

Univerzita Karlova v Praze

Filozofická fakulta

Ústav Českého národního korpusu

Matematická lingvistika

Teze disertační práce

Zkoumání hierarchizace pádů českého substantiva

v sémantických (kolokačních) třídách

Exploring case forms of Czech nouns and their ranking in semantic

(collocational) classes

Vedoucí práce: Prof. PhDr. Karel Kučera, CSc.

2015

Karin Schöne

Záměr a pozadí

Předkládaná práce si klade za cíl zjišťovat a analyzovat frekvenční distribuci substantivních pádových tvarů v současné češtině. Záměrem studie je identifikovat typické užití jednotlivých lemmat pomocí dat z Českého národního korpusu.

Námět práce vychází z analýzy textů nerodilých mluvčích češtiny a jejich frekventovaných chyb při produkci substantivních tvarů, a má tudíž především didaktické pozadí. Zajímalo nás, o které vzory (zástupce deklinačních tříd) nebo tvary se studenti opírají při produkci substantivních tvarů. Chybová analýza ukázala, že v 80 procentech nesprávného užití morfologické substantivní koncovky studenti volí tvar patřící sice ke stejnému paradigmatu, ale k jinému pádu. Nejčastěji se v takových případech vybírá nominativ singuláru¹, což je signálem jeho dominantního postavení v učebnicích a ve výuce. Navzdor tomu mohou však být i tvary akuzativu nebo lokálu upřednostňovány v určitých syntaktických konstrukcích, zejména v případě volného doplnění k určení směru nebo místa. Užití lokálních tvarů je studenty zjevně upřednostňováno u substantiv s lokální sémantikou, jako jsou názvy měst. U těchto slov je lokální tvar zřejmě lépe přístupný a zakotvený. Vzhledem k využití deklinačních vzorů lze shrnout, že studenti v případě nejistoty skloňovali substantiva přináležející k méně často zastoupeným vzorům spíše podle vzoru dominantního pro daný rod.

Poznatky bádání v oblasti osvojování jazyka ukazují, že učení je pevně spjato s vyhodnocováním informací o frekvenci a distribuci forem v textu. Zkušenost s frekvencí a distribucí jazykových jednotek a jejich struktur (či konstrukcí) tedy řídí jejich kognitivní reprezentaci. Student, který chce úspěšně aplikovat určitou konstrukci, určitý vzorec (*pattern*), se musí setkat s určitým množstvím odlišných příkladů (Bybee, 2008, str. 221). Velký význam pro zapamatování a zobecňování konstrukcí mají jejich prototypičtí zástupci (srov. např. Goldberg a Casenhiser, 2008; Ellis, 2014). Ellis (2014) postuluje, že čím větší je frekvence konkrétního exempláře, tím více přispívá k vymezení kategorie (tamtéž, str. 201).

Frekvence slovních tvarů jako kritérium výběru slov a jejich tvarů se proto přikládá v oblasti SLA² velkou závažnost. Dále proto diskutujeme v této práci roli frekvence substantivních tvarů a jejich kolokací při výběru a prezentaci substantivních lexémů.

¹ Přičemž u slov vzorů *hrad* a *město* je totožný s akuzativem.

² Second language acquisition = osvojování (prvního) cizího jazyka.

V učebnicích, jež se v této práci analyzují vzhledem k prezentaci pádů obecně a substantivních tvarů konkrétně, narážíme často na dominantní postavení paradigmatického pohledu na substantiva. Při vysvětlování substantivních morfologických paradigmat pádů a jejich funkcí v příslušných cvičeních je nápadný právě nedostatek příkladových konstrukcí. Vzhledem k omezenému času ve třídě se zdá rozumné, aby se vyučování koncentrovalo právě na frekventovaná slova a jejich kolokace (srov. např. Kennedy, 2008). Sinclair a Renouf (1988) navrhují, že by se student angličtiny měl soustředit na nejběžnější slovní tvary, na jejich centrální vzory užívání, a na jejich typické kombinace (tamtéž, str. 147). Zohledňování frekvence v inputu však není univerzálním lékem. Podle Schmidta (1990) je rovněž uvědomění a všímání si jevu předpokladem pro to, aby ho studenti později aktivně zvládli. Robinson a Ellis (2008) potvrzují pozitivní vlivy, jež má obrácení pozornosti studentů na formální jevy při zpracovávání a vyhodnocování textů.

Na tomto pozadí je závažným úkolem učebnic ukázat studentovi užité podstatných jmen v různých kontextech. Kurikula musí látku nutně parcelovat a vybírat. V počáteční výuce se nabízejí situace nejběžnější pro cílovou skupinu. Podstatná jména se zde mohou prezentovat v komunikativních kontextech, které jsou pro ně typické, a to rovněž v příslušných tvarech a kolokacích. Výběr textů a formulace úloh by měly studenty podporovat v tom, aby se soustředili jak na funkci tvaru v kontextu, tak na tvar samotný a na jeho roli v systému. Na základě výše řečeného považujeme za důležité, aby se funkce pádu vyjasnila na typických konstrukcích a na dostatku konkrétních příkladů, tedy pádových forem v kontextu. K tomuto účelu může být užitečné identifikovat typické chování substantiv v textu a rozložení jejich tvarů v současném úzu roditelých mluvčích.

Východiska a hypotézy

Analýza, kterou za tímto účelem provádíme, záměrně volí korpusový přístup k popisu daného jazykového jevu, tj. substantivní morfologie. Pomocí korpusů je možné ve velkém měřítku zkoumat nejen frekvenční distribuci ale i kontext slov. Korpusová lingvistika dokáže prokázat na konkrétních datech, že forma a význam jsou vzájemně propojené a že syntagmatické vztahy slova pomáhají vymezovat jeho význam. Sémantická pravidla, jež určují, jak se slova smysluplně kombinují, a zároveň pravidla, která řídí kombinace s gramatickými třídami, se považují za součásti jejich významů (srov. mj. Sinclair, 2014). Vzhledem k závažnosti gramatických a sémantických pravidel při tvoření syntaktických celků se má proto v korpusové lingvistice za to, že sémantická pravidla mají přednost před

gramatickými (srov. Čermák, 2006, str. 10; Čermák, 2010, str. 126). Vycházíme v této práci z toho, že kontext jazykového jevu významně přispívá k utváření jeho vlastností, včetně jeho gramatických kategorií nebo i pragmatických vlastností (Cvrček, 2013, str. 23) a že tudíž slovní tvar, jenž prokazuje konkrétní gramatické vlastnosti a zahrnuje v sobě možné funkce v textu, leccos napovídá o svém kontextu. Z výše řečeného spíše plyne, že právě slovní tvar je tou jednotkou, jež v sobě nese parametry, které řídí souvztažnosti s jinými jednotkami v textu.

V této práci se pokoušíme roztrždit substantiva na základě společné preference realizovat se s určitým slovním tvarem a popsat vzniklé třídy vzhledem k jejich sémantických charakteristik. Rovněž chceme na několika příkladech ověřovat, zda preference obdobných kontextů koreluje s upřednostňováním stejných slovních tvarů. Také si klademe otázku, zda se obecnější sémantické kategorie, tj. věci konkrétní a abstraktní, lidé, živočichové a místa³, projevují i v odlišných preferencích morfologických profilů.

Postup

Ve smyslu analýzy *corpus-driven*⁴ se pokoušíme nejprve pozorovat frekvenční distribuci zkoumaného jevu v korpusu na menším pilotním vzorku. Na základě výsledků a zjištěných podobností v chování postupujeme k hypotéze o možném roztržení. Tuto hypotézu potom ověřujeme na větším náhodném vzorku, který vytváříme ze základního souboru substantiv sestavovaného na základě didaktických a zároveň frekvenčních hledisek⁵.

Analýzu jak pilotního, tak vlastního náhodného vzorku provádíme v lemmatizovaném a morfologicky značkováném korpusu SYN2010. Za vhodné východisko pro naši analýzu považujeme korpus SYN2010, poněvadž je korpusem reprezentativním. Zrcadlí tedy

³ Srov. klasifikaci v Čermák, 2010, str. 165n.

⁴ Dle Tognini-Bonelli (2001) se postup *corpus-driven* na rozdíl od postupu založeného na korpusu vyznačuje právě empiričností: „The corpus-driven approach, on the other hand, fits well the definition in that it aims to derive linguistic categories systematically from recurrent patterns and the frequency distribution that emerges from language in context“ (tamtéž, str. 87).

⁵ Opíráme tuto studii o základní soubor dat, který by měl zajistit, aby byly poznatky závažné pro danou cílovou skupinu a na druhé straně pro výuku na vybrané úrovni jazykové kompetence (úroveň B1 dle SERR). Základní soubor podstatných jmen byl proto vytvářen ve třech krocích: 1. zjišťování vyjadřovacích potřeb studentů češtiny, 2. vymezení vyjadřovacích prostředků, které jsou potřebné pro řešení komunikačních úloh v identifikovaných komunikačních situacích na určité úrovni osvojovacího procesu za pomoci *Prahové úrovně* (Bischofová et al., 2001) a učebnic *Tschechisch Schritt für Schritt* (Holá, 2006) a *Česky krok za krokem 2* (Holá a Bořilová, 2009) a 3. úprava výběru substantiv na bázi frekvence výrazů v soudobém úzu rodilých mluvčích pomocí frekvenčního slovníku *Frequency dictionary of Czech. Core vocabulary for learners* (Čermák a Křen, 2011).

v dostatečné míře variabilitu v jazyce, popř. obsahuje v dostatečném množství texty všech variet⁶. Texty zahrnuté v korpusu SYN2010 odrážejí tudíž jazyk, kterým byl obklopen běžný čtenář v České republice v letech 2005–2010. Z tohoto hlediska ho považujeme i za reprezentativní pro jinojazyčného mluvčího, který buď pobývá v České republice na určitou (delší) dobu, nebo takový pobyt zamýšlí, a chce se proto učit češtině jako cílovému jazyku⁷. Na základě typů textů, s kterými se student musí umět vypořádat, jsme přizpůsobili korpus cílové skupině a vymezili Subkorpus SYNEK2010.

Při vyhledávání tvarů nutně využíváme danou lemmatizaci, jež se opírá o morfologický slovník o rozsahu 350 000 lemmat včetně lemmat propriálních (Jelínek a Petkevič, 2011, str. 159). Jsme si vědomi problému, na nějž již upozornil Sinclair (Sinclair 1991, str. 8⁸), že tvary nereflektují automaticky chování celého lemmatu. Podle něj jsou tedy svědky hlavně a jenom samy pro sebe (tamtéž). Postupujeme proto tak, že na základě pozorování slovních tvarů, které nám jsou nutně přístupné jen přes jejich lemma, činíme závěry o tom, zda odlišné užití neukazuje na zvýraznění určité sémantické složky uvnitř jednoho lemmatu.

Vyhodnocení a závěr

Pilotní sonda do korpusu na malém vzorku 50 slov ukázala, že se frekvence pádových tvarů pro konkrétní lexémy významně liší od průměrného rozložení pádových tvarů v korpusu. Četnost, s níž se určitý lexém vyjadřuje v konkrétním pádu, je do značné míry individuální jev, a to jak v celém korpusu tak v jeho subkorpusech.

Analýza substantiv náhodného vzorku o rozsahu 203 slov potvrdila hypotézu pilotní studie, že je možné roztřídit substantiva na základě jejich morfologických profilů, tj. jejich nejfrekventovanějších dvou až tří slovních tvarů. Porovnání výsledků pilotní a vlastní studie ukázalo, že objem typů, tedy počet substantiv, která se nejčastěji vyskytují v tom či onom pádu, je proměnlivý. Liší se v závislosti na povaze substantiv zvoleného vzorku. Nejfrekventovanějším pádem pro většinu slov byl genitiv a ve větším náhodném vzorku akuzativ. Celkově lze shrnout, že nominativ není a priori typickým pádem z hlediska užití konkrétních jednotlivých substantiv.

⁶ Viz Cvrček & Richterová, 'cnk:syn2010', 2015.

⁷ Cílovým jazykem rozumíme dle Šindelářové a Škodové (2013, str. 4) jazyk, „který jedinec-cizinec nabývá a jehož osvojení je cílem učebních aktivit, v němž chce být mluvčí schop komunikovat“.

⁸ „There is a good case for arguing that each distinct form is potentially a unique lexical unit, and that forms should only be conflated into lemmas when their environments show a certain amount and type of similarity.“

Roztřídění substantiv bylo založeno na dvou nejfrekventovanějších morfologických tvarech lemmat. Pomocí těchto profilů se nám podařilo zachytit substantiva v sedmi vysoko frekventovaných, osmi středně a devíti nízko frekventovaných typech. K sedmi nejčastějším typům lze přiřadit celkem 76,85 % zkoumaných substantiv. U jednotlivých typů bylo dále ověřováno, zda se substantiva seskupují po přiřazení další, třetí pozice frekvenčního pořadí slovních tvarů.

Největší skupina substantiv, typ <S2-S4>, obsahuje přirozeně nejvíce takových podtypů a zároveň i povahově celkem odlišné lexémy. Tento typ je však typický pro jména látková a názvy oborů. Typ <S4-S1> lze sémanticky těžko podchytit. Je to skupina slov převážně počítatelných a těch abstrakt, jejichž denotát může nabýt i konkrétnější povahy (např. u lemmat *návrh*, *oslovení*, *vzkaz*). Zřejmou přednost dávají názvy měsíců a s nimi i názvy jazyků typu <S2-S6>. Jinak je tento profil specifický pro substantiva s lokálním významem. Přiřazení třetí pozice jak u typu <S2-S6>, tak u typu <S6-S4> ukazuje, že substantiva těchto skupin upřednostňují singulárové tvary lokálu, nominativu a akuzativu, přičemž lokálový tvar je vždy buď prvním, nebo druhým nejčastějším tvarem. Kombinace tvarů lokálu a nominativu singuláru typu <S1-S6> je typická pro názvy měst, které mají i po přiřazení třetí pozice celkem výrazný profil <S1,S6,S2> nebo <S6,S1,S2>. Typ <S1-P1> nápadně hromadí nejvíce názvů osob, a to jak označení povolání, tak příslušníků národností.

V rámci omezeného náhodného vzorku se tedy podařilo identifikovat morfologické profily, které můžeme označit jako typické pro určitou sémantickou třídu substantiv. Zároveň by se dalo říci, že pro některé sémantické třídy substantiv jsou určité profily prototypické. Pro následující skupiny substantiv obsažené v základním souboru této studie je tomu takto:

- názvy oborů (lidské činnosti): <S2,S4,S1> a <S2,S1,S4>
- názvy měst: <S1-S6>, konkrétně <S1,S6,S2> a <S6,S1,S2>
- názvy zemí: <S6-S2>, konkrétně <S6,S2,S1>
- názvy měsíců: <S2-S6>, konkrétně <S2,S6,S1> a <S6,S2,S1>
- názvy jazyků: <S2-S6>, konkrétně <S2,S6,S4> a <S6,S2,S4>
- názvy příslušníků národností: <S1-P1>, konkrétně: <P1,S1,P2>
- obecné názvy míst: <S2-S6>
- obecné názvy osob: <S1-P1> a <S1-S2>
- substantiva látková: <S2-S4>, konkrétně: <S2,S4,S1> a <S4,S2,S1>

Pro zjišťování, zda substantiva sdílejí obdobné kolokační vzorce, které by odůvodňovaly častý výskyt právě těchto dvou preferovaných tvarů, jsme se podívali příkladově na kolokační profily některých substantivních lexémů s morfologickým profilem <S2-S4,S1>.

K tomuto účelu byly vyhledávány kolokáty dvou nejčastějších morfologických tvarů, tedy Gsg. a Asg., v levém a pravém kontextu od KWIC v rozsahu od -3 do 3. Nalezené kolokáty byly potom seřazené podle jejich absolutní frekvence v kontextu zkoumaného tvaru v sestupném pořadí. Po prozkoumání frekventovaných kolokací tvarů prvního typu lze obecně shrnout, že kolokace morfologického tvaru dávají dostatečně specifickou informaci, aby bylo možné srovnávat různé lexémy vzhledem k jejich chování v textu. Ve zkoumané třídě substantiv s profilem <S2-S4,S1> bylo možné identifikovat a popsat několik obecnějších kolokačních vzorců. Ty ovšem platí vždy jen pro několik zástupců třídy. Nejvíce výrazný kolokační profil mají jména látková a názvy oborů zahrnuté v této třídě. Jména abstraktní a konkrétní jinak nevynikají zvláštním profilem.

Vzhledem k distinkci abstraktum-konkrétum se ukázalo, že typy, k nimž se přiřklání nejvíce abstrakt, jsou identické s typy, které upřednostňují i konkréta, i když jsou v nich zastoupené v odlišných poměrech. Celkem je nominativ o něco častěji preferován věcmi konkrétními, mezi abstrakty je více lemmat nejčastěji zastoupených v akuzativu. U obou tříd najdeme v čele frekvenčního žebříčku i profily s plurálovými tvary. Zřejmě se třídění dle předložených morfologických profilů nehodí k tomu, aby odráželo odlišné chování konkrét a abstrakt. Spíš poukazuje na podobnosti v užití, které mohou být dány např. zvýrazněním konkrétního významu jinak abstraktních substantiv. Nápadné bylo například, že je několik málo abstrakt, jež se preferovaně vyskytují v textu v plurálových tvarech (např. *požadavek, přesčas, léto, koníček*), a že se ve skupině s profilem <S1-S4> nacházejí především abstrakta, jejichž význam se může v textu zkonkretizovat (např. *letecká linka, předpověď počasí, zpráva*).

Substantiva sémantické kategorie „člověk“ (HUM) se oproti tomu odlišují od ostatních kategorií hned třemi způsoby. Vytváří profil, totiž <S1-P1>, jenž slova jiných kategorií nemají. Je v něm zahrnuto nejvíce zástupců této kategorie. Jinak platí pro všechny názvy osob, že jeden z prvních dvou preferovaných tvarů je tvarem nominativním. Není divu, že jako jediná kategorie tvoří profil s vokativem. Z kategorií živočichů bylo ve vzorku příliš málo zástupců, aby bylo z jejich preferencí pádových tvarů možné odvodit závěry pro celou kategorii. Neudivuje výsledek, že se substantiva označující místo vyznačují lokálem singuláru jako nejčetnějším tvarem. Pro tuto kategorii byly identifikovány typy <S2-S6> a

<S1-S6> jako typické profily, přičemž poslední je preferován zejména názvy měst. Kromě lokálu, který však v některých profilech těchto substantiv mezi dvěma nejčastějšími tvary není (např. u lemmat *město*, *byt*, *domek*, *lokalita*), je důležitým pádem zejména genitiv.

Dalo by se říci, že právě kategorie ‚místo‘, a ‚člověk‘ ukazují preference pro určité morfologické profily. Celkem lze shrnout, že vzniklé typy naznačují preferenci pro určité sémantické složky, nejsou však vždy jasně a ostře vymežitelné vzhledem k sémantické povaze jejich členů.

Závislost frekvence pádů na funkčním stylu a sféře užívání

V kapitole 4 disertační práce ukazujeme, že obecná frekvenční distribuce pádů je závislá na funkčním stylu a sféře užívání, zvláště frekvence genitivu, akuzativu, ale i vokativu. Kolísání frekvence pádů ovlivňuje mnoho faktorů, zejména podíl určitých slovtvorných typů, užití deverbálních substantiv a pasivních konstrukcí.

Na pilotním i na větším náhodném vzorku jsme viděli, že četnost, s níž se určitý lexém vyjadřuje v konkrétním pádu, je do značné míry individuální jev. Obecnou tendenci, že by výskyt konkrétního pádu závisel významně i na funkčním stylu, jsme pozorovali v rámci pilotního průzkumu jen pro vokativ a dativ. Ovšem i tu záleží na sémantice slova, zda a v jakém počtu tyto pády hrají vůbec roli ve frekvenční distribuci tvarů. V celkovém přehledu frekvenční distribuce pádů se tyto pády ocitají ve všech stylech a také v mluvené češtině na spodním konci frekvenčního žebříčku. Jinak jsou změny ve frekvencích jednotlivých pádů v tom či onom subkorpusu významné vždy jen pro konkrétní slova. Ani v užití slov v mluvené češtině se nám nepodařilo identifikovat výraznější tendenci, která by byla platná pro všechna zkoumaná substantiva, než tu, již pozorujeme při souhrnném pohledu na frekvence pádů: tvary nominativu a akuzativu jsou v mluvené podobě češtiny celkem čteněji zastoupené než v psané podobě. Vliv mluvenosti na frekvenční pořadí morfologických tvarů je však silný. Jen u velmi málo lemmat zůstane frekvenční pořadí totožné. Jen 18 % slov pokrývá 85 % všech jejich užití stejnými morfologickými tvary. Ovšem alespoň u poloviny zkoumaných slov se nemění celkové složení tří nejfrekventovanějších tvarů. Jak bylo již naznačeno, stoupá u mnoha slov v pořadí akuzativ. Které pády se „v jeho prospěch“ posouvají na další pozice frekvenčního pořadí, je opět pro jednotlivé lexémy velmi různé. Ačkoliv tedy pády celkově (bez přihlídnutí k jednotlivým lexémům), obzvláště genitiv, akuzativ, vokativ a v omezené míře i dativ, reagují na změnu funkčního stylu a stylu projevu, mají změny ve frekvencích pádových tvarů konkrétních lexémů spíše individuální ráz.

V kapitole 5 bylo diskutováno, zda hypotéza, že jsou první tři nebo čtyři jevy na sestupně řazeném frekvenčním žebříčku obvykle relativně nejfrekventovanějšími, platí i pro zde zkoumaný výskyt morfologických tvarů (viz k hypotéze Těšitelová, 1990). Proto byla pro každé lemma náhodného vzorku spočítána kumulovaná frekvence jeho tvarů (seřazených podle frekvence). Výsledky dávají najevo, že 192 z 203 lemmat (95 %) pokrývá prvními třemi nejfrekventovanějšími tvary 50 % všech svých výskytů v textu, zbylých jedenáct lemmat k tomu potřebuje čtyři tvary. Bereme-li v potaz, že 90,6 % substantiv má celkem devět tvarů, 82,2 % z nich má deset a 72,4 % dvanáct tvarů, mohli bychom s opatrností říci, že se - v souladu se Zipfovým zákonem a hypotézami Těšitelové (srov. Těšitelová 1990 a kapitolu 4) - nejvíce tvarů nachází na nejvyšších třech (max. čtyřech) pozicích ve frekvenčním pořadí. Nevyřešíme v této práci ovšem otázku, zda můžeme zmíněných 50 % všech tvarů lemmatu považovat za jeho jádro – za to, co je pro něj typické.

Závěry pro prezentace substantiv v učebnicích

Podle našeho názoru mohou být identifikace morfologického profilu lemmatu, tj. jeho nejtypičtějších tvarů, a zjišťování častých kolokací vhodným způsobem vymezování náležitých případů jeho užití. Máme za to, že by se ve slovníkových částech učebnic měly uvádět mimo tvaru Nsg. také alespoň dva nejfrekventovanější tvary pokud možno s minimálním, zobecnovaným kontextem. U příslušníků sémantických tříd, jež spojuje podobný morfologický a sémantický profil, by se to mohlo uskutečnit pomocí odkazu na příkladový kolokační profil jednoho zástupce třídy. Předpokladem takového přístupu je samozřejmě analýza kolokačních profilů více slov než těch, která mohla být prozkoumaná v této práci. Příkladně v práci znázorníme navržený přístup na několika lexémech z náhodného vzorku. Je potřeba doplnit, že by se výběr příkladových kolokací měl řídit v ideálním případě frekvencí výskytu nejenom v korpusu psané, ale také mluvené češtiny.

Při sestavování látky pro učebnice nemůže být frekvence lemmat ani tvarů vždy předním kritériem výběru. Poněvadž texty, které se využívají v učebních materiálech, nabízejí vlastní konstrukce k probranému tématu, není vždy možné ani smysluplné texty frekvenčním profilům slov přizpůsobit. Avšak máme za to, že informace o morfologických profilech lze zohledňovat při sestavování slovní zásoby pro cvičení popř. pro tematické přehledy, při sestavování slovníkové části resp. doplněk učebnic, ale také při explikaci funkcí pádů a konstrukcí typických pro daný pád, resp. typických zástupců těchto konstrukcí.

Literatura

- BENEŠOVÁ-GRUNDOVÁ, D. 2006. Jak začlenit sociokulturní aspekty do výuky anglicky mluvících studentů : příklad české verze programu Euromobil. In ŠTINDLOVÁ, B.; ČEMUSOVÁ, J. (ed.) *Sborník Asociace učitelů češtiny jako cizího jazyka (AUČCJ)*, 2005-2006. Akropolis, 2006, s. 113-126. ISBN 80-86903-37-0.
- BISCHOFOVÁ, J.; H. CONFORTIOVÁ; J. CVEJNOVÁ; et al. 2001. *Prahová úroveň : čeština jako cizí jazyk*. Strasbourg : Council of Europe, 2001. ISBN 92-871-4761-2.
- BYBEE, J. 2008. Usage-based grammar and second language acquisition. In ROBINSON, P.; ELLIS N. C. (ed.) *Handbook of Cognitive Linguistics and Second Language Acquisition*. New York : Routledge, 2008, s. 216-236. ISBN 0805853510.
- ČERMÁK, F. 2006. Kolokace v lingvistice. In ČERMÁK, F.; ŠULC, M. (ed.) *Kolokace*. Praha : Nakladatelství Lidové noviny, Ústav Českého národního korpusu, 2006.
- ČERMÁK, F. 2010. *Lexikon a sémantika*. Praha : Nakladatelství lidové noviny, 2010. ISBN 978-80-7422-020-3.
- ČERMÁK, F.; KŘEN, M. 2011. *A Frequency Dictionary of Czech : Core Vocabulary for Learners*. New York : Routledge, 2011. ISBN 13: 978-0-415-57661-1.
- ČERMÁK, F., ADAMOVIČOVÁ, A.; PEŠIČKA, J. 2001. *Pražský mluvený korpus (PMK) : přepisy nahrávek pražské mluvy z 90. let 20. století* [online]. Ústav Českého národního korpusu FF UK. [cit. 2011-04-01]. Dostupný z WWW <<http://www.korpus.cz>>.
- ČERMÁK, F. et al. 2005. *SYN2005 : žánrově vyvážený korpus psané češtiny*. Ústav Českého národního korpusu FF UK, Praha 2005. Dostupný z WWW <<http://www.korpus.cz>>.
- CVEJNOVÁ, J. 2006. Úroveň B1 - Prahová úroveň. In ŠTINDLOVÁ, B.; ČEMUSOVÁ, J. (ed.) *Sborník Asociace učitelů češtiny jako cizího jazyka (AUČCJ)*, 2005-2006. Akropolis, 2006, s. 27-38. ISBN 80-86903-37-0.
- CVRČEK, V.; RICHTEROVÁ, O. 2015. 'cnk:syn2010'. *Příručka ČNK*. [Online], 17.02.2015 [cit. 12.04.2015]. Dostupný z WWW <<http://wiki.korpus.cz/doku.php?id=cnk:syn2010&rev=1424167381>>.
- CVRČEK, V. 2013. *Kvantitativní analýza kontextu*. Praha : Nakladatelství Lidové noviny, 2013. ISBN 978-80-7422-264-1.
- ELLIS, N.C. 2014. Frequency-based accounts of Second Language Acquisition. In GASS, S.M.; MACKEY, A. *The Routledge Handbook of Second Language Acquisition*. New York : Routledge, 2014, s. 193-210. ISBN 9780415709811.
- GOLDBERG, A.E.; CASENHISER, D. 2008. Construction learning and second language acquisition. In ROBINSON, P.; ELLIS, N.C. (ed.) *Handbook of Cognitive Linguistics and Second Language Acquisition*. London : Routledge, 2008, s. 197-215. ISBN 0805853510.
- HOLÁ, Lída. 2006. *Tschechisch Schritt für Schritt*. Praha : Akropolis, 2006. ISBN: 978-80-86903-35-4.

- HOLÁ, L. a BOŘILOVÁ, P. 2009. *Česky krok za krokem 2*. Praha : Akropolis, 2009. ISBN 978-80-86903-92-7.
- JELÍNEK, J.; BEČKA, J. V.; TĚŠITELOVÁ, M. 1961. *Frekvence slov, slovních druhů a tvarů v českém jazyce*. Praha : Státní pedagogické nakladatelství, 1961.
- JELÍNEK, T.; PETKEVIČ, V. 2011. Systém jazykového značkování korpusů současné psané češtiny. In PETKEVIČ, V.; ROSEN, A. (ed.). *Korpusová lingvistika Praha 2011 : 3 Gramatika a značkování korpusů*. Praha : Nakladatelství Lidové noviny, 2011, s. 154-170. ISBN 978-80-7422-116-3.
- KENNEDY, Graeme. 2007. Phraseology and language pedagogy : Semantic preference associated with English verbs in the British National Corpus. In MENUIER, F.;
- ROBINSON, P.; ELLIS, N.C. 2008. Conclusion - Cognitive Linguistics, Second language Acquisition and L2 instruction : issues for research. In ROBINSON, P.; ELLIS, N.C. (ed.) *Handbook of Cognitive Linguistics and Second Language Acquisition*. London : Routledge, 2008, s. 489-545. ISBN 0805853510.
- SCHMID, S. 1997. The Naturalness Differential Hypothesis : Cross-linguistic Influence and Universal Preferences in Interlanguage Phonology and Morphology. *Folia Linguistica*. XXX13-4, 1997, s. 331–348. ISSN 0165-4004.
- RADA EVROPY. COUNCIL FOR CULTURAL CO-OPERATION. MODERN LANGUAGES DIVISION. 2006. *Společný evropský referenční rámec pro jazyky : Jak se učíme jazykům, jak je vyučujeme a jak v jazycích hodnotíme (SERR)*. Překlad Ivanová, J. et al. 2. české vydání. Olomouc : Univerzita Palackého, 2006. ISBN 80-244-1425-2.
- SINCLAIR, J. 1991. *Corpus, concordance, collocation*. Oxford : Oxford University Press, 1991. ISBN 9780194371445.
- SINCLAIR, J. 2004. *Trust the text. Language, corpus and discourse*. London : Routledge, 2004. ISBN: 0-415-31768-1.
- SINCLAIR, J.; RENOUF, A. 1988. A lexical syllabus for language learning. In CARTER, R.; McCARTHY, M. *Vocabulary and Language Teaching*. Longman, 1988, s. 140-158. ISBN 0582553822.
- ŠINDELÁŘOVÁ, J. 2010. Čeština jako cizí jazyk dle SERR. *Metodický portál : Články* [Online]. 15.03.2010 [cit. 2015-04-14]. Dostupný z WWW <<http://clanky.rvp.cz/clanek/c/Z/8143/CESTINA-JAKO-CIZI-JAZYK-DLE-SERR.html>>. ISSN 1802-4785.
- ŠONKOVÁ, J. 2008. *Morfologie mluvené češtiny: Frekvenční analýza*. místo neznámé : Nakladatelství Lidové noviny, 2008. ISBN 978-80-7106-956-0.
- TĚŠITELOVÁ, M. 1990. On Semantics of Nouns from the Quantitative Point of View. *Prague Studies in Mathematical Linguistics*. Vol. 10, 1990, str. 9-24. ISSN 0165-7712.
- TOGNINI-BONELLI, E. 2001. *Corpus Linguistics at work*. Amsterdam : John Benjamins, 2001, ISBN 1588110613.