JAN CHROMÝ

Korpus a reprezentativnost[1]

Corpus and representativeness

This paper discusses the concept of representativeness in corpus linguistics. Representativeness is a concept used in empirical, quantitative science and it is a characteristic of the relationship between the sample and the population. It is argued that the population for the standard supposedly "representative" corpora of a whole language cannot be defined. The population could be reliably defined only for specialized corpora (e.g. corpora of newspaper texts), hence only this type of corpora could be truly statistically representative. The paper also discusses the idea that we could think about representativeness from the perspective of particular linguistic items instead of from the perspective of the whole language. It may be the case that the same corpus is representative for the use of one item and, at the same time, not representative for the use of another item.

Key words: corpus, representativeness, specialized corpora, population, inferential statistics

Klíčová slova: korpus, reprezentativnost, specializované korpusy, populace, inferenční statistika

V tomto článku se zaměříme na otázku reprezentativnosti v korpusové lingvistice. Východiskem přitom bude srovnání vymezení reprezentativnosti v empirických, kvantitativně orientovaných oborech s vymezením v korpusové lingvistice. Cílem bude poukázat na metodologické a teoretické problémy práce s tímto pojmem a na nesnáze lingvistické práce, které z toho vyplývají.

Pojem reprezentativnost a korpusová lingvistika

Reprezentativnost je pojem používaný ve statistice a v empiricky orientovaných vědách a je založen na vztahu mezi vzorkem, tedy reálně zkoumanou množinou jednotek, a populací, tedy množinou jednotek, na kterou se na základě poznatků o vzorku zobecňuje. Reprezentativnost lze v tomto smyslu chápat skalárně, přičemž čím více se vzorek ve zkoumaných charakteristikách podobá populaci, tím je tento vzorek reprezentativnější, a naopak. Chceme-li tedy mluvit o reprezentativnosti anebo o její míře, musíme nejprve vymezit, co je v našem případě vzorkem a co populací.

[1] Tato studie vznikla v rámci projektu GAČR 13-12973P *Sociolingvistická analýza užívání protetického /v/ v Čechách*.

V korpusové lingvistice můžeme snadno vymezit vzorek – představuje jej určitý korpus, případně nějaká jeho vybraná část. Problém však nastává, chceme-li určit populaci. Na tuto otázku upozorňuje ve svém již klasickém textu Douglas Biber (1993), když píše, že „vymezení populace je základní otázkou při tvorbě korpusu“ (s. 243). Biber konstatuje, že si při tvorbě korpusu musíme vždy určit faktické hranice populace (například všechny texty publikované v angličtině v roce 1961). Dále podle Bibera musíme zjistit, jak je tato populace stratifikovaná – jinými slovy musíme zmapovat, jaké žánry či registry se v ní používají. Z těchto žánrů či registrů následně náhodně vybíráme určitý počet textů o určité celkové délce. Biber přitom odmítá myšlenku, že by korpusy měly být tvořeny na základě demografických proporcí, což je běžné například v sociologickém výzkumu. Říká, že „jazykové korpusy potřebují jiné vymezení reprezentativnosti“ a že „badatelé potřebují vzorky, které jsou reprezentativní v tom smyslu, že zahrnují úplné spektrum jazykové variace, které v jazyce existuje“ (s. 247).

Na Bibera explicitně navazují a ve stejném smyslu o populaci a tvorbě korpusu mluví i McEnery s Wilsonem (2004, s. 77–81). Podobným způsobem nazírají reprezentativnost Cvrček a Kováříková (2011), pro které tento pojem představuje „vyváženost korpusu s ohledem na různé typy textů, žánry a témata“ (s. 130). O reprezentativnosti mluví rovněž ve svém článku o korpusu SYN2000 Králík a Šulc (2005, s. 358–359). Říkají, že „když vytváříme psaný korpus, nesnažíme se o reprezentativnost ze statistického úhlu pohledu. Nikdy nebudeme disponovat všemi jevy, které by mohly být v budoucnu zkoumány, a i kdybychom tyto jevy znali, nikdy bychom nezjistili jejich pravděpodobnostní rozdělení. Jinými slovy, žádné pravděpodobnostní konstrukce nám nemohou zaručit univerzální relevanci korpusu pro budoucí využití. Za těchto okolností není důležitá myšlenka statistické reprezentativnosti, ale naopak struktura textových zdrojů zastoupených v korpusu.“

Proti takto vymezené reprezentativnosti vystupuje Váradi (2001), který podotýká, že „reprezentativní vzorek vyžaduje znalost celé populace, která však není k dispozici“ (s. 590), a konstatuje, že Biberovo vymezení reprezentativnosti, jež je odlišné od vymezení tohoto pojmu ve statistice či jiných empirických vědách, než je korpusová lingvistika, vede k subjektivním a hodnotově založeným rozhodnutím. Říká, že „reprezentativní vzorek z populace můžeme získat pouze prostřednictvím charakteristik této populace, které jsou doloženy z nějakého nezávislého zdroje. Takovým zdrojem nezávislých poznatků jsou dostupná data o mluvčích“ (s. 590–591). Váradi tak populaci hledá v uchopitelné populaci mluvčích, nikoliv v neuchopitelné a nevymezitelné populaci textů.

Váradiho kritický pohled do určité míry sdílí Leech (2007). Klade si řečnickou otázku, zda vůbec v současnosti můžeme tvrdit, že by [deklarovaná] reprezentativnost korpusů, s nimiž pracujeme, měla zjevnou validitu (s. 135), a negativně

hodnotí to, že se reprezentativnosti korpusů věnuje poměrně málo pozornosti a že badatelé reprezentativnost spíše konstatují, než že by se zaměřovali na to, co tato reprezentativnost znamená a jak jí reálně dosahovat. Na rozdíl od Váradiho však Leech dochází k závěru, že to není až takový problém. Pokud podle něj uchopíme reprezentativnost skalárně, lze se spokojit s tím, že „i přesto, že je konečný cíl, tedy úplná reprezentativnost, z praktického hlediska nedosažitelný, můžeme se snažit se k tomuto cíli alespoň blížit“ (s. 140).

Pojmem reprezentativnost se v poslední době zabýval Křen (2013), který konstatuje, že se jedná o pojem terminologicky neujasněný. Přiklání se přitom k pojetí Sinclairově (2005) a říká: „Korpus je pouze více či méně věrným odrazem, vzorkem jazyka nebo některé jeho variety, ale nikdy ho nezastoupí celý. Není to ani nutné, je-li ovšem korpus reprezentativní, a odpovídá tak v mnohem menším měřítku celku jazyka (variety). Hlavním problémem, který se k reprezentativnosti váže, je však právě nemožnost objektivně měřit, nakolik korpus jazyku odpovídá.“ (s. 12) To samo o sobě však podle Křena není až takový problém: „Na reprezentativnost není možné rezignovat už proto, že bez ní nelze korpusová data interpretovat ve vztahu k jazyku. Musíme se tedy smířit s tím, že tento vztah není snadno kvantifikovatelný, korekce intuicí se zdá být nezbytná.“ (s. 13) Křen si je tedy vědom problémů s vymezením reprezentativnosti v korpusové lingvistice, avšak podobně jako pro Leeche je to pro něj pojem zásadní, který není – i přes jeho nejasnost – možné opustit.

Z uvedené literatury je zřejmé, že se pojem reprezentativnosti používá v korpusové lingvistice jinak, než je tomu běžné například v sociologii, respektive v jiných empiricky a kvantitativně orientovaných vědách. Důvodem je principiální nepřístupnost, respektive neznámost, či dokonce nezjistitelnost populace jako takové. Populaci lze sice verbálně definovat, například jako soubor všech textů v daném jazyce v určitém období, problém však je, že ve většině případů (obzvláště v případě mluveného jazyka) nedisponujeme a dost dobře nemůžeme disponovat celkovým soupisem těchto textů. Nastává tak problematická situace – základní množina jednotek, na kterou naše výsledky zobecňujeme, nám není známa. Nevíme tak, o čem činíme zobecnění, a nemůžeme ani ověřit, zda jsou tato zobecnění korektní, nebo ne. Platnost našich závěrů je tak opřena primárně o naši víru, že jsou tyto závěry platné.

V souvislosti s tím připomeňme ještě problém, na který upozorňuje v souvislosti s reprezentativností Kučera (2002, s. 246), totiž na problém autentičnosti textů: „korpus je reprezentativní [z hlediska autenticity], pokud věrně reprezentuje reálný jazyk, tj. pokud nebyly texty v korpusu ‚korigovány‘ či měněny jinak, než ryze formálně, což v dnešní době obvykle znamená – nemusí to tak však být v budoucnu – změny, jako jsou sjednocení fontů a stylů, odstranění obrázků či rušení dělení slov na konci řádků.“ Kučera se touto otázkou zabývá primárně

diachronně, z hlediska synchronního lze však upozornit na problematickou autentičnost u textů publikovaných, které byly podrobeny redakci a korektorským zásahům. Korpus sice může zachytit texty v podobě, ve které vyšly, ne však už v podobě, v níž skutečně vznikly – přičemž tato podoba může být značně odlišná, a to nejen po stránce pravopisné či tvaroslovné, ale třeba i po stránce syntaktické. Jistá neautentičnost textů tak může vést k zakrytí určitých aspektů toho, jak se jazyk skutečně užívá.

Hledání základní jednotky

To, že se v korpusové lingvistice statistické pojetí reprezentativnosti nepoužívá, neznamená, že ani není možné je používat. Domnívám se, že klíčovým je v tomto smyslu především stanovení základní zkoumané jednotky. Ty mohou být – z logiky věci – přinejmenším dvojího typu: jednotkou může být text, anebo člověk. Je přitom zřejmé, že tyto dva typy jednotek – samy o sobě – tvoří populace odlišného typu. V prvním případě se jedná o populaci textů, které vznikly, případně které byly recipovány v určité době. V případě druhém se jedná o okruh jedinců, kteří mají určitým způsobem internalizovaný jazyk. Z těchto rozdílů vyplývá i odlišná možnost popsat populaci – zatímco o populaci všech českých textů máme jen minimální informace, o populaci všech českých mluvčích víme díky sčítání lidu a dalším sociologickým průzkumům mnohé. Vytvořit statisticky reprezentativní vzorek českých mluvčích je možné (alespoň z hlediska sledovaných aspektů), u českých textů obecně to však nejde.

Mohli bychom se tak spokojit s tím, že za základní jednotku zvolíme mluvčího, a sestavíme reprezentativní korpus na tomto základě (využít by bylo možné například analogii kvótního výběru, s nímž pracují výzkumy Centra pro výzkum veřejného mínění). Intuitivně bychom však narazili na problém nezohlednění různých typů textů (žánrů či registrů). I pokud tedy vyjdeme od mluvčího jako od základní jednotky, musíme se vypořádat s otázkou stylové variace, tedy s otázkou, že se texty různého typu v nejrůznějších ohledech liší (viz např. Milroyová – Gordon, 2012). V praxi by to znamenalo, že bychom od každého mluvčího z reprezentativního vzorku museli sebrat více textů různých typů.

(Ne)užitečnost tzv. „reprezentativních“ korpusů jazyka

Uvedené zohlednění stylové i sociální variace je z pochopitelných praktických důvodů nerealizovatelné, tedy alespoň pokud bychom se snažili o popis nějakého jazyka jako celku. Tzv. reprezentativnost korpusu, jako je například SYN2010 apod., je tak pouze deklarovaná, není však doložitelná. To samo o sobě neznamená, že by nereprezentativní korpus tohoto typu nebyl lingvisticky využitelný. Je však potřeba se ptát, k jakému typu lingvistické práce je takto vytvořený korpus vhodný a pro jaké účely se naopak nehodí.

Pracovně zde rozliším tři funkce, které by korpus tohoto typu mohl v lingvistické práci naplňovat (toto rozlišení neaspiruje na úplnost, jistě bychom mohli jmenovat ještě další funkce): 1. funkce prosté zásobárny dat, 2. funkce zpravodaje o distribuci prostředků, 3. funkce řiditele výzkumu.

Funkci prosté zásobárny dat akcentuje Sinclair (1996) v souvislosti s referenčním korpusem, který definuje – poněkud odlišně, než je běžné v české lingvistice – jako „korpus, který je vytvořen s cílem poskytnout komplexní informace o jazyce“ a dále píše, že „má být tak velký, aby reprezentoval všechny relevantní variety jazyka a charakteristickou slovní zásobu a aby tak mohl být využit při tvorbě spolehlivých gramatik, slovníků, tezaurů a dalších jazykově referenčních materiálů.“ Jedná se tedy o funkci blízkou k funkci dřívějších lístkových katalogů: pro zpracování slovníků nám korpus poskytne kontexty, v nichž se dané slovo objevuje, pro zpracování gramatiky dostaneme řadu příkladů určitých typů konstrukcí, které nás zajímají. Vzhledem k velikosti takzvaně reprezentativních korpusů typu SYN2010 lze předpokládat, že pro velké množství prostředků tuto funkci naplňují poměrně dobře (alespoň co se psaného jazyka týče). Lze se však domnívat, že je to primárně důsledkem velikosti korpusu, a nikoliv jeho „reprezentativností“ – čím větší korpus bude, tím více různých dokladů určitého jevu v něm nalezneme.

Složitější je to s funkcí zpravodaje o distribuci prostředků, jíž mám na mysli využití korpusu pro základní statistické analýzy. Naprostá většina prací, které jsou založeny na korpusu, využívá korpus právě v této funkci. Může se jednat o prostá porovnání počtů výskytů různých variant jedné proměnné anebo může jít o složitější analýzy, například kolostrukční (Stefanowitsch – Gries, 2003) či kolexémovou analýzu (Stefanowitsch – Gries, 2005). V pracích tohoto typu se – implicitně či explicitně – předpokládá, že statistiky zjištěné v korpusu něco reálného odrážejí. Takový předpoklad je však vcelku vratký – jestliže víme, že se jedná o korpus ze statistického hlediska nereprezentativní, nemůžeme si být jisti, o čem naše výsledky vlastně vypovídají, k čemu je lze vztáhnout. Statisticky korektní je pouze říci, že naše výsledky informují o tom, jak dané jevy fungují v korpusu.

Třetí funkcí je funkce řiditele výzkumu, jejíž využívání spočívá v tom, že k datům přistupujeme bez předchozích teoretických předpokladů. V tomto smyslu lze mluvit o takzvaném „corpus-driven“ přístupu (viz o něm např. Cvrček – Kováříková, 2011, s. 122–124), jehož „cílem je odvozovat jazykové kategorie systematicky na základě opakujících se pravidelností a frekvenční distribuce, které vyvstávají z jazyka v kontextu“ (srov. např. Tognini-Bonelli, 2001, s. 87). I zde platí, že pokud není korpus statisticky reprezentativní, nemůžeme rozhodnout, zda naše výsledky skutečně odráží jazykovou realitu, anebo jsou nějakým způsobem vychýlené. To ostatně tvrdí i Křen (2013, s. 172), když píše: „Vzhledem k nejednoznačnému vztahu mezi korpusem a jazykovou realitou je zřejmé,

že zobecňování na korpusu založených závěrů na jazyk nebo jeho konkrétní varietu by mělo být opatrné i v případě, že jde o varietu, kterou by měl daný korpus přímo reprezentovat."

Celkově se domnívám, že využitelnost tzv. „reprezentativních" korpusů jazyka je mnohem omezenější, než se obvykle tvrdí, a to právě z toho důvodu, že jejich reprezentativnost je pouze deklarovaná, nikoliv dokládaná.

Rozsáhlost jako argument

Statistická nereprezentativnost korpusů je někdy vyvažována argumentem, že korpusy jsou natolik rozsáhlé, že jsou na jejich základě získané výsledky věrohodné. To tvrdí například Cvrček s Kováříkovou (2011, s. 117): „Ačkoli je reprezentativnost velkým tématem řady korpusových statí a polemik […], 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)."

Takováto argumentace dává smysl jen částečně. Obecně lze tvrdit, že čím větší je vzorek, tím více se shoduje s populací. Toto tvrzení však platí tehdy, je-li vzorek sestavován náhodným výběrem. Ten lze při tvorbě korpusu uplatnit jen zřídka, protože obvykle nemáme k dispozici všechny jednotky cílové populace. Většinou tak nemůžeme vyloučit, že jsou rozličné jevy v korpusu zastoupeny systematicky odlišně, než je tomu v populaci. Pokud například mluvíme o psané češtině, spadají sem mimo publicistiky, beletrie a odborné literatury (tedy tří základních kategorií zastoupených v českých „reprezentativních" korpusech) také e-maily, esemesky, blogy, jídelní lístky, nápisy na vývěsních štítech a vše možné, na co si vzpomeneme. Pokud jsou tyto složky populace v korpusu systematicky opomíjeny, je pravděpodobné, že nám určité aspekty užívání jazyka unikají, a to bez ohledu na velikost korpusu jako takovou.

Rovněž je potřeba si uvědomit, že existují rozdíly mezi různými jazykovými jevy v jejich variabilitě a v jejich frekvenci. Můžeme odůvodněně předpokládat, že na nižších jazykových úrovních (např. na morfologické) je variabilita nižší než na úrovních vyšších (například na textové). Čím nižší je variabilita, tím menší je pravděpodobnost, že se nereprezentativní stavba korpusu zásadně promítne do nesouladu našich výsledků s populací. Čím nižší je variabilita, tím větší je šance, že zachytíme všechny různé varianty v dostatečné míře na to, abychom mohli spolehlivě popsat jejich užívání. Z perspektivy jednotlivých prostředků bychom tak mohli mluvit o tom, že tentýž korpus může být reprezentativní pro určité prostředky (tyto prostředky jsou v něm zastoupeny v dostatečném množství a jejich distribuce odpovídá distribuci reálné) a zároveň nereprezentativní pro prostředky jiné.

Reprezentativní korpusy

Dosud jsme explicitně či implicitně mluvili o velkých, takzvaně reprezentativních korpusech celého jazyka. Došli jsme k tomu, že tyto korpusy ze statistického hlediska reprezentativní nejsou a že tak nevíme (a nemůžeme vědět), zda výsledky, které získáme na základě analýzy dat z těchto korpusů, skutečně odpovídají jazykové realitě, nebo nikoli.

Výše uvedené ovšem neznamená, že statistická reprezentativnost v korpusové lingvistice je nedosažitelná. Alternativou k „reprezentativním" korpusům celého jazyka mohou být v tomto smyslu specializované korpusy, u nichž je možné docílit solidní reprezentativnosti vzhledem k jasně omezené a sociologicky doložitelné populaci. Specializovanost korpusu nemusí nutně znamenat jeho textovou omezenost. Můžeme si například představit velice dobře reprezentativní korpus publicistiky určitého období. Domnívám se ostatně, že jedním z nejreprezentativnějších existujících českých korpusů je korpus SYN2009PUB, který obsahuje 700 milionů slov z publicistických textů z let 1995–2007. V případě korpusu publicistiky totiž můžeme ohraničit víceméně celou populaci (určité problémy by mohly nastat například tehdy, pokud bychom chtěli počítat periodika nemající ISSN, to by však mělo být zanedbatelné). Díky tomuto korpusu můžeme velice dobře poznat a popsat, jak funguje čeština v publicistických textech – tento korpus můžeme vcelku bez obav využít ve všech třech funkcích uvedených výše, avšak s tím omezením, že naše výsledky neplatí pro češtinu jako takovou, ale pouze pro česky psanou – a publikovanou – publicistiku.

Statisticky reprezentativní korpus může být vytvořen i pro mluvený jazyk, i když ani zde nemůže jít – z technických a finančních důvodů – v pravém slova smyslu o mluvenou češtinu jako takovou. Vzhledem k tomu, že reprezentativnost závisí na sociálních i stylových aspektech variace, můžeme předpokládat, že při snaze o reprezentativní korpus mluvené češtiny musíme zohlednit alespoň následující sociální proměnné (u nichž očekáváme vliv na užívání jazyka): věk, sociální status či třída, pohlaví, bydliště (respektive místo, kde osoba vyrůstala). Zároveň je nutné rozlišit různé styly (běžná mluva, formální projev apod.). Pokud bychom si – dosti hrubě – operacionalizovali věk na tři hodnoty (mladší, střední, starší generace), sociální status rovněž na tři (nízký, střední, vysoký), původ na 14 hodnot na základě krajů a kdybychom styly – velmi zjednodušeně – rozlišili na nízké, střední a vysoké, znamenalo by to celkem 756 kategorií mluvčích. Pokud bychom tak pracovali pouze s nahrávkou 1 člověka v každé kategorii, potřebovali bychom 756 nahrávek. Mezi mluvčími však existují individuální rozdíly, navíc potřebujeme relativně velký objem řeči, aby v něm byly v dostatečné míře zachyceny proměnné, které potenciálně chceme zkoumat. Při snaze o reprezentativnost pro mluvenou češtinu bychom se tak dostali k tisícům až desetitisícům hodin projevů. Ani to však nezaručí, že bude tento

korpus využitelný pro všechny možné účely, protože při výběru vzorku například nerozlišujeme téma hovoru, což může vyústit například v problém, že se v projevech neobjeví určitá část slovní zásoby.

Otázkou ovšem je, proč bychom vlastně měli usilovat o popis mluvené češtiny jako takové. Z tohoto hlediska by bylo podle mého názoru rozumným řešením – na základě technických a finančních možností – preferovat vytváření ohraničených mluvených korpusů, které by byly reprezentativní pro určitý výsek českých mluvčích a také pro určitý výsek řečových stylů. Například bychom se mohli zaměřit na běžnou mluvu mluvčích z Hradce Králové, jejichž rodiče pochází z Hradce Králové a okolí. Korpus tohoto typu může být realizovatelný, sociálně reprezentativní a můžeme na jeho základě dospět k relativně přesné představě o určitém výseku mluvčích a jejich jazyka. Různá dílčí zjištění z takovýchto ohraničených korpusů lze rovněž porovnávat mezi sebou, a docházet tak k obecnějším zjištěním.

Závěr

V tomto článku jsme se věnovali pojmu reprezentativnost, jak je užíván v korpusové lingvistice. Zjistili jsme, že reprezentativnost v korpusové lingvistice není chápána stejně jako reprezentativnost statistická, což je problematické zvláště v případě, že chceme z korpusů vyvozovat zobecnění pro širší populaci (například pro jazyk jako takový). Z naší úvahy vyplývá, že takzvané reprezentativní korpusy určitého jazyka jsou z technického hlediska nereprezentativní, a tedy nevyužitelné pro některé funkce, které lingvisté od korpusu očekávají. O reprezentativnosti lze mluvit pouze u specializovaných korpusů, které mají jasně ohraničenou populaci. Právě tyto korpusy mohou poskytnout lingvistům solidní oporu při poznávání toho, jak jazyk ve svých různých formách skutečně funguje.

LITERATURA

BIBER, D. (1993): Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8, s. 243–257.

CVRČEK, V. – KOVÁŘÍKOVÁ, D. (2011): Možnosti a meze korpusové lingvistiky. *Naše řeč*, 94, s. 113–133.

KRÁLÍK, J. – ŠULC, M. (2005): The Representativeness of Czech corpora. *International Journal of Corpus Linguistics*, 10, s. 357–366.

KŘEN, M. (2013): *Odraz jazykových změn v synchronních korpusech*. Praha: Nakladatelství Lidové noviny.

KUČERA, K. (2002): The Czech National Corpus: Principles, design, and results. *Literary and Linguistic Computing*, 17, s. 245–257.

LEECH, G. (2007): New resources, or just better old ones? The Holy Grail of representativeness. In: M. Hundt – N. Nesselhauf – C. Biewer (eds.), *Corpus Linguistics and the Web*. Amsterdam – New York: Rodopi, s. 133–149.

McENERY, T. – WILSON, A. (2004): *Corpus Linguistics. An Introduction*. 2nd edition. Edinburgh: Edinburgh University Press.
MILROYOVÁ, L. – GORDON, M. (2012): *Sociolingvistika: metody a interpretace*. Praha: Karolinum.
SINCLAIR, J. (1996): *EAGLES Preliminary Recommendations on Corpus Typology* [online]. Cit. 2014-06-19. <http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html>.
SINCLAIR, J. (2005): Corpus and text – Basic principles. In: M. Wynne (ed.), *Developing Linguistic Corpora: A Guide to Good Practice* [online]. Cit. 2014-06-19. <http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm>.
STEFANOWITSCH, A. – GRIES, S. (2003): Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8, s. 209–243.
STEFANOWITSCH, A. – GRIES, S. (2005): Covarying collexemes. *Corpus Linguistics and Linguistic Theory*, 1, 1–43.
TOGNINI-BONELLI, E. (2001): *Corpus Linguistics at Work*. Amsterdam – Philadelphia: John Benjamins.
VÁRADI, T. (2001): The linguistic relevance of Corpus Linguistics. In: P. Rayson – A. Wilson – T. McEnery – A. Hardie – S. Khoja (eds.), *Proceedings of the Corpus Linguistics 2001 Conference, Lancaster, 2001*. Lancaster: Lancaster University, s. 587–594.

Ústav českého jazyka a teorie komunikace FF UK
nám. Jana Palacha 2, 116 38 Praha 1
jan.chromy@ff.cuni.cz