Bohemica Olomucensia 2 – Philologica Juvenilia
2. ročník, 2010 (tomus)
ISSN 1803-876X

Korpusový pohled na postavení číslovek v systému slovních druhů

Václav Cvrček

1 Slovní druhy

Shrneme-li dosavadní přístupy, při klasifikaci slovních druhů se setkáváme se třemi druhy kritérií: 1) morfologickým (formálním), 2) sémantickým a 3) syntaktickým.[1] Morfologické kritérium, které preferuje např. Příruční mluvnice češtiny[2], se uplatňuje zejména při odlišení slov ohebných a neohebných, skloňovaných (jmen) a časovaných (sloves) a dále pak podle způsobu skloňování pro rozdělení slov v rámci jmen na jednotlivé slovní druhy (substantiva, adjektiva, zájmena a číslovky). Sémantické kritérium, které se snaží postihnout rozdíly ve významu jednotlivých skupin slov má také svoje „makroskupiny": autosémantika a synsémantika. O úroveň obecnosti níž se pak setkáme s charakteristikami typu: substantiva jsou názvy osob, zvířat, věcí, vlastností a dějů, které jsou pojímané jako nezávislé entity apod. Tímto způsobem jsme schopni rozlišit téměř všechny slovní druhy, až na slova neohebná. Zde vstupuje do hry syntaktické kritérium, které se vyjadřuje k syntaktické funkci jednotlivých slovních druhů (např. konjunkce spojují věty nebo větné členy). Inovativní přístup k syntaktickému kritériu představuje rozlišení částic a citoslovcí v rámci projektu PMK[3] aplikované i pro potřeby popisu v rámci Mluvnice současné češtiny[4], které za odlišující rys považuje schopnost tvořit samostatnou výpověď, která je vlastní pouze citoslovcím.

Problematických míst slovnědruhového zařazení bychom našli celou řadu. Ponecháme-li stranou číslovky, které jsou předmětem tohoto článku, můžeme jmenovat už dříve zmíněné „rozcestí" mezi částicemi a citoslovci (částečně do toho vstupují i adverbia). Problematická je i celá skupina tzv. predikativ a stavových adverbií, která můžou být interpretována jako adjektiva (např. ve spojení je možno, že lze tvar možno chápat jako krátký tvar adjektiva možný, čemuž nahrává i jasně adjektivní podoba tohoto spojení v mluvené češtině: je možný, že). Z jednotlivostí můžeme pak jmenovat frekventované slovo sám, které se může řadit k adjektivům (podle své formy) i k číslovkám (podle významu, na základě analogie dva – oba, proto jeden – sám); pouze zájmenná interpretace – užívaná v našich mluvnicích – není úplně opodstatněná.

1 Viz například Karlík et al.: *Encyklopedický slovník češtiny.* Praha. 2002, heslo Druh slovní, příp. Petr, J. et al.: *Mluvnice češtiny II.* Praha 1986, s. 13.
2 Karlík et al.: *Příruční mluvnice češtiny.* Praha 1995, s. 228.
3 Čermák, F. et al.: *Frekvenční slovník mluvené češtiny.* Praha 2007.
4 Cvrček, V. – Kodýtek, V. – Kopřivová, M. – Kováříková, D. – Sgall, P. – Šulc, M. – Táborský, J. – Volín, J. – Waclawičová, M.: *Mluvnice současné češtiny.* Praha 2010.

Zásadní pro pochopení celé problematiky je to, že navzdory způsobu organizace informací ve většině mluvnic, je slovnědruhová charakteristika součástí lexikologie (nikoli morfologie, která na této informaci už pouze staví). Zařazením slova do příslušného slovního druhu specifikujeme jeho nejobecnější význam na úrovni arbitrárně stanovených skupin. Jejich vytvořením se významně zjednodušuje popis, protože tyto skupiny reprezentují prototypy nejobecnějších významů, s nimiž se v rámci lexikonu můžeme setkat.

Dalším důležitým aspektem celé problematiky je fakt, že rozdělení na slovní druhy je jednou z nejtradičnějších oblastí celé lingvistiky (jako absurdní příklad můžeme poukázat na fakt, že v nejstarších latinských gramatikách přežíval slovní druh „člen" původně přejatý z řečtiny, ačkoli latina žádným členem nedisponovala; srov. i snahy o zachování stejného počtu slovních druhů v latině a řečtině).[5] Rozdělení na slovní druhy možná i z tohoto důvodu nebylo podrobeno výrazné revizi, která by byla ochotna úplně rezignovat na současné dělení a vytvořit systém zcela nový (dalším nesporným důvodem je fakt, že současné rozdělení je v mnoha ohledech funkční).

2 Korpusový pohled na slovní druhy

Všechna zmiňovaná kritéria se více či méně odrážejí v tom, jak se slovo užívá, do jakých vstupuje vztahů, která slova svojí přítomností ovlivňuje (ať už gramaticky nebo sémanticky). Úzus slova je přitom nejlépe pozorovatelný v rámci kontextu.

Z hlediska sémantiky můžeme konstatovat, že kontext je formativní charakteristikou lexikálního významu. Ne nadarmo proto většina korpusového výzkumu v této oblasti vychází z rčení, že „slovo poznáš podle toho, s čím se spojuje."[6] Uvažujeme-li o kontextu jako o kritériu slovnědruhové platnosti, rozšiřujeme tím pouze platnost uvedeného rčení z jednotlivých slov i na celé skupiny. Parafráze daného výroku pro slovní druhy by tedy zněla: slovní druh poznáš podle toho, s jakými skupinami slov se spojuje.

Syntaktické kritérium se z pohledu kontextu projevuje zejména prototypickými kombinacemi slov. Adjektivum, jehož prototypickou funkcí je atribut, se signifikantně častěji spojuje se jmény (předchází jim), než s jinými slovními druhy. Stejně tak adverbium bude v kontextu častěji mít sloveso než jiný slovní druh (nemluvě o předložkách, jejichž úzus ukazuje velmi silně na jejich funkci, např. absencí kombinace předložka – sloveso). Problematickým místem této úvahy se může zdát být relativně volný český slovosled. Jak se pokusím ukázat níže, vyhodnocujeme-li údaje z dostatečně velkého korpusu a navíc statistickými metodami, můžeme tento problém zanedbat s tím, že prototypické případy, o které nám tu jde především, jsou zároveň nejfrekventovanějšími způsoby užití.

Morfologické kritérium při pohledu na slovní druhy prizmatem kontextu můžeme opomenout z toho důvodu, že budeme pracovat s lemmatizovaným korpusem a slova nebudou do pokusu vstupovat jako samostatné slovní tvary, ale pouze jako lemmata. Tím je zaručená stejná slovnědruhová charakteristika pro slovní tvary jednoho lemmatu, a tedy jednoho způsobu ohýbání. Ovšem stejně jako syntaktická funkce i formálně morfologická platnost slova se v kontextu projevuje, např. kontext obsahující shodu v čísle a v pádě vyžaduje pouze slova určitého morfologického typu.

5 Viz Černý, J.: *Dějiny lingvistiky.* Olomouc 1996, s. 66.
6 Autorem tohoto výroku je slavný anglický lingvista J. R. Firth, viz Palmer, F. R. (ed.): *Selected papers of J. R. Firth, 1952–1959.* Bloomington / London 1968, s. 179.

Po takovémto úvodu můžeme tedy vyslovit první nesmělý předpoklad, že jedním z kritérií použitelných pro slovnědruhovou analýzu by mohl být úhrnný kontext skupiny slov, tvořící jeden slovní druh. Jinými slovy, slovní druh tvoří taková třída slov, jejichž kontext vykazuje nějakou míru shody.

3 Popis pokusu a výsledky

Pro zpřesnění představy o uvedeném obecném principu se pokusím o demonstraci na konkrétních datech. Pokusným slovním druhem jsou číslovky, o jejichž slovnědruhové platnosti můžeme vyslovit oprávněnou pochybnost. Na základě jejich formy, významu i funkce můžeme spekulovat o tom, zda nejde o specifický případ příslovcí, zájmen nebo adjektiv. Nemám zde přitom na mysli vágně vymezené číslovky neurčité, ale o číselné výrazy s konkrétním „vyčíslitelným" významem.

Můžeme tedy předpokládat, že číslovky řadové jsou jen specifickým případem adjektiv s číselným významem (první liga jako fotbalová liga). Analogicky jsou číslovky základní a druhové specifickým druhem zájmen (dvě jablka analogicky k ty jablka; jedny brejle jako ty brejle) a konečně číslovky násobné jsou specifickým případem adverbií (dvakrát zaslechl jako jasně zaslechl).

Testovat tuto hypotézu budeme na úhrnu kontextů, do nichž tyto skupiny číslovek vstupují. Pro účely pokusu použijeme největší dostupný korpus: půlmiliardový SYN (souhrn korpusů SYN2000, SYN2005 a SYN2006PUB),[7] který sice není reprezentativním korpusem, ale to nám v případě morfologického zkoumání, které se nezdá být tolik závislé na složení textů, nemusí vadit. Rozhodující výhodou tohoto korpusu je kromě jeho velikosti i fakt, že na něj byla aplikována nejnovější lemmatizace a morfologická analýza. Slovnědruhové zařazení tak vykazuje odchylku od trénovacích dat (tedy „chybovost") menší než 1 %.[8]

Z korpusu vyexcerpujeme, do jakých kontextů daný druh číslovek vstupuje a jak typický (frekventovaný) takový kontext je. Dostaneme následující sadu údajů (ukazujeme zde pouze výsek, prvních 20 řádek):

kontext	druhové	násobné	řadové	základní
A-A	101	453	6101	22723
A-C	11	194	1871	10995
A-D	54	805	557	2909
A-I	0	1	4	11
A-J	53	345	1380	8479
A-N	805	317	45889	221917
A-P	72	582	576	3267
A-R	102	1377	1375	10622
A-T	13	21	115	416
A-V	292	1580	1445	7569
C-A	24	320	3223	29997
C-C	10	401	1200	10740
C-D	15	886	2724	8478
C-I	0	5	3	60
C-J	22	518	1857	11825
C-N	296	457	16784	284581
C-P	22	874	673	5671
C-R	37	2024	1837	17305
C-T	2	27	74	841

V prvním sloupci je uvedena zkratka slovního druhu, který bezprostředně předchází dané číslovce a bezprostředně za ní následuje.[9] Čísla v dalších sloupcích ukazují, kolikrát se v daném kontextu vyskytne číslovka základní, řadová, druhová a násobná. Můžeme tak vidět, že v kontextu adjektivum – číslovka – adjektivum se 101× vyskytla číslovka druhová (např. obehnaný dvojí vysokou zdí), 453× číslovka násobná (např. mrňavej dvakrát zkříženej tchoř), 6101× číslovka řadová (např. možnou třetí světovou válkou) a 22723× číslovka základní (např. posledních devět čokoládových bonbónů).

Podobné údaje si zjistíme pro slovní druhy, s nimiž chceme číslovky porovnávat: adjektiva, zájmena, adverbia. Pro kontrolu můžeme do zkoumaných dat připojit i slovní druhy, u nichž žádnou strukturní nebo funkční podobnost s číslovkami neočekáváme, půjde o substantiva, slovesa, předložky a spojky.

V průběhu práce se ukázalo, že takto obecně vymezené slovní druhy představují příliš velké skupiny, proto jsem v některých případech přistoupil k dalšímu jemnějšímu dělení. Adjektiva jsou tak rozdělena na adjektiva „obyčejná" (v podstatě vzory mladý a jarní) a ostatní (vzor otcův/matčin a krátké tvary adjektiv), adverbia byla rozdělena na stupňovatelná (zhruba ta odvozená od adjektiv) a nestupňovatelná (zejména zájmenná příslovce).

Následující postup je už velmi prostý. Pro každou dvojici zkoumaných skupin slov vypočítáme Spearmanův korelační koeficient.[10] Pro tyto účely jsem používal statistický program R[11], který je zároveň schopen podávat informaci o statistické významnosti navrženého modelu. Pokud je korelace frekvencí kontextů mezi dvěma skupinami slov silná, znamená to, že se tyto dvě skupiny slov vyskytují v prototypicky stejných kontextech („chovají se podobně"). Tento způsob porovnání pomocí Spearmanova korelačního koeficientu je schopen kompenzovat jak rozdíly v celkové frekvenci slovních druhů (substantiva jsou nepoměrně frekventovanější než např. druhové číslovky, vstupují tedy častěji do téměř všech kontextů než číslovky), tak v nenormálnosti rozdělení, na kterou jsme u jazykových dat zvyklí.

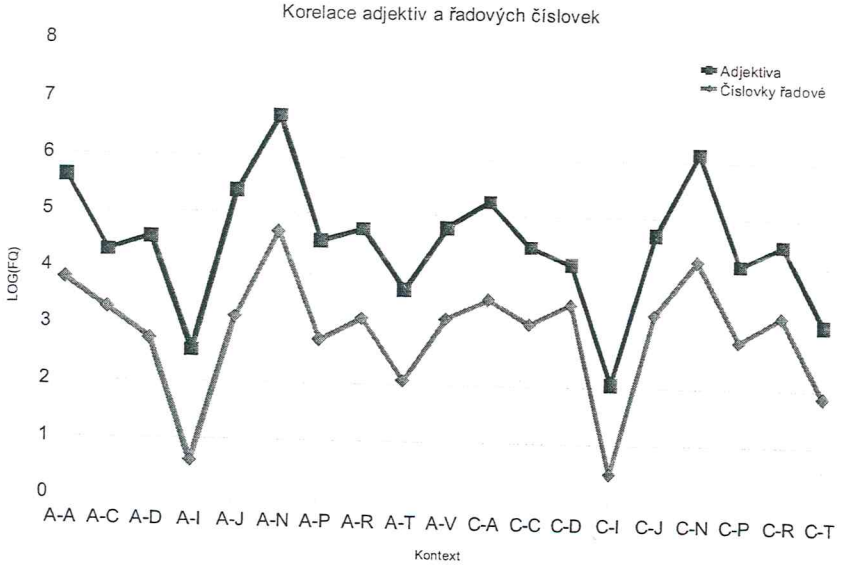
7 *Český národní korpus – SYN2000*. Ústav Českého národního korpusu FF UK, Praha 2000. Dostupný z WWW: <http://www.korpus.cz>; *Český národní korpus – SYN2005*. Ústav Českého národního korpusu FF UK, Praha 2005. Dostupný z WWW: <http://www.korpus.cz>; *Český národní korpus – SYN2006PUB*. Ústav Českého národního korpusu FF UK. Praha 2006. Dostupný z WWW: <http://www.korpus.cz>.

8 Spoustová, D. – Hajič, J. – Votrubec, J. – Krbec, P. – Květoň, P.: The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In: *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*. ACL, Prague 2007, s. 67–74.

9 Zkratky jsou převzaty z morfologické analýzy korpusů řady SYN, viz www.korpus.cz nebo Kocek, J. – Kopřivová, M. – Kučera, K.: *Český národní korpus – úvod a příručka uživatele*. Praha 2000.

10 Volín, J.: *Statistické metody ve fonetickém výzkumu*. Praha 2007. Oakes, M. P.: *Statistics for Corpus Linguistics*. Edinburgh 1998.

11 Baayen H. R.: *Analyzing Linguistic Data*. Cambridge 2008.



Graf ukazuje okem postřehnutelnou korelaci frekvencí kontextů řadových číslovek a adjektiv. Pro srovnatelnost jsou frekvence kontextů normalizovány pomocí logaritmu.

Výsledky pokusu sumarizuje následující tabulka. Z důvodů přehlednosti neuvádíme přímo Spearmanův korelační koeficient, ale determinační koeficient (vynásobený 100), který se počítá jako druhá mocnina koeficientu korelačního a který udává, kolik procent z variace jednoho souboru je vysvětlitelných variací v souboru druhém.

Číslovky	Substantiva	Adjektiva „obyčejná“	Adjektiva „ostatní“	Zájmena	Slovesa	Adverbia stupňovatel	Adverbia nestupňovatel	Předložky	Spojky
Základní	62,59 %	**82,82 %**	67,40 %	56,63 %	39,90 %	47,81 %	39,69 %	52,47 %	40,19 %
Řadové	68,30 %	**87,84 %**	76,49 %	70,81 %	46,64 %	54,90 %	48,07 %	55,98 %	46,55 %
Násobné	27,81 %	46,15 %	58,97 %	55,99 %	77,36 %	81,08 %	**85,71 %**	38,65 %	70,49 %
Druhové	65,85 %	83,19 %	85,05 %	**87,99 %**	63,21 %	70,89 %	65,16 %	43,82 %	60,98 %

Tučně vyznačené výsledky jsou nejvyšším dosaženým determinačním koeficientem dané skupiny číslovek. Tabulka tedy dokládá, že původní hypotéza se nezamítá u silné podobnosti řadových číslovek s adjektivy, násobných číslovek s adverbii a druhových číslovek se zájmeny. Problematické je jenom zahrnutí číslovek základních pod zájmena, kde se silná korelace neukázala. Za poměrně silnou korelací základních číslovek a adjektiv stojí nejspíš fakt, že základní číslovky fungují ve většině případů jako přívlastek před jménem (ačkoli můžou determinované jméno syntakticky řídit), což je funkce prototypická pro adjektiva.

4 Závěrem: obecný princip shlukování slov do slovních druhů

Tento postup nového vymezení slovních druhů se dá samozřejmě zobecnit. Na počátku takového pokusu, který je v současné chvíli skutečně pouze teorií, je každý různý slovní tvar, příp. každé lemma, samostatnou třídou, samostatným „slovním druhem“. Pro každou takovou třídu zjistíme, do jakých kontextů vstupuje a zjistíme, které dvě třídy spolu nejvíc korelují. Tyto dvě skupiny pak sloučíme do jedné a celý proces opakujeme, dokud nedostaneme takový počet tříd, který se nám zdá smysluplný.

Výhodou takovéhoto postupu je jeho 100% formálnost. Přistupujeme totiž na předpoklad, že významové charakteristiky (stejně jako charakteristiky tvarové a syntaktické) se dostatečně demonstrují v kontextu, který proto může být jediným vodítkem pro shlukování slov do skupin. Zároveň tento postup umožňuje fuzzy přístup ke slovnědruhové charakteristice, kdy slovní druh tvoří určité pevné jádro, které naplňuje všechny charakteristiky daného slovního druhu, zatímco periférie se více či méně liší. Takového popisu docílíme, budeme-li při každém spojování pozorovat, jak velký byl korelační koeficient, který je dostatečným kritériem „stejnosti/podobnosti“. Můžeme potom v rámci skupin diferencovat slova korelující svými kontexty silně (jádro) a slova, která tvoří periférii (jejich korelační koeficient je statisticky významně menší).

Naopak nevýhodami tohoto přístupu je vedle poměrně velké výpočetní složitosti zejména otázka hapaxů (a málo frekventovaných slov vůbec), které nemají kontext dostatečně bohatý na to, abychom mohli spolehlivě přistoupit k jejich zařazení. Zároveň není možné hapaxy z textů před analýzou vypustit, protože by tak vznikla „bílá místa“ a některé jednotky by tak zůstávaly bez kontextu.

Sporným momentem celé koncepce může být i to, že bere v úvahu pouze bezprostřední kontext. Korpusová pozorování v této oblasti ovšem ukazují, že z hlediska variability je bezprostřední kontext nejvíc ovlivněn zkoumaným slovem (zkoumanou skupinou slov), tudíž je oprávněný předpoklad, že v něm se nejvíc „odráží“ specifičnost daného slova. Tento poznatek se týká cca 85 až 95 % všech slov (vyšší procento najdeme u frekvenční špičky).[12] Dále by bylo možné poukázat na zbytečné nevyužití formálně morfologických charakteristik. Současná morfologie na tomto kritériu staví významně, je třeba ovšem mít na paměti, že velká část slov se neskloňuje, proto u nich tuto charakteristiku na rozdíl od kontextů, který je vlastní všem jednotkám, nelze použít.

Asi nejproblematičtější je ovšem otázka praktického uplatnění daného přístupu. V každém popisu musí lingvista zvažovat otázku tradice na jedné straně, aby mu veřejnost (i ta lingvistická) rozuměla, a otázku adekvátnosti popisu k popisovanému předmětu (jazyku). I když jsme tedy výsledky tohoto pokusu měli k dispozici při psaní Mluvnice současné češtiny,[13] rozhodli jsme se slovní druh číslovek ponechat; doplnili jsme k výkladu pouze krátkou rozpravu nad možností alternativního pohledu na slovnědruhové rozdělení, v němž by číslovky jako samostatný slovní druh neexistovaly. Soudíme, že problematická místa lingvistického popisu (ať

12 Cvrček, V: *Contextual approach to Parts of Speech*. Konference projektu InterCorp. 2010. (připravuje se)
13 Cvrček, V. – Kodýtek, V. – Kopřivová, M. – Kováříková, D. – Sgall, P. – Šulc, M. – Táborský, J. – Volín, J. – Waclawičová, M.: *Mluvnice současné češtiny*. Praha 2010.

už se jedná o formální morfologii nebo lingvistickou teorii) by měla být vysvětlována i širší veřejnosti tak, aby bylo zřejmé, že nejde o záležitost vyřešenou a černobílou.

Václav Cvrček
Odborné zaměření: korpusová lingvistika, morfologie

Ústav Českého národního korpusu
Filozofická fakulta
Univerzita Karlova
<vaclav.cvrcek@ff.cuni.cz>

K možnostem zpracování spojek a částic ve výkladovém slovníku
(K tzv. polyfunkčnosti spojek a částic a nejednoznačnosti při určování slovních druhů)

BARBORA ŠTĚPÁNKOVÁ

1

Příspěvek vznikl v souvislosti s koncepčními úvahami pro zpracování synsémantik při pracích na lexikální databázi češtiny počátku 21. století.

Základem jednojazyčného výkladového slovníku (a tudíž i lexikální databáze) jsou bezesporu autosémantika, zejména substantiva. Jejich zpracování je opřeno o výklad významu; exemplifikace pak tento výklad dokreslují a přikládají kolokace a typická užití. I další autosémantika jsou řešena podobným způsobem.

U synsémantických slovních druhů[1] je situace poněkud odlišná: jak už napovídá termín „synsémantika", je význam těchto slov „relační, vztahový, tj. závisí na spojení příslušné jednotky se substantivem, popř. i se slovesem (u předložek) a s větnou výpovědí nebo s větným členem (u spojek a částic)".[2] Při výkladu významu i při exemplifikaci je důležitou složkou popis jejich role v kontextu a doložení příkladem v kontextu.[3]

Každý nový slovník hledá vlastní způsob, jak přistupovat k výkladu významu, a to především v souvislosti se zaměřením na rozsah slovníku a na cílového uživatele. U autosémantik (hlavně u substantiv) je pak patrný vývoj slovní zásoby, který je ve slovníku potřeba zachytit – jedná se o posuny významů, zánik či změny v užívání „starých" a vznik nových významů, přesun z jiných jazykových rovin (např. z archaismů, slangů), posuny ve stylovém zařazení, vznik neologismů, nové přejímky atd.

I u synsémantik se slovní zásoba samozřejmě vyvíjí a obměnuje podobně jako u autosémantik, avšak hlavní změny (alespoň v poslední době) lze pozorovat v souvislosti s vývojem lingvistického chápání slovních druhů. Nejpatrnější je posun u částic, vlivem jejich širšího přijímání a ustanovování slovního druhu. U předložek a spojek se tyto změny týkají zejména

1 Jako synsémantika jsou v české/československé lingvistice označovány předložky, spojky a částice, někdy i citoslovce (srov. např. Peciar, Š.: Niektore problémy klasifikácie neohybných slov. In: Machek, V. (ed.): *Studie ze slovanské jazykovědy. Sborník k 70. narozeninám akademika Františka Trávníčka.* Praha 1958, s. 141–146.).

2 Viz Filipec, J. – Čermák, F.: *Česká lexikologie.* Praha 1985, s. 39. Srov. např. i Kopečný, F.: *Základy české skladby*, Praha 1958, s. 18.

3 Synsémantika samozřejmě mají určitý význam i bez kontextu, avšak pro uživatele slovníku je u synsémantik příklad v kontextu nezbytností. Srov. např. i články Šimková, M.: O lexikálnom význame častíc. *Slovenská reč*, 66, 2001, s. 37–51. a Šimková, M.: Z gramatiky a lexikografie tzv. malých slovných druhov. In: Ondrejovič, S. – Považaj, M. (eds.): *Lexicographica '99. Zborník na počesť Kláry Buzássyovej.* Bratislava 2001, s. 189–202.