ČASOPIS PRO MODERNÍ FILOLOGII 99, 2017, Č. 2, S. 167–180

Využití korpusových metod v popisu češtiny: příklad kolostrukční analýzy[1]

OPEN ACCESS

Eva Lehečková (Praha)

CORPUS METHODS AND THE DESCRIPTION OF CZECH: COLLOSTRUCTIONAL ANALYSIS

The paper presents a review of principles and applicability of *collostructional analysis* — a cluster of corpus methods developed by Anatol Stefanowitsch and Stefan Th. Gries since 2003. Collostructional analysis measures the strength of association by which lexemes are attracted to a particular position in a construction. It derives from the Construction Grammar concept of construction as conventionalized pairing of form and meaning. Collostructional analysis allows for systematic measurements of the degree of conventionalization of constructions. The paper introduces various types of collostructional analysis and provides an overview of research areas where collostructional analysis has been employed. It concludes with a case showing how collostructional analysis can be applied to the description of Czech.

KEY WORDS

collostructional analysis, association measures, construction grammar, variability, idiomatic expressions, Czech

KLÍČOVÁ SLOVA

kolostrukční analýza, asociační míry, konstrukční gramatika, variabilita, idiomy, čeština

Rozvoj korpusové lingvistiky po roce 1990 výrazně proměnil podobu výzkumu v četných odvětvích lingvistiky, ne-li již ve většině z nich. Korpusové přístupy k analýze jazykových dat jsou dnes považovány za jednu z reprezentativních metod, jejichž zjištění se obecně hodnotí jako platná a spolehlivá. Jedním z důsledků důvěry vkládané do této metodologie je rozšiřující se spektrum konkrétních kvantitativních metod, které umožňují korpusová data analyzovat různorodými způsoby a přicházet tak na nová zjištění o užívání jazyka. Jednou z těchto metod je takzvaná kolostrukční analýza (v původním znění *collostructional analysis*), jejíž představení je cílem této studie. Vznik kolostrukční analýzy se obvykle datuje k roku 2003, kdy byl publikován první text o této metodě (Stefanowitsch — Gries, 2003) — jedná se tedy o metodu poměrně novou, a proto pochopitelně v českém prostředí dosud málo známou. Z toho důvodu se jeví jako užitečné seznámit českou lingvistickou obec s podstatou kolostrukční analýzy (první oddíl), se způsoby a mezemi jejího uplatnění (druhý oddíl) i s ukázkou toho, jak ji lze využít při popisu českého jazykového materiálu (oddíl třetí).[2]

1 Tato studie vznikla za podpory projektu Univerzity Karlovy Progres č. 4, *Jazyk v proměnách času, místa, kultury*.

2 Původní podoba tohoto příspěvku zazněla na interním semináři pracovníků Ústavu Českého národního korpusu dne 1. 11. 2016. Všem účastníkům bych ráda poděkovala za cennou diskusi a podněty.

ISSN 0008-7386 © FILOZOFICKÁ FAKULTA, UNIVERZITA KARLOVA V PRAZE

OPEN ACCESS

1. VYMEZENÍ KOLOSTRUKČNÍ ANALÝZY

Autory kolostrukční analýzy (dále KolA) jsou původem němečtí badatelé Anatol Stefanowitsch (Freie Universität v Berlíně) a Stefan Th. Gries (Kalifornská univerzita v Santa Barbaře), kteří ji představili sérií čtyř článků v letech 2003–2005 (viz Stefanowitsch — Gries, 2003; Gries — Stefanowitsch, 2004a a 2004b, a Stefanowitsch — Gries, 2005), kolostrukční analýze se však věnují kontinuálně (viz například Stefanowitsch — Gries, 2008, nebo Gries — Wulff(ová), 2009).[3]

Kolostrukční analýza vychází z teoretického zázemí konstrukční gramatiky (*construction grammar*, viz například Hoffmann — Trousdale, 2013, v češtině Fried(ová), 2013). Zejména se opírá o kognitivně-funkční povahu tohoto přístupu, podle nějž jsou základními jednotkami jazyka konstrukce, ustálená spojení s určitými formálními vlastnostmi, významem a funkcí v komunikaci. Vzhledem k tomu, že konstrukční gramatika vychází z užívání jazyka (tj. je *usage-based*), předpokládá, že jazykové znaky vznikají a konvencionalizují se užíváním v komunikaci. Jejich podoba je do určité míry ovlivněna i kognitivními možnostmi a omezeními lidské mysli, v níž se jazykové konstrukce ukládají pravděpodobně ve formě nějaké komplexní uspořádané sítě jednotek a vztahů mezi nimi (podrobněji viz např. Diessel, 2015). Konstrukce jsou jednotky různé velikosti, od lexikálních konstrukcí (*krutopřísný*) přes valenční konstrukce (alternující větné vzorce u slovesa *naložit*, srov. *naložit auto zbožím* vs. *naložit zboží do auta*) nebo víceslovné frazémy (*cestou necestou*, *rád nerad*) po konstrukce větné (*Až naprší a uschne*). V tomto pojetí se tak — v souladu s obecnými principy kognitivní lingvistiky — stírá ostrá hranice mezi slovníkem a syntaxí a je nahrazena kontinuem konstrukcí od lexikálně specifických po otevřená konstrukční schémata, do nichž lze dosadit různé lexémy s náležitými vlastnostmi. Konstrukčněgramatický popis tím, že vychází z reálného užívání jazyka, věnuje pozornost variantnosti konstrukcí v komunikaci, přesto usiluje o identifikaci konstrukcí prototypických (s ohledem na určitý komunikační kontext), pokud existují. A právě k tomu slouží kolostrukční analýza.

Autoři KolA za svůj inspirační zdroj označují výzkum kolokací coby ustálených lexikálně-sémantických spojení slovních tvarů. Princip kolokací autoři přenášejí do vztahu mezi syntaktickou konstrukcí a lexikální jednotkou, odtud motivace pojmenování *kolo-strukce*, které odkazuje jednak ke kolokacím, jednak ke konstrukcím. Kolostrukční analýza je ve skutečnosti souborem několika metod, jejichž společným principem je to, že umožňují sledovat vztah (souvýskyt) mezi konkrétními syntaktickými konstrukcemi[4] a lexikálními jednotkami, které se v těchto konstrukcích vyskytují — těmto lexikálním jednotkám se pak obvykle říká kolexémy (*collexemes*). Vztah mezi konstrukcemi a kolexémy se vyjadřuje pomocí nějakého typu asociační míry.

3 Stefan Th. Gries je důsledným stoupencem volného přístupu k informacím, proto jsou všechny jeho časopisecké studie a kapitoly v monografiích, přinejmenším v rukopisných verzích, dostupné na jeho osobní stránce: http://www.linguistics.ucsb.edu/faculty/stgries/research/overview-research.html#PublicationsEditing [cit. 1. 3. 2017]

4 Lexiko-gramatické vztahy se někdy označují jako koligace, *colligations*, viz Nový encyklopedický slovník češtiny online, 2016, s. v.

OPEN ACCESS

Výsledkem kolostrukční analýzy je seznam lexémů, které signifikantně vůči jiným svým výskytům v korpusu tíhnou k výskytu v rámci určité syntaktické konstrukce. Například Stefanowitsch a Gries (2003) ve svém prvním textu na základě KolA uplatněné na data z Britského národního korpusu ukazují, že k pozici substantiva N v anglické konstrukci [N *waiting to happen*] typicky tíhnou substantiva *accident, disaster, earthquake* nebo *invasion*, ale že silnou asociační míru vykazují i substantiva *recovery*, *dream* nebo *event*, u nichž není inherentně přítomná negativní hodnota. Autoři výsledky KolA konfrontují se slovníkem *Collins COBUILD*, který jediný tuto konstrukci uvádí (pod heslem *accident* a v dokladech u hesla *disaster*), s následujícími závěry: na jednu stranu KolA u této konstrukce potvrdila, že je vhodné tuto konstrukci zmiňovat právě u těchto dvou substantiv, a že jde tedy o nenáhodný vztah. Na druhou stranu upozorňují, že je vhodné tuto anglickou konstrukci vymezovat neutrálněji jako děj, který téměř jistě nastane, jak je patrné v momentu promluvy, a který pouze často, nikoliv výhradně nese negativní konotace.

Gries a Stefanowitsch zdůvodňují potřebnost KolA tím, že zmíněné lexiko-gramatické vztahy mezi konstrukcemi a lexémy nelze efektivně sledovat běžnými kolokačními nástroji. Lze to doložit na příkladu substantiva *švestka* v korpusech Českého národního korpusu. Korpusový manažer KonText umožňuje vyhledat toto lemma a seřadit na základě zadaných kritérií jeho kolokáty. Pokud nás například v SYN2015 zajímají ustálené kolokáty na třech pozicích nalevo od hledaného slova, výsledek (zahrnující asociační míru logDice[5] a relativní frekvenci) řazený podle hodnot logDice vypadá tak, jak ukazují tabulky č. 1 (pro lemmata) a č. 2 (pro slovní tvary)[6]:

pořadí	kolokát — lemma	absolutní frekvence	logDice	relativní frekvence
1	sušený	104,00	10,46	6,44
2	nachytat	13,00	8,34	2,52
3	sbalit	21,00	8,21	1,38
4	hruška	13,00	8,07	1,66
5	hrušeň	8,00	7,98	3,35
6	jabloň	10,00	7,91	1,78
7	slivoň	9,00	7,81	1,77
8	meruňka	8,00	7,75	1,95
9	třešně	9,00	7,69	1,44
10	puchrovitost	5,00	7,66	71,43
11	pecka	9,00	7,51	1,10

TABULKA 1. Nejčastější kolokáty (lemmata) lexému *švestka* podle SYN2015

5 Podrobněji viz https://wiki.korpus.cz/doku.php/pojmy:asociacni_miry [cit. 1. 3. 2017].

6 Pro postižení kolokačního profilu substantiva *švestka* v co největší šíři uvádím nejčastější kolokace jak lemmat, tak slovních tvarů.

OPEN ACCESS

pořadí	kolokát — word	absolutní frekvence	logDice	relativní frekvence
1	sušených	37,00	10,07	11,42
2	sušené	36,00	9,78	6,69
3	sušenými	12,00	8,75	11,11
4	hrušky	10,00	8,24	3,60
5	sušená	7,00	7,99	7,07
6	sušenou	6,00	7,80	7,79
7	hrušně	5,00	7,56	7,69
8	sušené	5,00	7,54	6,58
9	meruňky	5,00	7,43	3,33
10	svejch	5,00	7,27	1,94
11	sbalil	5,00	7,23	1,75

TABULKA 2. Nejčastější kolokáty (slovní tvary) lexému *švestka* podle SYN2015

Podle nástroje SyD (Cvrček — Vondřička, 2011) kolokační profil lemmatu *švestka* na nejvýraznějších pozicích obsahuje lexémy *jablko*, *nachytat*, *jako*, *na*, *nebo* a *pudink*.

Výsledky vyhledávání kolokací nás informují o konkrétních výrazech, s nimiž se lemma *švestka* typicky pojí, a případně o jejich sémantických polích (např. peckovité ovoce, sušené ovoce, ovocné stromy, jejich nemoci). Zároveň je však zřejmé, že jednotlivé kolokáty obsazují různé pozice v konstrukcích, v nichž se spolu s lemmatem *švestka* vyskytují, a že se jedná o vyšší, ale neznámý počet konstrukcí. Informace o konstrukcích se z kolokační analýzy někdy dají vyvodit jednoznačně (viz modifikace substantiva *švestka* adjektivem *sušený*), někdy s velkou mírou pravděpodobnosti: ovoce a ovocné stromy se budou patrně vyskytovat s lemmatem *švestka* v koordinaci,[7] sloveso *nachytat* je patrně součástí frazému *nachytat na švestkách*, předložka *na* bude nejspíše souviset se způsobem přípravy pokrmu. Někdy je podobná úvaha obtížnější: *sbalit* může odkazovat k frazému *sbalit svých pět švestek*, ale švestky lze jistě i sbalit v doslovném významu — to vnáší další rovinu úvahy, která varianta je pravděpodobnější[8]; nebo spojka *jako* ukazuje na srovnávací konstrukci, ale není zřejmé, zda švestka bude srovnávaným pojmem, nebo srovnávacím standardem (nebo obojím?). Hlubší porozumění tomu, co vlastně ukazují uvedené přehledy kolokací, a vyřešení nejasných otázek o užívání lexému *švestka* tak vyžaduje navazující sérii zjišťování.

Gries a Stefanowitsch se podobnými příklady snaží ukázat, že necitlivostí k tomu, na jaké pozici a v jaké konstrukci se daný kolokát nachází, se snižuje výpovědní hodnota kolokačních přehledů, zejména v kontextu takových popisů jazyka, jako je konstrukční gramatika, které předpokládají, že reálná jazyková znalost je komplexní,

7 Poukazují na to i spojky *a* a *nebo* mezi nejobvyklejšími kolokacemi podle SyD — není však jasné, zda se *švestka* vyskytuje typicky na první pozici v koordinaci, nebo na druhé, nebo žádnou typickou pozici nemá.

8 Navíc v případě, že *sbalit* odkazuje k frazému *sbalit si svých pět švestek*, bude se k němu vázat i slovní tvar *svejch* v tabulce č. 2, tj. máme zde dva tvary odkazující ke dvěma různým pozicím v jedné konstrukci.

OPEN ACCESS

tj. zahrnuje jak morfosyntaktické, tak sémanticko-pragmatické informace o dané konstrukci. Zároveň si kladou otázku, nakolik je vůbec možné tyto kolokáty mezi sebou srovnávat, když by se měly spíše srovnávat s výrazy podobného typu, které se vyskytují ve stejné pozici v konkrétní konstrukci. Tato úvaha je zárodkem kolostrukční analýzy.

2. ZPŮSOBY A MEZE UPLATNĚNÍ KOLOSTRUKČNÍ ANALÝZY

Levshina (2015, s. 224n.) o KolA pojednává jako o jedné z asociačních měr, které jsou obvykle založeny na podobném principu, ať už měří vztah mezi slovy, nebo, jako KolA, vztah mezi slovy a konstrukcemi. Tento princip lze vyjádřit kontingenční tabulkou o dvou sloupcích a dvou řádcích, v nichž se uvádějí absolutní frekvence výskytu sledovaných jevů ve vzorku, viz schematicky v tabulce 3. Konkretizace jevů A a B i hodnot n v kontingenční tabulce závisí na konkrétní výzkumné otázce a zvolené metodě. V nejjednodušším případě například A představuje nějaký lexém, B konstrukci. Pro výpočet kolostrukční analýzy pak potřebujeme znát frekvenci (n) lexému A v dané konstrukci (nA a B), frekvenci jiných lexémů než A v dané konstrukci (n¬A a B), frekvenci A v jiných konstrukcích než B (nA a ¬B) a frekvenci jiných lexémů než A v jiných konstrukcích než B (n¬A a ¬B). Vzhledem k logickým vztahům v jednotlivých řádcích a sloupcích v tabulce 1 se typicky pro různé typy KolA (viz níže) používají hodnoty uvedené v řádcích a sloupcích nadepsaných *Celkem* v tabulce 1 (nA, n¬A, nB, ¬B a nA a B a ¬A a ¬B, přičemž poslední hodnota odpovídá velikosti korpusu).

	B	**¬B**	**Celkem**
A	nA a B	nA a ¬B	nA
¬A	n¬A a B	n ¬A a ¬B	n¬A
Celkem	nB	n¬B	nA a B a ¬A a ¬B

TABULKA 3. Schematická tabulka frekvencí pro jevy A a B a jejich souvýskyt[9]

Kolostrukční analýza je typem obousměrné asociační míry, tedy takové, která najednou pro množinu výrazů měří jak jejich tíhnutí k určité konstrukci nebo konstrukcím (*attraction*), tak odpor k ní/nim (*repulsion*). Tento rozdíl se vyjadřuje kladnou, respektive zápornou hodnotou výsledné asociační síly.

Již v úvodu bylo řečeno, že KolA je množinou metod umožňujících měřit vztah mezi konstrukcemi a lexémy. Gries a Stefanowitsch rozlišují tři základní typy:

1. Kolexémová analýza (*collexeme analysis*) — měří míru atrakce lexému k určité pozici v nějaké konstrukci; lze například zjišťovat, jaká substantiva se typicky vyskytují v konstrukci *jde o* N.
2. Distinktivní kolexémová analýza (*distinctive collexeme analysis*) — zjišťuje tíhnutí určitého lexému k nějaké konstrukci ve srovnání s jinou, funkčně srovnatelnou

9 Viz např. Stefanowitsch (2013).

OPEN ACCESS

konstrukcí; tímto typem lze například porovnávat lexémy v příslušných pozicích v konstrukcích *naložit auto zbožím* vs. *naložit zboží do auta* a přispět tak pomocí KolA k popisu jejich vzájemných funkčních rozdílů, které by se hypoteticky měly odrážet i v typické distribuci jednotlivých lexémů, vyjádřené asociační silou.

3. Kovariační kolexémová analýza (*covarying collexeme analysis*) — sleduje míru atrakce lexémů na jedné pozici v konstrukci s lexémy na jiné pozici v konstrukci; lze tak porovnávat například lexémy v pozici patientu a recipienta v ditranzitivních konstrukcích nebo slovesa na dvou pozicích v konstrukcích kontroly (*doporučit panovníkovi vyhlásit válku*).[10]

Například u kolexémové analýzy je nutné v kontingenční tabulce uvést absolutní frekvenci výskytu daného lexému L na určité pozici v konstrukci K (L v K), poté frekvenci výskytu lexému L v jiných konstrukcích v (sub)korpusu (L v nonK), dále frekvenci výskytu K v (sub)korpusu a nakonec velikost (sub)korpusu. U distinktivní kolexémové analýzy se uvádí absolutní frekvence lexému L v konstrukci 1 a v konstrukci 2 a poté frekvence jiných lexémů v jedné a druhé konstrukci. Díky tomu, že se zjišťuje vztah mezi hodnotami uvedenými v tabulce 3 výše, mohou mít i lexémy se stejnou absolutní frekvencí výskytu v dané konstrukci rozdílnou asociační sílu, což se chápe jako jedna z předností KolA. Očekává se přitom, že lexémy celkově velmi frekventované (často buď sémanticky vyprázdněné, nebo výrazně polysémní) budou ke specifické konstrukci přitahovány slabou silou, nebo i odpuzovány, a lexémy vyskytující se v obou porovnávaných konstrukcích (v případě distinktivní kolexémové analýzy) budou umístěny ve středu pole.

KolA lze provádět různými asociačními mírami, ale Gries a Stefanowitsch nejčastěji používají a doporučují Fisherův exaktní test s Yatesovou korekcí (FYE). Jedná se o jeden z testů statistické signifikance, kterým se (přesně, ne aproximativně jako u chí-kvadrátu) měří pravděpodobnost asociace lexému a konstrukce. Gries a Stefanowitsch jako výhody tohoto testu zmiňují to, že je na rozdíl od chí-kvadrátu využitelný i při malých výskytech lexémů ve sledovaných konstrukcích (tj. nižších než 5) a že nepředpokládá žádné (natož normální) rozdělení dat, což je pro jazyková data výhodné. Výsledkem KolA je uspořádaný seznam lexémů s konkrétní číselnou hodnotou udávající sílu asociace — pokud je kladná, jedná se o atrakci, pokud je záporná, jedná se o odpor. Tyto hodnoty lze převést na pravděpodobnostní škálu obvyklou v empirických studiích a zjistit, zda se jedná, nebo nejedná o signifikantní atrakci/odpor.

KolA se obvykle provádí ve volně dostupném počítačovém programu R umožňujícím statistickou analýzu dat.[11] S KolA je spojena průběžně aktualizovaná internetová

10 Autoři KolA později rozšířili distinktivní a kovariační kolexémovou analýzu i na více než dvě pozice v jedné konstrukci, popřípadě více než dvě konstrukce — tyto doplňky se označují jako mnohočetná kolostrukční analýza a z teoretického hlediska vhodně reflektují představu kognitivně-konstrukčních přístupů o tom, že jazyk je založen na síti mnohočetných a mnohaúrovňových vztahů mezi konstrukcemi.

11 R Development Core Team (2008): *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* Austria: Vienna. Dostupné z WWW: http://www.R-project.org. [cit. 1. 3. 2017]

OPEN ACCESS

stránka spravovaná Stefanem Griesem, která shromažďuje informace potřebné pro porozumění této metodě i pro její uplatnění. Z této stránky je také možné si stáhnout skript *Coll.analysis 3.5* vytvořený Griesem v programu R (Gries, 2014), který je možné použít přímo při zpracování jednotlivých typů KolA bez nutnosti tyto skripty samostatně vytvářet nebo kopírovat.[12] Výchozí asociační mírou je Fisherův exaktní test, avšak je možné z nabídky zvolit i jiné míry, například chí-kvadrát, MI/T-score či přirozený logaritmus poměru šancí (*log odds ratio*).

KolA se od svého vzniku využívá k různým účelům a oblasti, kde se uplatnila, se postupně rozšiřují. Typické je její využití při popisu částečně schematizovaných konstrukcí (tj. takových, které mají některé pozice otevřené, jako konstrukce *jde o* N), viz například Stefanowitschova analýza metaforické konstrukce *in the heart of* N (2005). Podobně KolA umožňuje sledovat tíhnutí slovesných lexémů k určitým valenčním konstrukcím nebo k určité hodnotě kategorií času, vidu a způsobu (viz kontrastivní výzkum kauzativních konstrukcí v angličtině a francouzštině ve studii Guilquin(ové), 2015).

Existují výzkumy zjišťující rozdíly v atrakci lexémů v závislosti na varietě (britská vs. americká angličtina, Wulff(ová) et al., 2007) nebo způsobu realizace promluvy (psaný/mluvený projev, Stefanowitsch — Gries, 2008). Posledně zmíněná studie je zajímavou extenzí původní verze KolA. Autoři na třech případových studiích zjišťují, zda budou výsledky KolA různé při změně komunikačního kanálu (vytvářejí tak třídimenzionální kolostrukční analýzu kombinující lexém, konstrukci a kanál). Jejich obecným zjištěním je, že výsledky třídimenzionální KolA zásadním způsobem výsledky původní KolA nemění, umějí však odhalit specifické asymetrie dané komunikačním kanálem: například v distribuci sloves v aktivu a pasivu existoval rozdíl mezi psaným a mluveným projevem v tom, že aktivní konstrukce v mluveném projevu přitahovaly nejsilněji anglická jednoslabičná slovesa germánského původu (například *do, get, say, want*), kdežto v psaném projevu se v aktivu vyskytovala typicky slovesa delší, původu románského (*enclose, provide, include, contain*) — tento rozdíl ale nebyl doložen u pasiva. Hilpert (2006 a 2012) ukazuje možnosti využití a způsoby interpretace KolA v diachronním výzkumu, v první studii mimo jiné na příkladu distinktivní kolexémové analýzy uplatněné na tři historická období, v nichž se v angličtině vyvíjelo užívání konstrukce *shall* + infinitiv. KolA se využívá i ve výzkumu osvojování druhého/cizího jazyka (Ellis — Ferreira-Junior, 2009; Gries — Wulff, 2005, 2009), a to zejména při kombinaci korpusových a experimentálních psycholingvistických výzkumů, v nichž se kupříkladu při hodnocení přijatelnosti nebo úlohách zaměřených na dokončování vět (*sentence completion tasks*) zjišťuje citlivost nerodilých mluvčích k tíhnutí konkrétních lexémů k určitým konstrukcím.

Od vytvoření KolA se objevily dvě její zásadnější kritiky (Bybee, 2010, a Schmid — Küchenhoff, 2013), které stojí za to zmínit. Na oba kritické texty rozsáhle reagoval Gries (2012 a 2015). V tomto textu uvádím jen dvě výhrady, které nemají jen technický[13], ale i teoretický rozměr. Za prvé je otázka, zda je adekvátnější vyjádřit nějaký

12 Skripty ke KolA uvádí například Levshina (2015).

13 Mezi techničtější otázky patří například fakt, že Schmid — Küchenhoff (2013) kritizují použitou asociační míru FYE a navrhují vlastní variantu měření, jež nepracuje s prav-

OPEN ACCESS

vztah asociační silou než absolutními frekvencemi. Bybee(ová) se dlouhodobě věnuje výzkumu vlivu tokenové a typové frekvence na ukládání, reprezentaci a vybavování jazykových jednotek, které se obvykle v těchto výzkumech udávají v absolutních frekvencích. Z tohoto hlediska jsou zajímavé psycholingvistické výzkumy, které realizoval Gries spolu s kolegy (Gries et al., 2010). Podle nich se zdá, že pro ukládání a vybavování konstrukcí je asociační míra lepší prediktor než absolutní frekvence. To je jistě cenný výsledek, který si vyžaduje další replikace i na jiných jazycích. Druhá námitka souvisí s otázkami kognitivní adekvátnosti — je možné provázat metodu KolA a její výsledky s některými z kognitivních schopností a procesů? Gries (2015) uvádí, že lze uvažovat o souvislosti s procesy ukládání (*entrenchment*) a vyhodnocování distinktivní hodnoty rysu (*cue validity and reliability*) při kategorizaci a vybavování informace. V reakci na kritické texty Gries (2015, s. 533) souhrnně konstatuje, že je nutné si být vědom toho, že KolA, tak jako jiné asociační míry, je založena na jednoduché metodě, která tak nutně komplexní multidimenzionální mentální prostor znalosti jazyka jeho mluvčími představuje ve zjednodušené formě.[14] Z toho důvodu je podle Griese zásadní v nejbližší budoucnosti rozvíjet sofistikovanější metody analýzy dat (smíšené modely zohledňující vliv jednotlivých lexémů a jednotlivých mluvčích, měření korelací mezi výsledky korpusových a experimentálních výzkumů apod.). Na druhou stranu není podle něj vhodné KolA pro její omezení zcela zatracovat: je vhodným a velmi snadným nástrojem pro zjištění systematických vztahů mezi lexémy a konstrukcemi, a hodí se proto pro první fáze výzkumů zaměřených na lexiko-gramatické vztahy. Navíc výzkumy poukazující na její prediktivní sílu činí její výsledky zajímavými i z kognitivního hlediska.

3. PŘÍPADOVÁ STUDIE: [V *(SI) (SVÝCH) PĚT/PÁR ŠVESTEK*]

Pro ilustraci uplatnění KolA na češtinu můžeme vyjít z výsledků analýzy kolokací lexému *švestka* v první části tohoto textu. Domnívali jsme se, že sloveso *balit* a tvar zvratného posesivního zájmena *svejch* mohou pocházet z frazému *sbalit si svých pět švestek*. Když se předběžně podíváme do psaných korpusů Českého národního korpusu (SYN ve verzi 4, viz Hnátková et al., 2014), vidíme, že konstrukce vykazuje poměrně velkou míru variability, jak ukazují příklady (1)–(6):

(1) Poté vzal svých „pět švestek", ženu zamkl a utekl.
(2) Sbalil jsem si svých pět švestek a pěkně po anglicku se vypařil.

---

děpodobnostními hodnotami. Gries namítá, že KolA je použitelná s různými asociačními mírami podle preferencí každého badatele, a uvádí, proč on sám preferuje FYE (viz výše v tomto textu). Dále u složitějších typů KolA je nutné v kontingenční tabulce uvádět počet konstrukcí odlišných od té sledované, v nichž se konkrétní lexém vyskytuje, což může v pouze lemmatizovaných korpusech (oproti syntakticky anotovaným) představovat metodologický problém. Gries (2015) navrhuje možné způsoby jeho řešení.

14 Vědomí této skutečnosti se ostatně u autorů KolA odráží i v tom, že vyvíjejí pokročilejší verze této metody, jako je mnohočetná nebo třídimenzionální KolA.

OPEN ACCESS

(3) Tak jsem si sbalil pět švestek a jelo se do Ostravy
(4) …v sešikmeném rohovém domě za vodárnou složil svých pět švestek.
(5) Přijdou, pak zase seberou svejch pět švestek a táhnou dál.
(6) Když jsme se vzali a každý rozbalil svých „pár švestek".

Jednak v šesti příkladech vidíme pět různých sloves (*vzít, sbalit, složit, sebrat, rozbalit*), ve větě může být přítomno zvratné zájmeno osobní (*si*) nebo přivlastňovací (*svých* nebo *svejch*) nebo obě najednou, kvantifikujícím výrazem může být buď číslovka *pět*, nebo *pár*. V prostoru, jaký poskytuje tato studie, se můžeme zaměřit na dvě otázky: 1. Jaká je variabilita na pozici řídícího slovesa a která z těchto sloves jsou ke sledované konstrukci přitahována? 2. Je variabilita na pozici kvantifikátoru (*pět/pár*) determinována lexémem na pozici predikátu? K získání odpovědi na první otázku použijeme jednoduchou kolexémovou analýzu, k druhé otázce použijeme kovariační kolexémovou analýzu.

Konstrukci *sbalit si svých pět švestek* chápeme jako částečně schematizovanou s následující strukturou [V *(si) (svých) pět/pár švestek*][15]. Podle toho můžeme formulovat korpusový dotaz: [tag="V.*"]{0,4}[lemma="pět|pár"][word="švestek"], který obsahuje otevřenou pozici slovesa, protože chceme sledovat lexikální variabilitu na této pozici v daném frazému; vzhledem k tomu, že může, nebo nemusí být přítomno zvratné osobní a/nebo přivlastňovací zájmeno, popřípadě ještě adverbium, na dané pozici umožňujeme přítomnost dalších výrazů — {0,4}, na další pozici pak očekáváme buď výraz *pár*, nebo *pět* a nakonec slovní tvar *švestek*[16]. Ze získaných 710 dokladů (i.p.m. 0.16) jsem ručně odstranila nefigurativní kontexty a duplicity. Zůstalo 605 konkordancí. U nich jsem doplnila slovesná lemmata, kterých bylo 66. Tam, kde se jednalo o polysémní slovesný lexém, nebo v příkladech, kde by slovesné lemma neodkázalo přesně k významu použitému v dané větě, jsem přidala nezbytné doplnění, viz například *dát_dohromady* nebo *koupit_za* níže. Následně jsem vytvořila kontingenční tabulku. Pro kolexémovou analýzu potřebujeme znát čtyři údaje: frekvenci slovesa ve sledované konstrukci, frekvenci slovesa v korpusu SYN (bez jeho výskytu ve sledované konstrukci), výskyt konstrukce v korpusu a velikost korpusu. Výsledná tabulka ve formátu .txt se vloží v programu R do skriptu vytvořeného S. Griesem (2014). Výsledkem je seznam sloves s udanou hodnotou jejich asociační síly. Asociační síla měřená Fisherovým exaktním testem se interpretuje tak, že pokud je kladná a větší než 3, je hodnota pravděpodobnosti $p < 0.001$, hodnota větší než 2 odpovídá $p < 0.01$ a hodnoty nad 1,3 odpovídají $p < 0.05$. Pokud je tedy kolostrukční síla vyšší než 1,3, daná slovesa jsou ke sledované konstrukci přitahována v míře větší, než vyplývá z náhodné distribuce, za dostatečně signifikantní se obvykle považují hodnoty vyšší než 2,0. Výsledky kolexémové analýzy pro prvních

15 Od možnosti postpozice slovesa za akuzativní NP v této sondě pro jednoduchost odhlížím, výsledky nejsou touto redukcí zásadně ovlivněny.

16 Hledání podle slovního tvaru *švestek*, nikoliv lemmatu *švestka* přispívá k eliminaci nerelevantních dokladů. Zároveň nepředpokládám, že výraz *švestek* bude bezprostředně premodifikován jiným výrazem ještě před modifikací kvantifikátorem *pět/pár*.

OPEN ACCESS

20 sloves ve zjednodušené podobě ukazuje tabulka č. 4, v níž se uvádí u každého lexému frekvence v korpusu F(C) a kolostrukční síla KS[17].

pořadí	V	F(C)	KS	pořadí	V	F(C)	KS
1.	sbalit	364	Inf[18]	11.	koupit_za	3	8.3374436
2.	stahnout_kata-ta_za	1	Inf	12.	popadnout	3	8.2203015
3.	sebrat	62	196.3118281	13.	mamit	2	8.0720433
4.	balit	25	86.1375902	14.	pakovat	2	7.1603187
5.	zabalit	24	77.0942039	15.	dat_dohro-mady	3	7.1563130
6.	spakovat	13	65.1182343	16.	odvezt	4	7.1241960
7.	vzit	20	36.3593781	17.	posbiravat	1	6.0784853
8.	posbirat	7	20.7701572	18.	zpakovat	1	6.0115385
9.	zbalit[19]	3	14.7625426	19.	rozbalit	2	5.9624222
10.	delat_za	3	10.1472622	20.	slozit	3	5.8677127

TABULKA 4. Výsledky kolexémové analýzy pro první dvacet sloves v konstrukci [V *(si) (svých) pět/pár švestek*]

Všech 66 sloves, popřípadě slovesných vazeb, vykazovalo kladnou hodnotu kolokační síly v dané konstrukci, tedy k ní bylo atrahováno. Avšak 6 posledních sloves mělo hodnoty na úrovni náhodné distribuce, tj. nižší než 1,3 (*nést*, *připomínat*, *mít*, *dát*, *přinést, dostat*). Zcela v souladu s obecnými předpoklady KolA se jedná o slovesa velmi frekventovaná, prototypická pro určité sémantické rámce, jako je vlastnictví (*mít*), transfer vlastnictví (*dát-dostat*) nebo determinovaný pohyb vyjadřující přesun něčeho (*nést-přinést*). Ačkoliv se tedy nejedná o signifikantní slovesa pro tuto konstrukci, vidíme zde některé sémantické komponenty (vlastnictví, transfer, přesun), které by mohly charakterizovat i slovesa silně tíhnoucí ke konstrukci [V *(si) (svých)*

17 V odborné literatuře o KolA se pojmy asociační a kolostrukční síla obvykle používají synonymně tam, kde se mluví o KolA. Obecně je kolostrukční síla podtypem asociační síly, specifikující vztah mezi lexémem a konstrukcí.

18 Hodnota Inf značí nejvyšší pásmo kolostrukční síly (v řádu stovek či tisíců, tedy výrazně významnou atrakci, či odpor) — objevuje se typicky při vysoké KS počítané na základě velkých korpusů (jako v případě této studie založené na korpusu SYN). S. Gries na vyžádání poskytuje skript ke KolA, který i pro tyto relativně vysoké hodnoty vrací konkrétní číselný údaj. V odborné literatuře se však výsledky KolA uvádějí běžně i s hodnotou Inf.

19 Z důvodu co nejmenší manipulace s výchozími daty byla slovesa s *sbalit/zbalit* a podobně i *spakovat/zpakovat* anotována zvlášť. Například podoba *zbalit* se v korpusu SYN vyskytuje 158krát, tedy ne zcela zanedbatelně, a odlišná grafika by mohla označovat nikoliv jen prostou pravopisnou chybu, ale postupné vznikání jiné konstrukce s posunutým významem.

OPEN ACCESS

pět/pár švestek]. Když se z této perspektivy podíváme na prvních dvacet sloves (a následně i na ta, která se vyskytují na pozicích 21–60), odhalíme dva sémantické rámce: jeden z nich označuje akční událost, při níž typicky subjekt po nějakém podnětu provádí shromáždění svého vlastního majetku s cílem ho přemístit. Pro tento význam je prototypické sloveso *sbalit* a patří sem i slovesa na 3.–9. místě nebo z druhé desítky například *popadnout* nebo *odvézt* (které implikuje předchozí shromáždění a sbalení). Dokončení tohoto transferu vyjadřují lexikální konstrukce *složit, rozbalit* nebo *dát dohromady* (používá se při vstupu do manželství, sestěhování partnerů apod.).

Druhý sémantický rámec — celkově na druhém, ale nikoliv zanedbatelném místě — představují mezi nejsilněji atrahovanými slovesy konstrukce *stáhnout kaťata za (pět/pár švestek)*, *dělat za* (ve významu ‚pracovat'), *koupit za*, *mámit (na někom pár/pět švestek)*. Zde výraz *pět/pár švestek* označuje specifický typ majetku, totiž peníze. Zatímco v prvním rámci *pět/pár* švestek může, ale nemusí obsahovat konotaci, že se jedná o nevelký majetek, u druhého rámce je tento významový rys vždy přítomný. Ve druhém rámci se ustaluje formální vyjádření předložkovou frází *za pět/pár švestek*, které se vyskytlo u třinácti slovesných lexémů.

Pokud jde o vlastní výsledky kolexémové analýzy, i na prvních dvaceti lexémech je dobře vidět, že absolutní frekvence nepredikuje dobře asociační sílu konkrétního lexému, viz různé asociační síly lexémů, které se v dané konstrukci vyskytují pouze jednou. Za zmínku stojí konstrukce *stáhnout kaťata za*, která má tak vysokou hodnotu, neboť se v korpusu SYN vyskytuje pouze jednou, a to právě v konstrukci s *pět/pár švestek*.[20] Těmto krajním případům je třeba při interpretaci kolostrukční analýzy věnovat pozornost, nicméně v našem případě i tato konstrukce dobře zapadá do jednoho z nalezených rámců.

Doložené sémantické rámce a jejich význam nás vedou k otázce, zda se na rozlišení významu nepodílí i různá distribuce číslovek *pět* a *pár*, tj. zda první není typická pro první rámec a druhá pro rámec druhý. A na tuto otázku může přinést odpověď kovariační kolexémová analýza. Při ní zjišťujeme asociační sílu pro dvě spolu související pozice, v našem konkrétním případě pro pozici řídícího slovesa a pozici kvantifikátoru *pět/pár*. Do kontingenční tabulky tak zadáváme následující údaje:

a) počet výskytů určitého slovesa (například *sbalit*) na první sledované pozici za souvýskytu jednoho ze sledovaných lexémů (například *pět*) na druhé pozici
b) počet výskytů slovesa *sbalit* na první pozici při souvýskytu s *pár* na druhé pozici
c) součet výskytů jiných sloves na první pozici při souvýskytu s lexémem *pět* na druhé pozici
d) součet výskytů jiných sloves na první pozici při souvýskytu s lexémem *pár* na druhé pozici.

Jak vidno, kovariační kolexémová analýza — na rozdíl od jednoduché kolexémové analýzy — odhlíží od celkové velikosti korpusu, s nímž pracujeme. Výsledky této analýzy opět ve zjednodušené formě uvádí tabulka č. 5: vidíme, že signifikantní vzájemnou atrakci vykazuje pouze číslovka *pět* a sloveso *sbalit*, které zároveň získalo výrazně

20 Doklad zní takto: *Dneska je moc mladejch holek, co stáhnou kaťata za pár švestek.*

OPEN ACCESS

zápornou hodnotu vztahu s výrazem *pár*. Obrácený výsledek jsme získali u slovesa *zabalit* (významná atrakce k *pár*, významný odpor k *pět*) a jen významnou atrakci k číslovce *pár* (nikoliv významný odpor k *pět*) projevilo pouze sloveso *odvézt*. Jelikož tato slovesa patří k prvnímu sémantickému rámci (a žádné další sloveso z něj nezískalo významnou hodnotu atrakce či odporu k jednomu nebo druhému kvantifikátoru), znamená to, že distribuce v užívání číslovek v prvním rámci může být dána pragmaticky, podle konkrétní komunikační situace nebo zvoleného jazykového vyjádření (prozodické důvody). Pokud jde o druhý sémantický rámec, u dvou slovesných konstrukcí (*dělat za* a *koupit za*) se ukázala atrakce k číslovce *pár*, jak jsme předpokládali, ale hodnoty u dalších sloves již nebyly významné. Hlubší porozumění distribuci číslovek v této konstrukci by tak bylo možné na základě komplexnější, multifaktorové analýzy, která by zahrnovala více rysů.

číslovka	**atrakce**	**odpor**
pět	sbalit	zabalit
pár	zabalit odvézt dělat za koupit za	sbalit

TABULKA 5. Výsledky kovariační kolexémové analýzy pro slovesa a kvantifikátory v konstrukci [V *(si) (svých) pět/pár švestek*]

Další analýzy již přesahují rámec této studie. Uvedu tedy pouze to, že problematiku reflexivních tvarů v dané konstrukci a jejich vztah k řídícímu slovesu by bylo možné zkoumat pomocí další kovariační kolexémové analýzy a na otázku, zda je konstrukce [V *(si) (svých) pět/pár švestek*] typická pro lexém *švestka*, by mohla odpovědět mnohočetná distinktivní kolexémová analýza, která by měřila atrakci lexému *švestka* k této konstrukci ve srovnání s jinými konstrukcemi, v nichž se vyskytuje (například ve frazému *nachytat na švestkách* nebo v koordinačních konstrukcích).

Cílem tohoto textu bylo představení kolostrukční analýzy jako zajímavé metody pro analýzu korpusových dat v širokém slova smyslu, tj. nejen dat z veřejně dostupných korpusů. Kolostrukční analýza může být obecně vhodná tam, kde je předmětem výzkumu vztah mezi syntaktickou konstrukcí a jejím lexikálním obsazením, což bývá označováno za hlavní přednost této metody. Vzhledem ke své jednoduchosti může rychle poskytnout první informace o pravidelnostech v užívání sledovaných konstrukcí a pomoci tak při formulování hypotéz pro navazující komplexnější analýzy. To, že lze tímto způsobem zkoumat české jazykové jevy, jsem ukázala ve třetím oddíle této studie. Nezbývá než doufat, že tento text přispěje k tomu, aby počet případových studií představujících kolostrukční analýzu českých konstrukcí v budoucnu narůstal.

OPEN ACCESS

LITERATURA

Bybee, J. (2010): *Language, Usage, and Cognition*. Cambridge: Cambridge University Press.

Cvrček, V. — Vondřička, P. (2011): *SyD — Korpusový průzkum variant.* Praha: FF UK. Dostupný z WWW: http://syd.korpus.cz [cit. 1. 3. 2017].

Diessel, H. (2015): Usage-based Construction Grammar. In: E. Dąbrowska — D. Divjak (eds.), *Handbook of Cognitive Linguistics*. Berlin — Boston: De Gruyter, s. 296–321.

Fried, M. (2013): Pojem konstrukce v konstrukční gramatice. *Časopis pro moderní filologii*, 95, 1, s. 9–27.

Gries, S. Th. (2012): Frequencies, probabilities, association measures in usage/exemplar-based linguistics: Some necessary clarifications. *Studies in Language*, 36, 3, s. 477–510.

Gries, S. Th. (2014): *Coll.analysis 3.5.* Skript pro kolostrukční analýzu v R. Dostupný z WWW: http://www.linguistics.ucsb.edu/faculty/stgries/teaching/groningen/ [cit. 28. 2. 2017].

Gries, S. Th. (2015): More (old and new) misunderstandings of collostructional analysis: on Schmid & Küchenhoff 2013. *Cognitive Linguistics*, 26, s. 505–536.

Gries, S. Th. — Hampe, B. — Schönfeld, D. (2010): Converging evidence II: more on the association of verbs and constructions. In: S. Rice — J. Newman (eds.), *Empirical and Experimental Methods in Cognitive/Functional Research.* Stanford: CSLI Publications, s. 59–71.

Gries, S. Th. — Stefanowitsch, A. (2004a): Extending collostructional analysis: A corpus-based perspectives on ‚alternations'. *International Journal of Corpus Linguistics*, 9, 1, s. 97–129.

Gries, S. Th. — Stefanowitsch, A. (2004b): Co-varying collexemes in the into-causative. In: M. Achard — S. Kemmer (eds.), *Language, Culture, and Mind.* Stanford, CA: CSLI, s. 225–236.

Gries, S. Th. — Wulff, S. (2005): Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics*, 3, s. 182–200.

Gries, S. Th. — Wulff, S. (2009): Psycholinguistic and corpus linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics*, 7, s. 163–186.

Guilquin, G. (2015): Contrastive collostructional analysis: causative constructions in English and French. *Zeitschrift für Anglistik und Amerikanistik*, 63, č. 3, s. 253–272.

Hilpert, M. (2006): Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory*, 2, s. 243–257.

Hilpert, M. (2012): Diachronic collostructional analysis: How to use it and how to deal with confounding factors. In: K. Allan — J. A. Robinson (eds.), *Current Methods in Historical Semantics.* Berlin — Boston: Mouton de Gruyter, s. 133–160.

Hnátková, M. — Křen, M. — Procházka, P. — Skoumalová, H. (2014): The SYN-series corpora of written Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavík: ELRA, s. 160–164.

Hoffmann, T. — Trousdale, G. (eds.) (2013): *Oxford Handbook of Construction Grammar.* Oxford: Oxford University Press.

Karlík, P. et al. (2016): *Nový encyklopedický slovník češtiny online.* Dostupný z WWW: https://www.czechency.org/slovnik [cit. 28. 2. 2017].

Křen, M. — Cvrček, V. — Čapka, T. — Čermáková, A. — Hnátková, M. — Chlumská, L. — Jelínek, T. — Kováříková, D. — Petkevič, V. — Procházka, P. — Skoumalová, H. — Škrabal, M. — Truneček, P. — Vondřička, P. — Zasina, A. (2015): *SYN2015: reprezentativní korpus psané češtiny*. Ústav Českého národního korpusu FF UK, Praha. Dostupný z WWW: http://www.korpus.cz [cit. 1. 3. 2017].

Levshina, N. (2015): Association measures: Collocations and collostructions. In: *How*

OPEN ACCESS

to do Linguistics with R. Data Exploration and Statistical Analysis. Amsterdam — Philadelphia: John Benjamins, s. 223–252.

Schmid, H.-J. — Küchenhoff, H. (2013): Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics*, 24, 3, s. 531–577.

Stefanowitsch, A. (2005): The function of metaphor: developing a corpus-based perspective. *International Journal of Corpus Linguistics*, 10, 2, s. 161–198.

Stefanowitsch, A. (2013): Collostructional analysis. In: T. Hoffmann — G. Trousdale (eds.), *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press, s. 290–306.

Stefanowitsch, A. — Gries, S. Th. (2003): Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, 8, 2, s. 209–243.

Stefanowitsch, A. — Gries, S. Th. (2005): Co-varying collexemes. *Corpus Linguistics and Linguistic Theory*, 1, 1, s. 1–43.

Stefanowitsch, A. — Gries, S. Th. (2008): Channel and constructional meaning: A collostructional case study. In: G. Kristiansen — R. Dirven (eds.), *Cognitive sociolinguistics*. Berlin — New York: Mouton de Gruyter, s. 129–152.

Wulff, S. — Stefanowitsch, A. — Gries, S. Th. (2007): Brutal Brits and persuasive Americans: variety-specific meaning construction in the into-causative. In: G. Radden et al. (eds.), *Aspects of meaning construction*. Amsterdam/Philadelphia: John Benjamins, s. 265–281.

Eva Lehečková | Ústav českého jazyka a teorie komunikace, Filozofická fakulta Univerzity Karlovy | nám. Jana Palacha 2, 116 38 Praha 1
eva.leheckova@ff.cuni.cz