds.) (2004): *Frekvenční slovník češtiny*. Praha: Nakladatelství Lidové noviny.
ČERMÁKOVÁ, A. (2009): *Valence českých substantiv*. Praha: Nakladatelství Lidové noviny.
Český národní korpus – ORAL2008 (2008). Praha: Ústav Českého národního korpusu FF UK. <http://www.korpus.cz>.
Český národní korpus – SYN2005 (2005). Praha: Ústav Českého národního korpusu FF UK. <http://www.korpus.cz>.
FILLMORE, Ch. J. – ATKINSOVÁ, S. B. T. (2000): Když začneme tam, kde slovníky končí: Výzva korpusové lexikografie. In: F. Čermák et al. (eds.), *Studie z korpusové lingvistiky*. Praha: Karolinum, s. 381–416.
GRIES, S. Th. (2009): *Quantitative Corpus Linguistics with R*. USA: Routledge.
HAJIČ, J. (2004): *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Vol. 1. Praha: Karolinum.
HRONEK, J. (1972): *Obecná čeština*. Praha: Univerzita Karlova v Praze.

CHROMÝ, J. (2009): Empirické zkoumání v lingvistice. Slovo úvodem. *Naše řeč*, 92, s. 225–226.
KRÁLÍK, J. (2001): Vyvážení zdrojů Synchronního korpusu češtiny SYN2000. *Slovo a slovesnost*, 62, s. 38–53.
KUČERA, H. – FRANCIS, N. W. (1967): *Computational Analysis of Present Day American English*. Providence: Brown University Press.
LEXIKO. Webové hnízdo o novodobé české slovní zásobě a výkladových slovnících (2005–2011) [online]. <http://www.lexiko.ujc.cas.cz/>
PALMER, F. R. (1968): *Selected papers of J. R. Firth, 1952–1959*. Bloomington/London: Indiana University Press.
PECINA, P. (2010): Lexical Association Measures and Collocation Extraction. In: P. Rayson – S. Piao – S. Sharoff – S. Evert – B. Villada Moirón (eds.), *Multiword expressions: hard going or plain sailing? Journal of Language Resources and Evaluation*. Netherlands: Springer.
PETKEVIČ, V. (2006): Automatické rozpoznání infinitivu. Případová studie jako příspěvek k automatické disambiguaci českých textů. In: F. Čermák – R. Blatná (eds.), *Korpusová lingvistika: Stav a modelové přístupy*. Praha: Nakladatelství Lidové noviny, s. 226–253.
SINCLAIR, J. – MAURANEN, A. (2006): *Linear Unit Grammar*. The Netherlands: John Benjamins Publishing Company.
STUBBS, M. (2007): Collocations and Semantic Profiles. In: W. Teubert – R. Krishnamurthy (eds.), *Corpus Linguistics. Critical Concepts in Linguistics (vol. I)*. London/New York: Routledge, s. 166–193.
ŠRAJEROVÁ, D. – KOVÁŘÍK, O. – CVRČEK, V. (2009): Automatic Term Recognition Based on Data-mining Techniques. *2009 World Congress on Computer Science and Information Engineering Proceedings*. CSIE, vol. 4, s. 453–457.
ŠRAJEROVÁ, D. (2009a): Automatické vyhledávání termínů a jeho dopad na definici termínu. *Časopis pro moderní filologii*, 91, s. 1–19.
ŠRAJEROVÁ, D. (2009b): Automatic Term Recognition as a Resource for Theory of Terminology [online]. *Corpus Linguistics Conference 2009 Proceedings*. Liverpool. <http://ucrel.lancs.ac.uk/publications/cl2009/>.
ŠTÍCHA, F. (ed.) (2006): *Možnosti a meze české gramatiky*. Praha: Academia.
ŠULC, M. (2001): Tematická reprezentativnost korpusů. *Slovo a slovesnost*, 62, s. 53–61.
TOGNINI-BONELLI, E. (2001): The Corpus-driven Approach. In: *Corpus Linguistics at Work*. Amsterdam: John Benjamins, s. 84–100.
VESELÝ, L. (2008): Práce o vidu založená na korpusu. *Naše řeč*, 2008, s. 213–216.
VOLÍN, J. (2007): *Statistické metody ve fonetickém výzkumu*. Praha: Epocha.
WACLAWIČOVÁ, M. (2009): Nový korpus mluvené češtiny ORAL2008. *Jazykovědné aktuality*, XLVI, s. 44–51.

Ústav Českého národního korpusu FF UK
Národní 416/37, 110 00 Praha 1
vaclav.cvrcek@ff.cuni.cz
dominika.kovarikova@ff.cuni.cz