Porovnání funkčních stylů v korpusu SYN2005 na základě frekvence syntaktických funkcí substantiv

Tomáš Jelínek
Ústav teoretické a komputační lingvistiky FF UK
tomas.jelinek@ff.cuni.cz

ABSTRACT: Large synchronic textual corpora of the Czech National Corpus are built as "representative": they contain a balanced quantity of texts of various styles, roughly divided into three "genres": fiction, technical/scientific literature and journalism. Comparisons of these genres have been performed on phonological and morphological level; in this paper, I deal with differences between genres on the surface-syntactic level.

I use an automatic syntactic annotation of the corpus SYN2005 in the formalism of the analytical layer of th e Prague Dependency Treebank. I compare the frequencies of syntactic functions of nouns in the three genres represented by the corresponding subcorpora of SYN2005. I also present a more detailed analysis of four syntactic phenomena: subtypes of the function of attribute in non-prepositional genitive; frequencies of agreeing groups of the type *pan Novák*; frequencies of the function of agent in passive constructions expressed by nouns in non-prepositional instrumental and the ratio of expression of the nominal part of verbal-nominal predicate by nominative and instrumental.

Significant differences found between genres in all the syntactic phenomena analyzed show that in comparing corpora one should carefully monitor their genre composition. *)*.

Key words: syntax – syntactic function – corpus – genre – representativeness

Klíčová slova: syntax –syntaktická funkce – korpus – žánr – reprezentativnost

1 Úvod

Rozsáhlé textové korpusy jako SYN2005 se sestavují z textů různých typů a žánrů tak, aby byly tzv. „reprezentativní", to znamená, aby představovaly jazykový systém, jak se skutečně používá. V korpusech ČNK jsou texty nahrubo rozděleny do tří funkčních stylů: beletrie, odborné literatury a publicistiky. Srovnáním těchto funkčních stylů na fonologické a morfologické rovině se již věnovala publikace Statistiky češtiny (Bartoň et al., 2009), tento pohled chci doplnit o porovnání z hlediska syntaxe. Vzhledem k omezenému prostoru této studie i potřebě rozsáhlé manuální práce se však věnuji pouze syntaktickým funkcím substantiv. Ukážu zde, jak se funkční styly liší ve frekvenci funkcí substantiv a v několika podrobněji analyzovaných syntaktických jevech.

2 Korpus SYN2005 a jeho syntaktická anotace

Pro srovnání funkčních stylů jsem zvolil korpus SYN2005 tak, aby byla data srovnatelná s citovanými Statistikami češtiny (použil jsem i stejné morfologické značkování). Korpus SYN2005, „synchronní reprezentativní korpus současné psané češtiny"[1], se skládá ze psaných textů publikovaných ponejvíce v letech 2000 až 2004, částečně i z textů starších; v korpusu jsou jak původní české texty, tak texty překladové.

1 http://korpus.cz/syn2005.php

2.1 Funkční styly (žánrové subkorpusy) v korpusu SYN2005

Korpus se skládá z textů rozdělených do tří základních funkčních stylů: beletrie, odborná literatura a publicistika. Beletrie (BEL) zahrnuje především romány a povídky, dále literaturu faktu, poezii, scénáře, některá populární periodika aj. Subkorpus odborné literatury (ODB) obsahuje texty „vědeckonaučné“, populárně naučné a texty z odborných a zájmových periodik. Publicistika (PUB) sestává převážně z novinových článků, částečně i z článků z časopisů. Podíl funkčních stylů na korpusu ukazuje následující tabulka.

Tabulka 1. Počet textových slov ve funkčních stylech korpusu

	BEL	ODB	PUB	SYN2005
textových slov	40 mil.	27 mil.	33 mil.	100 mil.

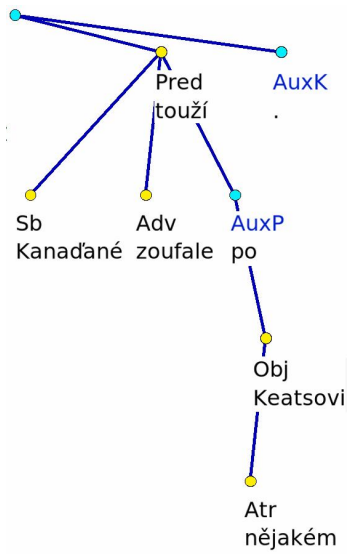
2.2 Závislostní syntaktické značkování korpusu

Žádnou podrobnější anotaci korpusu o rozsahu 100 miliónů slov nelze provádět „manuálně“. Automatická syntaktická anotace je však výrazně chybovější než anotace morfologická, aby byly výsledky dostatečně spolehlivé, bylo nutné zapojit i „manuální“ ověření dat na vzorcích.

2.2.1 Formalismus analytické roviny PDT

Aby bylo možné korpus syntakticky anotovat, bylo nutné zvolit formalismus syntaktické anotace. Účelům této studie nejlépe vyhovoval formalismus „analytické roviny“ Pražského závislostního korpusu (PDT), jak ho definuje anotační manuál PDT[2] (Hajič et al., 1999). Na analytické rovině je celá větná struktura včetně všech pomocných slov i interpunkce reprezentována závislostní strukturou (stromem), v níž má každý token právě jeden řídící element (větný člen nebo „technický“ kořen věty). Každému tokenu je přiřazena syntaktická funkce; používají se jednak „klasické“ syntaktické funkce víceméně podle Vl. Šmilauera

Obrázek 1. Příklad závislostní struktury v PDT



2 http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/a-layer/html/

(Šmilauer, 1966): podmět, přísudek, předmět aj., jednak „pomocné“ syntaktické funkce pro neplnovýznamová slova a interpunkci. Příklad takové struktury ukazuje obrázek 1 (AuxP a AuxK jsou pomocné funkce pro předložky a pro koncovou interpunkci).

2.2.2 Automatické syntaktické značkování korpusu

Pro označkování korpusu bylo nutné využít plně automatického značkování: použil jsem mírně vylepšené nastavení stochastického MST parseru (McDonald et al., 2005), některé opakující se chyby byly opraveny pomocí nově vytvořeného opravného programu, který nalezené chybné struktury opravuje na základě „lingvistických“ pravidel, více viz (Jelínek, 2011) a podrobněji v (Jelínek, 2012).

Úspěšnost výsledného syntaktického značkování korpusu je 86,3 % pro určení řídícího slova, cca 80,8 % včetně správného určení syntaktické funkce. Chybovost je tedy značná, asi 20 %. Naštěstí parser (a následné opravná pravidla) chybuje nejvíce u nejméně frekventovaných funkcí a nejméně častých realizací syntaktické funkce slovním druhem a morf. kategorií, např. téměř nikdy neurčí správně subjekt v předložkovém pádu, v datech PDT je takových případů jen asi 30, viz příklad z PDT (1).

(1) Do tří tříd soukromé obchodní akademie a hotelové školy **nastoupilo** včera v Turnově **po** 28 **žácích**/**Sb**.

U častějších případů je chybovost výrazně nižší, např. určení neshodného přívlastku v prostém genitivu nebo podmětu v nominativu je úspěšné ve více než 90 %.

2.2.3 Manuální ověření vzorků a přepočítání údajů

Kvůli vysoké chybovosti automatické anotace nebyla výsledná data dostatečně spolehlivá. Pro každou kombinaci syntaktické funkce a pádu substantiva jsem tedy ověřil náhodně vybraný vzorek dat o rozsahu 100 až 500 výskytů (podle frekvence jevu). Získané údaje jsem pak promítl do výsledků, uvádím tedy data takto přepočítaná.

3 Srovnání žánrových subkorpusů

U tří funkčních stylů (žánrových subkorpusů) v SYN2005 jsem popsaným způsobem zjistil celkové frekvence syntaktických funkcí substantiv. Pro lepší porozumění údajům, které uvedu, nejprve upřesňuji pojetí syntaktických funkcí. Pro srovnání žánrů jsem kromě celkové frekvence syntaktických funkcí zvolil také několik syntaktických jevů, které dobře ukazují rozdíly mezi funkčními styly, a podrobněji je analyzuji.

3.1 Syntaktické funkce

Syntaktické funkce substantiv až na drobné výjimky přebírám z PDT (jež, jak bylo řečeno, přibližně vychází ze syntaktické koncepce Vl. Šmilauera). Pracuji tedy s těmito funkcemi: podmět (Sb), jmenná část verbonominálního predikátu (Pnom), přívlastek (Atr), předmět (Obj), příslovečné určení (Adv) a doplněk (Atv). Samostatně jsem vyčlenil funkci „součást adordinační skupiny typu *pan Novák*“ (Adord). Funkci ExD, která v PDT označuje větné členy mimo větnou strukturu, do výsledků nezapočítávám. Následuje krátký přehled syntaktických funkcí, kde uvádím typické tvary substantiv a v případě potřeby upozorňuji na drobné odlišnosti oproti pojetí Vl. Šmilauera nebo PDT. Všechny příklady pocházejí

z korpusu SYN2005.

3.1.1 Podmět (Sb)

Podmět je obvykle v nominativu, zřídka může být i v genitivu (2) nebo v předložkovém pádu (3).

(2) Zatlačila ho do rohu a **nebylo úniku/Sb**.

(3) **Po** jednom **zástupci/Sb mělo** v této komisi ministerstvo zemědělství a ministerstvo životního prostředí.

3.1.2 Jmenná část verbonominálního predikátu (Pnom)

Tuto funkci může plnit substantivum v prostém nominativu (4) či instrumentálu (5), zřídka v genitivu. Předložkové pády se v koncepci PDT řadí (i z technických důvodů) mezi příslovečná určení: nelze spolehlivě automaticky rozlišit mezi příslovečným určením rozvíjejícím sloveso *být* a verbonominálním predikátem.

(4) Ale můj **obor/Pnom** to **není**.

(5) Soudím, že **příčinou/Pnom** všeho **byly** zase jen peníze.

3.1.3 Neshodný substantivní přívlastek (Atr)

Funkci neshodného přívlastku plní substantiva ve všech pádech, prostých i předložkových. Nejčastěji jsou v prostém genitivu (6) a předložkovém lokálu. Prostý akuzativ je v této funkci zcela výjimečný (7).

(6) Do **východu slunce/Atr** bylo ještě daleko.

(7) Anamnéza plošné bolesti s **trváním** nejméně tři **měsíce/Atr**.

3.1.4 Součást adordinační skupiny typu *pan Novák*, *řeka Mže* (Adord)

Jako samostatnou skupinu vyděluji součásti adordinační skupiny typu *pan Novák* nebo *řeka Mže*, které se v PDT řadí mezi přívlastky. Ve skupině shodných substantiv označujících osoby mají funkci Adord všechna substantiva kromě posledního, jež má funkci, kterou zastává celá skupina (8). V případě označení předmětů (krajinných útvarů, sídlišť, institucí apod.) je to naopak (9).

(8) Podle **soudkyně/Adord Vladimíry/Adord Hájkové** neexistuje jediný přímý důkaz

(9) Ve spořádaném **městě Portlandu/Adord** si připadal neužitečný.

3.1.5 Předmět (Obj)

Předmětem mohou být substantiva ve všech pádech kromě nominativu, prostých (10) i předložkových (11). Hranice mezi předměty a příslovečnými určeními je mnohdy obtížně definovatelná, zvláště v případě substantiv v předložkových pádech: zde rozlišuji ve shodě s PDT podle míry abstrakce. Naopak na rozdíl od PDT řadím určení původce děje mezi příslovečná určení.

(10) **Potřásla** modročernou **hřívou/Obj** a frkla

(11) Cosette se **do Marka/Obj zamiluje**.

3.1.6 Příslovečné určení (Adv)

Funkci příslovečného určení plní substantiva ve všech prostých (12) i předložkových (13) pádech. Ve shodě s PDT sem počítám i tzv. volný dativ (posesivní, prospěchový aj.).

(12) **Chvilku/Adv vypadal ztrápeně.**

(13) Domů se **vrátila** až **za tmy/Adv.**

3.1.7 Doplněk (Atv)

Syntaktická funkce doplněk se realizuje substantivem jen výjimečně, převážně v konstrukcích se spojkou *jako* (ne však ve srovnávacích spojeních).

(14) **Sloužil** jako **rezidence/Atv** významným osobnostem.

(15) **Dáváme** ti **darem/Atv** kouzelnou loutnu.

3.1.8 Substantiva mimo větné struktury (ExD)

Substantiva, která jsou v PDT označena funkcí ExD, tj. substantiva ve výpovědích bez slovesa, v názvech a titulcích, vokativy aj., do výsledků nezapočítávám.

(16) **Praha/ExD** – Václav Klaus může zůstat klidný:

(17) Milý **chlapče/ExD**!

3.2 Frekvence syntaktických funkcí substantiv

V tabulce uvádím zjištěný procentuální podíl jednotlivých syntaktických funkcí substantiv v celém korpusu a jeho třech žánrových subkorpusech. Barevně zvýrazňuji v rámci každého subkorpusu nejčastější syntaktické funkce.

Tabulka 2. Frekvence syntaktických funkcí substantiv

	Sb	Pnom	Atr	Adord	Obj	Adv	Atv	∑
BEL	24,1	2,7	16,6	2,3	26,5	26,5	1,3	100 %
ODB	20,8	2,2	31,2	1,7	21,5	21,6	1,1	100 %
PUB	23,7	2,0	25,9	4,8	21,2	21,7	0,8	100 %
SYN2005	22,9	2,3	24,5	3,0	23,0	23,2	1,1	100 %

Jak vyplývá z tabulky, největší rozdíly jsou mezi subkorpusy BEL a ODB, významně se liší v podílech téměř všech syntaktických funkcí. O něco menší jsou rozdíly mezi PUB a BEL, podíl podmětu mají dokonce téměř stejný.

Subkorpusy ODB a PUB si jsou bližší. Větší rozdíl mezi ODB a PUB je jednak ve frekvenci podmětu (kratší novinové texty vyžadují častější uvádění podmětu než delší odborné články), jednak ve frekvenci adordinačních spojení typu *pan Novák*, jimž se budeme věnovat později.

Nejčastější funkcí v celém korpusu a v subkorpusech ODB a PUB jsou přívlastky, kdežto v beletrii jsou dvě stejně zastoupené nejčastější funkce: Obj a Adv. Tento rozdíl vyplývá

především z odlišného rozložení slovních druhů v subkorpusech: v BEL je o polovinu více sloves než v ODB a naopak o třetinu méně substantiv, jak ukazuje tabulka 3. Je-li v subkorpusu méně sloves, musí tam být více větných členů rozvíjejících substantiva.

Tabulka 3. Podíl substantiv a sloves na všech slovech v korpusu

	substantiv	sloves
BEL	24,3 %	21,2 %
ODB	34,5 %	13,9 %
PUB	33,8 %	16,0 %
SYN2005	30,2 %	17,5 %

3.3 Frekvence delších nominálních skupin

Výše zmíněný rozdíl ve frekvenci sloves a substantiv ve funkčních stylech souvisí také s mírou užití dlouhých nominálních skupin v jednotlivých korpusech. V tabulce 4 představuji frekvenci „dlouhých“ nominálních skupin složených ze substantiva rozvitého dalšími substantivy v prostém genitivu nebo v předložkových pádech, substantiva mohou být také rozvita adjektivem. Jmenná skupina v příkladu z odborné literatury (18) má deset substantiv, jmenná skupina v příkladu z beletrie (19) substantiv devět.

(18) **Metodika sledování stavu hydratace cementu** pomocí **určení změn obsahu** volné **vody** v hydratující **směsi**

(19) **návrh krále Jiříka** na **vytvoření svazu panovníků** k **zabezpečení míru** v **Evropě**

Tabulka 4. Počet nominálních skupin o daném počtu substantiv na 1 mil. slov

substantiv	10	9	8	7	6	5
BEL	0,1	0,3	0,6	3,2	17,3	87,0
ODB	1,3	5,5	14,1	49,3	167,2	591,8
PUB	0,8	2,8	10,4	36,9	135,6	484,3
SYN2005	0,7	2,5	7,4	26,4	95,7	350,6

Podle očekávání jsou jmenné skupiny tím méně časté, čím jsou delší. Více se vyskytují v odborné literatuře než v beletrii. Zajímavý je ale průměrně desetinásobný rozdíl v počtu delších nominálních skupin mezi odbornou literaturou a beletrií.

3.4 Podíly jevů u neshodného přívlastku v genitivu

Funkce „neshodný přívlastek“ v prostém genitivu zahrnuje jevy, které spolu souvisejí jen velmi volně: kvantifikaci, valenční přívlastek deverbativ, široce přivlastňovací přívlastek aj. Některé z těchto jevů lze spolehlivě určit automaticky a kvantifikovat.

Frekvence přívlastku se mezi jednotlivými funkčními styly velmi liší a rozdílné jsou samozřejmě i frekvence přívlastku v prostém genitivu, jak ukazuje tabulka 5. Tyto rozdíly je třeba mít na zřeteli při interpretaci dále uváděných údajů: podílů jednotlivých jevů na celkové frekvenci přívlastku v prostém genitivu.

Tabulka 5. Podíl genitivních přívlastků na všech substantivech v korpusu

	BEL	**ODB**	**PUB**	**SYN2005**
podíl gen. Atr	13,0 %	25,9 %	20,6 %	19,7 %

Automaticky se podařilo identifikovat čtyři podtypy neshodného gen. přívlastku: přívlastek rozvíjející deverbativní substantivum s koncovkou *-ní/tí* (*řešení problému, vytí vlků, hnutí odporu*), přívlastek závislý na číslovce (*několik dní, deset metrů, 40 korun*), přívlastek rozvíjející substantivní kvantifikátor (*většina lidí, litr vody, hrstka soli*) a přívlastek závislý na substantivu s koncovkou *-ost,* s vyloučením nejfrekventovanějších substantiv, která nejsou deadjektivní (*rychlost světla, platnost karty, vážnost situace*). Okolo 70 % přívlastků se roztřídit nepodařilo, protože zkoumaný jev je příliš rozmanitý, i tak jsou zaznamenané rozdíly mezi subkorpusy zajímavé.

Oranžovou barvou jsou zvýrazněny nejvyšší podíly v dané kategorii, modrou nejnižší.

Tabulka 6. Podíl některých podtypů neshodného přívlastku v prostém genitivu

řídící slovo	**BEL**	**ODB**	**PUB**	**SYN2005**
deverb. substantivum (-ní/tí)	7,3 %	17,5 %	10,9 %	12,7 %
číslovka	8,9 %	4,2 %	11,3 %	8,0 %
substantivní kvantifikátor	7,3 %	6,3 %	6,0 %	6,4 %
vlastnost (-ost)	3,0 %	4,8 %	2,5 %	3,6 %
ostatní (nerozlišeno)	73,5 %	67,2 %	69,3 %	69,3 %
celkem	100 %	100 %	100 %	100 %

Zkoumané čtyři jevy jsou celkově nejméně zastoupené v beletrii. Beletrie se vyhýbá užití deverbativ rozvitých přívlastky, naopak zde mají nejvyšší podíl přívlastky závislé na substantivních kvantifikátorech.

Podle očekávání je v odborné literatuře nejvyšší podíl přívlastků závislých na deverbativech a na označeních vlastnosti ze všech tří žánrů. Překvapivý zde může být poměrně nízký podíl přívlastků závislých na číslovkách.

V publicistice je nejvyšší podíl přívlastků závislých na číslovkách, naopak nejméně jsou zde zastoupeny přívlastky rozvíjející názvy vlastností.

3.5 Frekvence součástí adordinačních konstrukcí typu *pan Novák*

Při syntaktické anotaci jsem jako samostatný jev vyčlenil typ *pan Novák* a *město Praha*, lze tak snadno srovnat, jak často se v tom kterém žánru používají. V tabulce uvádím frekvence tří podtypů tohoto jevu. Typ *město* ***Praha***: shodná propria v jedné nominální skupině s apelativy označujícími sídliště, krajinné útvary apod. Typ ***pan*** *Novák*: shodná apelativa jako *pan, doktorka, předseda, sopranistka,* která společně s následujícími apelativy či proprii označují osoby. Typ ***Jan*** *Novák*, kde shodná propria, křestní jména, spolu s následujícími propriem označují osoby. Všechny tři typy (červeně zvýrazněné) ukazuje příklad (20); modře zvýrazněné jsou „řídící“ substantiva skupiny (dle formalismu PDT), která k žádnému z podtypů nepočítám.

(20) Památník **olympioniku/Adord Emilu/Adord Zátopkovi/Atr** v rodném **městě/Adv Kopřivnici/Adord** odhalila u příležitosti jeho nedožitých 80. narozenin **manželka/Adord Dana/Adord Zátopková/Sb**.

V tabulce je uveden počet výskytů jednotlivých typů na 1 milión textových slov v (sub)korpusu. Oranžovou barvou jsou zvýrazněny nejvyšší podíly v dané kategorii, modrou nejnižší.

Tabulka 7. Počet výskytů členů adordinačních konstrukcí

	BEL	ODB	PUB	SYN2005
město **Praha**	52	72	171	97
pan Novák	1 385	502	2 537	1 527
Jan Novák	665	787	4 527	1 972
celkem	2 102	1 361	7 235	3 596

V publicistice se tento typ (všechny tři jeho varianty) používá několikanásobně častěji než v ostatních dvou žánrech. V beletrii jsou ještě relativně častá apelativa v této funkci *(paní, král, panna, generál...),* ostatní typy jsou zde zastoupeny velmi málo. Nejméně se tento jev vyskytuje v odborné literatuře.

3.6 Frekvence určení činitele děje

Substantivum v prostém instrumentálu ve funkci příslovečného určení činitele děje rozvíjí sloveso v pasivu (21) nebo deverbativní adjektivum na *-ný/tý* (22).

(21) A moji lidé že byli **napadeni** těmi **fanatiky**?

(22) O osudu města Villamedie, **napadené** nepřátelským **letectvem**, nemáme dosud bližších zpráv.

V tabulce uvádím frekvenci výskytu původce děje u opisného pasiva a u deverb. adjektiv, přepočítanou na 1 milión textových slov v (sub)korpusu. Oranžovou barvou jsou zvýrazněny nejvyšší podíly v dané kategorii, modrou nejnižší.

Tabulka 8. Frekvence určení činitele děje

	BEL	ODB	PUB	SYN2005
původce děje **V**	519	1 520	508	786
původce děje **A**	410	669	332	454
původce děje	929	2 189	840	1 240

Z tabulky je vidět, že určení činitele děje je zcela podle očekávání výrazně častější v odborné literatuře než v ostatních dvou žánrech. Překvapivé je, že v publicistice se tento jev vyskytuje ještě méně než v beletrii.

Vysvětlením je složení žánrových subkorpusů. Publicistika je poměrně homogenní (jen novinové a časopisecké články), ale beletrie i odborná literatura obsahují poměrně rozmanité texty (v BEL je např. také literatura faktu, v ODB je zastoupena i populárně naučná literatura). Abych ověřil tuto hypotézu, definoval jsem ještě dva menší subkorpusy: „jádro“ beletrie (pouze romány, novely a povídky) a „jádro“ odborné literatury (pouze „vědeckonaučná“ literatura, ne populárněnaučná aj.). Analýzu téhož jevu v těchto subkorpusech ukazuje tabulka 9.

Tabulka 9. Frekvence určení činitele děje v BEL a ODB a jejich „jádrech“

	BEL_j	BEL	ODB	ODB_j
původce děje **V**	397	519	1 520	1 606
původce děje **A**	370	410	669	709
původce děje	767	929	2 189	2 315

Ze srovnání tabulky 9 s tabulkou 8 je zřejmé, že v „jádru“ beletrie se určení činitele děje užívá méně než v publicistice a cca o 20 % méně než v celém žánru BEL. Naopak rozdíl mezi „jádrem“ ODB a celým subkorpusem ODB je jen velmi malý.

3.7 Realizace jmenné části verbonominálního predikátu

Posledním zkoumaným jevem je poměr realizace jmenné části verbonominálního predikátu prostým nominativem (23) a instrumentálem (24) substantiva:

(23) **Byla** to jen drobná **epizodka/Pnom**.

(24) **Předmětem/Pnom** sváru **je** primátorova přítelkyně,

V první tabulce uvádím frekvence jevu (nominativu, instrumentálu, celkově) v celém korpusu a jeho žánrových subkorpusech. V tabulce je uveden počet výskytů jednotlivých typů na 1 milión textových slov v (sub)korpusu. Oranžovou barvou jsou zvýrazněny nejvyšší podíly v dané kategorii, modrou nejnižší.

Tabulka 10. Frekvence jmenné části verbonominálního predikátu v nominativu a instrumentálu

	BEL	ODB	PUB	SYN2005
Pnom **1**	3 931	1 834	2 625	2 934
Pnom **7**	1 737	4 608	3 200	2 995
Pnom	5 668	6 442	5 825	5 929

Vidíme, že celkové frekvence substantiv ve funkci jmenné části verbonominálního predikátu jsou ve všech žánrech poměrně vyrovnané. Frekvence využití nominativu a instrumentálu se však výrazně liší: tento rozdíl je ještě lépe viditelný v tabulce 11, kde jsou uvedeny poměry výskytů nominativu a instrumentálu v procentech.

Tabulka 11. Podíl realizace jmenné části verbonominálního predikátu nominativem a instrumentálem

	BEL	ODB	PUB	SYN2005
Pnom **1**	69,4	28,5	45,1	49,5
Pnom **7**	30,6	71,5	54,9	50,5
Pnom	100 %	100 %	100 %	100 %

Poměr nominativu a instrumentálu v BEL a v ODB je, jak ukazuje tabulka, opačný: v BEL je přibližně 70 % substantiv v této funkci v nominativu, v ODB je jich 70 % v instrumentálu. Jen v publicistice je poměr vyrovnanější s mírně převažujícím instrumentálem.

Stejně jako u předcházejícího jevu jsem analyzoval jev také v „jádrech“ subkorpusů BEL a ODB, které obsahují pouze texty, které jsou pro dané funkční styly nejtypičtější.

Tabulka 12. Podíl realizace jmenné části verbonominálního predikátu v BEL a ODB a jejich „jádrech“

	BEL_j	BEL	ODB	ODB_j
Pnom **1**	74,8	69,4	28,5	27,4
Pnom **7**	25,2	30,6	71,5	72,6
Pnom	100 %	100 %	100 %	100 %

Stejně jako u předchozího jevu je v „jádru“ BEL trend (v tomto případě převaha nominativu) dále zvýrazněn, kdežto mezi „jádrem“ ODB a celým subkorpusem ODB je rozdíl jen malý: vědeckonaučná („jádro“) i populárněnaučná literatura si jsou v obou analyzovaných jevech podobné.

4 Závěr

Funkční styly v korpusu SYN2005 se ve všech zkoumaných syntaktických jevech významně liší. Rozdíly mezi žánry víceméně odpovídají očekávání, výzkum jen poskytl přesné údaje z rozsáhlého textového korpusu.

Značné rozdíly mezi žánrovými subkorpusy (a částečně i rozdíly v rámci jednotlivých žánrů), které výzkum konstatoval, ukazují na nutnost velké opatrnosti při interpretaci dat z „reprezentativního“ korpusu nebo při srovnávání „reprezentativních“ korpusů: zastoupení žánrů v korpusu i složení žánrů velmi ovlivňuje frekvenci jevů morfologických i syntaktických.

Z hlediska metodiky jazykového výzkumu se také ukázalo, že automatická syntaktická anotace je přes svou vysokou chybovost užitečným nástrojem pro výzkum jazyka, vyžaduje ale velké množství manuálního ověřování, které chyby neutralizuje. Před případným zveřejněním automaticky syntakticky anotovaného korpusu bude nutné zvýšit spolehlivost anotace: pro většinu uživatelů je procento chyb zatím příliš vysoké.

LITERATURA A DALŠÍ ZDROJE

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

Český národní korpus – SYN2005. Ústav Českého národního korpusu FF UK, Praha 2005. Dostupný z WWW: <http://www.korpus.cz>.

HAJIČ, J. – PANEVOVÁ, J. – BURÁŇOVÁ, E. – UREŠOVÁ, Z. – BÉMOVÁ, A. (1999): *Anotace Pražského závislostního korpusu na analytické rovině: pokyny pro anotátory (html)*.

JELÍNEK, T. (2011): Systém pro syntaktické značkování velkých textových korpusů. In: V. Petkevič, A. Rosen (eds), *Korpusová lingvistika Praha 2011, sv. 3: Gramatika a značkování korpusů.* Praha: Nakladatelství Lidové noviny / Ústav českého národního korpusu, 123–142.

JELÍNEK, T. (2012): *Forma a funkce u substantiv v češtině: vztah pádu a syntaktické funkce. Na materiálu korpusu současné psané češtiny (SYN2005).* Disertační práce. Praha: FF UK.

MCDONALD R. – PEREIRA F. – RIBAROV K. – HAJIČ J. (2005): Non-Projective Dependency Parsing using Spanning Tree Algorithms. In: *Proceedings of HLT-EMNLP 2005*. ACL, Vancouver, 523–530.

ŠMILAUER V. (1966): *Novočeská skladba*. Praha: SPN.

GRANTOVÁ PODPORA
Tento příspěvek byl podpořen z grantu GAČR P406/10/0434.