UNIVERZITA PALACKÉHO

FILOZOFICKÁ FAKULTA

Katedra bohemistiky

Miroslav Kubát

Funkční styly z hlediska lexikální statistiky

diplomová práce

Vedoucí práce: Mgr. Darina Hradilová, Ph.D.

Olomouc 2012

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně na základě uvedené literatury.

V Olomouci 8. května 2012 .....

Rád bych poděkoval především Mgr. Darině Hradilové, Ph.D. za velice vstřícné vedení mé diplomové práce. Poděkování patří také Mgr. Martině Benešové, Ph.D. za přínosné rady při statistickém zpracování dat a PhDr. Petru Pořízkovi, Ph.D. za pomoc při práci s ČNK.

1 ÚVOD 6
2 FUNKČNÍ STYLY 8
2.1 Diferenciace funkčních stylů 10
3 BOHATSTVÍ TEXTU 14
4 METODOLOGIE LEXIKÁLNÍ ANALÝZY 21
4.1 Kvantitativní lingvistika 21
4.2 Stylistika a kvantitativní lingvistika 22
4.3 Základní prvky statistické analýzy 24
4.4 Vytvoření výběrového souboru 25
4.5 Český národní korpus jako zdroj dat 26
4.6 SYN2010 26
4.7 PMK 29
5 ROZSAH, ROZPTÝLENÍ A KONCENTRACE LEXIKA 31
5.1 Metodologie přiřazení subkorpusů k funkčním stylům 32
5.1.1 Metodologie vytvoření výběrového souboru 35
5.1.2 Kvantitativní hledisko 35
5.1.3 Kvalitativní hledisko 37
5.1.4 Jak spojit SYN2010 a PMK? 40
5.1.5 Lemma vs. slovoforma 41
5.1.6 Metodologie analýzy 42
5.2 Rozsah lexika 43
5.2.1 Umělecký styl 44
5.2.2 Odborný styl 47
5.2.3 Publicistický styl 49
5.2.4 Administrativní styl 52
5.2.5 Hovorový styl 52
5.2.6 Komparace funkčních stylů 54
5.3 Rozptýlení lexika 57
5.3.1 Umělecký styl 58
5.3.2 Odborný styl 59
5.3.3 Publicistický styl 60
5.3.4 Administrativa 60

5.3.5 Hovorový styl ... 61
5.3.6 Komparace funkčních stylů ... 62
5.4 Koncentrace lexika ... 63
5.4.1 Umělecký styl ... 63
5.4.2 Odborný styl ... 65
5.4.3 Publicistický styl ... 67
5.4.4 Administrativní styl ... 71
5.4.5 Hovorový styl ... 71
5.4.6 Komparace funkčních stylů ... 73
5.5 K problematice délky textu ... 74
6 LEMMA VS. SLOVOFORMA ... 78
6.1 Wilcoxonův test ... 78
6.2 Postup výpočtu WSRT ... 78
6.3 Vypočet WSRT ... 79
6.4 Vyhodnocení testu ... 79
6.5 Závěr ... 80
7 ZÁVĚR ... 82
8 ANOTACE ... 85
9 SUMMARY ... 86
10 ZDROJE ... 88
10.1 Literatura ... 88
10.2 Internetové odkazy ... 89
10.3 Použitý software ... 89
11 PŘÍLOHA ... 90

1 ÚVOD

Cílem této práce je podat popis funkčních stylů na základě lexikostatistických parametrů, jež spadají do základních veličin bohatství textu v českém jazykovědném prostředí. Funkční styly přitom nebudou podrobeny jen vzájemné komparaci mezi sebou, ale pozornost bude také směřována k rozdílům uvnitř jednotlivých funkčních stylů. Kruciální otázkou bude, jak se diferenciace funkčních stylů v současných stylistikách odráží obecně v jejich lexiku.

Paralelně vedle primárního šetření bude také sekundárně zkoumáno, zda tradiční pojetí základní jednotky lexikostatistických výpočtů je pro češtinu správné. Tedy zda je pro češtinu vhodnější pracovat se slovem systémovým (lemmatem), nebo slovem textovým (slovoformou). Všechny výpočty tak budou paralelně vždy počítány v obou jednotkách a bude sledováno, zda a do jaké míry se výsledné hodnoty liší. Výsledkem bude statistické vyjádření relevance užití těchto jednotek.

V úvodní části textu bude pojednáno obecně o teorii funkčních stylů v českém lingvistickém prostředí, a to jak z pohledu vývoje, tak současného pojetí v moderních stylistikách. Následně bude stanoveno, jak jsou funkční styly a jejich rozdělení chápány v této práci. V další části bude pojednáno obecně o problematice bohatství textu, jehož definice je velmi komplikovaná zejména z důvodu nejednoznačnosti chápání tohoto pojmu mezi lingvisty. Proto na pozadí různých přístupů stanovíme náš pohled na problematiku a vysvětlíme, které parametry aplikujeme v našem šetření a proč. V následujících částech pak budou definovány metodologické postupy a východiska, která budou při výzkumu uplatněna.

Na počátku kvantitativní analýzy bude sestaven výběrový soubor, jenž bude excerpován z databáze Českého národního korpusu. Následně budou sestaveny frekvenční seznamy, jež budou základním materiálem pro výpočty. Z výsledných hodnot pak budou generovány grafy, jež umožní snadnější komparaci. V rámci každé podkategorie každého funkčního stylu bude uveden komentář výsledných dat, následně budou komparaci a interpretaci podrobeny jednotlivé funkční styly jako celky. Přestože bude průběžně přihlíženo k paralelně uváděným výsledným hodnotám výpočtů na základě obou základních jednotek, teprve další část provede

samostatné vyhodnocení užití těchto jednotek, a to pomocí statistického ověření závislosti obou sad výsledných hodnot.

Nedílnou součástí předkládaného textu bude také příloha, jež bude obsahovat získané frekvenční slovníky, jejichž rozsah bude stanoven na prvních 50 slov. Vzhledem k zachování koncepce celého výzkumu budou uvedeny slovníky všech sledovaných kategorií, stejně tak budou podány vedle sebe seznamy slov systémových i textových. Tyto frekvenční slovníky tak budou sloužit jednak jako logické doplnění celého textu, jednak jako potenciální materiál pro další lingvistická šetření.

Tento výzkum do značné míry navazuje na bakalářskou diplomovou práci *Rozsah, rozptýlení a koncentrace lexika v psané a mluvené češtině*. Charakter této práce se však bude lišit jednak zkoumaným objektem (funkční styly), jednak několikanásobně obsáhlejším výběrovým souborem, jež by měl zvýšit relevantnost celého výzkumu. Vzhledem k tomu, že na zmíněný text budeme v této práci často odkazovat, budeme pro zjednodušení užívat zkratku „*RRKLPMČ*“.

2 FUNKČNÍ STYLY

Jak je již patrné z názvu této práce, primárním předmětem pozornosti budou funkční styly, proto považujeme za nezbytné hned na začátku vymezit naše chápání funkčních stylů, neboť česká stylistika není zcela jednotná, zejména v diferenciaci základních funkčních stylů, ať už co do jejich kvality či kvantity. Je zřejmé, že pouze tato problematika si rozsahem vyžaduje samostatnou práci, proto již v úvodu deklarujeme, že tento text nemá ambice široce teoreticky zpracovat problematiku funkčních stylů či vůbec stylu jako takového. Přesto nemůžeme alespoň v rámci možností této práce nepojednat o elementárních otázkách, jejichž jasné zodpovězení je klíčové pro správné a celistvé pochopení celé práce.

Než přejdeme k výkladu termínu „funkční styl“, podáme nejdříve význam samotného výrazu „jazykový styl“. Konsenzuální definici stylu nabízí autoři *Stylistiky současné češtiny*, pro něž je jazykový styl „způsob cílevědomého výběru a uspořádání (organizování) jazykových prostředků, který se uplatňuje při genezi textu; v hotovém komunikátu se pak projevuje jako princip organizace jazykových jednotek, který z částí a jednotlivostí tvoří jednotu vyhovujícím komunikačnímu záměru autora“[1] Velmi důležitou pro naši práci shledáváme poznámku následující bezprostředně po uvedené definici, v níž autoři tvrdí, že jazykový styl existuje v textu v souvislosti s jeho obsahovou stránkou, ale z výzkumných a didaktických důvodů se při rozborech můžeme soustředit jen na výrazovou složku.

Než se dostaneme k vymezení našeho chápání termínu „funkční styl“, ukážeme, jak jej prezentují slovníky. Nejdříve uvedeme definici Edvarda Lotka, který funkční styl chápe jako „druh objektivního stylu, který je určen základní promluvovou funkcí, např. funkční styl prostě sdělovací, publicistický, odborný, umělecký aj. (dříve též funkční jazyk).“[2] Lotko tedy vymezuje funkční styly jako podkategorii stylů objektivních, přičemž objektivní styl definuje jako „opřený o normu a úzus, do něhož patří typy prostředků podle adresáta, situace, kontaktu, funkce apod. (např. styl adresní, zakotvený, psaný, odborný).“[3]

---

[1] Čechová, M . a kol.: *Stylistika současné češtiny*. Praha 1997, s. 9.

[2] Lotko, E.: *Slovník lingvistických termínů pro filology*. Olomouc 2005, s. 110.

[3] Tamtéž.

Zatímco Lotkův slovník je především základním přehledem pro studenty, *Encyklopedický slovník* češtiny (dále jen ESČ) představuje konsenzuální východisko ke klíčovým termínům češtiny, a je tak pro naše účely ideálním výchozím zdrojem informací o chápání funkčních stylů v českém lingvistickém prostředí. Stejně jako Lotko i Milan Jelínek, jenž je autorem hesla „styl funkční", hned na začátku vymezuje tento termín na pozadí objektivních stylů, jejichž je jedním z typů. Funkční faktory, jež determinují funkční styl, pak Jelínek označuje za nejvýraznější ze všech objektivních faktorů. Hned v úvodu také autor zmiňuje specifické postavení funkční stylistiky v české jazykovědě ve vztahu k Pražské škole, jejíž vklad k funkčnímu pojetí jazyka by bylo nepochybně redundantní detailněji rozepisovat.

Dle autorů *Stylistiky současné češtiny* při „charakteristice funkční stylové diferenciace spisovného jazyka vychází současná česká lingvistika z Jedličkova pojetí tří rovin stylových jevů."[4] Těmito základními jevy jsou:

a) stylová sféra,
b) stylová vrstva,
c) stylový typ.

Autoři dále vysvětlují součinnost zmíněných stylových jevů: „Vymezení jednotlivých stylů vychází z existence specifických funkcí, kterou plní jazykové projevy v určité sféře komunikace – v tzv. stylové sféře (oblasti). K plnění této funkce se v jazyce mj. vytváří i specifická vrstva výrazových (jazykových) prostředků, stylová vrstva, a celková stavba textu se podřizuje více nebo méně pevné stylové normě. Existencí specifických norem nebo celým jejich souborem platným pro určitou sféru se liší od sebe jednotlivé funkční stylové typy."[5]

Základní terminologický problém tedy nastává při vymezení jednotlivých funkčních stylů, jejichž výčet je určen zejména dvěma faktory:

a) výčtem komunikačních funkcí, které zřetelně ovlivňují slohotvorné procesy;
b) stupněm zobecnění.[6]

Je zřejmé, že stupeň zobecnění má negativní korelaci k počtu jednotlivých funkčních stylů, uvnitř publicistického stylu můžeme například rozlišit styl

[4] Čechová, M . a kol.: *Stylistika současné češtiny*. Praha 1997, s. 30.

[5] Tamtéž, s. 31.

[6] Karlík, P. – Nekula, M. – Pleskalová, J. (eds.).: *Encyklopedický slovník češtiny*. Praha 2002, s. 450.

zpravodajský a komentářový. Narážíme tedy jednak na problematické kvantitativní i kvalitativní rozlišení jednotlivých stylů, jednak také na jejich vzájemné prolínání. Jelínek navrhuje, že „jako kritérium pro klasifikaci funkčních stylů může sloužit na prvním místě šíře prostoru, který poskytuje mluvícímu (píšícímu) subjektu příslušná komunikační funkce.“[7] Jako další možná kritéria Jelínek dále uvádí:

- otevřenost nebo zavřenost funkčních stylů pro aktualizované výrazové prostředky,
- možnost užívání expresivních prostředků,
- míra terminologizovanosti výraziva,
- míra kondenzovanosti textů,
- podíl frazémů a idiomů.[8]

Přestože bychom mohli tento výčet dále rozšiřovat a teorii diferenciace funkčních stylů dále prohlubovat, máme za to, že bude vhodnější vrátit se na začátek, tedy formulovat definici termínu „funkční styl“ závaznou pro koncepci této práce. Styl chápeme jako registr určitých výrazových prostředků (stylémů), v případě funkčních stylů se obsah jednotlivých registrů formuje na základě primární komunikační funkce dané množiny projevů. O šíři a počtu jednotlivých množin pojednáme v další kapitole, neboť se jedná o velmi komplikované téma, které bude pro naši práci velmi závažné.

2.1 Diferenciace funkčních stylů

Jak bylo nastíněno v minulé kapitole, je značně nesnadné definovat jednotlivé funkční styly, neboť záleží na několika parametrech (viz výše), které zásadně ovlivní výsledný počet i charakter jednotlivých funkčních stylů. Stanovit tedy nějaké závazné kodifikační rozdělení je velmi svízelné a v podstatě nemožné, nehledě k tomu, že i při zachování jednotných parametrů v průběhu času v závislosti na společenských změnách se nutně musí měnit i povaha funkčních stylů. Stejně na situaci nahlíží i Milan Jelínek: „Vzhledem k tomu, že se v našich i zahraničních stylistikách uvádí

[7] Karlík, P. – Nekula, M. – Pleskalová, J. (eds.).: *Encyklopedický slovník češtiny*. Praha 2002, s. 450.

[8] Tamtéž.

nestejný počet základních funkcí a že kolísá také vymezování jejich vlastností, a také vzhledem k tomu, že s vývojem společnosti přibývají nové funkce, nemůžeme podat definitivní klasifikaci funkčních stylů.“[9]

Proto považujeme za vhodné podívat se nejdříve na vývoj diferenciace funkčních stylů v české lingvistice. Na přelomu 20. a 30. let Pražská škola v čele s Bohuslavem Havránkem stanovila čtyři základní funkční styly (tehdy označované jako funkční jazyky):

a) hovorový (konverzační),
b) pracovní (věcný),
c) vědecký,
d) básnický.

Postupně se rozvinuly do dnes široce akceptovaných funkčních stylů, přičemž pracovní a vědecký splynuly v jeden a přibyl publicistický styl, který byl jako novinářský uváděn v některých pracích i dříve, avšak do konečných přehledů zpravidla zahrnut nebyl. Zformulovaly se tak tyto čtyři základní styly:

a) prostěsdělovací,
b) odborný,
c) publicistický,
d) umělecký.

Je nutno dodat, že dnes se velmi často k těmto stylům začleňují styly administrativní a řečnický. Jelínek navrhuje jiné členění funkčních stylů, a to do třech základních skupin:

1. „Styly věcně informační (jejich funkcí je podávat adresátům informace v různé míře objektivnosti a úplnosti.
2. Styly persvazivní (přesvědčovací, získávací, v nichž funkce informační slouží jen jako východisko k ovlivňování adresáta).
3. Styly umělecké (prozaické, dramatické, poetické, esejistické, v nichž má dominantní postavení funkce estetická.“[10]

V rámci těchto tří obecných stylů pak lze vydělit široké množství specifických stylů. Z výše uvedeného vyplývá kardinální otázka, zda má vůbec význam

[9] Grepl, M. a kol.: *Příruční mluvnice češtiny*. Praha 2003, s. 724.

[10] Karlík, P. – Nekula, M. – Pleskalová, J. (eds.).: *Encyklopedický slovník češtiny*. Praha 2002, s. 450n.

vymezovat nějaké základní funkční styly a v případě, že ano, jakým způsobem? Je vhodnější formulovat čtyři, pět popř. šest základních funkčních stylů nebo jít cestou tří obecných množin a v podstatě ponechat možnost rozmanitého množství konkrétních stylů, tedy jakési dvoustupňově dělení? Odpověď samozřejmě není snadná, nicméně v této práci budeme konzervativní a vydáme se cestou tradičnějšího rozdělení do šesti zmíněných základních funkčních stylů, přičemž jsme si vědomi složité problematiky klasifikace. K tomuto problému se nepochybně vrátíme níže, až se budeme zabývat tím, které texty zahrnout do analýzy jednotlivých funkčních stylů.

Pro demonstraci toho, jak široké jsou možnosti klasifikace funkčních stylů, nám dobře poslouží následující přehledová tabulka.[11]

Komunikační funkce	**Funkční styl**
Estetickoaktualizační • ve vyprávění, řidč. popisu • v dialozích • s důrazem	Styl umělecký • prozaický • dramatický • poetický
Konverzační • v prostředí soukromém • v prostředí veřejném	Styl konverzační (kolokviální) • soukromý • veřejný
Epistolární (korespondenční) • v prostředí soukromém • v prostředí veřejném	Styl epistolární • soukromý • veřejný
Odborná • teoretická • praktická • popularizační	Styl odborný • vědecký • odborněpraktický • odborněpopularizační
Právně-administrativní (úřední)	Styl právně-administrativní (úřední)
Ekonomická (hospodářská)	Styl ekonomický

[11] Grepl, M. a kol.: *Příruční mluvnice češtiny*. Praha 2003, s. 725.

	(hospodářský)
Propagační	Styl propagační (reklamní)
Získávací pro ideologii	Styly ideologické (např. náboženský)
Publicistická (novinářská) • zpravodajská • úvahová • interviewová • přesvědčovací (persvazivní)	Styl publicistický • zpravodajský • úvahový • interviewový • přesvědčovací
Esejistická	Styl esejistický
Direktivní	Styl direktivní
Orientační	Styl orientační

3 BOHATSTVÍ TEXTU

Tato práce sestává ze dvou hlavních komponent: teorie funkčních stylů a bohatství textu, o první problematice jsme pojednali v předchozí kapitole, nyní vymezíme druhou. Jak již bylo zmíněno v úvodu, mezi lingvisty v současné době neexistuje absolutní shoda na definici termínu „bohatství textu“, navíc v českém lingvistickém prostředí se takový termín téměř ani neužívá. Proto považujeme za nezbytné alespoň stručně pojednat o tomto tématu, neboť jasné vymezení východisek našeho výzkumu je pro správné chápání celého textu nezbytné.

Frekvence slov, tedy četnost výskytu konkrétního slova či jeho tvaru, je vůbec nejstarším oborem kvantitativní lingvistiky, neboť prvky takového bádání můžeme hledat již v 17. století např. v díle Jana Amose Komenského. Tento zájem o nestejnou distribuci slov v textu však nepramenil ani tak z teoretického zájmu o jazyk, motivaci nalezneme spíše v praktickém uplatnění takového zkoumání. Frekvenci slov využívala a stále využívá zejména lexikografie při hledání jádra slovní zásoby jednotlivých jazyků, tzn. určit jejich centrum a periferii. V rámci jazykovědy je bohatství textu také důležitým parametrem v oblasti určování autorství textu. Postupně studium četnosti slov nacházelo uplatnění i v dalších oblastech jako např. v informačních technologiích, stenografii, pedagogice nebo psychologii. Do teoretického zkoumání jazyka, zejména v lexikálním plánu, vstoupil zájem o matematické a statistické zkoumání slovníku teprve až s průkopnickou prací Georgea Kingsleyho Zipfa v pol. 20. stol. Tento americký lingvista, který jako první soustavně sledoval vztahy mezi frekvencí slov a jejich rankem, stanovil několik lexikostatistických vztahů, jež jsou známy jako Zipfovy zákony.

Slovenský kvantitativní lingvista Gejza Wimmer také datuje počátky bádání v oblasti bohatství textu do pol. 20. století, ale připisuje je práci britského statistika George Yulea, kterého však zajímala spíše charakteristika textů než rozdělení slov v textu. Wimmer tvrdí, že bohatství textu je „rozhodne nejaká funkcia počtu odlišných slov, ale zachytiť túto funkciu pre charakterizačné, porovnávacie a teoretické účely je pomerne zložité,“[12] přičemž dodává, že přesná definice

[12] Wimmer, G.: *Úvod do analýzy textov*. Bratislava 2003, s. 124.

je komplikovaná jak v matematice, tak v jazykovědě. Wimmer vysvětluje tento pojem na analogické problematice v biologii, kde se „skúma napríklad vzťah počtu odlišných rastlín ku všetkým rastlinám na vymedzenej ploche, alebo v medicíne ako počet odlišných ochorení ku všetkým ochoreniam v meste a pod.“[13] Wimmer dále upozorňuje na problematický fakt, že stejně jako bohatství rostlin je závislé na velikosti zkoumané plochy, bohatství slovníku musí být nutně závislé na velikosti textu, neboť platí, že s narůstající délkou textu přibývá stále méně slov.

Bohatství textu v podstatě odpovídá konceptu horizontálního rozložení slov slovenského lingvisty Jozefa Mistríka, který tvrdí, že při horizontálním rozložení je „dôležitá miera striedania slov, miera opakovania slov, rozsahu slovníka v texte, výskyt okazionálnych a tematických slov, koncentrovanosť lexiky atď. Ide o vlastnosti, ktoré charakterizujú zloženie a povahu textu.“[14] Zároveň Mistrík dodává, že i údaj o horizontálním rozložení slov je jen jedním indikátorem vlastností textu, který nemůže podat komplexní povahu konkrétního komunikátu. Na druhou stranu Mistrík chápe slovník jako samostatnou entitu, která tvoří vlastní systém, a to jak na úrovni parole, tak i langue.

Obecně můžeme říct, že bohatství textu představuje základní charakteristiku textu v rámci lexikální statistiky, popř. kvantitativní lingvistiky vůbec, jež je v českém jazykovědném prostředí označována také jako bohatství slovníku nebo rozsah lexika. Primárně je bohatství textu definováno jako vztah mezi slovníkem (*V*), jímž je myšlen počet lexikálních jednotek (lemmat), a délkou textu (*N*), která představuje celkový počet slov, přičemž za slovo je považováno písmeno či řada písmen v textu mezi mezerami, někdy označované jako slovoforma nebo textové slovo. Největšího bohatství slovníku tak dosáhneme, pokud se každé slovo v textu bude vyskytovat právě jednou. Je zřejmé, že tato situace je extrémní a můžeme se s ní setkat jen u velmi krátkých textů, a to zejména kvůli častému opakování gramatických slov. Existuje mnoho způsobů, jak vypočítat bohatství textu, základní rovnice vyjadřující tento vztah je následující:

$$R = \frac{V}{N}$$

[13] Wimmer, G.: *Úvod do analýzy textov*. Bratislava 2003, s. 124.

[14] Mistrík, J.: *Frekvencia slov v slovenčině*. Bratislava 1969, s. 65.

Vzhledem k tomu, že mezi délkou textu a počtem různých slov není vztah přímé úměry, nebylo na základě uvedené rovnice možné podrobit vzájemné komparaci texty odlišné délky. Z tohoto důvodu byla tato rovnice mnohokrát modifikována, jednu z prvních úprav provedl francouzský lingvista Pierre Guiraud, jenž se touto problematikou dlouhodobě zabýval. Pro výpočet stanovil tyto dvě formule:

$$R = \frac{V}{\sqrt{N}}$$

$$R = \frac{V}{\sqrt{2N}}$$

V druhém případě Guiraud počítá pouze s plnovýznamovými slovy, neboť pro francouzštinu předpokládá, že tato slova představují polovinu textu. Způsob, jakým vyjadřujeme bohatství textu, se úzce váže na konkrétní jazyk, který podrobujeme výpočtu. V českém lingvistickém prostředí se dané problematice systematicky věnovala Marie Těšitelová, která zjistila, že poměr mezi plnovýznamovými a formálními slovy je asi 80:20, přičemž za plnovýznamová slova autorka považuje substantiva, adjektiva, pronomina, numeralia, verba a adverbia. Těšitelová proto modifikovala rovnici pro výpočet bohatství slovníku pro češtinu takto:

$$R = 100 \ \frac{V}{\frac{8}{10} N}$$

Bohatství textu je však mnohem širší pojem, než by mohla vyjádřit jediná rovnice, pro relevantní analýzu textu je třeba aplikace více výpočtů zachycujících vztahy uvnitř lexika jednotlivých textů. Teprve kombinace více výpočtů nám může poskytnout širší představu o lexiku konkrétního textu. K bohatství slovníku (rozsahu) postupně přibývaly další parametry a výraz bohatství textu se stal jejich zastřešujícím pojmem, přičemž rozsah lexika je zcela dominantní a mnozí jej považují za synonymní výraz k bohatství textu. Těšitelová na základě svých šetření zjistila, že k „náležitému posouzení dané problematiky – z hlediska kvantitativního – nestačí jednotlivé charakteristiky samy o sobě, nýbrž že je třeba tuto vlastnost slovníku hodnotit komplexně, a to alespoň třemi charakteristikami, které respektují:

a) globální charakteristiku opakování slov v textu, která se projevuje v rozsahu slovníku, porovnáme-li délku textu *N* a jeho slovník *V*;
b) sílu pásma slov s frekvencí 1–10, tzv. rozptýlení slovníku, které má významné postavení v slovníku kteréhokoli textu a ukazuje nejen specifiku jazyka stylu funkčního, ale i individuálního;
c) koncentraci slovníku, která ukazuje, jaký podíl slovníku textu, popř. autora připadá na slova nejfrekventovanější, tj. na prvních 10 nejfrekventovanějších slov.“[15]

Rozptýlení (*D*) je veličina, jež ukazuje, jaký podíl v textu připadá na pásmo slov s nejnižší frekvencí (V_1), tj. slova s četností 1–10. Tato veličina tak úzce souvisí s bohatstvím slovníku, neboť lze předpokládat, že právě nejméně užívaná slova stojící na periferii dosahují nejvyšší entropie, a tudíž by rozptýlení mělo pozitivně korelovat s rozsahem lexika. Hypotéza je tedy taková, že čím má text bohatší slovník, tím větší zastoupení mají jednotky z periferie slovní zásoby. Formule pro výpočet pak vypadá následovně:

$$D = 100\,\frac{V_1}{V}$$

Koncentrace (*K*) se naopak soustředí na slova s nejvyšší četností a vyjadřuje poměr délky textu odpovídající prvním deseti nejfrekventovanějším slovům. Tato veličina završuje ústřední trojici, jež podává základní charakteristiku textu z hlediska lexikální statistiky, a nutně tedy souvisí s dvěma předešlými. Zatímco rozsah lexika vystihuje nejkomplexnější údaj o slovníku a rozptýlení se pak soustředí na periferní oblast lexika, koncentrace se zaměřuje na opačný pól slovní zásoby – její samotné centrum. Tento parametr vystihuje vztah mezi délkou textu odpovídající deseti nejfrekventovanějším jednotkám a celkovou délkou textu. Přestože tedy nejde o přímý protiklad koncentrace lexika, je logické mezi těmito veličinami očekávat určitou míru negativní korelace. Vzorec pro výpočet koncentrace vypadá takto:

[15] Těšitelová, M.: *Kvantitativní lingvistika*. Praha 1987, s. 71.

$$K = 100\ \frac{N_1}{N}$$

Těšitelová přitom předpokládá, že mezi těmito veličinami existuje závislostní vztah negativní korelace, tedy čím větší je rozptýlení nějakého textu, tím nižší je jeho koncentrace. Pro celkové hodnocení bohatství textu Těšitelová klade největší důraz na rozptýlení lexika. Je třeba si však uvědomit, že ve vzájemném vztahu není jen rozptýlení a koncentrace, ale také samotný rozsah. Právě rozsah a rozptýlení lexika při srovnávání výsledných hodnot různých textů zpravidla vykazují téměř stejné výsledky, což je dáno vesměs totožnými činiteli, mezi nejvýraznější patří:

a) forma komunikátu (psanost-mluvenost),
b) funkční styl,
c) téma textu,
d) připravenost-nepřipravenost,
e) charakter podavatele,
f) oficiálnost-soukromost,
g) explicitnost a určitost,
h) míra abstrakce,
i) organizovanost.

Bohatství textu se dlouhodobě věnoval také slovenský jazykovědec Jozef Mistrík, který definoval vlastní čtyři charakteristiky umožňující základní deskripci lexika konkrétního textu. První veličinou je index opakování slov (*Ii*), který vystihuje základní vztah mezi délkou textu (*N*) a slovníkem (*L*). Index opakování je definován touto rovnicí:

$$Ii = \frac{N}{L}$$

Jakkoliv je index opakování výmluvný, naráží na základní metodologický problém, a to fakt, že s délkou textu narůstá i míra opakování slov. Je zřejmé, že i při analýze textů stejného autora získáme v závislosti na délce konkrétních textů vždy jinou hodnotu opakování slov. Tento vztah však nemá podobu přímé úměry, ale spíše logaritmické křivky. Kromě délky textu je míra opakování slov závislá také

na dalších činitelích, a to zejména stylové povahy popsané výše. Jozef Mistrík zjistil, že nejmenšího indexu opakování dosahují žurnalistické texty a básně, středních hodnot dosahují umělecké texty, největší mírou opakování se pak vykazují texty mluvené, vědecké a odborné.

Druhou charakteristikou bohatství textu u Mistríka je zatíženost slovníku (I_g), kde je pozornost soustředěna na slova, jichž autor textu užil tzv. suverénně, tj. kterých bylo užito více než jednou ($L_{f1<}$). Mistrík definuje index zatíženosti tak, že je třeba si jej představit jako „stupeň frekvenčnej aktivity tých slov, s ktorými sa v texte pracuje ako so základnými. Pomocou týchto opakujúcich sa slov objasňuje sa hlbšie problematika, ktorá je predmetom prejavu.“[16] Vzhledem k tomu, že Mistrík počítá s tím, že formální slova představují asi polovinu textu, násobí tato slova 2. Základní rovnice pro výpočet zatíženosti slovníku je tato:

$$I_g = \frac{2\,L_{f1<}}{\frac{N}{10}} = \frac{20\,L_{f1<}}{N}$$

Dalším Mistríkovým parametrem je index exkluzivnosti textu (I_e), který představuje protiváhu indexu zatíženosti. Pozornost je totiž zaměřena na slova, která se neopakují, tedy slova s frekvencí 1 (L_{f1}). Mistrík těmto okazionálním slovům přikládá značný vliv na charakter projevu. „Sú prejavom najmä autorovho vkusu a schopnosti prejav ‘štylizovať” a ozvláštniť ho pomocou exkvizitných výrazov.“[17] Mistrík však index exkluzivnosti nevnímá jako protiváhu indexu zatíženosti, ale jako jeho doplnění. Výpočet se uskutečňuje podle následujícího vzorce:

$$I_e = \frac{2\,L_{f1}}{\frac{N}{10}} = \frac{20\,L_{f1}}{N}$$

Čtveřici charakteristik podávajících obraz o lexiku textu Mistrík uzavírá veličinou zvanou variabilnost textu (V), která vychází ze dvou předešlých parametrů, vyjadřuje totiž poměr počtu slov s frekvencí 1 k počtu slov, která se opakují. Mistrík variabilonost textu odvozuje takto: „Ak striedanie slov s f_1 so slovami, ktoré

[16] Mistrík, J.: *Frekvencia slov v slovenčině*. Bratislava 1969, s. 74.

[17] Tamtéž, s. 76.

sa opakujú, spôsobuje lexikálnu variabilnosť prejavu, potom pomerom medzi L a L_{f1} možno v percentách vyjadriť i lexikálnu variabilnosť textu.“[18] Rovnice pro výpočet vypadá takto:

$$V = \frac{2\,L_{f1}\,100}{L}$$

Výše jsme uvedli dva základní náhledy na možnost vyjádření bohatství textu, které jsou v českém, popř. slovenském jazykovědném prostředí již tradiční. Z podstaty věci je však zřejmé, že bohatství textu představuje v rámci kvantitativní lingvistiky vlastní disciplínu, jež zahrnuje množství různých pojetí, jak zkoumat rozložení slov v textu. Do bohatství textu tak můžeme zahrnout jak míru opakování slov, tak například entropii nebo nejrůznější type-token indexy. Výsledné hodnoty mohou být jak numerické povahy, tak grafické, a to zejména pomocí bodových grafů.

[18] Mistrík, J.: *Frekvencia slov v slovenčině*. Bratislava 1969, s. 77.

4 METODOLOGIE LEXIKÁLNÍ ANALÝZY

4.1 Kvantitativní lingvistika

Lingvistika patří mezi společenské vědy, neboť předmětem jejího zkoumání je jazyk, u něhož platí axiom, že jde o společenský jev. Současně také platí, že jazyk je velmi složitým útvarem, jehož zákonitosti nejsme dosud s to plně popsat. Mnoho obecně platných hypotéz, které většina lingvistické obce přijímá jako fakt a mají tak mnohdy axiomatický charakter, jsou stále jen hypotézami.

Samozřejmě lze namítnout, že odlišné metody zkoumání přirozeně vycházejí z odlišného charakteru zkoumaného předmětu. Mnozí také tvrdí, že jazyk je něco tak složitého, že ani matematicky uchytit nelze. Taková tvrzení shledáváme pravdivými, ale… Než začneme s vysvětlováním všech „ale“, uvedeme náhled Luďka Hřebíčka na situaci v jazykovědě: „Způsob uvažování o jazyce, způsob jeho zkoumání od počátku po dnešek má charakter zkoumání předvědeckého. […] Různé obory, přírodní i humanitní, dospěly ve druhé polovině 20. století k jakési obecné představě, co je to věda. Nauka o jazyce se nemůže tvářit, jako by se jí to netýkalo.“[19]

Hřebíčkova slova musíme samozřejmě brát s nadsázkou, ani my nehodláme v žádném případě kritizovat současnou jazykovědu. Jde nám spíše o snahu implementovat matematické metody do lingvistiky jako jednu z elementárních lingvisty respektovaných metod. Přestože zájem o tento přístup roste, stále je okrajovou záležitostí, jíž se zabývá úzká skupina badatelů.

Pokud máme stručně vysvětlit základní přínos matematického přístupu ke zkoumání jazyka, vyjdeme z jeho odlišného charakteru. Jak již bylo zmíněno výše, mnohé lingvistické teze vycházejí spíše z intuice bez relevantních důkazů, jejichž doklad je často považován za nemožný. Právě zde spatřujeme jednu z klíčových rolí matematických metod v lingvistice, které jsou ideálním způsobem, jak ověřit vyřčené hypotézy. Mnohé teze jsou obecně přijímány jako pravdivé, a i my s nimi můžeme souhlasit, z našeho hlediska však považujme vědecky za nesprávné alespoň se nepokusit tyto hypotézy podložit konkrétními daty. Mnohé práce z oblasti

[19] Hřebíček, L.: Vprávění o lingvistických experimentech s textem. Praha 2002, s. 12n.

kvantitativní lingvistiky ani nemají za cíl objevit zcela nové poznatky o jazyce, ale pokusit se nějakým způsobem kvantifikovat již známé. Příkladem může být naše práce zkoumající rozdíly psané a mluvené češtiny, kde nám šlo primárně o kvantifikaci analyzovaných jevů. Například víme, že v mluveném projevu se užívá mnohem více kontaktových prostředků než v psaném; nezabývali jsme se však pravdivostí tohoto jevu, kladli jsme si otázky o kolik je jich skutečně více než v psaných textech, kolik místa opravdu kontaktové prostředky zaujímají v textu, jaká je jejich četnost, které jsou nejužívanější apod.[20] Na takové otázky právě kvantitativní výzkum může odpovídat konkrétními jednoznačnými daty, nikoli domněnkami.

4.2 Stylistika a kvantitativní lingvistika

Nyní navážeme na předchozí kapitolu a nastíníme, jaké možnosti nabízí kvantitativní metody při výzkumu v oblasti stylistiky. Ze všech jazykovědných disciplín se právě stylistika může jevit jako nejvzdálenější exaktním matematickým přístupům. Abychom neprezentovali jen naše názory či názory kvantitativních lingvistů, uvedeme nejdříve pohled stylistky Jany Hoffmanové: „Otvírá se tu cesta ke zpracování velkých souborů dat a k získání explicitních údajů, k reprezentativnímu porovnávání jednotlivých souborů, ke zjišťování frekvence určitého jevu i jeho distribuce v textu, k výpočtu průměrných hodnot a měření odchylek a k pořizování výstižných diagramů, ke grafickému znázorňování všech těchto výsledků. Tento vývoj nabízí i stylistice možnost rozložit styl na limitované množství exaktně měřitelných charakteristik a na základě těchto měření formulovat přesný, explicitní popis autorského stylu (ale i jednotlivého díla, skupiny autorů nebo děl atd.) místo obvyklých značně subjektivních a intuitivních soudů a impresionisticky zabarvených vágních formulací.“[21]

[20] Kubát M.: *Rozsah, rozptýlení a koncentrace lexika v psané a mluvené češtině* (diplomová práce). Olomouc 2010, s. 45n.

[21] Hoffmannová J.: *Stylistika a… Současná situace stylistiky*. Praha 1997, s. 64.

Hoffmanová tak v podstatě vystihuje podstatu přínosu kvantitativního zkoumání jazyka tak, jak bylo pojednáno v předchozí kapitole, kde jsme mluvili zejména o exaktní kvantifikaci určitých jevů. Kvantitativní lingvistika nabízí rozmanité množství konkrétních charakteristik, které mohou být uplatněny ve vztahu ke stylové analýze, Hoffmanová uvádí sedm základních:

- délka slova;
- délka věty;
- frekvence slov, slovních druhů a tvarů;
- frekvence gramatických kategorií,
- frekvence slov v textu v poměru k celkovému bohatství, variabilitě slovní zásoby autora či textu;
- skladba souvětí: frekvence určitého typu souvětí, typu vedlejší věty apod.;
- frekvence a distribuce jednotlivých typů konektorů a dalších prostředků mezivětného navazování.[22]

Marie Těšitelová formuluje předmět stylistické statistiky takto: „Chápeme-li jazykový styl jako charakteristický způsob organizace jazykového projevu, který vychází od autora (mluvčího nebo pisatele) a směřuje k adresátovi (posluchači nebo čtenáři) a zakládá se na výběru a využití jazykových prostředků, potom aplikace kvantitativních metod, zejména statistických, má za úkol zjišťovat příslušné jazykové charakteristiky.“[23] Autorka dále tvrdí, že pomocí těchto charakteristik, které označuje jako objektivní, je možné odhalovat jazykový styl, a to jak individuální, tak funkční.

Jak jsme již naznačili v úvodu kapitoly, stylistika má jistá specifika, která značně komplikují užití kvantitativních metod. Podstatné však je, že dle našeho názoru určité aspekty stylistiky statistickou analýzu sice znesnadňují, avšak neznemožňují její aplikaci vůbec. Pokud máme zmínit základní výtky, jež jsou adresovány těmto metodám, můžeme vyjít opět z Hoffmannové, která se také touto problematikou zabývala. K základním pochybnostem se vyjadřuje takto: „Týkaly se hlavně toho, aby se takto získané výsledky nepovažovaly za samoúčelné a aby se autoři těchto výzkumů neomezovali pouze na jejich efektní prezentaci, tj. aby byla dostatečně

[22] Hoffmannová J.: *Stylistika a… Současná situace stylistiky*. Praha 1997, s. 65n.

[23] Těšitelová, M.: *Kvantitativní lingvistika*. Praha 1987, s. 126.

zastoupena erudovaná a všestranná interpretace dat.“[24] S těmito námitkami bezezbytku souhlasíme, jsme si plně vědomi toho, že samotné výsledné hodnoty jakkoliv efektně zpracované do tabulek, grafů apod. nejsou ničím jiným než pouhými čísly, jež samy o sobě mají nulovou vypovídací hodnotu. Již jsme výše uvedli, že taková data jsou teprve základním materiálem pro vůbec nejdůležitější část každého statistického šetření, a to relevantní kvalifikovanou interpretaci.

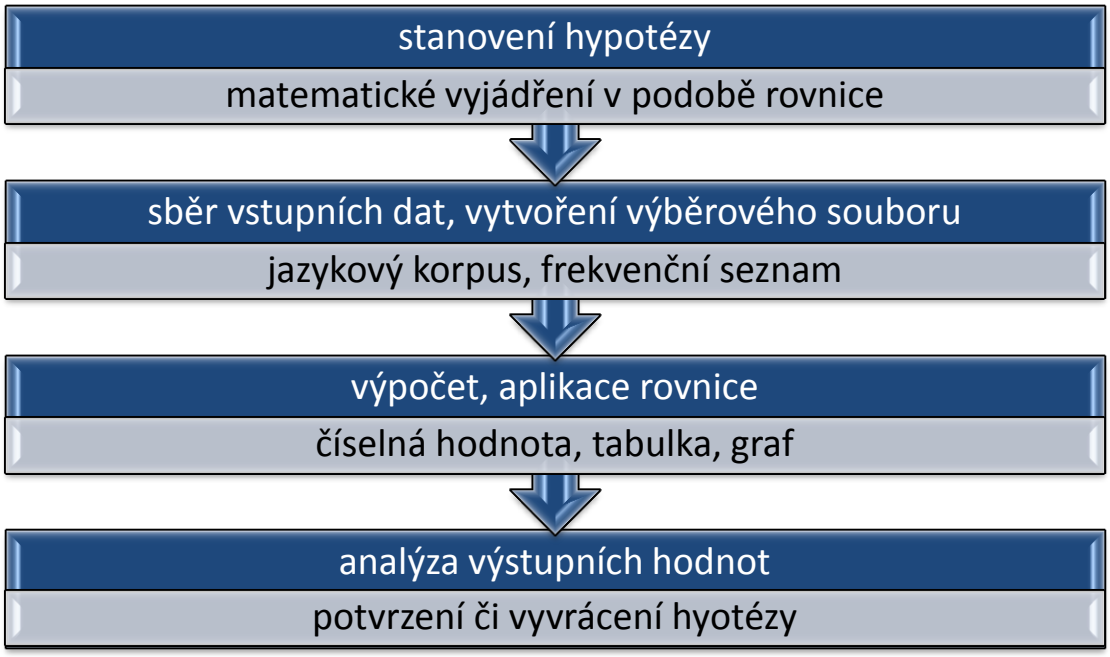
4.3 Základní prvky statistické analýzy

Jakákoliv statistická analýza, lexikální nevyjímaje, má jednotnou základní strukturu. Proto nejdříve v kapitole věnované metodologii nastíníme pro přehlednost onu základní strukturu. Zcela na začátku každého statistického zkoumání stojí určitá hypotéza, jejíž pravdivost chceme ověřit. Na základě toho je třeba vytvořit reprezentativní vzorek, na němž bude prováděn rozbor. Tato část je elementárním aspektem úspěšného výzkumu, důležité je zejména stanovení dostatečného rozsahu a jasné vymezení jednotky. Dalším bodem je aplikace určité rovnice na vzorek, jejíž výpočet poskytne výsledné hodnoty. Následuje interpretace získaných dat, jež je samotným cílem celé práce. Například vydávané frekvenční slovníky nemají samy o sobě žádnou hodnotu, jsou pouhým (ale nezastupitelným) podkladem k samotnému výzkumu. Pro přehlednost uvádíme jednoduché čtyřstupňové schéma statistické analýzy:

Lexikální statistická analýza se skládá z vytvoření souboru, kterým rozumíme množinu jednotek (textových slov), které převedeme do určité podoby (frekvenčního seznamu). Dále je proveden výpočet na základě určité rovnice, získaná data mohou

[24] Hoffmannová J.: *Stylistika a… Současná situace stylistiky*. Praha 1997, s. 64.

mít formu číselné hodnoty, tabulky či grafu. Nakonec jsou získané hodnoty interpretovány. Jednotlivé části budou dále rozvedeny do konkrétní podoby našeho výzkumu, nyní pro přehlednost podáváme schéma, jehož složky budou korespondovat s následujícími kapitolami.



4.4 Vytvoření výběrového souboru

Základním prvkem každé analýzy je shromáždění vhodného materiálu, který bude zkoumán. V případě statistiky je zásadní vytvořit takový vzorek, kde nejdůležitější je jeho maximální míra podobnosti se skutečností. Statistika nám umožňuje zkoumat jevy, které nejsme s to uchopit vcelku, kde není možné zahrnout do analýzy úplnou množinu sledovaného materiálu. Proto využíváme model takové množiny, která je rozsahem mnohem menší, ale ve všech ostatních parametrech se mu co nejvíce podobá. Právě správné vytvoření výchozího materiálu je nezbytným předpokladem pro další kroky statistického rozboru a především pro relevantní výstupní hodnoty.

Využití statistických metod v oblasti stylistiky se věnoval Jozef Mistrík, k výběru zkoumaného vzorku uvádí: „Matematická štatistika na základe zákonitostí existujúcich v reprezentatívnom výbere usudzuje s využitím analógie na neznáme zákonitosti v základnom súbore prvkov. Dôležitý je preto výber, ktorý sa využije ako

východisko kvalifikácie jazykových javov a na základe toho na čo najpravdivejší opis základného súboru, ktorý býva spravidla fiktívny alebo nekonečný.“[25]

Ideálním stavem tedy je obsáhnout 100 % zkoumané oblasti, kdy už samozřejmě nejde o model, toho můžeme dosáhnout například analýzou konkrétního románu. Vzhledem k tomu, že naším záměrem je výzkum funkční stylů a de facto celého jazyka, tedy všech realizovaných projevů současné češtiny, budeme zcela odkázáni na model, jehož rozsah bude naprosto klíčový.

Zatímco v *RRKLPMČ*[26] zkoumající současnou psanou a mluvenou češtinu jsme sestavovali vlastní korpusy, nyní použijeme subkorpusy Českého národního korpusu (ČNK). Při vytváření vlastních korpusů jsme totiž narazili na základní problém – kvantitativní omezení. Je zřejmé, že vytváření zcela nových korpusů je značně pracné a zejména časově náročné, proto jsme se v této práci rozhodli vydat cestou využití již existujících korpusů. Zatímco při sestavovaní vlastního výchozího materiálu jsme se rozsahem pohybovali maximálně v tisících jednotek, při využití ČNK můžeme pracovat teoreticky i s miliony. Míra reprezentativnosti analyzovaných textů tak získá zcela jiný rozměr, a tím samozřejmě i relevantnost výstupních hodnot.

4.5 Český národní korpus jako zdroj dat

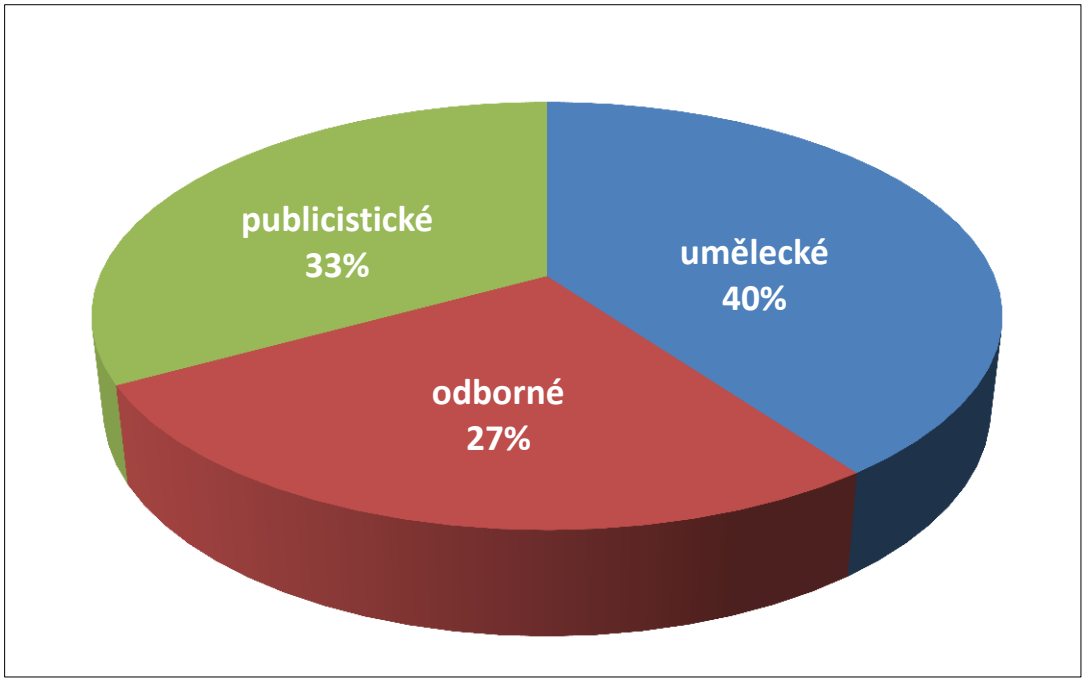
Vzhledem k náročnosti vytváření vlastní korpusů, zejména z důvodu lemmatizace, jsme se rozhodli využít již existující zdroj dat, a to největší dostupnou databázi jazykového materiálu – Český národní korpus. Tento zdroj je svým rozsahem v současné době bezkonkurenční a i z kvalitativního hlediska nabízí široké možnosti.

4.6 SYN2010

[25] Mistrík, J.: *Štylistika*. Bratislava 1989, s. 37.

[26] Kubát, M.: *Rozsah, rozptýlení a koncentrace jazyka v psané a mluvené češtině* (diplomová práce). Olomouc 2010.

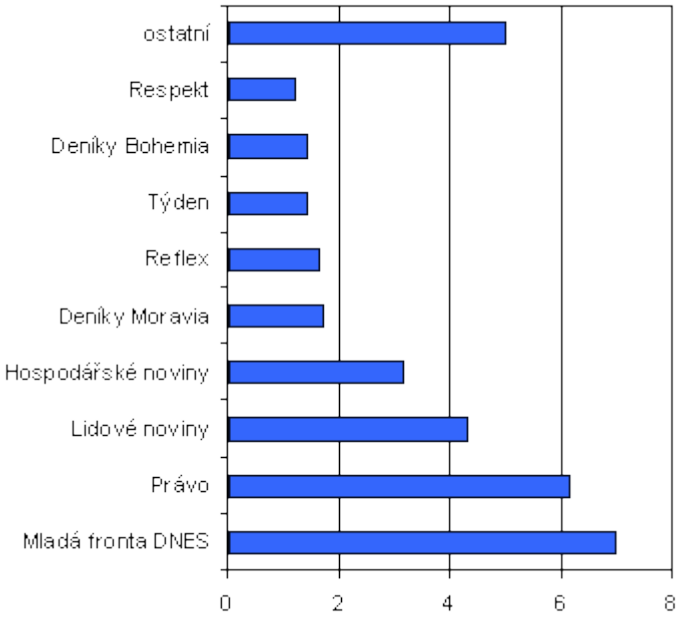
Z mnoha dostupných korpusů ČNK jsme pro náš výzkum zvolili SYN2010, neboť se jedná o synchronní korpus českého jazyka, který zachycuje široké rozpětí textů (beletrie, odborné texty, publicistika), jeho cílem tak je zachytit téměř celou současnou psanou češtinu. Tento korpus navazuje na řadu synchronních korpusů SYN, jejími předchůdci jsou SYN2000 a SYN2005. SYN2010 jsme vybrali samozřejmě zejména pro jeho aktuálnost, publicistické texty byly excerpovány v letech 2005–2009 při konstantním množství. Pro beletrii a odborné texty pak platí, že musí být napsány po roce 1989, obecně však platí, že oproti předchozím korpusům řady SYN starších textů ubývá ve prospěch novějších. Celkový objem jednotek dosahuje v korpusu SYN2010 100 milionů, přičemž za jednotku je považováno textové slovo. Základní složení korpusu se člení do tří základních oblastí: umělecké, odborné a publicistické texty, jejich poměrné zastoupení demonstrujeme na následujícím grafu.[27]



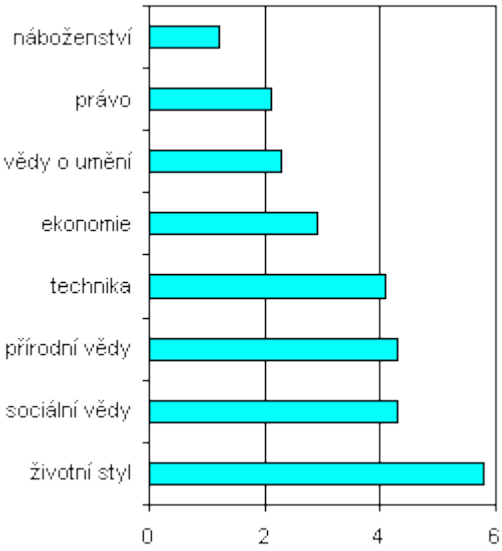
V následující tabulce uvádíme zastoupení jednotlivých periodik, kde čísla na vodorovné ose udávají počet textových slov v milionech.[28]

[27] http://ucnk.ff.cuni.cz/co_je_korpus.php [cit. 8. 1. 2012]

[28] Tamtéž.



Obdobný graf uvádíme i u odborných textů, kde je vyjádřeno zastoupení různých tematických oblastí, na vodorovné ose hodnoty opět uvádí počet excerpovaných textových slov v milionech.[29]



Korpus SYN2010 jsme však nevybrali jen kvůli jeho aktuálnosti, i když tento aspekt byl rozhodně nejdůležitější. Přihlédli jsme také k tomu, že při jeho tvorbě autoři použili pokročilejší lemmatizaci, jež je pro lexikální statistiku zásadní.

[29] http://ucnk.ff.cuni.cz/co_je_korpus.php [cit. 8. 1. 2012]

4.7 PMK

Vzhledem k tomu, že korpus SYN2010 zahrnuje jen psané texty, potřebovali jsme pro úplnost ještě korpus obsahující texty mluvené. Zatímco korpusů psaného jazyka je mnoho, u mluvených projevů je tomu naopak. Je zřejmé, že tento stav způsobují především dva důvody. Prvním je skutečnost, že lingvisté tradičně tendují k výzkumu spíše psaných textů než mluvených, druhým důvodem pak jsou technická omezení. Je samozřejmě otázkou, do jaké míry druhý faktor determinuje první. Jakkoli by bylo nepochybně velmi hodnotné tuto otázku zodpovědět, zaměření a rozsah této práce to neumožňuje.

Jako jediný dostupný mluvený korpus vhodný pro naše účely jsme zvolili Pražský mluvený korpus, který je prvním rozsáhlým mluveným lemmatizovaným korpusem mluvené neformální češtiny. PMK obsahuje celkem 780 322 pozic (konkordančních řádků), po odstranění interpunkce zbyde 644 909 slovoforem. Vzhledem ke způsobu započítávání jednotek, je jejich výsledný počet ještě nižší. František Čermák problematiku vysvětlují takto: „Protože řada slovních tvarů je zároveň komponenty víceslovných jednotek, je počet slovních tvarů nevázaných nižší a odpovídá číslu 548 091 (v tomto čísle je tedy každá víceslovná jednotka započítána jen jednou, a to vždy skrze její první komponent, zatímco ostatní komponenty zde zahrnuty nejsou)“[30] Pro úplnost můžeme ještě uvést, že PMK se skládá z 296 souborů. Celkový počet jednotlivých promluv pak autoři udávají 15 015, přičemž promluvu definuje Čermák jako: „Souvislý projev jednoho mluvčího (nepřerušovaný jiným mluvčím); v dialogu (označený jako N na posledním místě čtyřmístného sociolingvistického kódu) je takto pojatých promluv vztahujících se k jednomu mluvčímu pochopitelně vždy víc.“[31]

Protože při tvorbě výběrového souboru nejsou důležité jen kvantitativní parametry, ale také kvalitativní, zmíníme některé faktory, jež byli při vytváření PMK zohledněny. Ze sociologického hlediska autoři PMK stanovili několik binárních opozic, u nichž se snažili o vyváženost. Jedná se o:

[30] Čermák, F. a kol.: *Frekvenční slovník mluvené češtiny* (CD-ROM příloha). Praha 2007.

[31] Tamtéž.

a) pohlaví;
b) věk (do výzkumu byli zařazeni jen mluvčí starší 20 let, specifický jazyk dospívajících tak do korpusu zařazen nebyl, hranice binární opozice mladší-starší je stanovena na 35 let);
c) vzdělání (za nižší vzdělání je považována ZŠ a SŠ, za vyšší pak VŠ);
d) formálnost (o této kategorii pude podrobněji pojednáno níže).

Samotný název korpusu dává tušit ledasco o teritoriálním vymezení pořízení nahrávek, jde tedy o Prahu a její okolí. Čermák vysvětluje specifikum takového omezení na jednu lokalitu: „Vzhledem k centrálnímu a jedinečnému postavení Prahy tu jazykově dochází k velkému míšení lidí ze všech oblastí země a obraz jejího jazyka má tudíž do vysoké míry celonárodní povahu.“[32] Přestože je Praha nepochybně centrem, do kterého se stěhují lidé z celé republiky, zvláště my Moravané jen těžko můžeme přistoupit na zmíněné zobecnění, že jazyk obyvatel Prahy má do vysoké míry celonárodní povahu. Souhlasit s takovým tvrzením můžeme pouze za předpokladu, že atribut „celonárodní“ se vztahuje k národu českému, nikoli moravskému či slezskému. Na tomto poli se nabízí prostor pro samostatný výzkum, jež by zkoumal právě rozdíly mezi jednotlivými nářečími na základě lexikostatistických metod. Co se týká časového hlediska, všechny nahrávky byly pořízeny mezi lety 1988–1995. Jedná se tedy o poněkud starší texty, otázkou je jak a zda vůbec se nějak stáří textu projevuje v lexikostatistických parametrech.

[32] Čermák, F. a kol.: *Frekvenční slovník mluvené češtiny* (CD-ROM příloha). Praha 2007.

5 ROZSAH, ROZPTÝLENÍ A KONCENTRACE LEXIKA

Jak již bylo zmíněno, páteří této práce bude lexikální analýza textů, kde dominantními veličinami budou rozsah, rozptýlení a koncentrace lexika. Navazujeme tak na *RRKLPMČ*, kde tomu bylo stejně. Rozdíly jsou především dva, jednak nyní využíváme již existující korpusy, jednak naším předmětem zájmu není ani tak diference mezi psanou a mluvenou češtinou, ale diference jednotlivých funkčních stylů. Záměrně začínáme těmito veličinami, které jsou elementárními složkami kvantitativní lingvistiky v oblasti výzkumu slovníku. Právě zde očekáváme nejviditelnější rozdíly mezi jednotlivými funkčními styly, přičemž v této fázi ještě nemáme zcela jasno, jakým směrem se bude tato práce ubírat po této stěžejní kapitole věnované uvedených třem základním lexikostatickým charakteristikám. Vycházíme tak opět ze zkušeností z práce *RRKLPMČ*, kde teprve při analýze a interpretaci dat určených k výpočtům zmíněných veličin, se nám otevírala široká škála poznatků, jež jsme později uplatnili při interpretaci různých jevů v lexikální rovině, jež měly konkrétnější ráz. Přestože tyto rozbory tvořily spíše jakousi nadstavbu naší práce, významem patrně předčily primární komparaci sledovaných parametrů rozsahu, rozptýlení a koncentrace lexika v psané a mluvené češtině.

Klíčovým jevem, který je pojítkem a zároveň determinujícím faktorem zmíněných sledovaných parametrů, je opakování slov. Právě opakování jazykových jednotek je důležitým signifikantním rysem při diferenciaci funkčních stylů. Jozef Mistrík popisuje místo principu opakování ve stylistice takto: „Človek je citlivý na javy, ktoré sa opakujú. Uvedomuje si opakovanie prvkov v reči. Spôsobom opakovania sa rečový prejav formuje, štylizuje. Niekedy je charakter textu daný výlučne disperziou a opakovaním výrazových prostriedkov.“[33] Právě vzhledem důležitosti principu opakování výrazů v textech shledáváme rozsah, rozptýlení a koncentraci lexika jako ideální výchozí veličiny, jež jsou schopny podat relevantní výstupní hodnoty pro komparaci funkčních stylů.

[33] Mistrík, J.: *Štylistika*. Bratislava 1989, s. 145.

Důležitou veličinou je tedy samotná velikost slovní zásoby, neboť je zřejmé, že s její vzrůstající tendencí se snižuje míra opakování slov. Mluvčí s menší slovní zásobou musí nutně více jednotlivé výrazy opakovat. Přestože velikost slovníku mluvčího/pisatele je základním faktorem určujícím právě míru opakování slov, důležitou roli zde hrají i další činitelé. Těšitelová uvádí tři následující:

a) druh projevu (psaný vs. mluvený),
b) téma,
c) funkční styl.[34]

5.1 Metodologie přiřazení subkorpusů k funkčním stylům

Základem naší analýzy tedy bude kvantitativně vyjádřit sledované parametry a následně získaná data interpretovat a zasadit do kontextu teorie funkčních stylů. Výše jsme již poměrně zevrubně pojednali o našem výchozím zdroji jazykových dat – ČNK, přesněji pak SYN2010 a PMK. Tento výchozí materiál považujeme pro naše účely za nejlepší dostupný zdroj dat, neznamená to však, že by byl bezchybný. Elementárním problémem se ukázalo přiřazení subkorpusů k jednotlivým funkčním stylům. Jestliže máme zkoumat rozdíly či podobnosti mezi funkčními styly, musíme si nejdříve jasně stanovit, které funkční styly do našeho bádání zahrneme. O značné variantnosti možných základních funkčních stylů jsme již pojednali výše, nebudeme se k této tematice tedy nijak široce vracet. Naším stěžejním cílem je popsat na základě určitých parametrů rozdíly mezi základními funkčními styly, musíme tedy jasně stanovit, které funkční styly považujeme za základní, jde o tuto pětici:

a) umělecký,
b) odborný,
c) publicistický,
d) administrativní,
e) hovorový (prostěsdělovací).

[34] Těšitelová, M.: *Kvantitativní lingvistika*. Praha 1987, s. 69.

Původně jsme ještě chtěli do výzkumu zařadit i řečnický styl, museli jsme od toho však upustit vzhledem k tomu, že korpusy SYN2010 a PMK takové texty nezahrnují. Původně jsme uvažovali o využití formálních mluvených projevů, které jsou obsaženy v PMK. Při zjištění, že atribut formálnosti však nekoresponduje s charakteristikou řečnického stylu, jsme museli řečnický styl vyřadit. Musíme však zmínit, že nezařazení řečnického stylu nemělo ráz teoretický, ale čistě technický.

Než začneme s vysvětlováním, které texty jsme zařadili ke kterému funkčnímu stylu a proč, považujeme za vhodné nejdříve podat přehled, jaké typy textů korpus SYN2010 nabízí, a to formou tabulky:

BELETRIE - román - soubor povídek, jednotlivá povídka - literatura faktu - jiné imaginativní texty - básně - písně - dramatický text, scénář
ODBORNÁ LITERATURA - vědeckonaučná literatura - populárněnaučná literatura, též profesní a zájm. časopisy - učebnice - abecedně, systematicky a jinak uspořádaná díla - administrativa
PUBLICISTIKA - rozmanité (efemera) - publicistika (noviny a neodborné časopisy)

Korpus SYN2010 je strukturován do tří základních částí (beletrie, odborná literatura, publicistika), mohlo by se tedy zdát, že tyto kategorie analogicky přetransformujeme do funkčního stylu uměleckého, odborného a publicistického. Pohled na podkategorie těchto tří částí však dává tušit, že tak jednoduché to nebude. V rámci beletrie nepovažujeme za vhodné ze zřejmých důvodů zahrnovat literaturu faktu do uměleckého stylu. Poněkud komplikovanější je zařazení písňových textů, které mají značně specifický charakter, a proto jsme se rozhodli tyto texty do uměleckého stylu nezařazovat. Ostatní položky již dle našeho názoru zjevně s uměleckým funkčním stylem úzce korespondují.

V případě odborné literatury je situace celkem jednoduchá, do odborného funkčního stylu můžeme zahrnout všechny uvedené podkategorie, ovšem s jednou

výjimkou – administrativními texty. Takové texty v našem pojetí samozřejmě tvoří samostatnou sekci, a to administrativní styl. V publicistice jsme extrahovali pouze texty z kategorie „publicistika“, jež zahrnuje noviny a neodborné časopisy, tedy to, co primárně očekáváme od publicistického funkčního stylu. Podkategorii „rozmanité“ jsme do našeho výběru nezahrnuli, jednak proto, že není zcela zřejmé, o jaké texty se vlastně jedná, jednak proto, že těchto textů je v korpusu SYN2010 minimum, a tedy nestačí pro naše účely.

Patrně nejsložitější byla situace při výběru textů pro analýzu hovorového funkčního stylu. Je evidentní, že texty tohoto rázu korpus SYN2010 nezahrnuje, protože jsme však nehodlali od tohoto stylu ustoupit, rozhodli jsme se využít jiný korpus, a to již zmíněný PMK. Máme totiž za to, že texty patřící do prostěsdělovacího funkčního stylu mají a priori mluvenou formu, navíc jde zpravidla o komunikáty neformálního charakteru, proto v rámci hovorového stylu budeme analyzovat neformální texty z PMK. Přestože jsme původně měli v úmyslu formální mluvené projevy použít pro analýzu řečnického stylu, který by měl teoreticky mluveným formálním komunikátům odpovídat, situace je jiná. Význam výrazu „formální“ je v PMK do jisté míry zavádějící, formálnost zde totiž neznamená oficiálnost a připravenost, ale fakt, že tyto texty byly pořízeny formou otázek obecnějšího rázu konkrétnímu respondentovi, takové texty tedy nemohou patřit zcela do kategorie zcela spontánních dialogů, avšak rozhodně ne ani do registru řečnického funkčního stylu. Protože se tyto texty pohybují na určité hranici, zařadili jsme je do analýzy hovorového funkčního stylu, přičemž jsme provedli opatření, která nejasné interferenci těchto formálních a neformálních textů zabrání (viz dále).

Výše jsme naznačili problematiku určitého prolínání několika druhů textů v rámci jednoho funkčního stylu. Považujeme za nesprávné a nepřesné jednoduše zahrnout výše uvedené podkategorie jednotlivých funkčních stylů a podat nějakou získanou hodnotu, aniž by nebyly zřejmé rozdíly mezi těmito podkategoriemi. Rozhodli jsme se tedy vytvořit subkorpusy, které budou odpovídat jednotlivým podkategoriím, a tyto separovaně analyzovat, přičemž z výsledného aritmetického průměru vypočtených hodnot získáme údaj validní pro celý funkční styl. Jinými slovy, rozhodli jsme se zkoumat nejen rozdíly mezi jednotlivými funkčními styly, ale také mezi jejich podkategoriemi, jejichž hodnoty mohou dosahovat značných amplitud.

Jako příklad nám může posloužit publicistický funkční styl, z něhož by někteří lingvisté chtěli vytvořit dva: žurnalistický a analytický (publicistický v užším smyslu). Získané hodnoty by teoreticky mohly naznačit, které podkategorie jednotlivých funkčních stylů nejvíce vybočují, a tedy tendují k vydělení se v samostatný funkční styl. Tím samozřejmě nemělo být řečeno, že pokud se některé texty například v bohatství slovníku výrazně odlišují od ostatních v rámci jednoho funkčního stylu, automaticky do tohoto stylu nepatří. Jde nám o to podívat se na často sporné rozdělení jednotlivých funkčních stylů z perspektivy lexikální statistiky, která je jen součásti mnoha jiných hledisek, ledacos však může napovědět.

5.1.1 Metodologie vytvoření výběrového souboru

Než přejdeme k popisu metod, jakými jsme rozdělili podkategorie jednotlivých funkčních stylů, uvedeme nejdříve obecněji metodologii vytvoření výběrového souboru. Vzhledem k tomu, že není možné získat základní soubor – tedy veškeré realizované texty ve sledovaných oblastech – musíme se jako u většiny společenskovědních analýz uchýlit k vytvoření výběrového souboru, který by měl být samozřejmě maximálně reprezentativní. Právě reprezentativnost takového souboru determinuje osud veškerých dalších postupů a především výstupů celé práce. Proto je velmi důležité nyní pojednat o našich krocích, jež nás vedli ke konečnému rozhodnutí, jak náš výběrový soubor vytvořit.

Obecně pro vytvoření každého výběrového souboru, jenž má sloužit jako východisko pro statistickou analýzu, platí základní dva faktory, které musíme zvážit. Jde o faktory:

a) kvantitativní,
b) kvalitativní.

5.1.2 Kvantitativní hledisko

Nejdříve se zaměříme na problematiku kvantity výběrového souboru, jejíž pozitivní korelace s reprezentativností je zcela zřejmá. Vzhledem k tomu, že ideálním stavem je základní soubor, jenž obsahuje veškerá data ve sledované

oblasti, je logické, že čím větší soubor vytvoříme, tím validnější získáme výsledky. Přestože ČNK nabízí velmi rozsáhlý jazykový materiál v řádu desítek milionů jednotek, nebylo možné SYN2010 a PMK využít v plném rozsahu. Vzhledem k tomu, že jsme z těchto korpusů potřebovali excerpovat jen určité typy textů (viz výše) a vytvořit z nich několik vlastních subkorpusů, museli jsme vycházet ze sumy nejmenšího souboru obsahujícího jednu vybranou kategorii. Jinými slovy, rozložení textů není v SYN2010 rovnoměrné, zatímco umělecké texty tvoří asi třetinu celého korpusu SYN2010, což představuje přes 30 milionů jednotek, zastoupení textů spadajících do administrativního funkčního stylu zahrnuje jen kolem 100 tisíc jednotek. Právě tento aspekt tvořil základní problém při sestavování výběrového souboru.

Každá statistická analýza založená na komparaci vyžaduje nezbytné dodržení dvou základních podmínek výběrových souborů, jejichž nedodržení musí nutně generovat systémovou chybu celého výzkumu. Abychom mohli srovnávat hodnoty získané statistickými výpočty, musí výchozí výběrové soubory být naprosto identické co do:

a) velikosti,
b) vymezení základní jednotky.

Není tedy možné, abychom například srovnávali hodnoty získané na základě statistické analýzy korpusů s odlišným rozsahem. Tento fakt nám zamezil využít plně kvantitativní potenciál korpusu SYN2010. Abychom dodrželi nezbytný požadavek identického rozsahu, museli jsme stanovit velikost na základě nejmenšího subkorpusu, v našem případě se jedná o již zmíněný korpus textů spadajících svým charakterem do administrativního funkčního stylu. Po zjištění jeho možností jsme rozsah každého zkoumaného subkorpusu stanovili na 90 000 jednotek (slovoforem).

Přestože uvedenou velikost považujeme za dostatečnou, rozhodli jsme se vzhledem k možnostem korpusů SYN2010 a PMK vytvořit systém, který by zvýšil množství analyzovaných jednotek a přitom neporušoval podmínku identického rozsahu srovnávaných souborů. Jakkoli zní takové tvrzení rozporuplně, vysvětlení je poměrně jednoduché. Víme, že základním cílem naší práce je zjistit ve sledovaných parametrech rozdíly mezi funkčními styly, zároveň však také víme o našem rozhodnutí sekundárně zjišťovat rozdíly mezi různými typy textů uvnitř jednotlivých

funkčních stylů. Například v rámci uměleckého stylu budeme separovaně také zkoumat podkategorie (román, povídka, báseň apod.). Jestliže každá jedna podkategorie bude mít stanovených 90 000 jednotek, aritmetický průměr výsledných hodnot těchto podkategorií uvnitř jednoho funkčního stylu nám poskytne validní hodnotu, která bude mít několikanásobně vyšší míru reprezentativnosti. Tento postup bezprostředně souvisí s kvalitativními faktory při vytváření výběrového souboru, o kterých pojednáme dále.

5.1.3 Kvalitativní hledisko

Jakkoli je kvantitativní složka výběrového souboru základním faktorem jeho reprezentativnosti, ani při její sebevětší velikosti nezískáme relevantní výsledky, pokud vhodně nezohledníme kvalitativní hledisko, jež je striktně determinováno cílem celého výzkumu. Výše jsme již uvedli, jak jsme přiřadili různé typy textů k jednotlivým funkčním stylům. Konkrétní vymezení, které texty jsou kam zařazeny a z jakého důvodu, popíšeme v této podkapitole. Než tak učiníme, považujeme za vhodné uvést fakt, že zatímco v *RRKLPMČ* jsme veškeré kvalitativní parametry zpravidla určovali sami, nyní jsou naše možnosti omezeny danými parametry korpusů SYN2010 a PMK. Shledáváme nadbytečným zde široce popisovat různá kvalitativní hlediska, která tvůrci avizovaných korpusů stanovili při jejich vytváření. Níže uvedeme jen nejnutnější parametry, jejichž znalost je nezbytná pro recepci celé naší práce, detailnější informace o kvalitativních parametrech užitých při tvorbě jednotlivých korpusů ČNK jsou dostupné jednak v mnoha publikacích, jednak on-line na webových stránkách ČNK.[35]

Do rámce uměleckého funkčního stylu jsme zařadili celkem pět podkategorií:

a) román
b) povídka
c) jiné imaginativní texty
d) báseň
e) drama, scénář

[35] http:/www.korpus.cz

Každá podkategorie reprezentuje jeden vytvořený subkorpus, jehož rozsah je vždy uvedených 90 000 jednotek. Celkem je tedy umělecký funkční styl tvořen téměř půl milionem jednotek, přičemž připomínáme, že získané hodnoty v žádném případě nebudeme sčítat, ale budeme pří výpočtu výsledných hodnot pro celý umělecký funkční styl pracovat výhradně s aritmetickým průměrem získaným z výsledných hodnot zmíněných subkorpusů. O tom, proč jsme do tohoto funkčního stylu nezařadili některé další typy textů, které byly obsaženy v korpusu SYN2010 v rámci beletrie, již bylo pojednáno výše.

Odborný funkční styl reprezentují čtyři podkategorie textů:

a) vědeckonaučné,
b) populárněnaučné,
c) učebnice,
d) jiná uspořádaná díla.

O poznání komplikovanější bylo rozhodnout, které podkategorie vytvořit uvnitř publicistického funkčního stylu. Abychom předešli jednoduchému rozlišení do dvou nabízejících se sekcí (noviny, časopisy), rozhodli jsme se jich vytvořit celkem šest. Vzhledem k tomu, že jsme si vědomi značných rozdílů mezi jednotlivými periodiky, vytvořili jsme šest subkorpusů skládajících se z textů jednoho konkrétního média, přičemž jsme přihlédli k zastoupení seriózních deníků, bulváru a analytických časopisů. Zvolené subkorpusy jsou tyto:

a) MF Dnes,
b) Lidové noviny,
c) Právo,
d) Blesk,
e) Reflex,
f) Respekt.

Považujeme za nezbytné v případě rozdělení publicistického stylu zmínit fakt, že na rozdíl od ostatních funkčních stylů zde nesrovnáváme jednotlivé žánry. Důvodů, které nás k takovému rozčlenění vedly, je více, nicméně dva dominují. Za prvé jsme reflektovali současnou situaci v tisku, kde dochází k interferenci žánrů, a je tak takřka nemožné přiřadit jednotlivé texty ke konkrétním žánrům. Za druhé jsme při vydělování podkategorií publicistického funkčního stylu vycházeli z primárních

funkcí textů. V současnosti můžeme z hlediska funkce vydělit tři základní kategorie textů:

a) zpravodajské (informační funkce),
b) komentářové (persvazivní funkce),
c) bulvární (zábavní funkce).

Jednotlivá námi zvolená periodika tak vždy zastupují jednu z uvedených kategorií. Domníváme se, že takové rozdělení mnohem lépe reflektuje skutečný stav v médiích a také mnohem lépe odpovídá funkčnímu pojetí.

Nejjednodušší situace nastala při sestavování výběrového souboru pro administrativní funkční styl. Jak vyplývá z výše uvedeného, nebylo možné zde vytvořit více podkategorií, neboť SYN2010 obsahuje pouze jeden set textů s atributem administrativní. Netřeba již dodávat, že právě velmi malé zastoupení těchto textů uvnitř korpusu SYN2010 určilo základní velikost každého námi tvořeného subkorpusu na 90 000 jednotek. Považovali jsme však za nesprávné jen kvůli tomuto relativnímu nedostatku ustoupit od analýzy celého administrativního stylu, jehož absence by výrazně ublížila kvalitě celé této práce.

Jestliže vytváření výběrového souboru pro zkoumání administrativního stylu bylo bezproblémové, v případě hovorového stylu je tomu zcela naopak. První problém vychází ze skutečnosti, že na rozdíl od ostatních souborů nevycházíme z korpusu SYN2010, ale PMK. Přestože oba korpusy patří do množiny ČNK, PMK vykazuje určité rozdíly ve zpracování dat, jež jsou dány především tím, že se jedná o korpus mluvených textů, což s sebou přináší určitá specifika. Základní informace o obou korpusech jsme již uvedli v předchozích kapitolách, problematice definování jednotky se pak budeme věnovat dále.

PMK se skládá ze základních dvou zhruba stejně velkých částí – formálních a neformálních textů. Otázkou bylo, které korespondují s charakteristikou prostěsdělovacího funkčního stylu. Vzhledem k tomu, že právě neformálnost je jedním ze základních slohotvorných faktorů hovorového funkčního stylu, tyto texty jsme do výběrového souboru bez dlouhého rozvažování zařadili, složitější bylo rozhodování o textech formálních. Ona formálnost, jak již bylo uvedeno výše, je v PMK chápána takto: „Formální promluva je monolog vytvářený sledem odpovědí na otázky kladené nahrávajícím (pro zamezení ovlivnění odpovědí, ať už

kódem spisovným či nespisovným, měly smíšenou povahu nespisovně-spisovnou). Týkaly se takových širokých témat jako škola, mládež, zaměstnání ap. a nahrávány ani přepisovány nebyly (byly ve všech případech stejné).“[36] Abychom úplnosti učinili zadost, nabídneme i autorskou definici neformálních textů: „Neformální promluva je vlastně dialogický soubor promluv dvou mluvčích, kteří se znají; téma jejich rozhovoru nebylo nijak předem určováno, volili ho sami. Nahrávky usilovaly o proporční vyváženost desítek sociolingvistických kombinací (typu MIBF, MIAF, MIAN apod.) takto vzniklých a jsou tedy v tomto smyslu reprezentativní pro všechny proměnné. Nepřipravenost odpovědí a dialogů zaručuje maximální možnou spontánnost užitého jazyka.“[37]

5.1.4 Jak spojit SYN2010 a PMK?

Výše popsané metodologii lze nepochybně vytknout, že nejprve striktně požadujeme kvantitativně i kvalitativně identické texty a posléze chceme komparovat hodnoty vzešlé ze zcela jiných korpusů. Taková námitka je samozřejmě zcela na místě a jsme si tohoto faktu zcela vědomi se všemi jeho potenciálními důsledky. Tato kapitola tedy bude mít za cíl vysvětlit, resp. ospravedlnit, naše rozhodnutí.

Prvně by bylo dobré zodpovědět nabízející se otázku, proč jsme se do takového lehce napadnutelného experimentu pustili, zvlášť s přihlédnutím k faktu, že jde o školní práci. Hlavní důvod byl ten, že jsme nechtěli v naší komparaci funkčních stylů vynechat styl hovorový, jež se značně odchyluje od ostatních. Prostěsdělovací styl má mnohá specifika, jež z něj tvoří z našeho pohledu pozoruhodný exemplář. Jinými slovy nezahrnout jedinečnost hovorového funkčního stylu by dle našeho názoru velmi citelně uškodilo významu celé práce, proto jsme přes všechna možná rizika zařadili do našeho výzkumu i PMK.

Z výše uvedeného tedy vyplývá, že pro správnou recepci komparace je nezbytně nutné přesně znát rozdíly ve zpracovávání obou korpusů. Tyto informace můžeme nalézt v našich největších dvou frekvenčních slovnících, jež jsou základním

[36] http://ucnk.ff.cuni.cz/pmk.php [cit. 25. 1. 2012]

[37] Tamtéž.

publikovaným výstupem obou korpusů, jde o *Frekvenční slovník češtiny* (FSČ) a *Frekvenční slovník mluvené češtiny* (FSMČ). Vzhledem k tomu, že zde není prostor na detailní popis všech těchto pravidel a zároveň by ani nemělo žádný smysl zdlouhavě parafrázovat informace obsažené ve zmíněných publikacích, odkazujeme na výše zmíněné publikace.

Závěrem této kapitoly tedy znova deklarujeme, že jsme si v plném rozsahu vědomi všech potenciálních rizik spojených s komparací dvou různých korpusů. Přes všechna úskalí však jsme naprosto přesvědčeni, že je možné hodnoty získané ze SYN2010 a PMK podrobit vzájemné komparaci, jež je schopna přinést validní výsledky.

5.1.5 Lemma vs. slovoforma

Relativně dlouhá tradice kvantitativních výzkumů se opírá o tezi, že ve slovanských flexivních jazycích počítáme základní statistické parametry na základě lemmat, a nikoli slovoforem. Problematiku reflektuje i Marie Těšitelová, která poukazuje na komplikované rozlišení těchto dvou opozic: „V jazycích s bohatou morfologií, jako jsou jazyky slovanské, se tyto ‚dvě podoby' v lexikální statistice většinou jasně diferencují […]. V jazycích s morfologií ‚chudou', jako jsou jazyky germánské, románské aj., nejsou tyto termíny-pojmy slovo – různé slovo vždy jednotně diferencovány."[38] Autorka se zabývá tím, jak vhodně lemmatizovat, nebere však již v potaz, zda má vůbec smysl lemmatizaci provádět. Proč bychom tak měli činit, je skutečně lemma základem lexikální statistiky? Není lemmatizace jakýmsi lingvistickým „znásilňováním" přirozeného jazyka? Tyto a podobné otázky si klademe a rozhodli jsme se je integrovat paralelně do našeho výzkumu. Ostatně značně výmluvný je samotný fakt, že je mnohdy velmi složité či dokonce nemožné rozhodnout, které tvary patří ke kterému lemmatu. Při tomto přístupu vždy nutně narážíme na skutečnost, že žádné univerzální pravidlo pro lemmatizaci neexistuje, každý tvůrce frekvenčních seznamů se tak vždy musí na počátku své práce rozhodnout ad hoc, jak a co bude považovat za jednu lexikální jednotku.

[38] Těšitelová, M.: *Kvantitativní lingvistika*. Praha 1987, s. 61.

Pokud máme definovat naše pojetí základní jednotky, musíme vyjít z možností ČNK. V našem výzkumu budeme paralelně pracovat s dvěma základními jazykovými jednotkami, a to lemmatem a slovoformou. Slovoformou rozumíme písmeno či řadu písmen mezi mezerami bez jakéhokoli ohledu k jejich lexikálnímu významu. Značně složitější je definovat, co chápeme pod výrazem lemma, zejména proto, že neurčujeme tuto jednotku sami, ale zcela přejímáme definici obou zkoumaných korpusů (SYN2010, PMK).

5.1.6 Metodologie analýzy

Nyní pojednáme konkrétněji o jednotlivých krocích, jež jsme aplikovali při zpracovávání výchozích materiálů. Půjde nám tedy o metodologický popis spíše technického rázu.

Nejdříve bylo třeba extrahovat data z ČNK, tedy vytvořit z korpusů SYN2010 a PMK vlastní subkorpusy o požadované velikosti 90 000 jednotek obsahující vybrané texty. V prostředí korpusového manažeru Bonito jsme zadáním příslušného dotazu vyhledali požadovaný typ textů a pomocí funkce „N-filtr“ následně odstranili z textů interpunkci. Dalším krokem bylo omezit nalezené texty na požadovaný rozsah 90 000 jednotek, toho jsme dosáhli pomocí funkce „redukce“. Následujícím úkolem pak bylo z tohoto subkorpusu, z nějž již byla odstraněna interpunkce a u něhož byl omezen rozsah na požadovanou hodnotu, získat frekvenční seznam. To umožňuje funkce „frekvenční distribuce“, která po zadání požadovaných parametrů generuje odpovídající frekvenční seznam. Takový seznam jsme již mohli exportovat na harddisk našeho počítače. V této fázi jsme tedy ukončili práci v korpusovém manažeru Bonito.

Další operace vedla přes editor PSPad, ve kterém jsme mohli otevřít získaný frekvenční seznam, jež byl generován jako prostý text (plain text). Skrze PSPad jsme data dále exportovali do formátu tabulkového procesoru Excel, přičemž jsme používali novější formát xlsx. Je třeba zmínit, že při vytváření frekvenčních seznamů jsme narazili na systémovou chybu manažeru Bonito. Jedná se o podstatnou funkci, která automaticky při sestavování frekvenčního seznamu ignoruje velikost písmen. Zjistili jsme, že právě tato funkce nepracuje správně, resp. vůbec. Proto jsme museli

získané frekvenční seznamy podrobit úpravě v prostředí programu Excel pomocí funkce „sumif“. V prostředí Excelu jsme potom již jednoduše zadali příslušné matematické operace a program nám generoval požadované výsledné hodnoty. Tato data jsme následně v témž programu zpracovali do grafů, jež se staly základním podkladem pro samotnou interpretaci.

5.2 Rozsah lexika

Jde o jednu z nejdéle sledovaných veličin v rámci cele kvantitativní lingvistiky, jež vystihuje vztah slovníku a délky textu. Slovníkem (V) je myšlen počet lexikálních jednotek, tedy lemmat, délka textu (N) potom představuje celkový počet slov, přičemž za slovo je považováno písmeno či řada písmen v textu mezi mezerami, v této práci budeme takovou jednotku zpravidla označovat jako slovoformu. V paralelních výpočtech ve slovoformách pak budeme dosazovat za V počet slovoforem, zatímco N zůstane stejné.

Rozsah slovníku je jednou z nejzákladnějších veličin, jež nám může jednoduchým výpočtem ukázat na elementární charakteristiku textu v lexikálním plánu. Tento lexikostatistický parametr nachází široké uplatnění například při určování autorství, neboť každý autor má svůj charakteristický idiolekt, který se projevuje ve všech jazykových plánech, slovník nevyjímaje.

Je zřejmé, že pokud rozsah lexika je poměrně úspěšný při určování autorského stylu, měl by být analogický stejně relevantní při rozlišování funkčních stylů. Předpokládáme, že mezi jednotlivými styly musejí být rozdíly nehledě na autora. Tudíž nejen autor, ale také určitá oblast textů musí vykazovat určité charakteristiky, včetně bohatství slovníku. Například očekáváme, že básnické texty mají mnohem širší slovní zásobu než administrativní úřednické spisy. V tomto případě je situace celkem jasná, ale mnohem složitější je již odhadnout jaké rozdíly jsou například mezi uměleckými a publicistickými texty, nebo jaké jsou diference uvnitř samotného publicistického stylu. Právě rozsah lexika by v kombinaci s rozptýlením a koncentrací lexika měl ukázat tyto rozdíly a exaktně je kvantifikovat.

Základní rovnici pro výpočet bohatství slovníku stanovil francouzský lingvista Pierre Guiraud:

$$R = \frac{V}{N}$$

Tato základní rovnice pro výpočet bohatství slovníku Těšitelová modifikuje: „[…] je třeba – pro češtinu – počítat jen s 80 % textu, tj. pouze se slovy plnovýznamovými (substantiva, adjektiva, zájmena, číslovky, slovesa a adverbia), popř. se 70 % textu, pokud za slova plnovýznamová pokládáme jen substantiva, adjektiva, slovesa a adverbia.“[39]

Takto modifikované rovnice pak vypadají následovně:

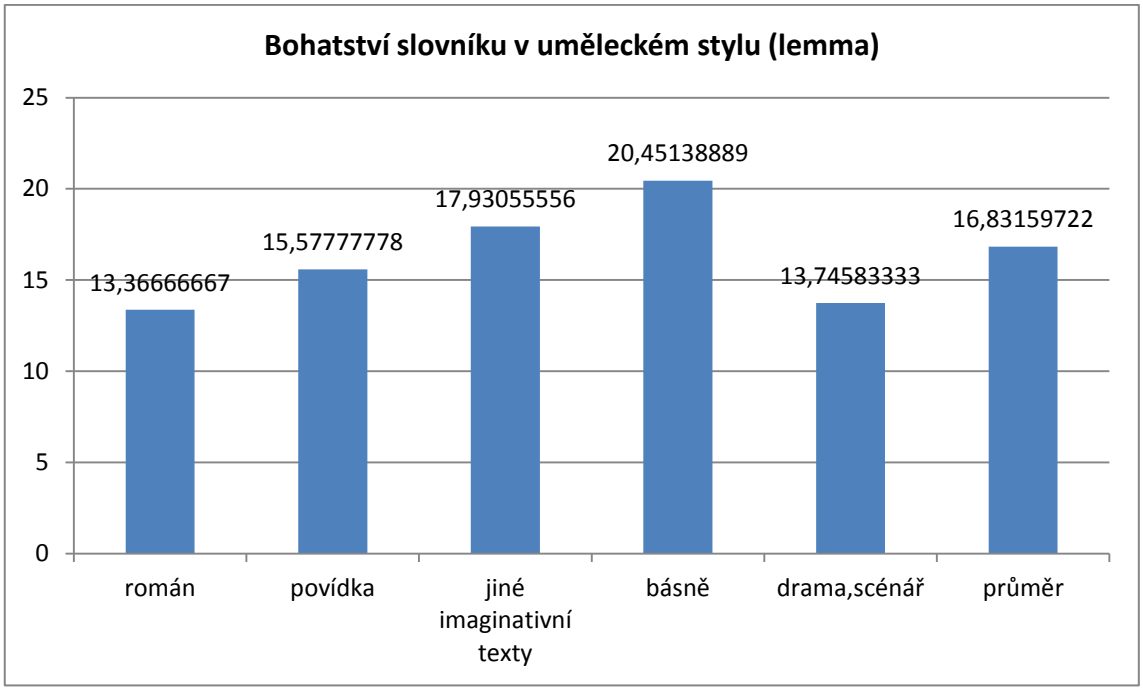
$$R = 100\ \frac{V}{\frac{8}{10}\ N}$$

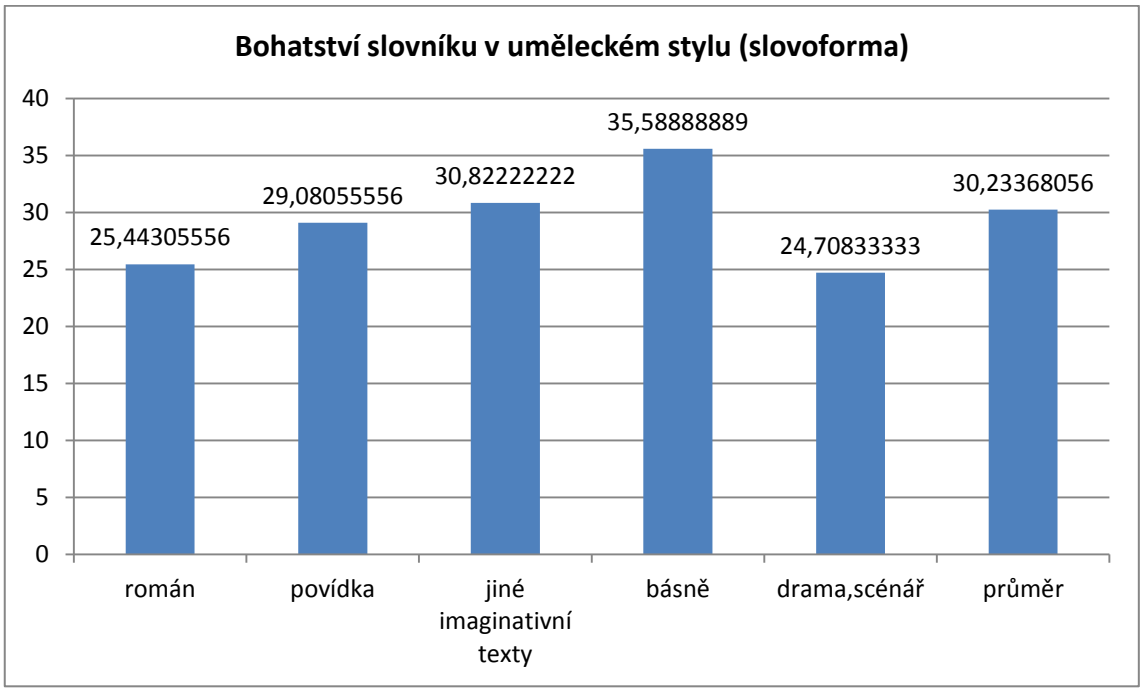
$$R = 100\ \frac{V}{\frac{7}{10}\ N}$$

Stejně jako Těšitelová budeme primárně pracovat s první rovnicí, která počítá s 80 % plnovýznamových slov.

5.2.1 Umělecký styl

[39] Těšitelová, M.: *Kvantitativní lingvistika*. Praha 1987, s. 71.





Hned první pohled na grafy napovídá, že naše rozhodnutí diverzifikovat každý funkční styl do více podkategorií byl správný. Jednak aritmetický průměr získaný z několika stejně velkých subkorpusů poskytuje mnohem objektivnější hodnotu, jednak můžeme sledovat různé hodnoty u jednotlivých podkategorií, které tak můžeme také podrobit vzájemné komparaci.

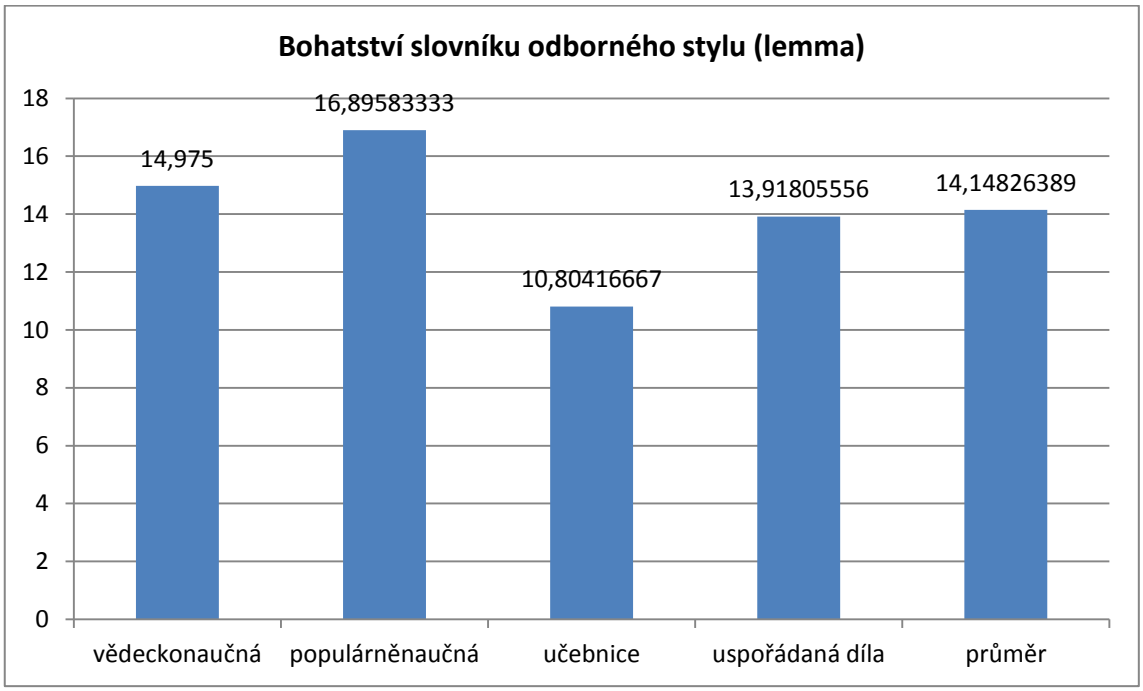
Nejvýraznější amplituda, kterou se vykazují básnické texty, je patrně způsobena několika faktory. Prvním bude fakt, že v básních se užívá více substantiv a adjektiv než v ostatních uměleckých textech, což je dáno jednak nutností popisovat jednu skutečnost různými výrazy, jednak podstatně ochuzenou gramatickou složkou. Básnický jazyk nepotřebuje tolik gramatických slov jako ostatní texty, není zde potřeba skládat delší komplikovanější větné útvary, také není třeba dodržovat jinak závazná pravidla. Přitom právě gramatická slova jsou samozřejmě jádrem slovní zásoby, stojí v jejím samotném centru. Důležitou roli zde hraje také fakt, že v básnických textech se předpokládá barvité neotřelé vyjadřování umocněné i v lexikální oblasti, a to užíváním stylisticky velmi rozmanitých výrazů, od zcela hovorových až k archaickým, nadto samozřejmě ještě užitím zcela specifické množiny slovní zásoby – poetismů. Značný rozdíl spočívá rovněž ve faktu, že básnické texty jsou nesrovnatelně kratší, i v rámci jedné básnické sbírky každá jednotlivá báseň je často značně obsahově rozdílná. Při aplikaci statistických metod se tento faktor musí projevit, otázkou je, kdybychom do korpusů zařadili stejné množství úryvků románů totožné délky, zda a jak by se to odrazilo ve výsledných hodnotách. Při této úvaze se však rovněž musíme ptát, zda by takový zásah do výzkumu měl význam, neboť bychom narušili přirozené celky koherentních textů, a de facto bychom tak narušili právě ty struktury, jejichž jedinečnost zkoumáme.

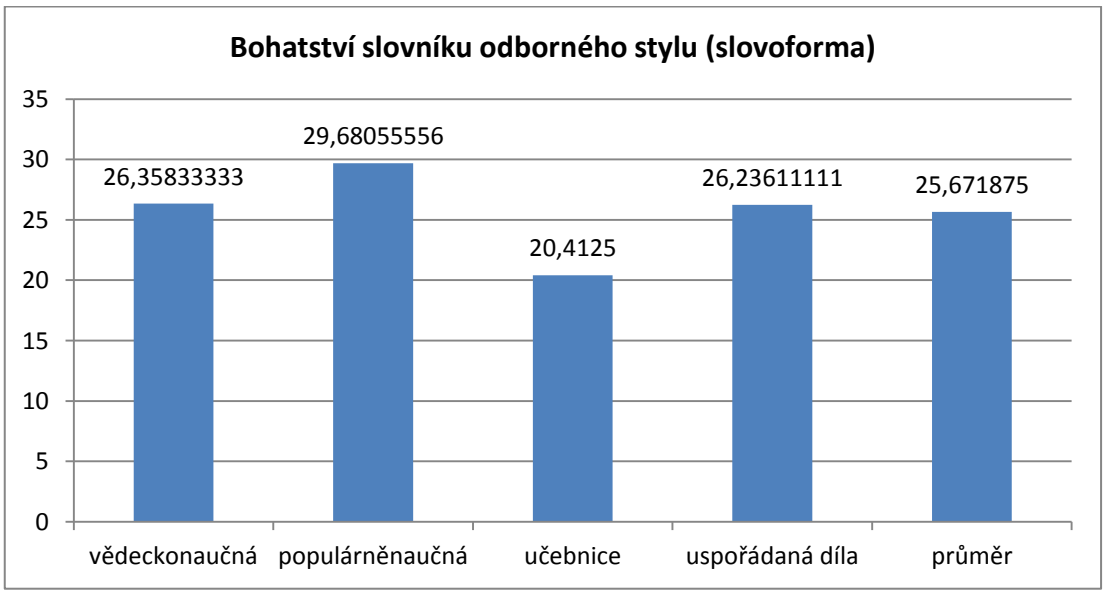
Posledně výše zmíněný aspekt pravděpodobně hraje roli i ve vyšší hodnotě bohatství slovníku v povídkách oproti románům. Co se týká imaginativních textů, jejich samotné označení dává tušit, že půjde o umělecké texty odlišující se od tradičních románových a povídkových texů tím, že směřují do volnější gramatické stavby a dodržování norem směrem k básnickým textům. Tento předpoklad by získaná data jen potvrzovala.

Poměrně překvapivé se zdají téměř shodné výsledné hodnoty románů a scénářů a dramat. Dramatické a scénáristické texty mají velmi specifický ráz odlišující je od ostatních uměleckých textů. Jejich odlišnost můžeme sledovat v několika jevech, přičemž nejdůležitějším je snaha produkovat mluvený jazyk, avšak zatímco filmoví autoři usilují o maximální simulaci běžného mluveného jazyka, v divadelních hrách jde spíše o velmi umělý stylizovaný jazyk. Důležitá je také velmi silně zastoupená dialogická forma projevů, která se podepisuje na výrazném směřování těchto textů

k hovorovému stylu. Na nízké hodnotě rozsahu lexika se podílí rovněž úzce související potřeba opakování a redundance vzhledem k formě produkce těchto textů, pro diváka filmových či divadelních představení je mnohem obtížnější recepce textu než pro čtenáře ostatních uměleckých děl.

5.2.2 Odborný styl

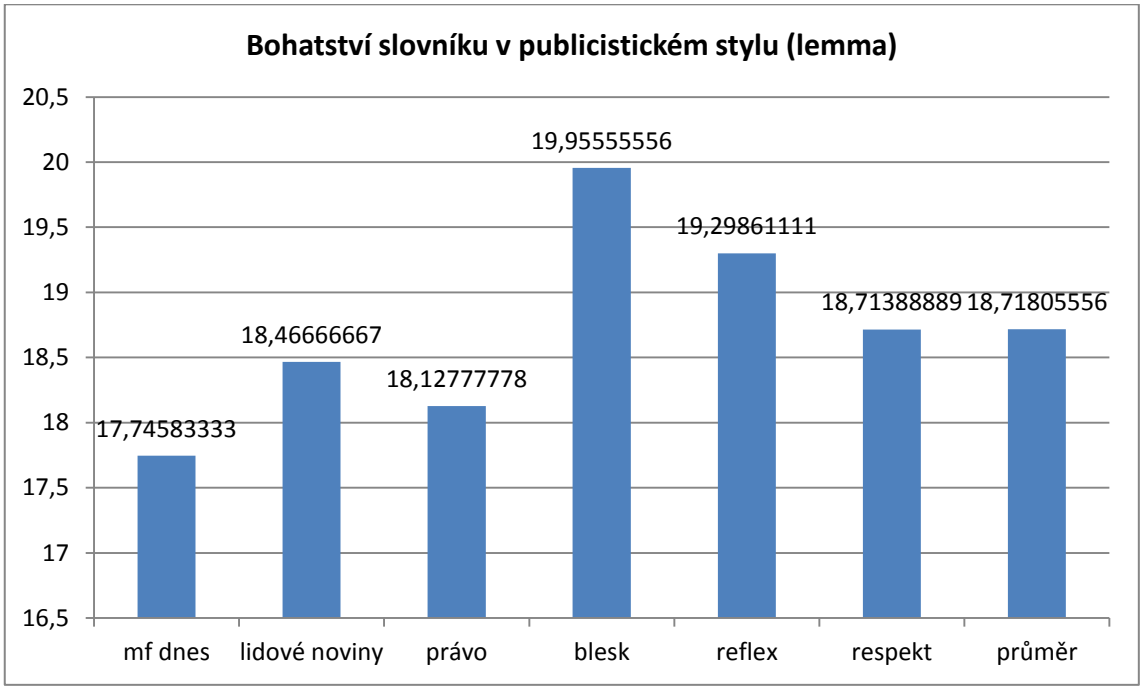


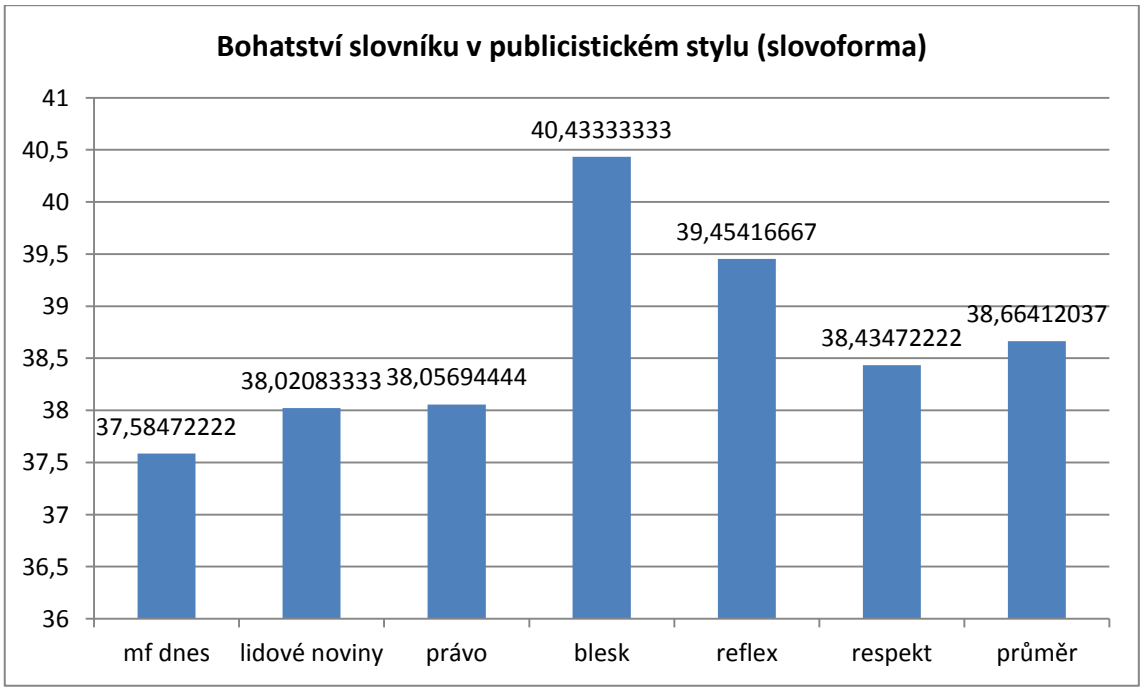


V rámci odborného stylu nejsou amplitudy relativně příliš vysoké, výjimku tvoří učebnice, jež vykazují značnou odchylku od ostatních textů. Na opačném konci stojí populárně-naučné texty, jejichž hodnoty jsou nejvyšší, nejsou však tolik vzdáleny ostatním. V případě učebnic je nízká hodnota rozsahu lexika pravděpodobně dána jejich specifických charakterem, který je značně odlišuje od ostatních odborných textů. Přestože determinant zde bude více, základní je fakt, že většina učebnic je určena ještě nezletilým čtenářům, jejichž jazyková kompetence a zejména jejich slovní zásoba musí být nutně nižší než u dospělé populace. Významný je rovněž účel těchto publikací, který se odráží v častém opakování, svou roli by také mohla hrát monotematičnost, na druhou stranu u ostatních odborných textů tomu není jinak.

Zatímco vysvětlení nejnižšího slovníkového bohatství bylo relativně jednoduché, jako obtížnější shledáváme interpretaci opačného pólu stupnice – populárně-naučné literatury. Za kardinální otázku zde považujeme, proč rozsah lexika v případě těchto textů dosahuje vyšší hodnoty než u vědecko-naučné literatury. Abychom dospěli k zodpovězení nastolené otázky, musíme vyjít z odlišností obou množin. První je různá délka textů, předpokládáme, že populárně naučné články jsou průměrně kratší, což se musí v našem 90 000 korpusu projevit větší rozmanitostí. Na druhou stranu odbornější texty by se měly vykazovat širší slovní zásobou obohacenou například mnohými termíny a slovy cizího původu; s celkově vyšší složitostí těchto textů však zase souvisí užívání komplikovanějších větných celků, což implikuje vyšší zastoupení nejfrekventovanějších gramatických slov, jejichž podíl se tak zvyšuje na úkor ostatní slovní zásoby. Jedním z nejrelevantnějších faktorů však patrně bude jistá svázanost odbornějších textů, které značně omezuje škálu výběru jazykových prostředků, přísné dodržování normy je pro takové texty zcela charakteristické. Oproti tomu populárně naučné texty mohou využít mnohem pestřejší paletu výrazových prostředků, jež může obsahovat například i, v čistě odborných textech zcela nepřípustnou, frazeologii. Populárně naučná literatura také obsahuje větší míru fatických částí, neboť atraktivnost takových textů je zásadní. Naopak vědecké texty určené velmi úzkému jasně specifikovanému publiku umožňuje vypuštění mnohého, neboť se očekává od čtenáře široká znalost tématu.

5.2.3 Publicistický styl





V rámci publicistických textů suverénně nejvyšší hodnoty bohatství slovníku dosáhl nejprodávanější český bulvární deník Blesk. Toto zjištění je docela překvapující, neboť tento typ publicistických textů je zaměřen na nejširší okruh čtenářů se spíše nižším vzděláním. Od toho by se měla teoreticky odvíjet i celkově

jednodušší výstavba textu, včetně užšího lexika, aby se i méně vzdělaným čtenářům usnadnila recepce textu. V textech bulvárních deníků se rozhodně neužívá odborných termínů nebo slov cizího původu, stylistické využití archaismů či různé frazeologie také není podporováno, na druhou stranu autoři těchto textů často užívají výrazu spadajících do hovorové či nespisovné oblasti. V potaz musíme brát také jednoznačně kratší útvary oproti seriózním deníkům, což může částečně výslednou hodnotu vysvětlovat. Pravděpodobně další možná vysvětlení poskytnou frekvenční seznamy, které uvedeme v příloze.

Dalším překvapivým výsledkem je fakt, že týdeník Reflex má větší rozsah lexika než jeho konkurent Respekt. Tyto časopisy jsou jak formátem, tak obsahem takřka totožné; rozdíl mezi těmito periodiky tkví ve větší bulvárnosti Reflexu oproti zaměření Respektu na nejnáročnější čtenáře v dané oblasti všeobecných analytických magazínů. Tento fakt by spíše napovídal bohatšímu slovníku Respektu než Reflexu, námi získaná data však ukazují, že co do náročnosti recepce textů, alespoň v lexikální rovině, Reflex poněkud svého konkurenta převyšuje.

Jestliže se podíváme na hodnoty tří celostátních seriózních deníků, zjistíme, že MF Dnes dosahuje nejnižšího bohatství slovníku. Takový výsledek je patrně očekávatelný oproti Lidovým novinám, které mají analytičtější charakter a tématům se věnují více do hloubky. Otázkou je, proč deník Právo dosahuje v podstatě totožných hodnot jako Lidové noviny, neboť nepředpokládáme, že by zmíněná hlubší analytičnost platila i pro Právo. Přestože rozdíly mezi zmíněnými třemi seriózními deníky nejsou velké, určitou tendenci z nich patrně odečíst můžeme. Teoretickým vysvětlením může být například proporčně větší zastoupení sportovní rubriky v MF Dnes, takové texty jsou poměrně lexikálně chudší, zpravidla jde o takřka aplikaci určitých schémat, kde se jen dosazují proměnné, podobně jako v administrativním stylu, možnosti kreativního užití jazykových prostředků má v této oblasti novinář značně omezený.

Pro správné pochopení výše uvedeného je nezbytné uvést, že popsané rozdíly je třeba chápat v kontextu míry diferenciace. Je zřejmé, že přes značné odchylky, jež poskytují oba grafy, je nutné vnímat také číselné hodnoty, jež v tomto případě ony odchylky značně relativizují. Maximální rozdíl totiž dosahuje pouze hodnoty 2,22 v případě lemmat a 2,86 u slovoforem. Pro srovnání můžeme uvést rozdíl mezi

minimálními a maximálními hodnotami v případě uměleckého (7,08 a 10,88) a odborného stylu (6,1 a 9,27). Rozdíly vyjádřené v grafu tak opticky zkreslují skutečný stav, vždy tak musíme brát v potaz škálu, na jaké je graf generován.

Pokud se podíváme na získaná data uvnitř publicistického stylu v obecnější rovině, můžeme tvrdit, že časté diverzifikace tohoto funkčního stylu na zpravodajský a analytický (též komentářový či publicistický v užším smyslu) má v rovině bohatství slovníku své opodstatnění. Hned však musíme uvést, že takové tvrzení platí jen u seriózních periodik, výsledné hodnoty nejprodávanějšího bulvárního deníku ukázaly, že bychom měli publicistický funkční styl dále rozšířit minimálně o jednu další kategorii, a to právě bulvární periodika. Pokud bychom totiž považovali seriózní a bulvární deníky jako jeden celek v opozici k analytickým týdeníkům či měsíčníkům, dopouštěli bychom se patrně značného pochybení, protože seriózní a bulvární deníky jsou si mnohem vzdálenější než seriózní deníky a časopisy.

Samozřejmě výše zmíněné předpoklady vycházely čistě z hlediska výsledných hodnot bohatství slovníku, na druhu stranu však považujeme za vhodné alespoň nastínit, zda diferenciace funkčního stylu do nejméně tří kategorií má opodstatnění i v jiných rovinách. Přestože jsou bulvární deníky zpravodajskými, je zřejmé, že jejich agenda je poněkud odlišná. Události, které jsou na prvních stránkách seriózních deníků, se vyskytují v bulváru v pozadí a samozřejmě v mnohem menším rozsahu; naopak témata na titulkách bulvárních deníků se do seriózních periodik dostanou na okraj společenské rubriky nebo vůbec. Základní rozdíl mezi těmito tiskovinami tak je zejména v odlišném obsahu, naopak zpravodajský charakter mají společný, což na jednu stranu podporuje myšlenku zahrnovat oba typy do jedné zpravodajské kategorie, na druhou stranu je patrné, že v obsahové rovině jsou si velmi vzdálené. Týdeníky jako Reflex, Respekt nebo Týden v podstatě přebírají agendu seriózních deníků s tím, že je uvádí do širšího kontextu a provádí hlubší analýzu, obsahově jsou si tak velmi podobné. Pokud tedy chceme nějak diverzifikovat publicistický funkční styl, je třeba zvolit hledisko, podle kterého budeme jednotlivé charakteristiky třídit. Stávající stav je takový, že za primární považujeme opozici zpravodajství-komentáře, otázkou je, zda takové dělení odpovídá dnešní realitě, kdy bulvární periodika dosahují nejvyšších nákladů a tvoří značně specifickou oblast v rámci publicistiky. Z našeho pohledu je styl bulvárních

periodik v českých stylistikách neoprávněně opomíjen a nevěnuje se jim adekvátní prostor, což dle našeho názoru vyplývá z tradice a částečně snad i z neochoty lingvistů zabývat se tímto „nízkým“ stylem.

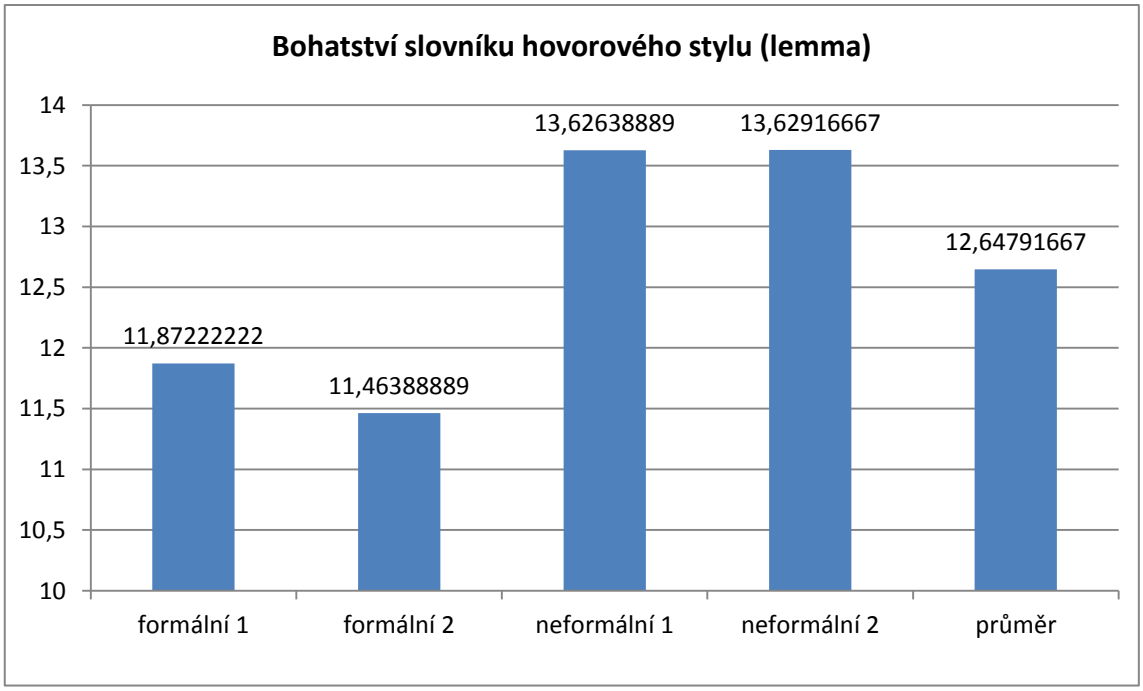
5.2.4 Administrativní styl

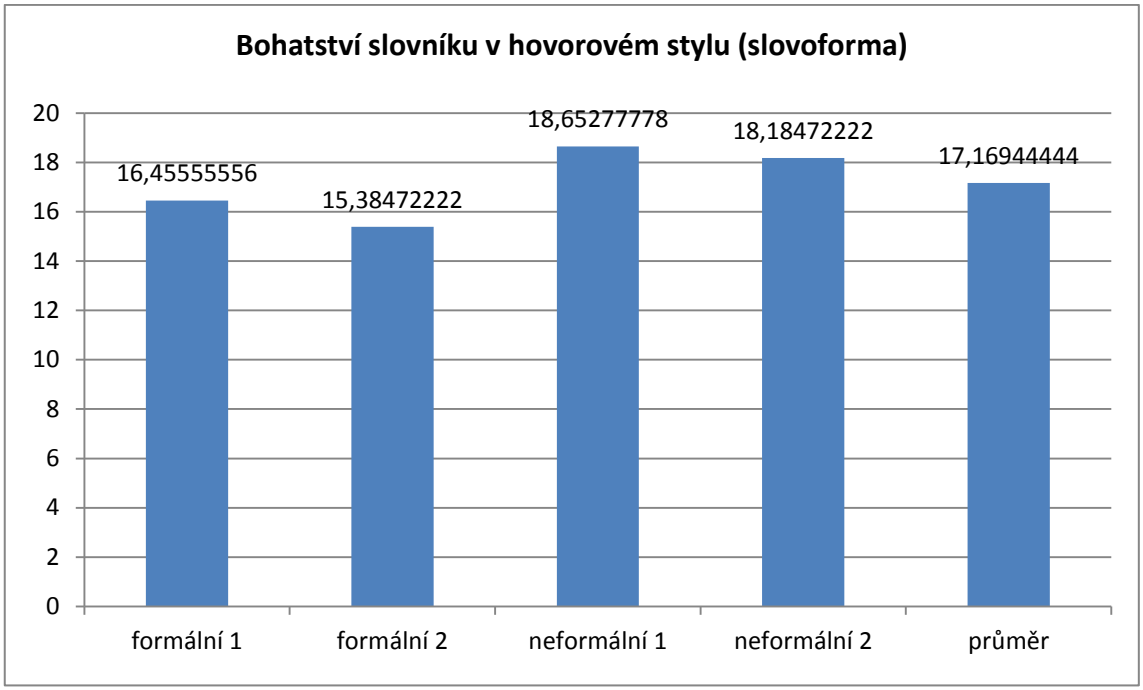
R (lemma) = 9,090278

R (slovoforma) = 17,29444

Vzhledem k tomu, že jsme neměli možnost administrativní styl diverzifikovat do více kategorií, je zřejmé, že zde nemůžeme provádět žádnou komparaci uvnitř tohoto funkčního stylu, tudíž ani nemá význam generovat na základě získaných dat grafy jako u ostatních funkčních stylů. V této fázi by byla jakákoliv interpretace předčasná, data budou vyhodnocena až v závěru této kapitoly v rámci komparace průměrných hodnot rozsahu lexika jednotlivých funkčních stylů.

5.2.5 Hovorový styl

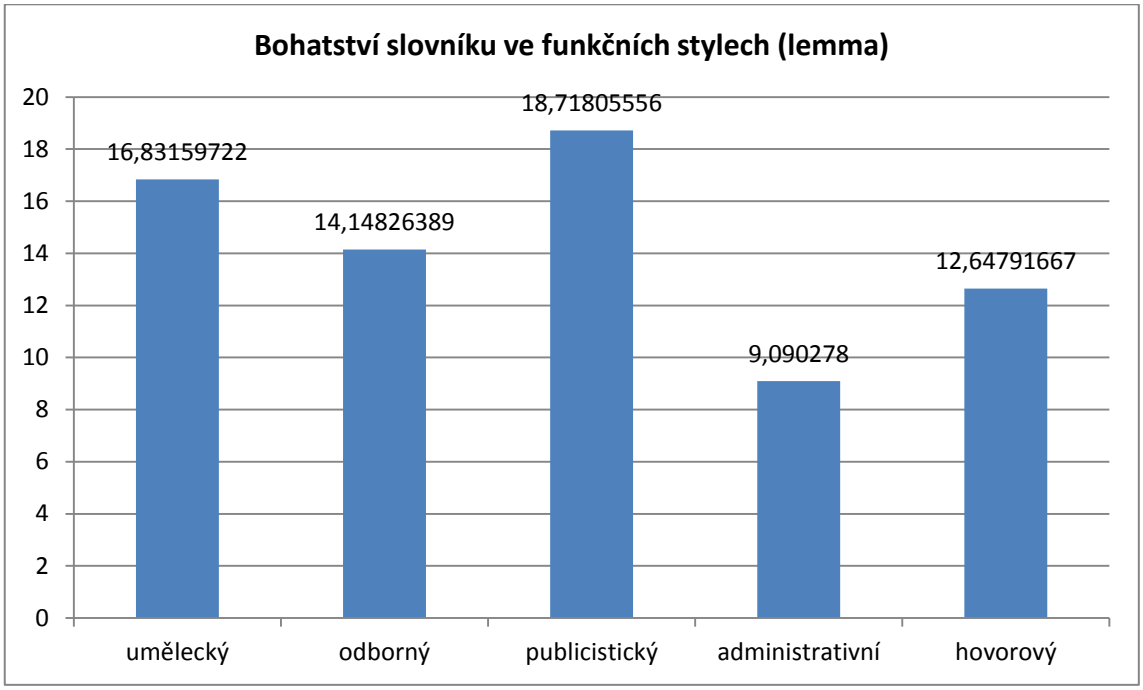


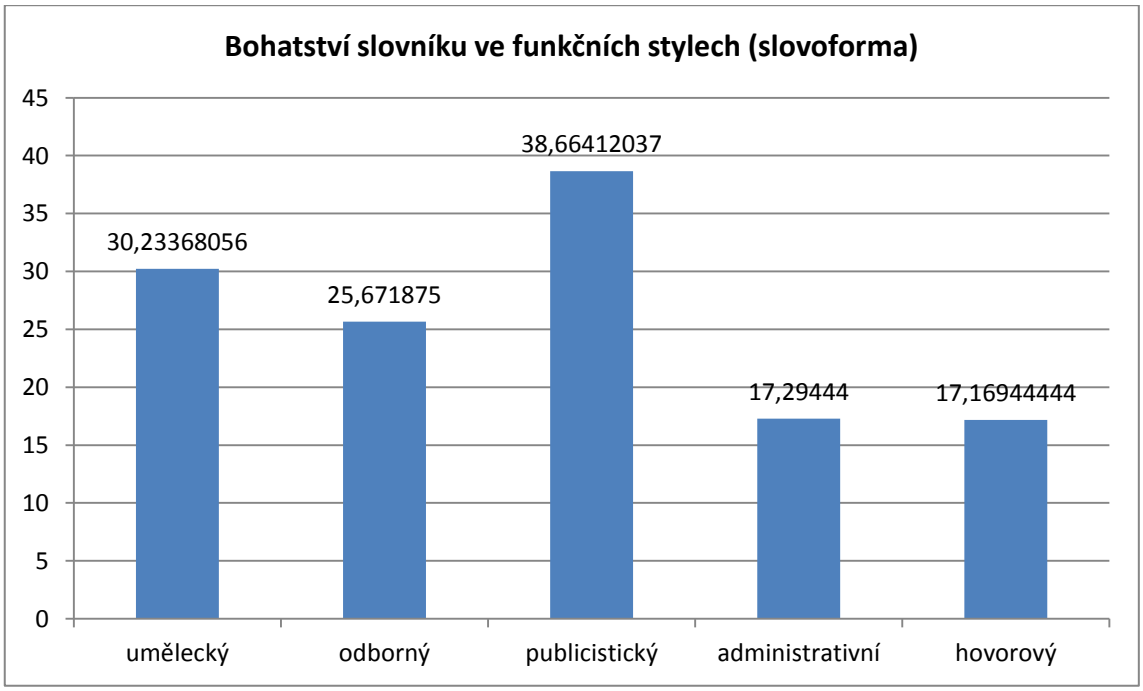


V případě hovorového stylu jsme neměli příliš široké pole, jak tuto oblast diferenciovat. Našemu srovnání uvnitř tohoto stylu tak podléhá opozice formálnost-neformálnost. O podstatě tohoto rozdělení bylo již pojednáno výše, proto by bylo opětovné vymezování těchto charakteristik redundantní. Z obou kategorií jsme vybrali dva různé vzorky o stejné délce, kde přes poněkud odlišné hodnoty v rámci jedné kategorie je patrné, že neformální texty dosahují přece jen vyšších hodnot než formální.

Interpretace takových výsledků vyplývá samozřejmě z charakteru analyzovaných textů. V první řadě je třeba si uvědomit, že v případě formálních mluvených textů šlo do značné míry o monologické projevy, neboť respondent dostával určité předem stanovené otázky obecného rázu. Jednalo se tak spíše o kratší monologické texty než dialogy. Na opačné straně stojí neformální texty, jež představují zcela přirozené dialogy na různorodá témata, do nichž nebylo nijak zasahováno. Z toho vyplývá, že již širší tematické pole a více účastníků konverzace přispívá k bohatšímu lexiku, na druhou stranu právě dialogická forma vyžaduje vyšší zastoupení fatických výrazů a také opakování slov, témata těchto dialogů navíc také zpravidla neměla příliš složitá témata, šlo spíše o velmi obecné texty nasycené zejména fatickou funkcí. Je tedy patrné, že zde proti sobě stojí několik protichůdných faktorů, jejichž výsledkem je poněkud vyšší hodnota rozsahu lexika u neformálních projevů.

5.2.6 Komparace funkčních stylů





Při prvním pohledu na grafy je patrná zásadní odchylka administrativního stylu oproti ostatním, proto začneme s interpretací právě těchto hodnot. Samotný fakt, že administrativní texty mají nejužší slovník, není nijak neočekávaný, hodnota této odchylky od uměleckého, odborného a publicistického stylu však může být

překvapivá, neboť například oproti publicistice administrativa dosahuje polovičních hodnot. Vysvětlení těchto dat vychází ze specifického charakteru těchto textů, především jde velmi úzce zaměřenou roli ve společnosti, kde administrativa tvoří prostředek neosobní zcela nezbytné komunikace buď mezi institucemi a jednotlivcem, nebo mezi institucemi vzájemně. Tyto texty jsou tedy přísně citově nezabarvené a podstatné je, že obsahují jen a pouze nezbytné údaje, každá klauze má své pevné nenahraditelné místo, jejíž vypuštění by mohlo zcela porušit platnost a obsah sdělení. Je nutné si uvědomit, že administrativní texty nemají pouze informační funkci, ale také (mnohdy především) funkci perlokuce. Z toho vyplývá další specifický rys administrativního stylu, a to jejich značná schematičnost, dodržování daných formulací a výstavby textu je nezbytné pro úspěšné splnění požadované funkce. Můžeme říct, že produkce administrativních textů má často charakter doplňování proměnných do předem daného schématu. Z těchto tvrzení je zřejmé, jak se tato specifika musí nutné projevit i v lexikální rovině, včetně samotného rozsahu slovníku.

Poněkud překvapivý může být fakt, že bohatství slovníku hovorového stylu převyšuje styl administrativní, to platí v případě lemmat, ve slovoformách jsou hodnoty takřka totožné. Toto zjištění by poněkud narušovalo hypotézu o jednoznačně širším lexiku psaného projevu oproti mluvenému. Jakkoliv jsme předpokládali velmi nízké hodnoty u administrativních textů, neočekávali jsme, že se dostanou na stejnou, nebo dokonce pod úroveň hovorového stylu, v našem případě zcela reprezentovaného mluvenými projevy. V kapitole věnované vývoji funkčních stylů jsme zmínili pozdější vydělení administrativního funkčního stylu, který byl dříve zahrnut v odborném stylu (stejně tak jsou v korpusu SYN2010 zařazeny tyto texty mezi odborné). Na základě získaných dat tedy musíme jen potvrdit správnost takového vydělení, neboť z hlediska rozsahu lexika jsou tyto texty naprosto specifické a odlišné od textů oborného funkčního stylu. Odlišnost administrativního stylu samozřejmě není dána pouze lexikální rovinou, ale to již přesahuje rámec naší práce.

Na opačném pólu, tedy mezi styly s nejvyššími hodnotami rozsahu lexika je situace snad ještě složitější, na první pozici stojí publicistický styl následovaný stylem uměleckým. Abychom mohli správně interpretovat tyto výsledky, musíme

vycházet z faktu, že tyto hodnoty byly determinovány zařazením a poměrným rozložením jednotlivých kategorií uvnitř každého výběrového souboru, jenž reprezentoval určitý funkční styl. Jen stěží tak můžeme tvrdit, že publicistický styl má nejširší lexikum, jakkoliv tomu tak skutečně může být. Zde se projevuje skutečnost, na základě které jsme odmítli srovnávat funkční styly bez ohledu na jejich značnou vnitřní diferencovanost. Ze získaných dat například víme, že značně jiných výsledků dosahují bulvární a seriózní deníky, přesto jsou uvnitř jednoho funkčního stylu, v rámci beletrie podobná situace nastává například mezi básněmi a romány. Klíčovým bodem ke správnému zobecnění tedy musí být dostatečná reprezentativnost srovnávaných výběrových souborů. I když jsme se při tvorbě zkoumaných vzorků snažili maximálně reflektovat realitu, jde stále jen o model, který naráží na mnohá technická omezení, například nemožnost analýzy textů realizovaných v rozhlasu a televizi v rámci publicistického funkčního stylu.

Přes všechna možná úskalí jsme však přesvědčeni, že získaná data jsou validní a mají slinou vypovídací hodnotu, ale samozřejmě s přihlédnutím k výše nastíněným úskalím. Při zkoumání jazyka neustále narážíme na tyto překážky, při jakémkoliv výzkumu je třeba vždy jasně stanovit rámec, z něhož vychází všechny hypotézy a zjištění. I my jsme tedy sestavili jasně definovaný model při maximální snaze o simulaci reality a relevantnost výstupních dat, všechny naše závěry jsou nutně tímto modelem determinovány. Při zvážení širokého spektra různých teoretických překážek si dovolujeme tvrdit, že publicistické texty mají obecně bohatší slovník, i když rozdíl nebude příliš markantní.

Zatímco mezi publicistikou a beletrií nebyly získané hodnoty příliš odlišné, odborné texty se vydělily od těchto dvou mnohem jasněji. Ostatně již při komparaci hodnot uvnitř odborného funkčního stylu jsme zjistili poměrně nízké amplitudy jednotlivých kategorií, což jen potvrzuje značné specifické postavení těchto textů oproti ostatním. Odborný styl vyžaduje jednoznačné explicitní vyjádření, přípustná je pouze spisovná varieta češtiny; citové zabarvení, frazeologie a fatičnost jsou zapovězeny. Příznačná je také schematičnost užívání termínů, pro autora je primárním cílem, zcela nadřazeným ostatním, maximální přesnost vyjádření sledované problematiky, čtenáře takřka vypouští ze zřetele. Jak již bylo uvedeno výše, dochází zde ke kumulaci a interferenci několika jevů, jež mají protichůdné

tendence. Výsledkem je potom nižší hodnota rozsahu slovníku oproti publicistice a beletrii, ale vyšší než v případě administrativního a hovorového stylu.

5.3 Rozptýlení lexika

Rozptýlení lexika je druhou základní veličinou slovníku textu, ukazuje na zastoupení pásma jednotek s nejnižší frekvencí. Tato veličina tak úzce souvisí s bohatstvím slovníku, neboť lze předpokládat, že právě nejméně užívaná slova stojící na periferii dosahují nejvyšší entropie, a tudíž by rozptýlení mělo pozitivně korelovat s rozsahem lexika. Hypotéza je tedy taková, že čím má text bohatší slovník, tím větší zastoupení mají jednotky z periferie slovní zásoby.

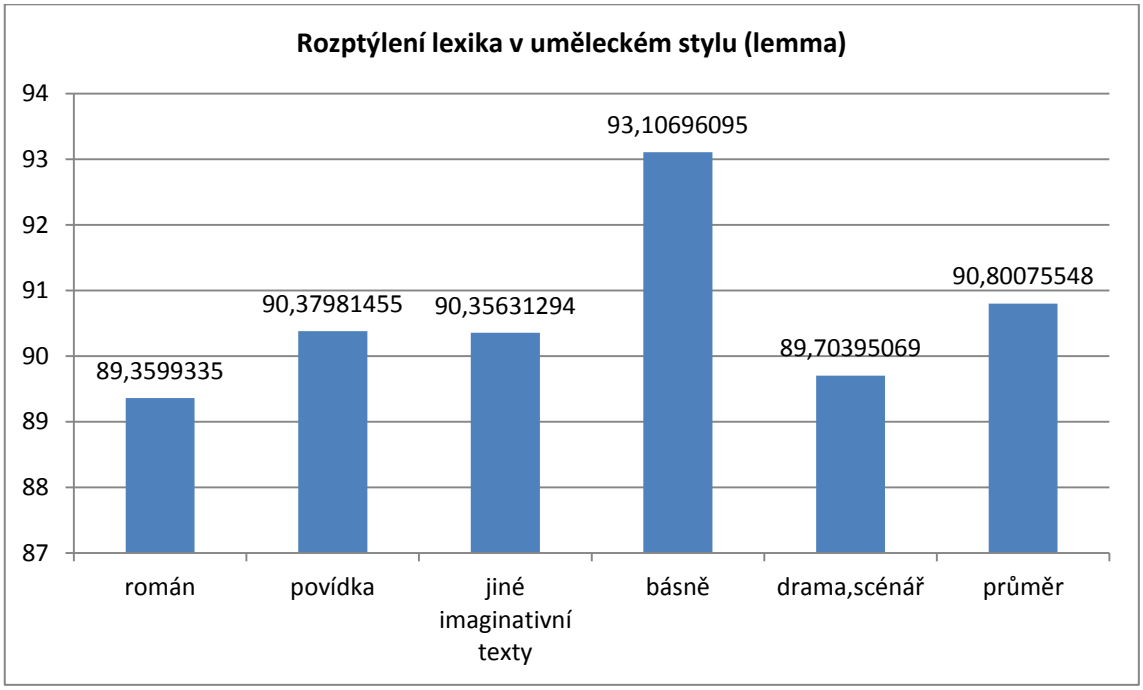
Po komparaci průběhu grafů s jejich protějšky z rozsahu lexika, jsme zjistili, že jsou takřka totožné. Tento fakt tedy potvrzuje naši hypotézu, dokonce tak můžeme říct, že obě veličiny vyjadřují totéž, není tomu však bezvýhradně. Následující grafy nebudeme komentovat, neboť předpokládáme, že faktory, které determinovaly získaná data, byly stejné jako v případě bohatství slovníku, bylo by tedy redundantní opakovat stejné závěry.

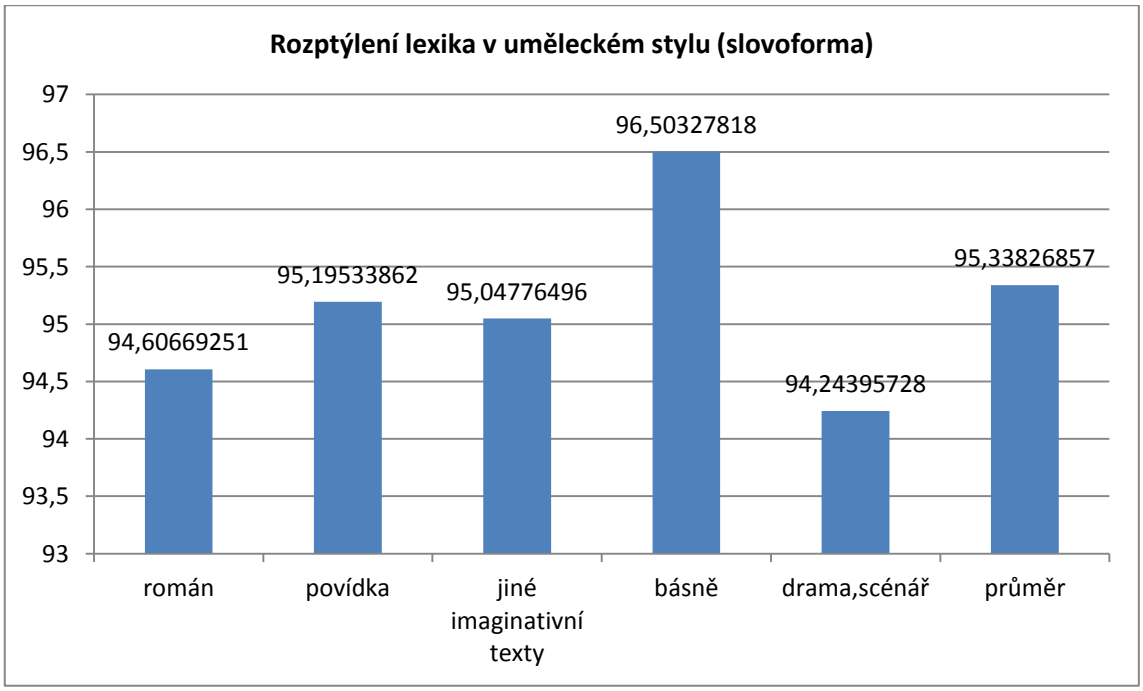
Zatímco v případě průběhu grafů uvnitř jednotlivých funkčních stylů byly stejné s rozsahem lexika, při závěrečné komparaci průměrných hodnot funkčních stylů jsme zjistili určité odchylky. Tuto skutečnost budeme reflektovat na konci této kapitoly.

Rozptýlení lexika (D) vyjadřuje poměr počtu plnovýznamových slov s frekvencí 1–10(V_1) k slovníku (V). Základní rovnice pro výpočet:

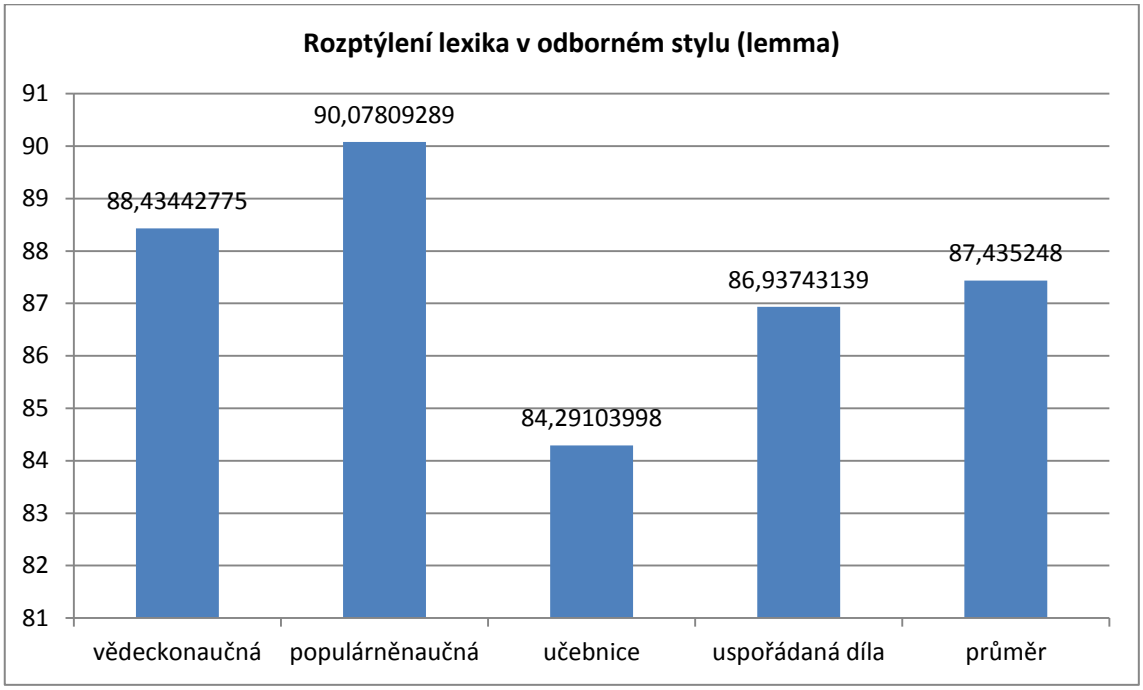
$$D = 100\ \frac{V_1}{V}$$

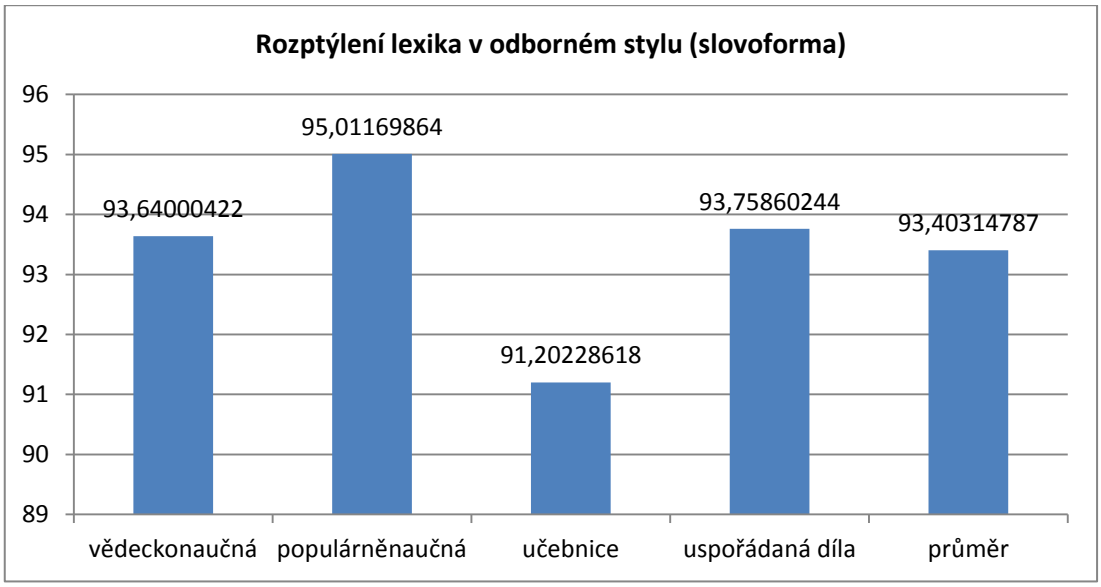
5.3.1 Umělecký styl



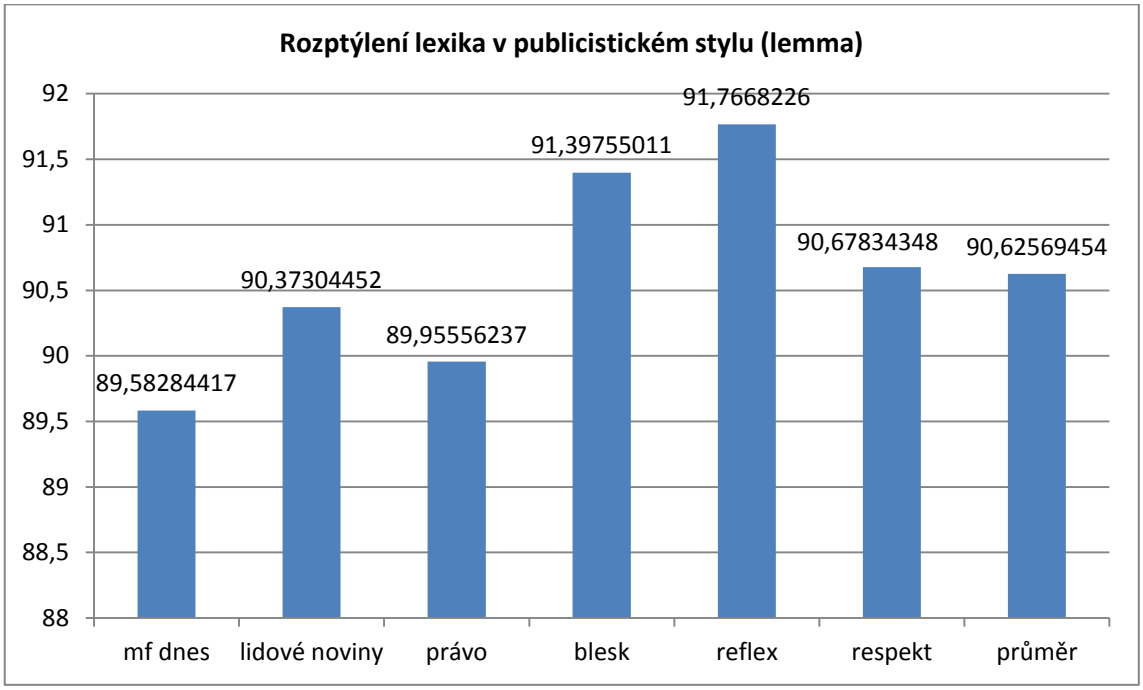


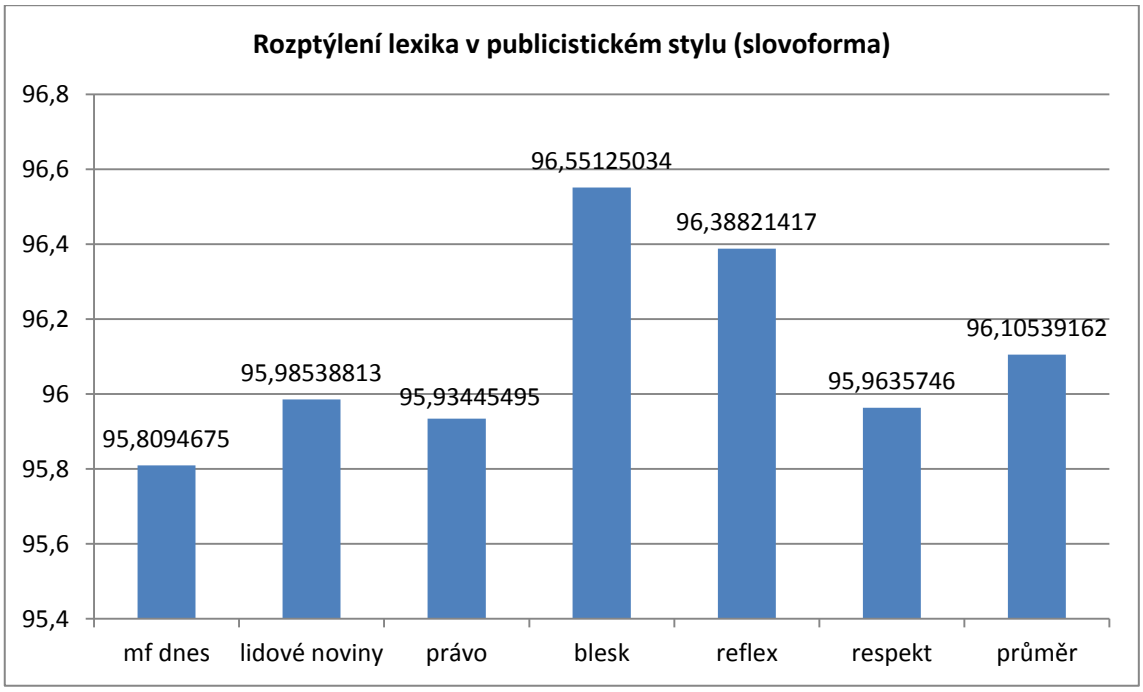
5.3.2 Odborný styl





5.3.3 Publicistický styl



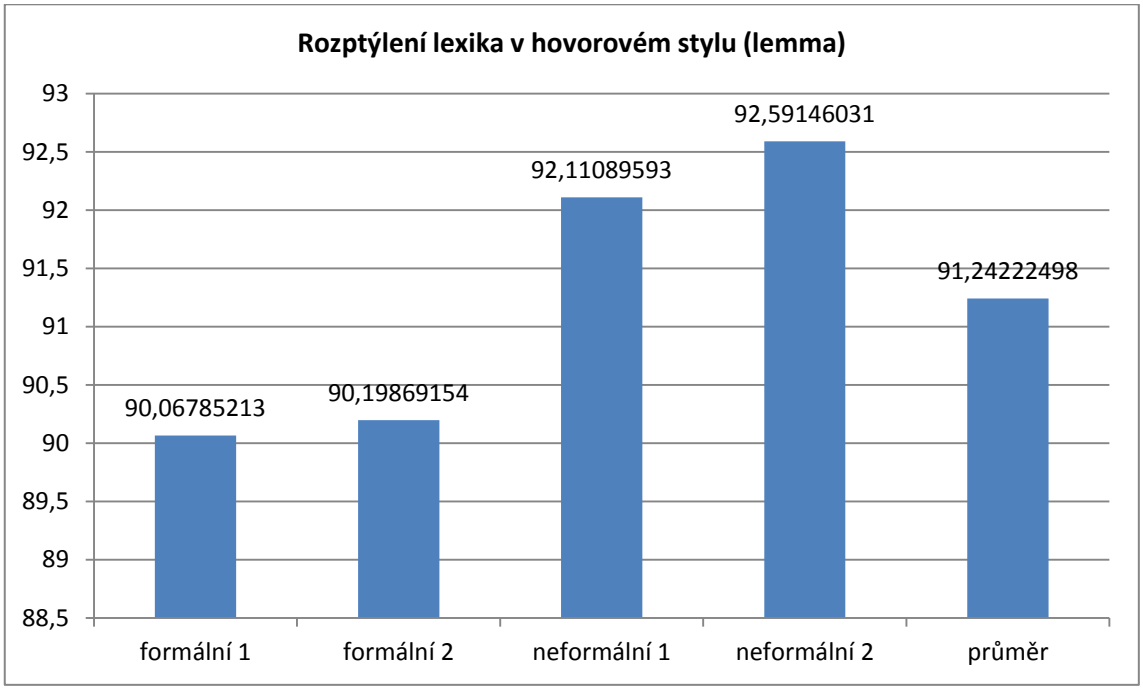


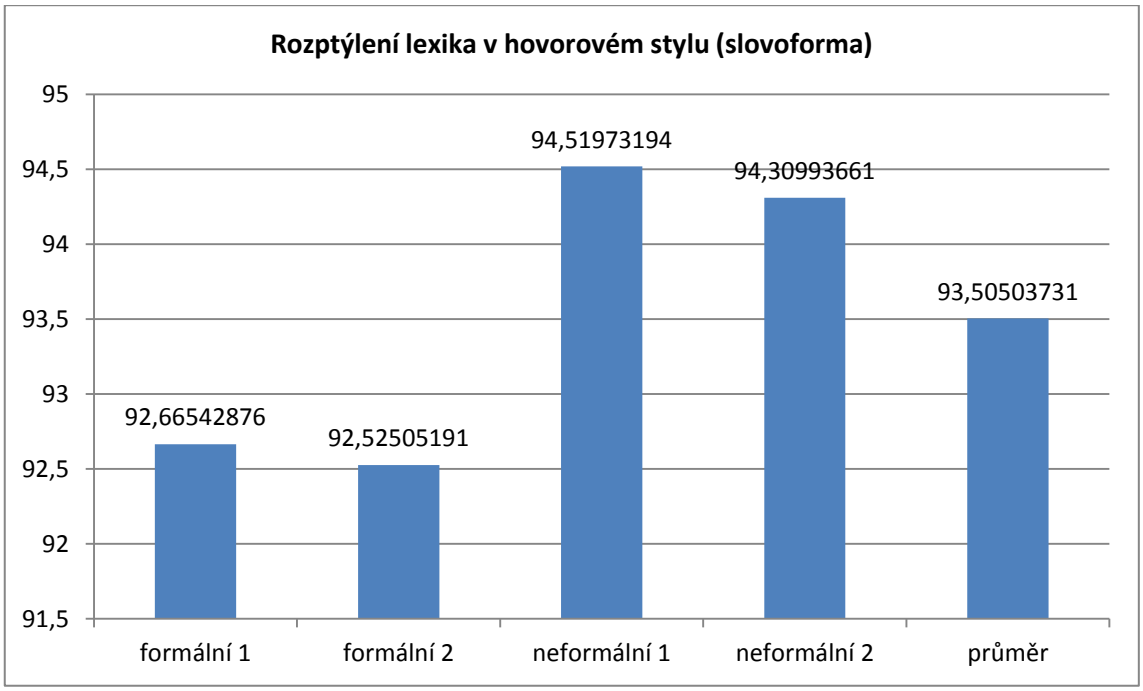
5.3.4 Administrativa

D (lemma) = 82,39878

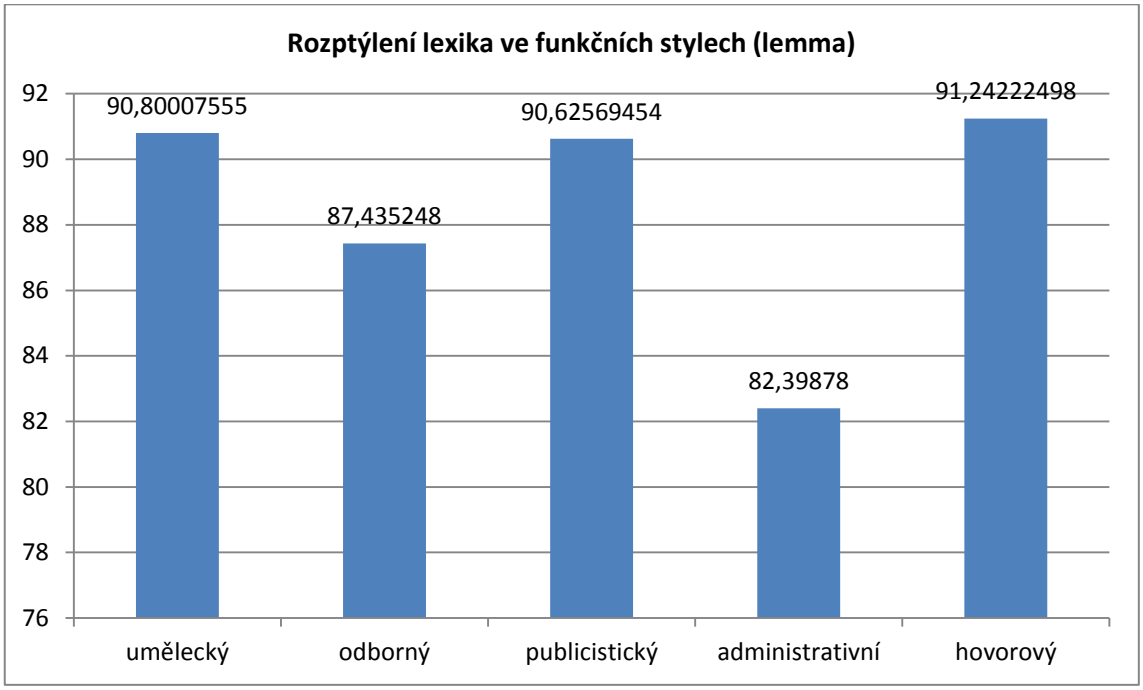
D (slovoforma) = 89,49566

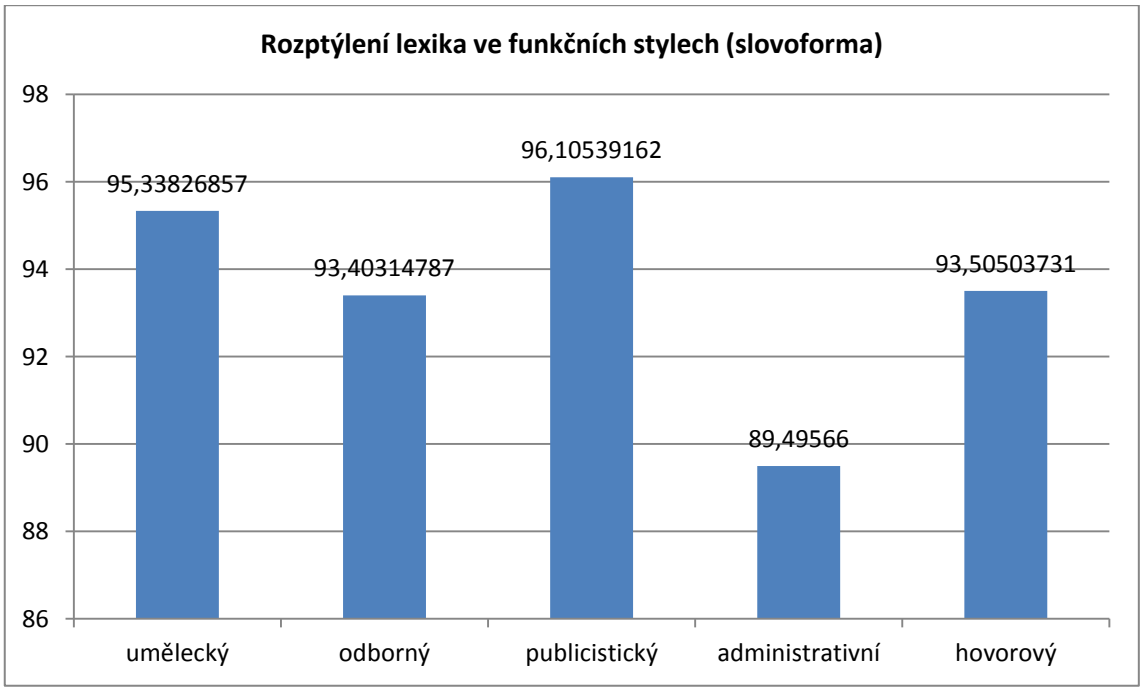
5.3.5 Hovorový styl





5.3.6 Komparace funkčních stylů





Přestože jsme z důvodu téměř stejných průběhů grafů jako v případě rozsahu lexika upustili od komentářů a interpretací jednotlivých funkčních stylů, při komparaci celých funkčních stylů je situace jiná. Při výpočtech s lemmaty dokonce hovorový styl vykazuje nejvyšší hodnotu, což zcela odporuje výsledkům v bohatství

slovníku. Vzhledem k tomu, že při výpočtech ve slovoformách tomu již tak není, můžeme tuto odchylku částečně přisoudit poněkud odlišnému způsobu lemmatizace korpusu PMK.

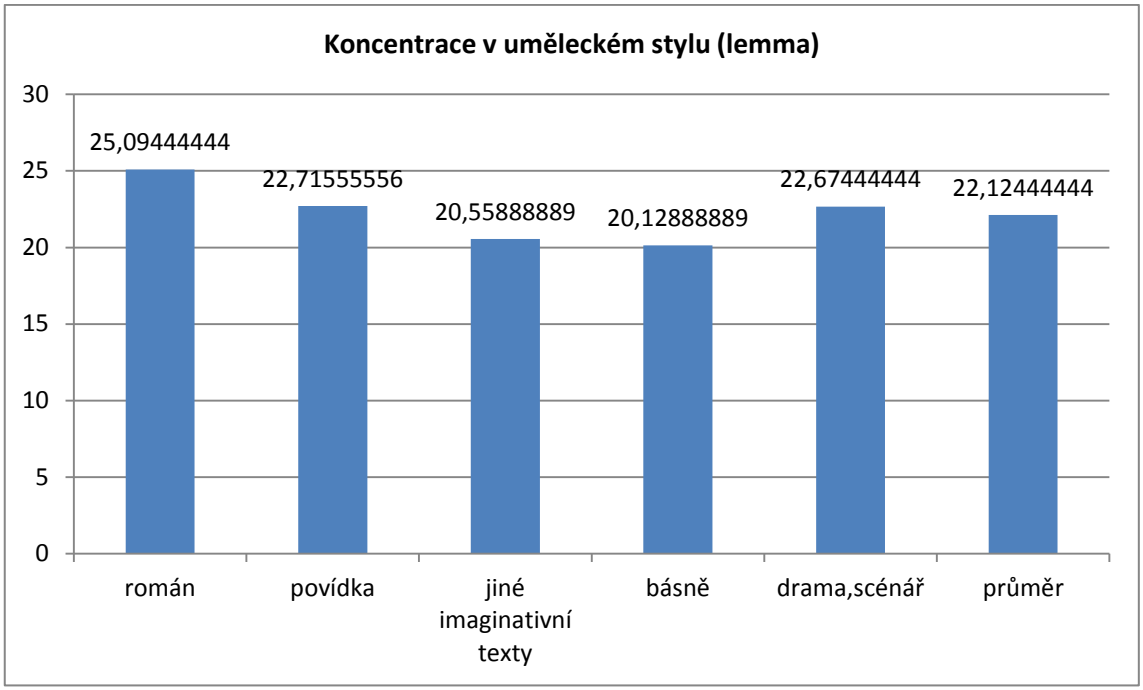
5.4 Koncentrace lexika

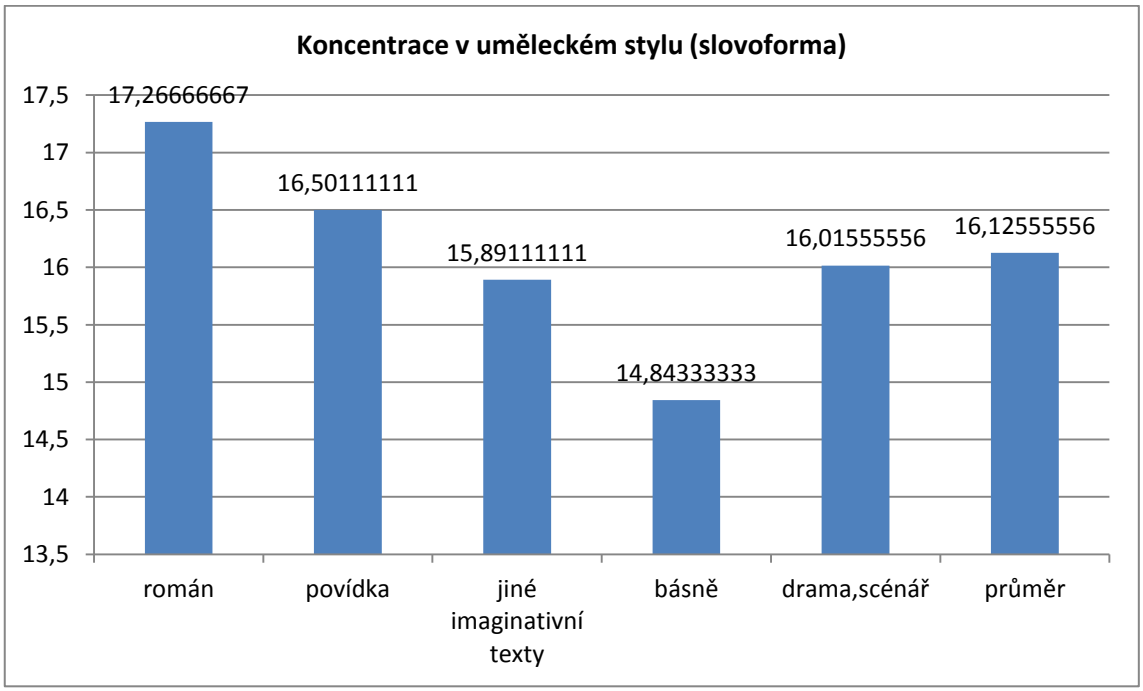
Tato veličina završuje ústřední trojici, jež podává základní charakteristiku textu z hlediska lexikální statistiky, a nutně tedy souvisí s dvěma předešlými. Zatímco rozsah lexika vystihuje nejkomplexnější údaj o slovníku a rozptýlení se pak soustředí na periferní oblast lexika, koncentrace se zaměřuje na opačný pól slovní zásoby – její samotné centrum. Tento parametr vystihuje vztah mezi délkou textu odpovídající deseti nejfrekventovanějším jednotkám a celkovou délkou textu. Přestože tedy nejde o přímý protiklad koncentrace lexika, je logické mezi těmito veličinami očekávat určitou míru negativní korelace.

Koncentrace lexika (K) vyjadřuje poměr délky textu odpovídající prvním 10 nejfrekventovanějším slovům (N_1) k celkové délce textu (N). Základní rovnice pro výpočet:

$$K = 100\ \frac{N_1}{N}$$

5.4.1 Umělecký styl





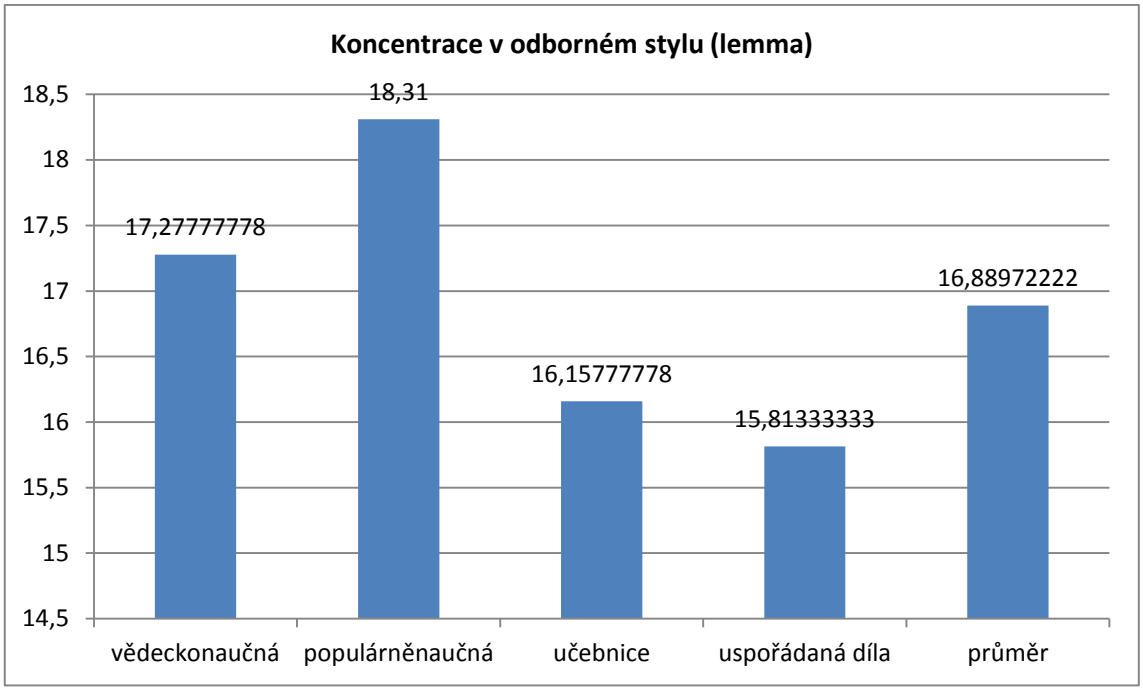
Nejnižších hodnot dosahují básnické texty, což odráží jejich mnohem vyšší disperzi. Centrum slovní zásoby tvoří gramatická slova, jejichž dominantnost se právně v těchto textech oslabuje, neboť není třeba striktně dodržovat normu, navíc se v básních tvoří zpravidla velmi jednoduché syntaktické útvary, jež si nevynucují tak časté užití těchto slov. Pro tyto texty je charakteristické užívání širokého spektra

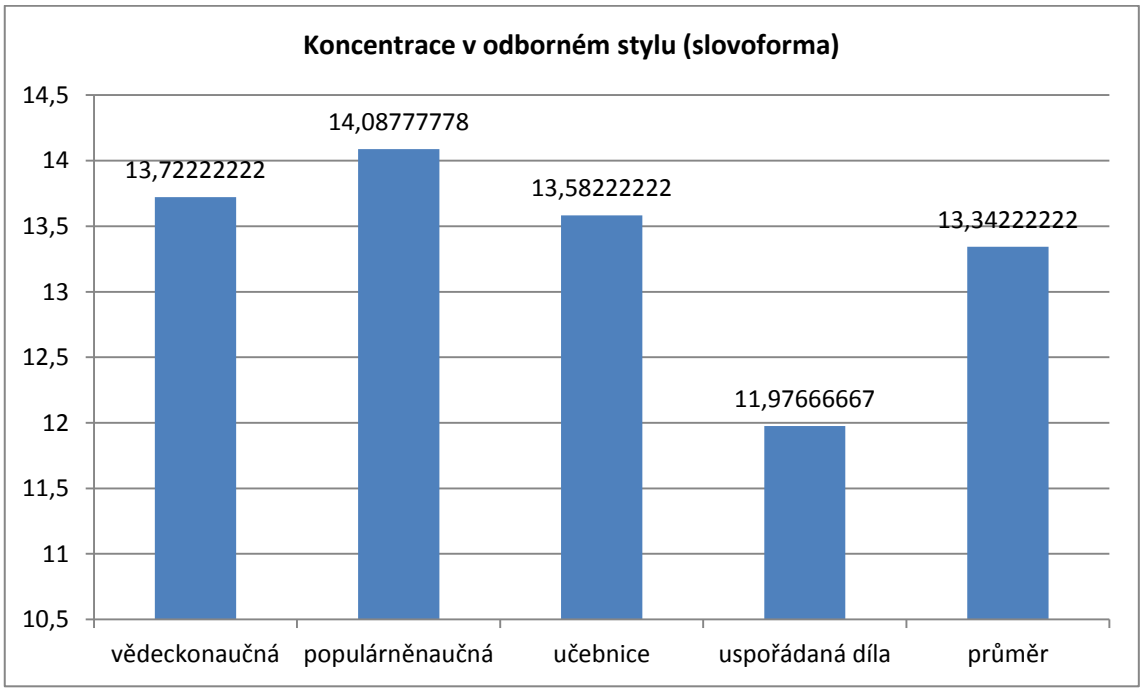
slovní zásoby, zejména těch částí, které v ostatní komunikaci nejsou příliš četná. Milan Jelínek hovoří o poetických textech jako o esteticky vůbec nejúčinnějších v rámci uměleckého funkčního stylu. K lexikální rovině těchto útvarů se vyjadřuje takto: „Lyričnost poezie a lyrizovanost epických poetických textů má ovšem své důsledky ve výběru výrazových prostředků: v poezii vzrůstá podíl lexikálních prostředků označujících city a prožitky, zvyšuje se účast postupů významově aktualizačních a ze všech uměleckých stylů se klade největší důraz na novost stránky vyjadřovací.“[40]

Na opačném pólu stojí romány, které naopak vykazují nejsložitější syntaktické celky, jež logicky vyžadují velkou četnost gramatických prostředků, navíc tyto texty zpravidla pevně dodržují normu. Významnou roli zde patrně také hraje skutečnost, že romány jsou vůbec nejdelší texty v rámci uměleckého stylu, tvořené jedním autorem, který má svůj vlastní styl a centrum užívaných výrazových prostředků se tak musí v komparaci s ostatními uměleckými díly mnohem více projevit. Povídky jakožto v podstatě kratší obdoby románu pak logicky obsadily druhou pozici v koncentraci lexika. Jiné imaginativní texty a dramata se scénáři pak dosahují poněkud nižších hodnot, jejich odchylka však není příliš velká. Základním zjištěním tak je opozice román-báseň, v tomto rozmezí se pak pohybují ostatní kategorie v závislosti na jejich charakteru blížícímu se jednomu či druhému útvaru.

5.4.2 Odborný styl

[40] Karlík, P. – Nekula, M. – Pleskalová, J. (eds.).: *Encyklopedický slovník češtiny*. Praha 2002, s. 456.





Nejvyšších hodnot dosahuje v rámci odborného funkčního stylu populárně naučné texty, jež se svou povahou liší od vědeckých, které představují v podstatě čistou formu odborného stylu. Populárně naučné texty se obsahově shodují s ostatními odbornými texty, avšak formálně využívají prvků publicistického, popř. uměleckého stylu. Tento útvar je tak charakteristický svou interferencí několika různých stylů,

přičemž tento aspekt patrně hraje důležitou roli ve výsledných hodnotách. Milan Jelínek tvrdí, že „aby autor dosáhl srozumitelnosti textu, musí často obětovat jeho přesnost a volit větší míru textové explicitnosti. Někdy se u textů tohoto druhu uplatňuje i funkce získávací (autor chce čtenáře přesvědčit o poznatcích, které sděluje). Vedle průniku se stylem publicistickým dochází i k průniku se stylem uměleckým.“[41]

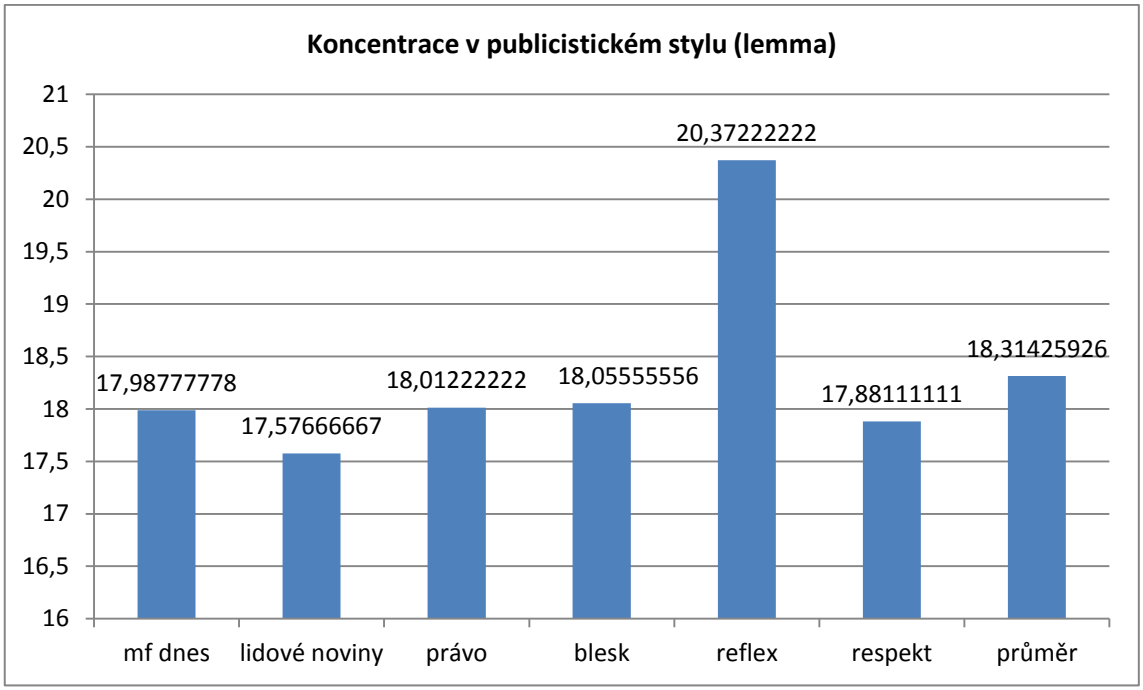
Značně specifické místo uvnitř odborného stylu zaujímají učebnice, Marie Čechová upozorňuje na záměrné působení na adresáta, které tvoří vedle samotné informativní funkce důležitý aspekt konstituující učební styl. Dále autorka dodává, že „stylizace učebních textů je ovlivněna i tím, že učebnice, především školní, tvoří systém, jenž musí obsahem i formou splňovat nejen požadavky odborné, ale i didaktické.“[42] Je třeba si uvědomit, že tyto texty musí vzhledem ke své funkci splňovat určité parametry. Učebnice obsahují různé útvary, jež se v jiných odborných textech zpravidla nevyskytují, jedná se zejména o různá cvičení, doplňující otázky, různé shrnující výčty důležitých pojmů apod. Všechny uvedené aspekty se projevují i do lexika, které má úzký rozsah slovníku (viz výše) a zároveň nižší koncentraci.

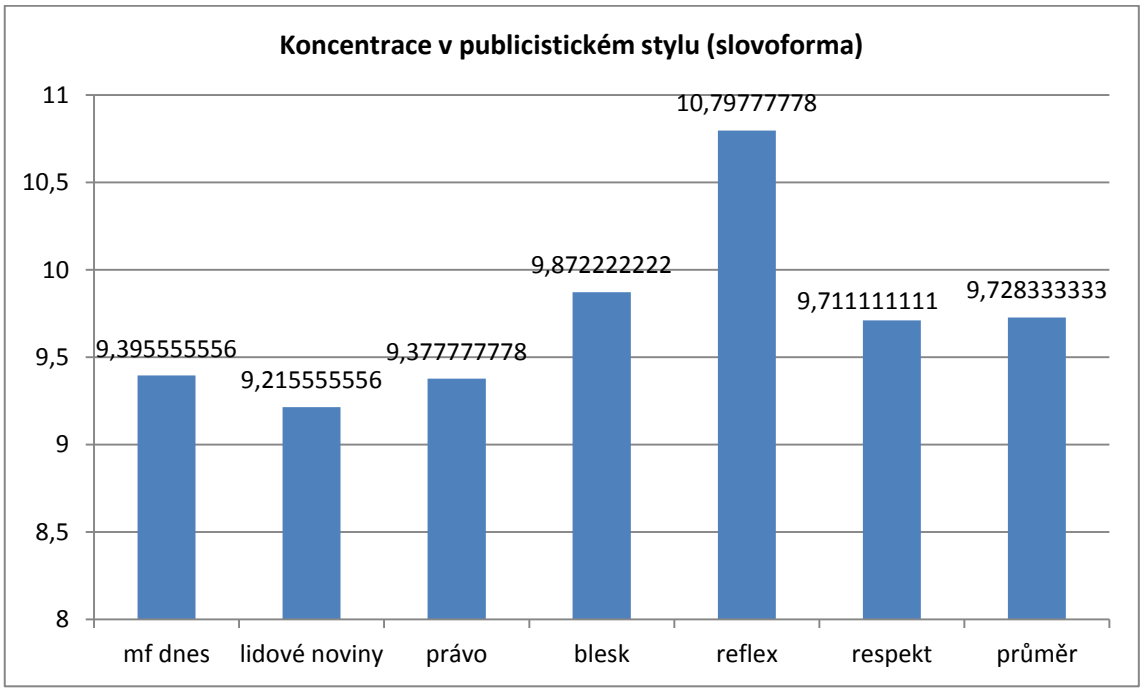
Nejnižších hodnot dosáhla kategorie „uspořádaná díla“, její umístění můžeme přisuzovat zejména faktu, že jde o nejrůznější díla sdružující více kratších textů různých autorů. Právě tento fakt je patrně klíčový, neboť vycházíme z hypotézy, že počet autorů jednoho textu negativně koreluje s koncentrací lexika. Každý autor má svůj charakteristický styl, včetně nejužívanějších slov, která tvoří centrum jeho slovní zásoby, pokud kompilujeme texty více autorů, musí se to v námi sledovaných hodnotách projevit.

5.4.3 Publicistický styl

[41] Karlík, P. – Nekula, M. – Pleskalová, J. (eds.).: *Encyklopedický slovník češtiny*. Praha 2002, s. 455.

[42] Čechová M. a kol.: *Současná česká stylistika*. Praha 2003, s. 189.



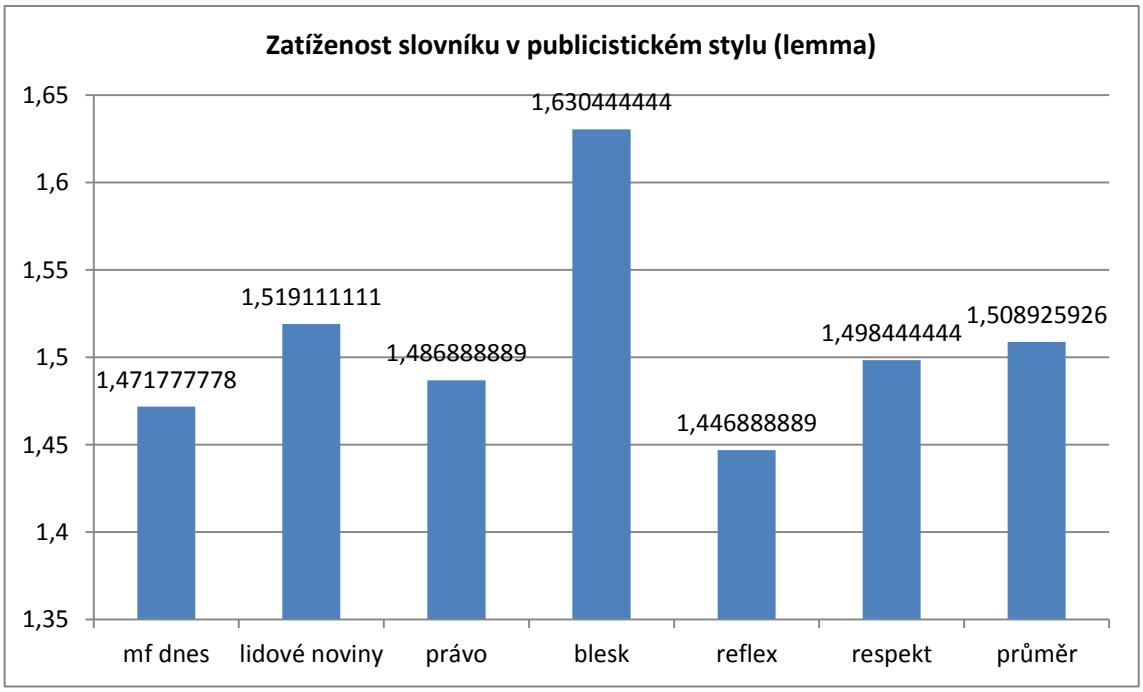


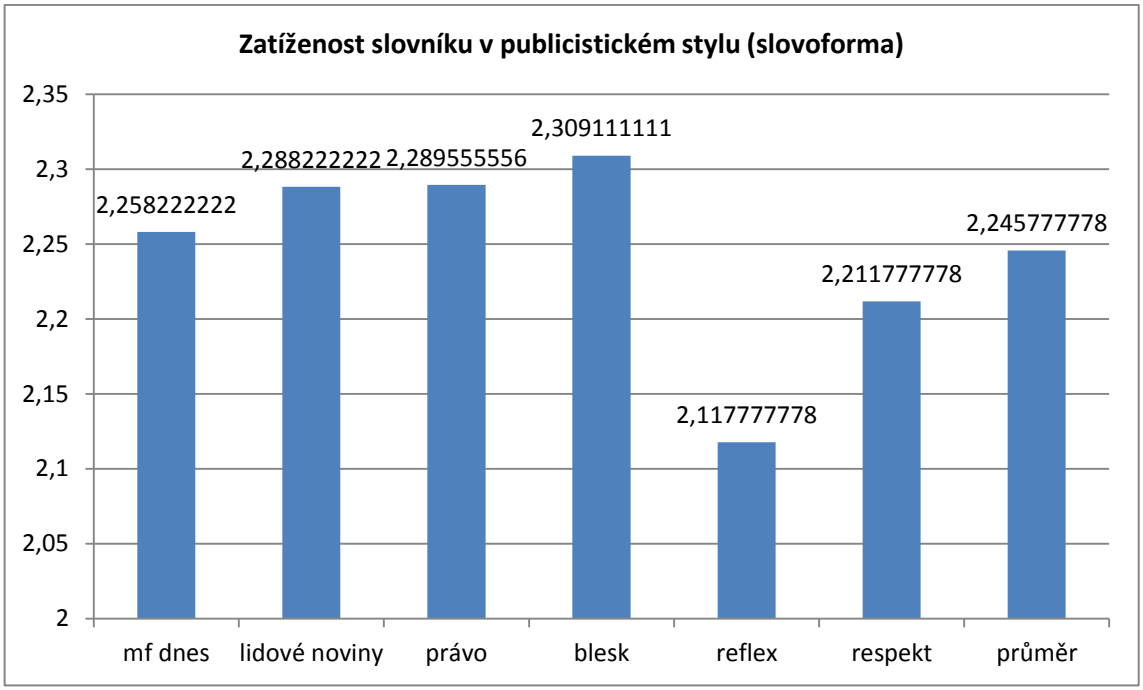
Průběh grafu uvnitř publicistického stylu má značně překvapivý průběh, jednotlivá periodika dosahují podobných hodnot, kromě jedné – týdeníku Reflex. Jen stěží hledáme vysvětlení pro tuto skutečnost, neboť obdobný týdeník Respekt dosahuje podstatně nižší hodnoty, velmi podobné bulvárnímu deníku Blesk. Kruciální otázkou je, čím se právě Reflex tolik odlišuje od ostatních periodik. Výše

jsme již uvedli, že tento týdeník oproti svému hlavnímu konkurentu inklinuje spíše k bulvárnosti, tím jsme částečně vysvětlili jeho bohatší slovník, neboť se ukázalo, že čistě bulvární deník Blesk měl v rámci publicistického funkčního stylu suverénně nejvyšší hodnotu rozsahu lexika, v tomto případě tomu však tak není.

Získané hodnoty jsme se tedy rozhodli ověřit pomocí parametru blízkého charakteru koncentrace lexika, jedná se o zatíženost slovníku (I_g). Tuto veličinu stanovil Jozef Mistrík jako základní lexikostatistickou charakteristiku. Do výpočtu jsou zahrnuta slova, která autor užil tzv. suverénně, což znamená, že se v textu objevili více než jednou ($L_{f1<}$). Rovnice pro výpočet je následující:

$$I_g = \frac{2\,L_{f1<}}{\frac{1}{10}\,N}$$





Grafy zatíženosti slovníku potvrzují zvláštní disproporci uvnitř textů Reflexu. Při zahrnutí všech slov s frekvencí vyšší než 1 vzhledem k délce textu ostatní periodika vykazují srovnatelné hodnoty, včetně Blesku, který má vedoucí pozici. Zdá se tedy, že ona disproporce vyplývá z hodnot frekvencí v pásmu s nejvyšší četností. Pro ověření této hypotézy bude nejprůkaznější tabulka s deseti nejužívanějšími slovy.

Deset nejfrekventovanějších slov (lemma)												
	MF DNES		LIDOVÉ NOVINY		PRÁVO		BLESK		REFLEX		RESPEKT	
1	BÝT	2931	BÝT	3139	BÝT	3193	BÝT	2911	BÝT	3995	BÝT	2970
2	V	2811	V	2397	V	2642	SE	2393	A	2649	A	2536
3	SE	2260	SE	2168	A	2137	V	2348	SE	2452	SE	2190
4	A	2028	A	2117	SE	1990	A	2061	V	2161	V	2140
5	NA	1896	NA	1556	NA	1713	NA	1915	TEN	2010	NA	1635
6	TEN	1110	TEN	1246	TEN	1189	TEN	1360	NA	1565	TEN	1278
7	Z	894	ŽE	862	ŽE	946	Z	901	ŽE	1140	ŽE	891
8	S	791	Z	814	Z	813	S	833	S	805	Z	884
9	KTERÝ	789	KTERÝ	776	S	812	ON	782	Z	795	KTERÝ	799
10	I	679	S	744	KTERÝ	776	ŽE	746	MÍT	763	S	770
		16189		**15819**		**16211**		**16250**		**18335**		**16093**

Deset nejfrekventovanějších slov (slovoforma)												
	MF DNES		LIDOVÉ NOVINY		PRÁVO		BLESK		REFLEX		RESPEKT	
1	SE	2047	SE	1964	SE	1868	SE	2115	SE	2069	SE	2056
2	NA	1800	NA	1550	NA	1723	NA	1904	NA	1565	NA	1650
3	JE	819	ŽE	937	ŽE	973	JE	805	ŽE	1213	JE	1027
4	TO	673	JE	918	JE	908	ŽE	796	JE	1153	ŽE	979
5	ŽE	672	TO	617	DO	650	TO	752	TO	993	TO	672
6	DO	635	DO	559	TO	625	DO	671	JSEM	698	DO	549
7	SI	526	ALE	463	VE	450	SI	572	SI	575	ALE	530
8	ALE	465	SI	443	ALE	429	ALE	475	ALE	532	ZA	438
9	VE	464	BY	427	BY	413	ZA	439	DO	523	BY	428
10	ZA	355	VE	416	ZA	401	VE	356	TAK	397	SI	411
		8456		**8294**		**8440**		**8885**		**9718**		**8740**

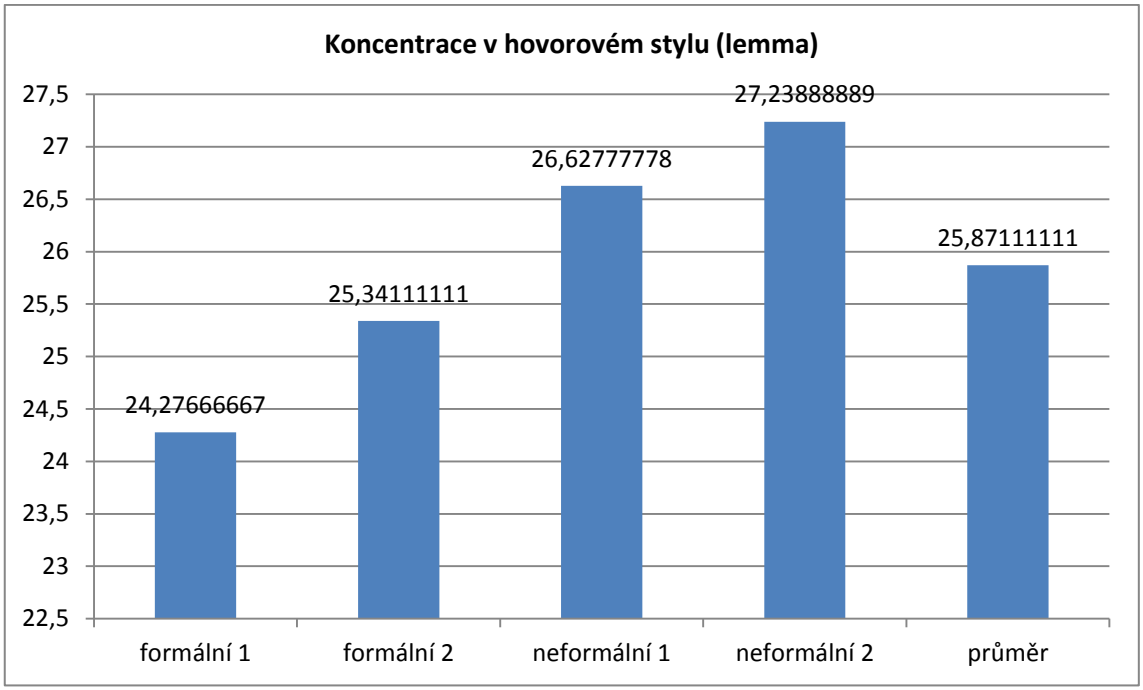
Zejména z výsledných hodnot sum prvních deseti nefrekventovanějších slov je zcela patrná odchylka u Reflexu, a to jak v případě lemmat, tak u slovoforem.

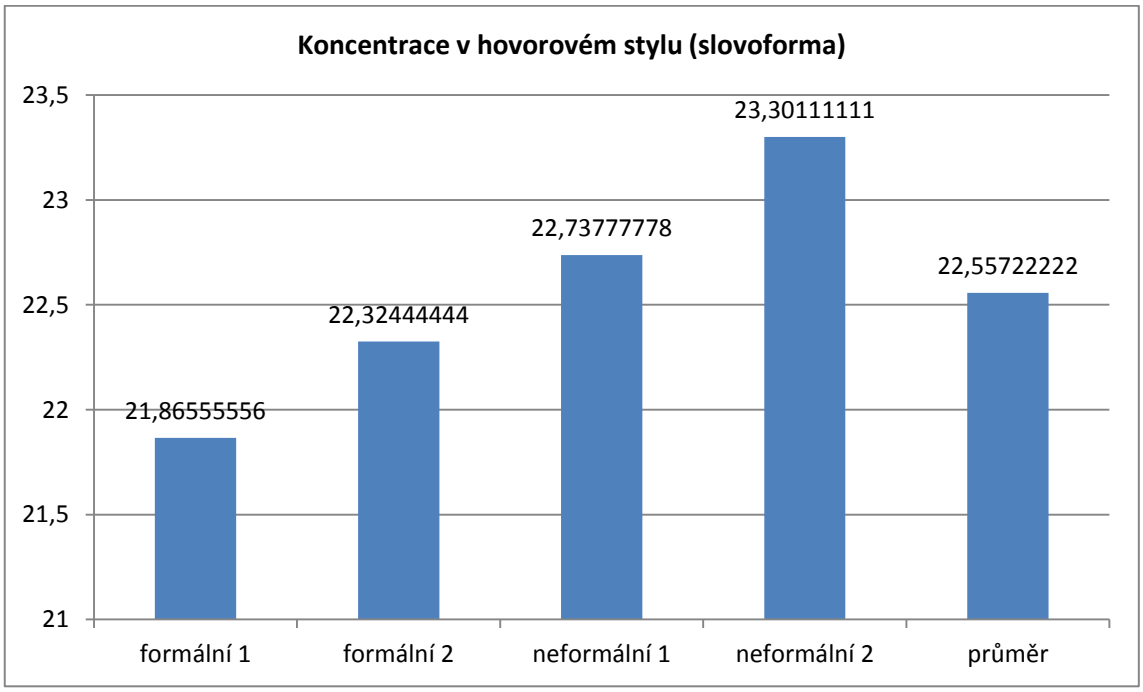
5.4.4 Administrativní styl

K (lemma) = 17,09667

K (slovoforma) = 14,04778

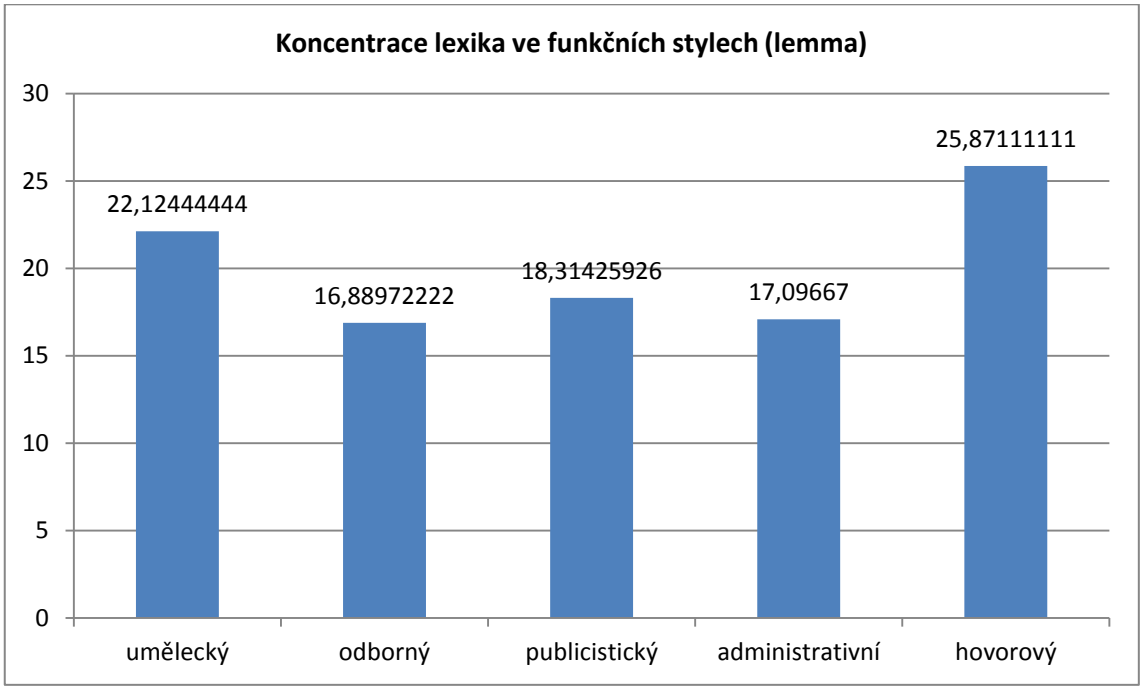
5.4.5 Hovorový styl

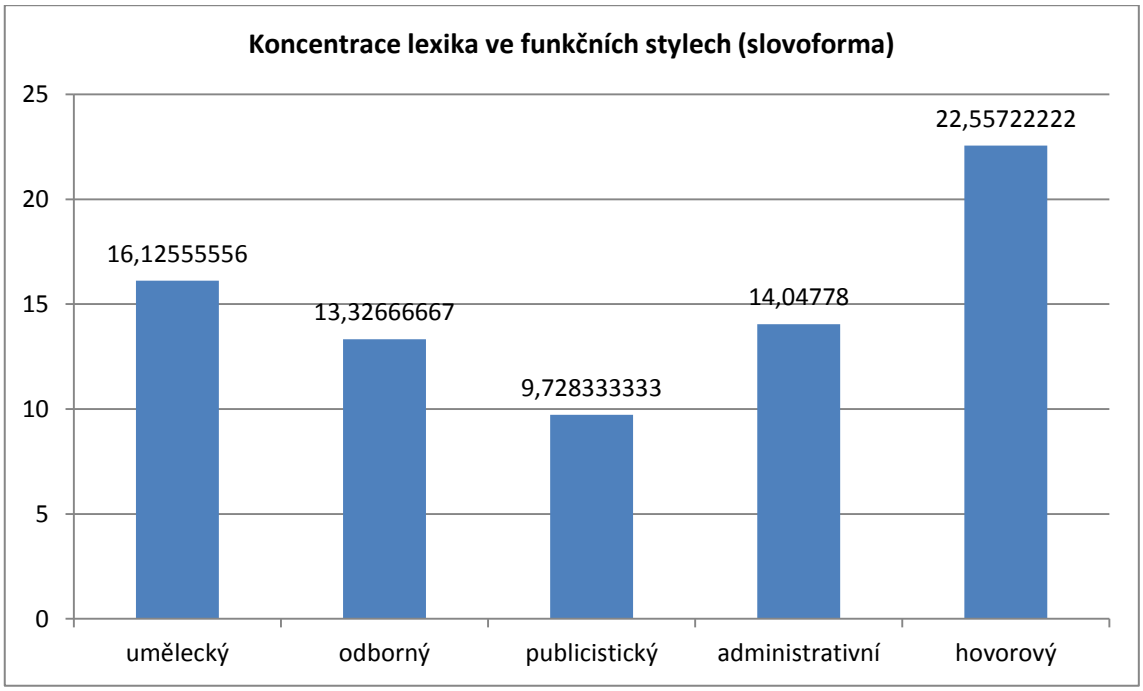




Grafy nasvědčují tomu, že hovorový styl je v případě koncentrace lexika poměrně rozkolísaný. Avšak přes tuto skutečnost můžeme přeci jen poměrně jasně sledovat rozdíl mezi námi stanovenými dvěma kategoriemi určitý rozdíl, formální promluvy evidentně dosahují nižší koncentrace lexika.

5.4.6 Komparace funkčních stylů





5.5 K problematice délky textu

Při interpretacích výsledných hodnot jsme často narazili na fakt, že se do sledovaných parametrů musí projevit nestejná délka zkoumaných textů. Nemáme teď na mysli velikost výběrových souborů v rámci jednotlivých kategorií, ty jsme vytvořili co do celkové délky naprosto totožné (90 000 textových slov). Narážíme na rozličnou délku a počet jednotlivých textů uvnitř těchto stejně dlouhých výběrových souborů. Je patrné, že nám například vyjdou zcela jiné hodnoty bohatství slovníku, pokud bude jeden výběrový soubor sestávat z jednoho koherentního textu, například románu, a zcela jiný výsledek bude generovat výběrový soubor o stejné celkové délce, avšak tvořen například deseti úryvky různých románu. V druhém případě můžeme předpokládat podstatně širší lexikum.

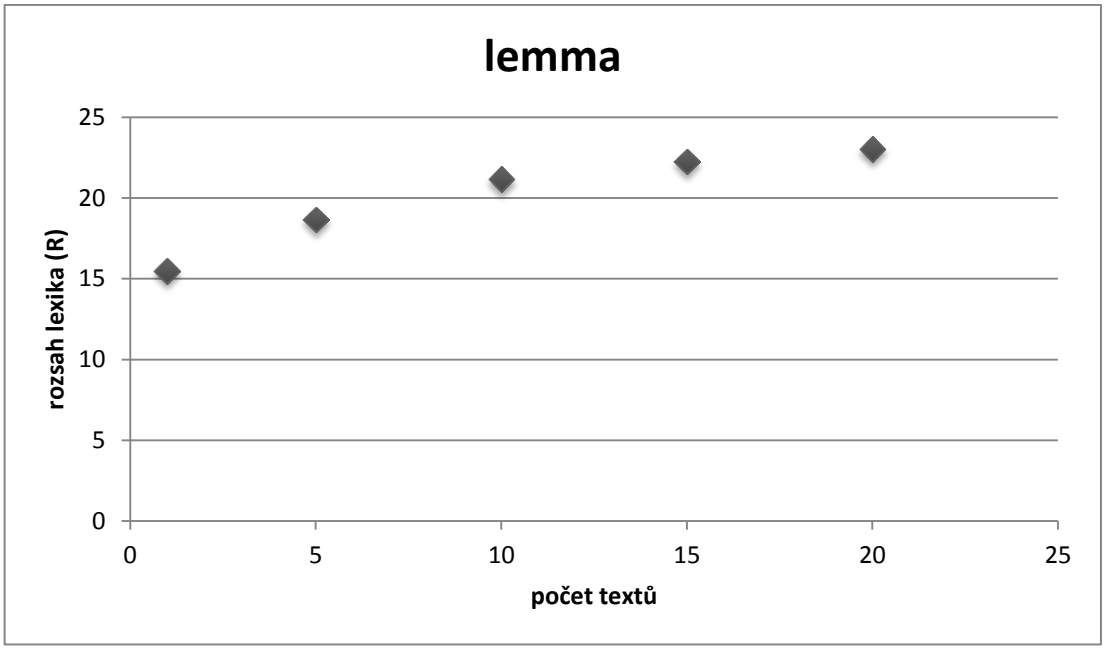
Dostáváme se tak do určitého, možná zdánlivého, rozporu s danými zásadami správného statistického šetření, neboť do vzájemné komparace proti sobě sice stavíme stejně dlouhé výběrové soubory, ale s jinou strukturou, například dlouhé románové texty proti krátkým novinovým článkům. Jakkoliv považujeme hypotézu o pozitivní korelaci počtu různých textů uvnitř jednoho výběrového souborů s hodnotou rozsahu lexika za pravdivou, předpokládáme, že přímá úměra zde nebude absolutní. Kruciální otázkou v tomto případě proto není ani tak samotná existence popsaného vztahu, ale spíše jeho kvantifikace, zjištění jak moc tato skutečnost ovlivňuje výsledné hodnoty našeho výzkumu. Rozhodli jsme se proto ověřit a zároveň matematicky vyjádřit, jaký průběh tento vztah ve skutečnosti má.

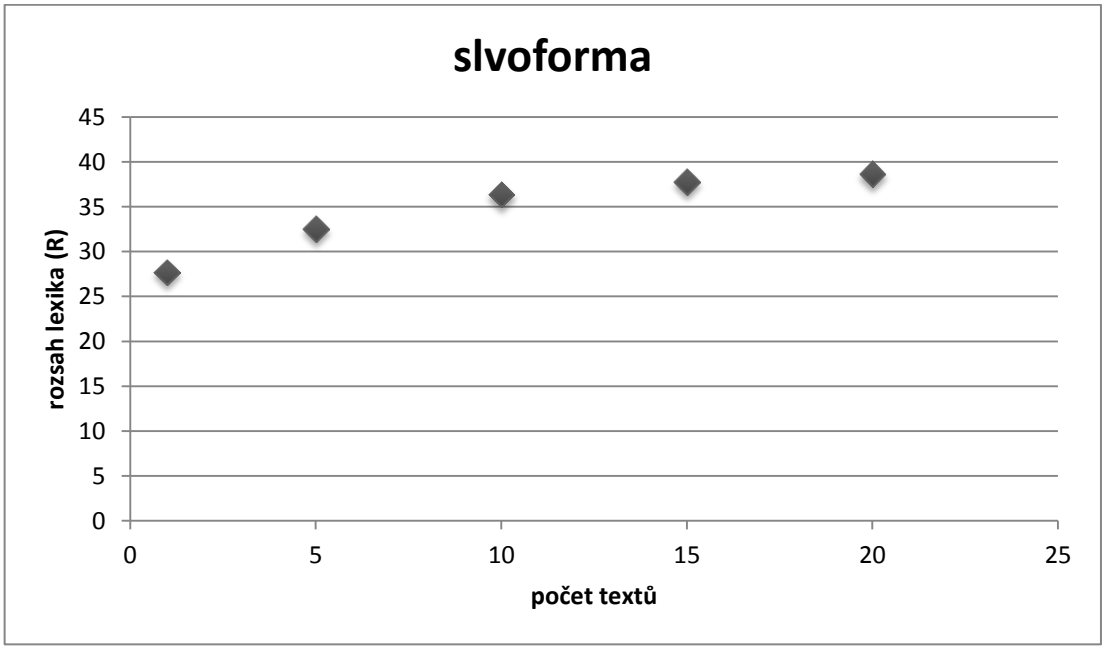
Pro analýzu jsme vybrali 20 románů z databáze korpusu SYN2010, které se nevyznačují žádnou specifickou tematikou či uspořádáním s cílem maximálně zabránit různým odchylkám daných specifickými texty, přitom jsme vycházeli zejména z třídění SYN2010, kde všechny vybrané texty spadali do kategorie [NOV,X,B]. Nejdříve jsme stanovili velikost výběrových souborů na 50 000 textových slov, poté jsme vypočítali rozsah lexika (R) pro text, jenž byl sestaven pouze z jednoho románu. Dále jsme již postupně rozšiřovali strukturu výběrových souborů o větší počet úryvků z románů, a to na 5, 10, 15 a 20, celková velikost výběrového souboru však zůstala samozřejmě vždy zachována. Pro jednotnost celé práce jsme opět paralelně počítali jak v lemmatech, tak ve slovoformách. Výsledná

data jsme nejdříve pro přesnost uvedli v číselných hodnotách a následně tyto hodnoty vynesli do grafu, ze zřejmých důvodů jsme rozhodli pro jednoduchý bodový graf.

LEMMA		
POČET TEXTŮ	**DÉLKA JEDNOTLIVÝCH TEXTŮ**	***R* CELÉHO VÝBĚROVÉHO SOUBORU**
1	50 000	15,46
5	10 000	18,6525
10	5 000	21,135
15	3 333	22,2475
20	2 500	23,03

SLOVOFORMA		
POČET TEXTŮ	**DÉLKA JEDNOTLIVÝCH TEXTŮ**	***R* CELÉHO VÝBĚROVÉHO SOUBORU**
1	50 000	27,675
5	10 000	32,5025
10	5 000	36,325
15	3 333	37,7325
20	2 500	38,,62





Ze získaných dat je zcela zřejmé, že sledovaný vztah má poměrně silnou stoupající tendenci, o absolutní přímou úměru se však v žádném případě nejedná. Přesto můžeme vidět, jak výrazně nám narůstá rozsah lexika, a získaná data tak ukazují, že rozhodně nejde o zanedbatelné změny, právě naopak. Jestliže mezi jedním a dvaceti použitými texty je již rozdíl 7,57 v případě lemmat a 10,945 v případě slovoforem, je patrné, že srovnávání např. románů a novinových článků je

velmi problematické. Nyní víme, do jaké míry přibližně (přibližně proto, že jsme si samozřejmě všech okolností, jež mohli náš experiment ovlivnit) se tento vztah promítá do výsledků. Dostáváme se tak k základní otázce, kterou jsme již naznačili v kapitole 5.2.1, totiž zda by takový zásah do výzkumu měl význam, neboť bychom porušili přirozené celky koherentních textů, a de facto bychom tak narušili právě ty struktury, jejichž jedinečnost zkoumáme.

6 LEMMA VS. SLOVOFORMA

6.1 Wilcoxonův test

Výše jsme uvedli výsledky všech měření vždy paralelně v lemmatech a slovoformách, přičemž bylo možno sledovat značnou podobnost mezi každou dvojicí grafů. Nyní provedeme statistické šetření pro zjištění závislosti výsledků získaných ze slov systémových a textových. Pro analýzu jsme zvolili Wilcoxonův test, který umožňuje porovnávat výsledky dvou různých měření provedených na stejném výběrovém souboru. Wilcoxonův signed-rank test (WSRT) je v podstatě neparametrickým ekvivalentem nezávislého t-testu. Obvykle se WSRT používá pro vyhodnocení měření před a po zásahu, pro náš účel však také vyhovuje.

6.2 Postup výpočtu WSRT

Postup výpočtu pomocí WSRT pro přehlednost nejdříve znázorníme v následujícím schématu:

1. • Stanovíme alternativní (H_1) a nulovou (H_0) hypotézu.
2. • Zjistíme rozdíly mezi párovými hodnotami (Z) , nulové rozdíly z dalšího hodnocení vyřadíme.
3. • Rozdíly seřadíme vzestupně v abosulutních hodnotách (n – počet párů s nenulovým rozdílem).
4. • Rozdílům přiřadíme pořadové číslo.
5. • Vypočítáme testovací kritérium (W), které získáme takto: W = min (W+, W-), kde W+ znační součet pořadí odpovídajících kladným rozdílům a W- součet pořadí odpovídajících záporným rozdílům.
6. • Testovací kritérium (W) srovnáme z tabulkou kritických hodnot pro významovou hladinu 0,05 a potvrdíme či vyvrátíme nulovou hypotézu (H_0).

Pro srovnání použijeme výsledky všech měření rozsahu lexika získaných z korpusu SYN2010. Nyní stanovíme naše hypotézy:

- H_0: Není žádný vztah mezi R_{lem} a R_{slo}.
- H_1: Existuje vztah mezi R_{lem} a R_{slo}.

(R_{lem} – hodnota rozsahu lexika měřená v lemmatech, R_{slo} – hodnota rozsahu lexika měřená ve slovoformách.)

6.3 Vypočet WSRT

		R_{lem}	R_{slo}	*Z*	pořadí
1	román	13,36667	25,44306	-12,0764	-5
2	povídka	15,57778	29,08056	-13,5028	-9
3	imaginativní	17,93056	30,82222	-12,8917	-8
4	báseň	20,45139	35,58889	-15,1375	-10
5	drama	13,74583	24,70833	-10,9625	-3
6	vědeckonaučná	14,975	26,35833	-11,3833	-4
7	populárněnaučná	16,89583	29,68056	-12,7847	-7
8	učebnice	10,80417	20,4125	-9,60833	-2
9	uspořádaná díla	13,91806	26,23611	-12,3181	-6
10	mf dnes	17,74583	37,58472	-19,8389	-13
11	ln	18,46667	38,02083	-19,5542	-11
12	právo	18,12778	38,05694	-19,9292	-14
13	blesk	19,95556	40,43333	-20,4778	-16
14	reflex	19,29861	39,45417	-20,1556	-15
15	respekt	18,71389	38,43472	-19,7208	-12
16	administrativa	9,090278	17,29444	-8,20416	-1

W- = 136

W+ = 0

W = min (0, 136) = 0

6.4 Vyhodnocení testu

Dle tabulky[43] pro n_{16} platí při významové hladině 0,05 kritická hodnota 29,9. Protože $W = 0 < 29{,}9$ nemůžeme vyvrátit nulovou hypotézu, tedy že není žádný vtah mezi R_{lem} a R_{slo}. Vzhledem k tomu, že i výpočty v rozptýlení a koncentraci lexika by byly nepochybně stejné, nebudeme je již počítat.

6.5 Závěr

Přestože se nám nepodařilo prokázat na dané hladině (0,05) vztah mezi výsledky získanými při výpočtech s lemmaty a slovoformami, jsme přesvědčení o určité souvislosti. Mnohem zásadnější otázkou však je, zda je vhodné následovat dogmatické počítání s lemmaty. Skutečnost, že získaná data nejsou zcela totožná, neznamená chybnost počítání s textovými slovy, neboť takový pohled by byl zcela dogmatický. Není totiž zřejmé, který způsob vykazuje validnější data.

Výpočty v lemmatech jsou potvrzovány skutečností, že čeština je flektivní jazyk, měli bychom proto při analýzách reflektovat tuto skutečnost. Opozice slova systémového a textového by tedy měla být zahrnuta. Základním problémem značně komplikujícím takový přístup však je absence jednoznačných pravidel pro lemmatizaci. Navíc elektronické zpracování rozsáhlých korpusů vždy generuje jistou chybu, úspěšnost u korpusu SYN2000 například dosahovala 94 %.[44] Je zřejmé, že tato skutečnost musela do značné míry ovlivnit jak výsledky jednotlivých parametrů v lemmatech, tak samotné testování pomocí WSRT.

Pro výpočty ve slovoformách naopak svědčí fakt, že segmentace má 100% úspěšnost, díky jasným pravidlům zde neexistuje prostor pro chyby. Při výpočtech v lemmatech tedy musíme vždy počítat s dvojím zkreslením dat daným jednak nepřesnou lemmatizací, jednak problematickým srovnáváním různých korpusů, a to i například korpusů řady SYN. Naopak výpočty ve slovoformách zaručují vždy přesná data, která mohou být bez problémů podrobena komparaci s daty získanými

[43] http://cit.vfu.cz/stat/FVL/Teorie/tabulky.htm#Wilcoxon [cit. 8. 5. 2012]

[44] KOPŘIVOVÁ, M.: *Český národní korpus na přelomu tisíciletí*. Dostupné z WWW: <http://korpus.cz/doc/2002_cnk.rtf> [cit. 8. 5. 2012].

z jiných korpusů. Absence procesu lemmatizace také znamená odstranění bariéry, která vždy omezovala bádání, především v rozsahu výběrového souboru.

Na základě uvedených skutečností konstatujeme, že pro výpočty parametrů bohatství textu (konkrétně rozsahu, rozptýlení a koncentrace lexika) považujeme za vhodnější a přesnější pracovat se slovoformami. Při tomto tvrzení jsme si zcela vědomi skutečnosti, že bude třeba tuto hypotézu ověřit dalšími výzkumy. Považujeme za vhodné také dodat, že naše preference výpočtů ve slovoformách nijak neodporuje pojetí slova ve dvou rovinách (systémové a textové), které je pro flexivní jazyky zásadní.

7 ZÁVĚR

Závěrem můžeme konstatovat, že všechny cíle avizované v úvodu byly splněny. Jednalo se především o popis funkčních stylů a jejich vnitřní diferenciace na základě lexikostatistické analýzy. Dále byla ověřena relevance základní jednotky lexikální statistiky, tedy zda je vhodnější pracovat s lemmaty nebo slovoformami. Třetím nejdůležitějším cílem bylo zjistit, do jaké míry a zda vůbec je ČNK vhodným zdrojem dat pro tvoření výběrových souborů pro lexikální statistiku.

Při snaze popsat rozdíly mezi jednotlivými funkčními styly na základě lexikostatistických metod jsme zvolili tři základní parametry spadající do bohatství textu, které umožňují základní charakteristiku ve sledované oblasti, šlo o rozsah, rozptýlení a koncentraci lexika. Při výpočtech jsme se opírali o rovnice stanovené Marií Těšitelovou. Již při teoretickém konstruování analýzy jsme zjistili, že není vhodné srovnávat jednotlivé funkční styly bez přihlédnutí k jejich vnitřní diferenciaci. Každý funkční styl je totiž tvořen z více či méně rozličných textů, jejichž charakteristika se může v různých oblastech, včetně bohatství textu, značně lišit. Došli jsme tak k závěru, že je zcela nezbytné zkoumat nejdříve jednotlivé funkční styly a jejich podkategorie separovaně a teprve na základě znalostí jejich charakteristik je možné podrobit komparaci funkční styly jako celky.

Při analýze uměleckého funkčního stylu jsme zjistili značně specifickou pozici básnických textů, které se ve všech sledovaných parametrech značně odchylovaly od ostatních. Při interpretaci tohoto jevu jsme došli ke zjištění, že determinant, jež odlišují básně od ostatních beletristických projevů, je více. Mezi nejdůležitější patří vyšší četnost substantiv a adjektiv, chudší gramatická složka a potřeba bohatého vyjadřování zahrnující všechny vrstvy slovní zásoby. Na opačném konci v rámci uměleckého stylu stojí filmové a divadelní texty, důvody jejich poměrně chudého slovníku je třeba hledat především ve skutečnosti, že se jedná v podstatě o simulace mluveného projevu.

Analýza odborného stylu ukázala, že největší odchylku tvoří učebnicové texty, což vyplývá zejména z jejich didaktické funkce, jež se musí nutně projevit i v lexikální rovině. Ostatní texty uvnitř odborného stylu nevykazovali příliš velké amplitudy, nejvyšších hodnot bohatství slovníku dosáhli populárně-naučné texty, jež vyžadují na rozdíl od ostatních odborných textů fatický charakter, širší slovní zásobu

a důležitou roli patrně hraje také fakt, že tyto texty jsou průměrně kratší. Ukázalo se, že vědecké komunikáty, které jsou teoreticky produkovány podavateli s nejvyšší slovní zásobou, nejsou z hlediska bohatství slovníku nejpestřejší. Předpokládáme, že důvod tkví zejména ve značných omezeních čistě odborného stylu.

Poměrně překvapivá data byla získána uvnitř publicistického stylu, kde jsme zjistili značně specifickou pozici bulvárního deníku Blesk, jenž vykazoval největší bohatství slovníku. Jeho hodnoty nejen značně převyšovaly seriózní noviny, ale také analytické všeobecné týdeníky. Při zvážení ostatních charakteristik bulvárních periodik jsme poukázali na jejich specifické místo uvnitř publicistického stylu a navrhli jsme tradiční rozdělení na texty zpravodajské a komentářové rozšířit o bulvární. Celkový rozptyl získaných hodnot však nebyl v publicistickém stylu příliš velký, ve srovnání s ostatními dokonce několikrát menší. Právě tento aspekt může do určité míry relativizovat zjištěné odchylky.

Při konečné komparaci jednotlivých funkčních stylů jsme zjistili, že v oblasti bohatství slovníku nejvyšších hodnot dosahuje styl publicistický, následuje umělecký, odborný, administrativní a hovorový. Nejvýrazněji se odlišovaly právě poslední dva uvedené styly, které dosahovaly podobných hodnot. Fakt, že administrativní texty se svými specifiky dosahují malého bohatství textu, není překvapivý, ovšem skutečnost, že je srovnatelný se stylem hovorovým, jenž je tvořen výhradně mluvenými projevy, je poměrně nečekaný, neboť narušuje představu o tom, že psané texty mají vždy bohatší lexikum než texty mluvené.

Při komparacích jsme narazili na metodologický problém, a to skutečnost, že jsme srovnávali sice stejně velké výběrové soubory, avšak s různým počtem a délkou jednotlivých textů. Proto jsme se pokusili ukázat, do jaké míry se tento vztah projevuje, výsledné hodnoty jsme vynesli na bodových grafech. Nastolili jsme tak kardinální otázku, zda je z hlediska lexikální statistiky správné takto strukturně rozdílné výběrové soubory srovnávat. Stojí zde proti sobě dvě hlediska: matematické a lingvistické. Není totiž zcela zřejmé, zda by bylo správné skutečně srovnávat jen stejně dlouhé texty, neboť krácení např. románu na úroveň novinových článků by narušovalo právě ty struktury koherentních textů, které jsou hlavním předmětem našeho zkoumání.

Výsledná data získaná z paralelních výpočtů v lemmatech a slovoformách jsme podrobili statistickému šetření pomocí Wilcoxonova testu. Přestože se nám nepodařilo potvrdit vztah mezi oběma sadami dat, upozornili jsme na základní problémy lemmatizace, které mohly ovlivnit Wilcoxonův test, ale zejména způsobují určitou nepřesnost jakýchkoliv výpočtů na základě lemmat. Proto jsme konstatovali, že pro lexikostatistickou analýzu textů je přesnější pracovat se slovoformami, což také umožňuje srovnávat různé výběrové soubory.

Důležitou komponentou naší práce také bylo zjistit, zda je ČNK vhodným zdrojem pro lexikostatistické analýzy. Domníváme se, že pro češtinu v současné době neexistuje lepší zdroj jazykového materiálu než právě ČNK, a to z mnoha důvodů. Jedná se co do rozsahu o bezkonkurenční soubor českých textů (psaných i mluvených), je volně dostupný a kvalita zpracování (zvlášť lemmatizace) je garantována renomovaným pracovištěm, kde neustále probíhá vývoj užívaného softwaru. Rozsah a dostupnost však nejsou jedinou výhodou, propracovaná struktura tagování a promyšlené rozložení různých typů textů dělá z ČNK skutečně ojedinělý zdroj i z kvalitativního hlediska, neboť můžeme separovaně zkoumat takřka libovolnou kategorii komunikátů, od jednotlivých funkčních stylů či konkrétní publikace až po projevy různých sociálních skupin.

8 ANOTACE

Příjmení a jméno autora: Kubát Miroslav

Název katedry a fakulty: Katedra bohemistiky Filozofické fakulty Univerzity Palackého v Olomouci

Název diplomové práce: Funkční styly z hlediska lexikální statistiky

Vedoucí diplomové práce: Mgr. Darina Hradilová, PhD.

Počet znaků: 112 194 (včetně mezer a pozn. aparátu)

Počet příloh: 10

Počet titulů použité literatury: 23

Klíčová slova: funkční styly, kvantitativní lingvistika, lexikální statistika, bohatství textu, frekvence slov, lemma, slovoforma, výběrový soubor, jazykový korpus

Tato práce popisuje na základě lexikostatistické analýzy elementární distinkce mezi funkčními styly a jejími podkategoriemi v oblasti bohatství textu, primárně jsou sledovány hodnoty rozsahu, rozptýlení a koncentrace lexika. Výsledné hodnoty jsou graficky znázorněny a interpretovány. Výběrové soubory jsou excerpovány z databáze ČNK. Analýza je paralelně prováděna při započítávání dvou různých základních jednotek – slova systémového (lemmatu) a slova textového (slovoformy), cílem je ověřit relevanci obou jednotek v rámci lexikální statistiky. Součástí práce jsou také frekvenční slovníky jednotlivých funkčních stylů.

9 SUMMARY

This thesis describes the functional styles in Czech based on lexical statistics. The survey contains all the basic functional styles: artistic, scientific, journalistic, administrative and colloquial. The functional styles are not only examined as units, but due to their large internal differentiation are also subjected to analysis of their subcategories. Only the arithmetic mean of the resulting values may prove truly relevant information about each functional style.

All analyzed texts were obtained from the database of the Czech National Corpus and subsequently were created frequency lists. These lists were samples whose size was set at 90,000 words (word-form). In total, there were included texts with a total length of 1,800,000 words in the research. These samples were then subjected to calculations of text richness, namely extent, dispersion and concentration of the lexicon. The calculations were based primarily on the concept of Marie Těšitelová. The obtained values were then generated by simple bar graphs, which allowed easy comparison.

After obtaining the graphs, the subcategories within each functional style were subjected to a mutual comparison. These data were interpreted linguistically and set in a wider context. Subsequently, obtained average values were subjected to comparison among the functional styles. Here we confirmed that the functional styles cannot be rightly compared without taking into account their considerable internal differentiation. The main aim of these analyses was exactly describe characteristics of the functional styles in terms of text richness and offer perspective on the issue from a different perspective.

The final comparison of the functional styles showed this rank of vocabulary richness: the highest value reached journalistic style, followed by artistic, technical, administrative and colloquial. The last two styles markedly differed from others and reached very similar values. The fact that the administrative texts reached low value of vocabulary richness it is not surprising due to their amount of specifics, but it is quite unexpected that the results are comparable to the colloquial style, which consists solely of spoken speeches. This finding undermined the assumption that written texts are always richer of vocabulary than spoken texts.

During the comparison we ran into a methodological problem, namely the fact that although we compared samples of the same size, the samples consisted of different number and length of the articles. Therefore, we tried to show to what extent this relationship influenced the survey, the resulting values we brought to the point charts. We have therefore established the cardinal question of whether it is possible to compare so structurally different samples from the lexical statistics point of view. It is questionable if such reduction for example of novel texts to the level of newspaper articles would undermine the very structure of coherent texts, which are the main subject of our investigation.

The resulting data obtained from the parallel computations were tested by the Wilcoxon signed rank test. Although we could not confirm the relationship between the two data sets, we discovered that problems of lemmatization are able to affect the Wilcoxon test and also cause some inaccuracy of any calculation based on lemmas. Therefore, we proposed to calculate lexical analysis only in word-form, which is more accurate and also allows comparing different samples.

10 ZDROJE

10.1 Literatura

- ANDĚL, Jiří: *Statistické metody*. Praha: Matfyzpress, 2007, 299 s.
- BERTÓK, Imrich – JANOUŠEK, Ivo: *Počítače a umenie*. Bratislava: Slovenské pedagogické nakladateľstvo, 1989, 168 s.
- ČECHOVÁ, Marie a kol.: *Současná česká stylistika*. Praha: ISV 2003, 344s.
- ČERMÁK, František a kol.: *Frekvenční slovník češtiny*. Praha: Nakladatelství Lidové noviny, 2004, 596 s.
- ČERMÁK, František a kol.: *Frekvenční slovník mluvené češtiny*. Praha: Karolinum, 2007, 512 s.
- GREPL, Miroslav a kol.: *Příruční mluvnice češtiny*. Praha: Nakladatelství Lidové noviny, 2003, 800 s.
- HOFFMANOVÁ, Jana: *Stylistika a… Současná situace stylistiky*. Praha: Trizonia 1997, 200 s.
- HŘEBÍČEK, Luděk: Vyprávění o lingvistických experimentech s textem. Praha: Academia, 2002, 195 s.
- JELÍNEK, Jaroslav – BEČKA, Josef V. – TĚŠITELOVÁ, Marie: *Frekvence slov, slovních druhů a tvarů v českém jazyce*. Praha: SPN, 1961, 588 s.
- KOPŘIVOVÁ, Marie. *Český národní korpus na přelomu tisíciletí*. Článek dostupný na <ucnk.ff.cuni.cz/doc/korplex.rtf> [cit. 8. 5. 2012].
- KARLÍK, Petr – NEKULA, Marek – PLESKALOVÁ, Jana (eds.): *Encyklopedický slovník češtiny*. Praha: Nakladatelství Lidové noviny, 2002. 604 s.
- Kol. autorů: *Mluvnice češtiny (2)*. Praha: Academia 1986, 536 s.
- KUBÁT, Miroslav: *Rozsah, rozptýlení a koncentrace lexika v psané a mluvené češtině* (diplomová práce). Olomouc 2010, 62 s.
- LOTKO, Edvard: *Slovník lingvistických termínů pro filology*. Olomouc: Univerzita Palackého, 2005, 132 s.
- MISTRÍK, Jozef: *Frekvencia slov v slovenčině*. Bratislava: Slovenská akadémia vied, 1969, 728 s.

- MISTRÍK, Jozef: *Frekvencia tvarov a konštrukcií v slovenčině*. Bratislava: Veda, 1985, 319 s.
- MISTRÍK, Jozef: *Štylistika*. Bratislava: SPN , 1989, 584 s.
- POPESCU, Ioan-Iovitz: *Word Frequency Studies*. Berlin – New York: Mouton de Gruyter, 2009, 278 s.
- TĚŠITELOVÁ, Marie a kol.: *O češtině v číslech*. Praha: Academia, 1987, 208 s.
- TĚŠITELOVÁ, Marie a kol.: *Psaná a mluvená odborná čeština z kvantitativního hlediska*. Praha: Ústav pro jazyk český ČSAV, 1983, 147 s.
- TĚŠITELOVÁ, Marie: *Kvantitativní lingvistika*. Praha: SPN, 1987, 187 s.
- TĚŠITELOVÁ, Marie: *Otázky lexikální statistiky*. Praha: Academia, 1974, 292 s.
- WIMMER, Gejza: *Úvod do analýzy textov*. Bratislava: Veda, 2003, 344 s.

10.2 Internetové odkazy

- http://ucnk.ff.cuni.cz
- http://www.ujc.cas.cz

10.3 Použitý software

- Microsoft Excel 2010 (verze 14.0.6112.5000)
- Bonito (verze 1.49)
- Bonito (verze 1.80)
- PSPad (verze 4.5.4 /2356/)

11 PŘÍLOHA

FREKVENČNÍ SLOVNÍK UMĚLECKÉHO STYLU (LEMMA)								
	ROMÁN		POVÍDKA		IMAGINATIVNÍ		BÁSEŇ	
1	BÝT	4773	BÝT	3388	BÝT	4298	SE	3941
2	A	3268	A	3296	A	2856	BÝT	3907
3	SE	3174	SE	2937	SE	2426	A	2881
4	TEN	1968	V	2569	V	2099	TEN	2276
5	V	1881	NA	1661	TEN	1587	NA	1679
6	NA	1323	TEN	1231	NA	1492	ON	1615
7	ŽE	1233	S	945	Z	869	JÁ	1310
8	JÁ	1000	DO	871	JÁ	860	V	1135
9	S	943	Z	871	S	855	S	838
10	ON	881	ON	734	DO	774	ŽE	825
11	KTERÝ	816	ROK	629	I	578	K	759
12	MÍT	750	JÁ	603	ŽE	571	DO	755
13	Z	689	ŽE	588	SVŮJ	563	PRINC	714
14	SVŮJ	617	K	532	KTERÝ	555	CO	675
15	O	586	SVŮJ	514	MŮJ	536	MÍT	653
16	DO	565	JAKO	470	JAKO	533	Z	611
17	ALE	564	KTERÝ	468	ON	504	ALE	594
18	K	533	ALE	455	MÍT	500	TY	548
19	I	502	TY	424	K	491	TAK	510
20	ČLOVĚK	476	I	420	ALE	457	VY	451
21	JAKO	420	O	415	O	453	JAKO	448
22	PO	400	MÍT	408	JAK	448	ANDERSON	439
23	KDYŽ	393	PO	391	CO	422	ZA	433
24	VŠECHEN	383	TAK	319	ONI	363	KRÁL	432
25	TENTO	375	JEHO	304	UŽ	357	PÁN	405
26	CO	369	TENTO	297	TAK	354	JÍT	390
27	TAK	366	ZA	282	JEN	344	JAKUB	374
28	MOCI	363	MOCI	277	ZA	335	VŠECHEN	349
29	ZA	353	CO	262	VŠECHEN	329	UŽ	343
30	ONI	344	U	262	MY	328	KTERÝ	319
31	PRO	337	VŠECHEN	241	KDYŽ	326	BAJAJA	315
32	MY	291	UŽ	238	JENŽ	316	O	315
33	ABY	287	PRO	235	PO	314	SVŮJ	313
34	ŽIVOT	286	MŮJ	234	ČAS	295	ONI	302
35	JAK	282	OD	215	CHTÍT	276	MOCI	294
36	ROK	276	JAK	212	PRO	256	PO	285
37	CHTÍT	268	JEŠTĚ	204	JEŠTĚ	248	CHTÍT	282
38	CÍRKEV	255	JENŽ	203	TVŮJ	226	JAK	277
39	UŽ	232	KDYŽ	191	NÁŠ	214	MŮJ	277
40	OD	215	PAK	189	MOCI	209	VĚDĚT	273
41	JEDEN	206	JAN	173	NAD	207	PRINCEZNA	256
42	NÁŠ	204	FRANTIŠEK	168	KDE	200	KŮŇ	248
43	VĚDĚT	203	MUSET	167	VĚDĚT	196	ABY	246
44	STÁT	200	TEĎ	164	TY	192	NE	243
45	MUSET	197	ONI	163	U	192	POPELKA	240
46	MŮJ	190	NEBO	161	JÍT	191	KDYŽ	237
47	PANÍ	188	AŽ	159	DEN	189	OLGA	235
48	PAN	187	PRAHA	154	ŽENA	179	ŘÍCI	233
49	ŽENA	186	JOSEF	153	TENTO	174	STÁT	229
50	ŘÍCI	182	PRVNÍ	150	ABY	173	MY	215

FREKVENČNÍ SLOVNÍK UMĚLECKÉHO STYLU (SLOVOFORMA)								
	ROMÁN		POVÍDKA		IMAGINATIVNÍ		BÁSEŇ	
1	A	3266	A	3280	A	2850	SE	3124
2	SE	2536	SE	2360	SE	1973	A	2879
3	V	1534	V	2147	V	1773	NA	1679
4	NA	1323	NA	1661	NA	1492	TO	1450
5	JSEM	1259	JSEM	969	JE	1357	JE	1332
6	ŽE	1233	DO	871	JSEM	877	V	935
7	TO	1214	JE	860	TO	837	ŽE	819
8	JE	1125	S	733	DO	774	DO	755
9	S	748	TO	713	Z	727	SI	735
10	SI	613	Z	708	S	699	S	706
11	O	586	ŽE	587	I	573	CO	628
12	DO	565	SI	553	ŽE	570	PRINC	602
13	ALE	564	JAKO	470	JAKO	533	JSEM	598
14	Z	555	ALE	455	SI	483	ALE	594
15	I	480	K	427	ALE	457	K	585
16	K	437	I	412	O	451	TAK	506
17	JAKO	420	VE	410	JAK	448	Z	494
18	PO	400	O	393	K	418	JÁ	454
19	KDYŽ	393	PO	391	CO	385	JAKO	448
20	TAK	366	TAK	319	UŽ	357	ZA	433
21	MI	356	JEHO	304	TAK	354	ANDERSON	384
22	ZA	353	ZA	282	JEN	344	KRÁL	375
23	VE	346	U	262	MI	337	MI	362
24	PRO	337	ROKU	249	ZA	335	UŽ	343
25	MĚ	323	CO	239	KDYŽ	326	PÁN	324
26	CO	315	UŽ	238	VE	325	O	305
27	JAK	282	PRO	235	PO	314	BY	286
28	JSME	266	BYL	227	BY	261	PO	285
29	BYL	247	JAK	212	PRO	256	JAKUB	279
30	BYLA	247	OD	209	JEŠTĚ	248	JAK	276
31	BY	240	JEŠTĚ	204	JSME	229	BAJAJA	266
32	UŽ	232	KDYŽ	191	TU	226	TY	256
33	KTERÉ	223	PAK	189	JSOU	212	MU	255
34	BYLO	219	BY	188	BYL	203	JI	251
35	JSOU	215	MĚ	187	NAD	201	HO	244
36	OD	209	MI	174	KDE	200	NE	243
37	KTERÝ	197	KTERÝ	169	U	192	KDYŽ	236
38	ABY	191	TEĎ	164	JÁ	189	MĚ	220
39	PANÍ	188	ZE	161	BÝT	187	JSTE	219
40	JEHO	184	NEBO	161	BYLO	186	PRINCEZNA	210
41	NEBO	180	AŽ	159	BYCH	178	OLGA	201
42	JEN	178	JÁ	159	NENÍ	173	VE	198
43	SVÉ	176	TI	158	STÁLE	172	TI	197
44	MÁ	169	MU	152	AŽ	167	NENÍ	196
45	JÁ	169	JSI	147	MĚ	165	DINA	191
46	U	161	HO	143	OD	164	TU	190
47	JEŠTĚ	159	KDE	137	POD	163	I	190
48	ANI	156	TAM	132	KTERÉ	161	JEŠTĚ	180
49	PROTOŽE	153	SVÉ	131	SVÉ	161	U	177
50	TÍM	150	ROCE	130	NÁS	157	PAK	177

FREKVENČNÍ SLOVNÍK ODBORNÉHO STYLU (LEMMA)								
	VĚDECKONAUČNÁ		POPULÁRNĚNAUČNÁ		UČEBNICE		USPOŘÁDANÁ DÍLA	
1	A	2909	BÝT	3469	BÝT	2787	BÝT	3242
2	V	2892	SE	2665	A	2314	V	2374
3	BÝT	2440	A	2327	V	2235	A	2192
4	NA	1408	V	1500	SE	1865	NA	1444
5	S	1142	NA	1478	NA	1166	SE	1261
6	Z	1083	TEN	1260	KTERÝ	941	S	885
7	SE	1045	ZÁKON	1103	K	904	KOSTEL	830
8	ROK	1023	VÍNO	1064	S	818	ROSTLINA	689
9	ULICE	903	Z	808	Z	768	Z	672
10	TENTO	705	KTERÝ	805	O	744	K	643
11	KTERÝ	704	S	708	2	739	JAKO	633
12	ARCHITEKT	686	ŽE	615	ROZTOK	688	I	530
13	DŮM	671	NEBO	606	TENTO	669	TEN	496
14	ČP	636	MÍT	570	PŘI	659	ŽE	496
15	I	586	O	545	REAKCE	653	DO	491
16	K	541	DO	460	ATOM	558	JENŽ	460
17	JAKO	438	LI	422	C	502	TENTO	454
18	O	423	K	422	DO	492	KTERÝ	442
19	STAVBA	383	CO	413	LÁTKA	467	SV	432
20	TEN	341	PRO	410	TEN	466	MÍT	428
21	OD	339	MOCI	404	KYSELINA	460	MĚSTO	420
22	ANTROPOLOGIE	338	I	374	ŽE	454	MOCI	417
23	PRO	335	ON	355	PRO	439	O	412
24	MÍT	334	TENTO	327	I	430	ROK	395
25	DO	325	U	322	MOLEKULA	421	NEBO	369
26	OBJEKT	316	ONI	292	VODA	412	PŘI	346
27	BUDOVA	296	PŘI	291	JAKO	393	PRO	342
28	ČLOVĚK	284	ALE	291	IONT	389	JEHO	317
29	U	283	PRAVIDLO	277	ENERGIE	379	STOLETÍ	310
30	FASÁDA	280	JAKO	275	MOCI	359	OD	301
31	SVŮJ	263	PRÁCE	270	NEBO	335	LIST	297
32	KULTURA	262	KDYŽ	267	ELEKTRON	326	CHÓR	270
33	JENŽ	247	DŘEVO	261	MÍT	322	STAVBA	259
34	STAVITEL	246	TAK	257	H	311	LOĎ	246
35	Č	246	VELKÝ	254	3	307	PO	244
36	STOLETÍ	239	ZA	241	U	300	VŠAK	240
37	DALŠÍ	237	VŠECHEN	238	NAPŘ	281	VELKÝ	239
38	SPOLEČNOST	235	ABY	223	KONCENTRACE	277	LZE	238
39	VĚDA	235	NEŽ	222	P	275	NOVÝ	233
40	MEZI	228	PO	214	OBR	271	TAK	230
41	STYL	218	LZE	207	1	267	KAPLE	222
42	TAKÉ	216	VŮNĚ	204	DVA	246	U	220
43	JEHO	215	SVŮJ	203	CHEMICKÝ	243	TEDY	212
44	ŽE	213	VELMI	195	JEDEN	242	DŮM	210
45	SECESE	212	PRVNÍ	194	HODNOTA	240	STARÝ	202
46	ČÁST	211	DOBRÝ	190	MOŽNÝ	239	ALE	201
47	PODLE	203	POKUD	187	ORBITAL	223	SVŮJ	198
48	PRŮČELÍ	203	PAK	180	VAZBA	223	ČÁST	195
49	SOCIÁLNÍ	195	JEDEN	179	4	223	KLENBA	183
50	DOBA	193	MUSET	176	MEZI	221	POKOJOVÝ	182

FREKVENČNÍ SLOVNÍK ODBORNÉHO STYLU (SLOVOFORMA)								
	VĚDECKONAUČNÁ		POPULÁRNĚNAUČNÁ		UČEBNICE		USPOŘÁDANÁ DÍLA	
1	A	2830	SE	2349	A	2294	A	2184
2	V	2248	A	2321	SE	1815	V	1919
3	NA	1408	JE	1884	V	1765	NA	1444
4	SE	1128	NA	1478	JE	1601	SE	1264
5	JE	1031	V	1208	NA	1166	JE	941
6	Z	957	ZÁKON	888	K	770	S	737
7	S	940	Z	685	2	739	JAKO	633
8	ČP.	636	TO	645	O	725	Z	571
9	VE	593	ŽE	615	S	690	K	560
10	ROKU	579	NEBO	606	PŘI	659	I	526
11	I	567	S	566	Z	656	ŽE	496
12	ULICI	489	O	545	JSOU	618	DO	491
13	JAKO	438	VÍNA	508	DO	492	VE	452
14	K	414	DO	460	C	480	SV.	432
15	O	403	JSOU	430	ŽE	454	BYL	413
16	OD	338	LI	422	VE	443	O	412
17	PRO	335	PRO	410	PRO	439	NEBO	369
18	DO	325	SI	391	KTERÉ	430	JSOU	352
19	ULICE	298	I	373	I	424	PŘI	346
20	BYL	296	VÍNO	354	JAKO	393	PRO	342
21	JSOU	295	K	351	REAKCE	382	JEHO	318
22	U	283	U	322	ROZTOKU	355	STOLETÍ	306
23	DŮM	282	KTERÉ	292	NEBO	335	KOSTELA	306
24	ARCHITEKTA	275	ALE	291	3	307	BYLA	305
25	ANTROPOLOGIE	260	PŘI	291	H	302	ROSTLINY	305
26	Č.	246	VE	285	U	298	OD	301
27	STOLETÍ	238	JAKO	275	NAPŘ.	278	KOSTEL	287
28	ARCHITEKT	236	CO	268	OBR.	271	BYLO	269
29	KTERÉ	235	KDYŽ	267	1	267	PO	244
30	MEZI	228	TAK	257	P	264	VŠAK	240
31	BYLA	227	ZA	241	ENERGIE	257	TAK	230
32	TÉTO	216	NEŽ	222	KYSELINY	242	ROKU	223
33	TAKÉ	216	PO	214	4	223	U	220
34	JEHO	215	PRAVIDLO	205	MEZI	221	BÝT	219
35	ŽE	213	MÁ	197	MOŽNÉ	220	BY	214
36	PODLE	203	VELMI	195	ZA	218	TEDY	212
37	KTERÝ	202	POKUD	187	VODY	201	ROSTLIN	209
38	STYLU	192	NENÍ	184	PO	199	ALE	201
39	SECESE	192	BÝT	182	TAK	199	KTERÉ	198
40	MÁ	187	PAK	180	LÁTKY	196	TO	189
41	PO	184	KTERÝ	179	MOL	188	LZE	189
42	OBDOBÍ	173	TÍM	176	KTERÝ	181	BYLY	183
43	NÁMĚSTÍ	172	ABY	175	-1	178	JAK	181
44	ZA	170	MŮŽE	172	IONTŮ	176	AŽ	180
45	AŽ	169	BUDE	165	IONTY	176	14	178
46	NEBO	167	PRVNÍ	162	L	172	JEJICH	178
47	SPOLEČNOSTI	167	OD	160	MNOŽSTVÍ	172	JEN	171
48	VŠAK	157	PRÁCE	159	JEJICH	171	MEZI	166
49	PRŮČELÍ	157	JAK	157	PROTO	170	MĚSTA	165
50	DOMŮ	156	BY	153	PH	166	CHÓRU	158

FREKVENČNÍ SLOVNÍK PUBLICISTICKÉHO STYLU (LEMMA)												
	MF DNES		LIDOVÉ NOVINY		PRÁVO		BLESK		REFLEX		RESPEKT	
1	BÝT	2931	BÝT	3139	BÝT	3193	BÝT	2911	BÝT	3995	BÝT	2970
2	V	2811	V	2397	V	2642	SE	2393	A	2649	A	2536
3	SE	2260	SE	2168	A	2137	V	2348	SE	2452	SE	2190
4	A	2028	A	2117	SE	1990	A	2061	V	2161	V	2140
5	NA	1896	NA	1556	NA	1713	NA	1915	TEN	2010	NA	1635
6	TEN	1110	TEN	1246	TEN	1189	TEN	1360	NA	1565	TEN	1278
7	Z	894	ŽE	862	ŽE	946	Z	901	ŽE	1140	ŽE	891
8	S	791	Z	814	Z	813	S	833	S	805	Z	884
9	KTERÝ	789	KTERÝ	776	S	812	ON	782	Z	795	KTERÝ	799
10	I	679	S	744	KTERÝ	776	ŽE	746	MÍT	763	S	770
11	MÍT	663	MÍT	702	MÍT	736	MÍT	713	O	668	O	677
12	DO	632	O	700	O	730	KTERÝ	674	ALE	621	MÍT	613
13	ŽE	623	I	589	I	632	DO	669	ON	598	ROK	552
14	O	537	DO	568	DO	616	O	538	JÁ	595	DO	540
15	ALE	468	ALE	497	K	501	I	527	KTERÝ	588	ALE	539
16	ON	441	ROK	460	ROK	447	ALE	489	I	564	I	501
17	ROK	426	K	432	ALE	443	ZA	457	DO	514	ON	478
18	ZA	368	ON	427	ZA	414	PO	355	K	422	ZA	440
19	MOCI	346	PRO	415	ON	411	K	328	TAK	412	K	429
20	K	345	ZA	390	PRO	369	UŽ	325	ROK	397	SVŮJ	420
21	PO	321	SVŮJ	357	MOCI	321	SVŮJ	323	SVŮJ	394	JAKO	386
22	SVŮJ	321	TENTO	317	SVŮJ	315	ROK	317	JAKO	388	PRO	361
23	ČLOVĚK	305	PODLE	309	TENTO	310	JÁ	272	CO	365	PO	339
24	PRO	301	MOCI	304	PODLE	304	PRAHA	271	ZA	357	MOCI	310
25	ONI	299	TAK	279	PO	292	TAK	267	PRO	353	ONI	293
26	UŽ	298	JAKO	271	ONI	278	ONI	257	MOCI	343	UŽ	290
27	OD	272	PO	265	ČLOVĚK	255	MOCI	255	ONI	326	JEHO	288
28	JEN	247	ONI	239	ŘÍCI	234	PRO	237	JENŽ	288	TAK	286
29	TAK	220	ČESKÝ	239	JAKO	221	CHTÍT	230	KDYŽ	288	ČLOVĚK	255
30	ČESKÝ	217	UŽ	233	ČESKÝ	217	CO	229	ČLOVĚK	278	TENTO	250
31	MUSET	214	ČLOVĚK	232	UŽ	216	JAKO	227	UŽ	271	OD	245
32	TENTO	210	JEN	216	OD	215	KDYŽ	221	VŠECHEN	262	CO	241
33	JAKO	202	OD	213	MUSET	211	U	219	JAK	259	STÁT	233
34	KDYŽ	202	JEDEN	210	STÁT	208	JEHO	218	PO	241	JENŽ	219
35	DVA	201	VELKÝ	210	TAK	204	DVA	213	TENTO	236	VELKÝ	214
36	NOVÝ	201	JEHO	209	CHTÍT	202	ŘÍCI	212	JEN	223	VŠECHEN	198
37	JÁ	199	CO	204	NOVÝ	196	PŘED	205	JEHO	220	JEN	192
38	PRVNÍ	199	STÁT	203	JÁ	191	JEN	202	JEDEN	200	JEDEN	190
39	CHTÍT	198	VŠECHEN	196	PŘI	184	VŠECHEN	194	OD	194	ŘÍKAT	179
40	JEDEN	196	CHTÍT	193	CO	184	MUŽ	192	MUSET	193	NOVÝ	176
41	MY	193	JÁ	190	ABY	183	OD	190	CHTÍT	183	JAK	174
42	AŽ	188	NOVÝ	186	JEHO	182	KORUNA	181	JÍT	179	ČESKÝ	173
43	VELKÝ	187	DALŠÍ	184	VELKÝ	181	JEDEN	180	NEBO	175	NEBO	172
44	ŘÍCI	182	JENŽ	178	PŘED	178	MUSET	174	ČESKÝ	175	VLÁDA	170
45	TAKÉ	177	JAK	178	DALŠÍ	177	AŽ	173	U	173	NEŽ	170
46	DALŠÍ	175	MUSET	174	JAK	177	ČLOVĚK	165	STÁT	171	KDYŽ	168
47	KRAJ	175	LN	171	JEDEN	175	DALŠÍ	163	NEŽ	170	PŘED	166
48	U	174	DVA	168	DVA	175	PODLE	162	VELKÝ	165	MY	165
49	ABY	172	KDYŽ	164	JENŽ	173	ABY	160	PAK	161	JÍT	165
50	PŘED	171	ABY	160	VŠECHEN	171	PŘI	159	AŽ	160	PODLE	162

FREKVENČNÍ SLOVNÍK PUBLICISTICKÉHO STYLU (SLOVOFORMA)												
	MF DNES		LIDOVÉ NOVINY		PRÁVO		BLESK		REFLEX		RESPEKT	
1	SE	2047	SE	1964	SE	1868	SE	2115	SE	2069	SE	2056
2	NA	1800	NA	1550	NA	1723	NA	1904	NA	1565	NA	1650
3	JE	819	ŽE	937	ŽE	973	JE	805	ŽE	1213	JE	1027
4	TO	673	JE	918	JE	908	ŽE	796	JE	1153	ŽE	979
5	ŽE	672	TO	617	DO	650	TO	752	TO	993	TO	672
6	DO	635	DO	559	TO	625	DO	671	JSEM	698	DO	549
7	SI	526	ALE	463	VE	450	SI	572	SI	575	ALE	530
8	ALE	465	SI	443	ALE	429	ALE	475	ALE	532	ZA	438
9	VE	464	BY	427	BY	413	ZA	439	DO	523	BY	428
10	ZA	355	VE	416	ZA	401	VE	356	TAK	397	SI	411
11	BY	304	PRO	397	PRO	382	PO	316	JAKO	380	JAKO	389
12	PRO	291	ZA	376	SI	368	UŽ	316	BY	371	PRO	366
13	UŽ	280	TAK	280	KTERÝ	276	JSEM	297	PRO	356	PO	315
14	PO	277	JAKO	270	PO	276	KTERÝ	269	ZA	349	VE	292
15	JSME	270	JSOU	255	BUDE	233	TAK	250	VE	336	UŽ	291
16	JSEM	262	KTERÉ	252	OD	221	BY	241	JSOU	277	TAK	279
17	OD	245	PO	239	JAKO	221	BUDE	238	UŽ	262	KTERÝ	263
18	KTERÝ	240	UŽ	228	TAK	216	PRO	236	CO	244	JEHO	253
19	JEN	235	JEN	227	UŽ	213	BYL	228	PO	233	OD	244
20	NA	225	KTERÝ	225	KTERÉ	208	HO	221	JEN	221	KTERÉ	228
21	AŽ	220	PODLE	215	JSEM	207	JAKO	214	JAK	217	JEN	211
22	TAK	219	OD	209	PODLE	198	PŘED	210	KDYŽ	206	JSOU	210
23	KTERÉ	218	BUDE	206	BYL	188	MU	201	JEHO	205	NEBO	191
24	JAKO	212	JEHO	200	MÁ	184	JEHO	196	BYL	205	NEŽ	186
25	JSOU	198	JSEM	195	JSOU	181	PRAHA	191	NEŽ	195	CO	179
26	BUDE	192	VŠAK	182	JSME	177	JEN	190	MI	191	ŘÍKÁ	174
27	BYL	180	AŽ	175	PŘED	176	OD	186	OD	188	LET	165
28	TAKÉ	172	LN	172	PŘI	171	KORUN	176	MÁ	186	BUDE	161
29	PŘED	168	NEŽ	169	JEHO	169	AŽ	175	MĚ	183	MÁ	159
30	VŠAK	162	MÁ	163	KORUN	167	CO	173	KTERÝ	181	VŠAK	157
31	KDYŽ	156	BYL	161	AŽ	163	JSME	169	NENÍ	177	PŘED	154
32	JEŠTĚ	154	KTERÁ	159	ABY	162	NA	167	NEBO	176	ANI	154
33	ZE	153	KORUN	157	NA	158	VČERA	167	TO	175	BYL	154
34	ABY	153	NA	156	KTERÁ	151	MÁ	167	JSME	174	ABY	153
35	KORUN	150	ZE	152	ZE	151	ZE	163	AŽ	157	JAK	152
36	JEHO	148	CO	151	ŘEKL	151	KDYŽ	158	PAK	154	JEŠTĚ	151
37	MÁ	141	PŘED	148	BÝT	150	JEŠTĚ	151	TOM	147	JSEM	150
38	ŘEKL	140	ABY	148	JEN	144	KTERÉ	150	ANI	142	NENÍ	146
39	KTERÁ	139	PODLE	146	JEŠTĚ	140	JSOU	148	PROTOŽE	141	ZE	144
40	BYLO	138	PŘI	141	TAKÉ	140	TO	144	JEJICH	140	JEJICH	144
41	ANI	138	ANI	141	CO	139	PŘI	136	PŘED	137	NA	142
42	CO	133	TAKÉ	140	NEŽ	137	ABY	129	KTERÉ	136	SVÉ	133
43	PRVNÍ	132	NEBO	137	VŠAK	136	TOTIŽ	126	JEŠTĚ	135	TAKÉ	132
44	NAD	130	BÝT	131	JAK	136	PAK	124	JE	133	PODLE	129
45	NEBO	127	KDYŽ	130	NEBO	132	KTERÁ	123	TOHO	131	KDYŽ	125
46	KTEŘÍ	124	ČI	129	PODLE	128	ANI	121	NA	131	BÝT	125
47	JAK	124	JEJICH	128	KDYŽ	126	ŘEKL	119	ZE	131	ROCE	124
48	DNES	119	JAK	126	BUDOU	119	NEŽ	116	BYLO	130	KTERÁ	122
49	SOBOTU	118	JEŠTĚ	124	BYLO	118	TISÍC	116	ALE	129	AŽ	115
50	PŘI	114	MEZI	124	ROKU	117	PODLE	115	BUDE	128	KDE	115

FREKVENČNÍ SLOVNÍK ADMINISTRATIVNÍHO STYLU (LEMMA)		
1	BÝT	2691
2	V	2578
3	A	2358
4	NA	1534
5	O	1326
6	SE	1151
7	S	956
8	K	946
9	KTERÝ	944
10	ZÁKON	903
11	TENTO	887
12	SLUŽBA	868
13	NEBO	798
14	Z	742
15	PRO	618
16	TEN	590
17	VEŘEJNÝ	580
18	CIVILNÍ	558
19	ZAKÁZKA	511
20	ŽE	476
21	PODLE	469
22	ŘÍZENÍ	469
23	ZA	461
24	VÝKON	448
25	ZADAVATEL	446
26	MOCI	424
27	DO	412
28	OBČAN	401
29	MÍT	386
30	BYT	365
31	PŘI	347
32	Č	341
33	1	340
34	I	333
35	OSOBA	321
36	DEN	309
37	PRÁVNÍ	308
38	OBEC	283
39	ORGANIZACE	278
40	SB	278
41	LI	271
42	ÚŘAD	265
43	OBECNÍ	256
44	ODST	256
45	PŘÍPAD	256
46	PO	251
47	POKUD	248
48	ROK	248
49	DOBA	243
50	JENŽ	243

FREKVENČNÍ SLOVNÍK ADMINISTRATIVNÍHO STYLU (SLOVOFORMA)		
1	A	2348
2	V	2064
3	NA	1534
4	O	1314
5	SE	1160
6	JE	1140
7	K	837
8	S	830
9	NEBO	798
10	PRO	618
11	Z	573
12	SLUŽBY	571
13	CIVILNÍ	558
14	ZÁKONA	521
15	VE	503
16	ŽE	476
17	PODLE	469
18	ZA	461
19	ŘÍZENÍ	457
20	DO	412
21	KTERÉ	376
22	PŘI	347
23	Č.	341
24	VÝKONU	340
25	1	338
26	I	326
27	TO	306
28	LI	271
29	ODST.	253
30	PO	251
31	POKUD	248
32	SB.	246
33	VEŘEJNÉ	244
34	MŮŽE	232
35	ZAKÁZKY	223
36	ZADAVATEL	217
37	JSOU	215
38	BY	213
39	U	205
40	BÝT	204
41	ZÁKON	204
42	2	191
43	ORGANIZACE	188
44	MUSÍ	182
45	TOHOTO	182
46	OD	180
47	VEŘEJNÝCH	180
48	KTERÝ	179
49	PRÁVNÍ	179
50	OBECNÍ	178

FREKVENČNÍ SLOVNÍK HOVOROVÉHO STYLU (LEMMA)								
	FORMÁLNÍ 1		FORMÁLNÍ 2		NEFORMÁLNÍ 1		NEFORMÁLNÍ 2	
1	TEN	6317	TEN	6292	TEN	6834	TEN	6755
2	BEJT	3215	A	3070	BEJT	3576	BEJT	3558
3	A	2974	BEJT	3051	A	3246	A	3520
4	ŽE	2680	ŽE	2664	JÁ	1875	JÁ	1989
5	V	1357	TEDA	1356	NO	1855	NO	1874
6	JAKO	1134	MÍT	1353	ŽE	1598	ŽE	1652
7	MÍT	1116	V	1347	JAKO	1294	TAM	1444
8	JÁ	1088	JAKO	1305	TAM	1240	MÍT	1327
9	NA	1040	JÁ	1283	MÍT	1240	NA	1279
10	TAK	928	NA	1086	NA	1207	TAK	1117
11	ALE	917	MYSLET_SI	986	TAK	1131	V	1063
12	PROSTĚ	873	TAK	961	ALE	1056	ALE	1044
13	NO	783	ALE	948	V	1055	JAKO	995
14	NEBO	700	NO	942	JO	904	VON	850
15	TAM	648	TŘEBA	723	VON	861	VONA	724
16	MYSLET_SI	646	TAM	677	PROSTĚ	697	TEDA	672
17	TEDA	641	PROSTĚ	627	UŽ	626	UŽ	600
18	DÍTĚ	598	DÍTĚ	626	S	582	JO	556
19	TŘEBA	595	NEBO	559	TY	554	S	544
20	BÝT	591	PROTOŽE	555	JAK	542	DO	538
21	MOCT	576	ASI	526	TAKOVEJ	536	TO	519
22	PROTOŽE	541	MOCT	500	CO	534	TY	504
23	S	532	Z	467	VONI	519	MY	503
24	ASI	518	ŽE_JO	463	VONA	500	CO	499
25	LIDI	480	TAKŽE	461	TAKY	488	TAKOVEJ	468
26	ŽE_JO	457	BÝT	459	VĚDĚT	473	JAK	468
27	KTEREJ	426	DĚLAT	448	PROTOŽE	446	VONI	466
28	UŽ	415	S	446	Z	445	ŘIKAT	466
29	TAKŽE	393	PRÁCE	439	TEDA	441	VĚDĚT	463
30	I	391	I	420	MY	415	TAKY	448
31	TAKOVEJ	388	LIDI	405	DO	414	PROTOŽE	444
32	CO	381	KTEREJ	405	MOCT	395	EŠTĚ	406
33	K	365	VONI	399	NEBO	374	JÍT	389
34	ŽENA	364	ABY	395	ŘIKAT	371	MUSET	387
35	MUSET	363	TAKOVEJ	393	MUSET	363	Z	382
36	ČLOVĚK	360	K	392	NE	353	ŽE_JO	376
37	Z	358	UŽ	388	TADY	349	TAKŽE	373
38	DO	349	MUSET	372	CHTÍT	348	MOCT	362
39	ABY	342	VŠECHEN	345	ŇÁKEJ	335	PROSTĚ	362
40	ŠKOLA	324	VĚDĚT	341	TO	335	TADY	359
41	VONI	316	ŇÁKEJ	338	PAK	323	NEBO	340
42	JO	315	ŽENA	335	TŘEBA	322	DĚLAT	317
43	SVŮJ	312	DO	335	NÓ	321	NE	317
44	E	308	ZASE	333	DĚLAT	314	ASI	315
45	JAK	307	TAKY	330	EŠTĚ	308	TEĎ	307
46	ŇÁKEJ	304	CO	328	HM	306	ŇÁKEJ	307
47	VŠECHEN	295	JO	311	TAKŽE	297	ZA	307
48	DĚLAT	286	ŠKOLA	287	ASI	295	NÓ	305
49	VĚDĚT	285	VO	279	ŽE_JO	281	TŘEBA	302
50	TAKY	278	ČLOVĚK	274	JÓ	273	ABY	295

FREKVENČNÍ SLOVNÍK HOVOROVÉHO STYLU (SLOVOFORMA)								
	FORMÁLNÍ 1		FORMÁLNÍ 2		NEFORMÁLNÍ 1		NEFORMÁLNÍ 2	
1	TO	3511	TO	3617	TO	4551	TO	4627
2	ŽE	3025	ŽE	3045	A	2984	A	3159
3	A	2700	A	2782	NO	2152	NO	2158
4	SE	2101	SE	2007	JE	1881	ŽE	2004
5	JE	1992	JE	1905	ŽE	1792	TAK	1784
6	TAK	1571	TAK	1737	SE	1732	JE	1711
7	V	1283	SI	1429	TAK	1722	SE	1614
8	SI	1221	JAKO	1216	JO	1286	TAM	1342
9	NA	1204	NO	1179	JÁ	1203	JÁ	1289
10	JAKO	1071	V	1175	NA	1161	NA	1283
11	ALE	875	NA	1170	TAM	1139	SEM	1062
12	PROSTĚ	784	TEDA	999	JAKO	1135	SI	1002
13	JÁ	768	JÁ	972	SEM	1055	ALE	961
14	NO	759	ALE	899	V	996	JAKO	926
15	NEBO	711	MYSLIM	771	ALE	968	JO	902
16	JO	624	SEM	748	SI	964	V	859
17	TY	608	TY	720	TY	679	TY	721
18	TEDA	583	JO	706	CO	620	SME	612
19	TAM	568	TŘEBA	682	SME	593	TEDA	598
20	MYSLIM	560	BY	638	UŽ	579	UŽ	570
21	BY	547	NEBO	616	PROSTĚ	577	DO	533
22	SEM	513	TAM	614	JAK	566	CO	530
23	SOU	499	PROSTĚ	578	TEN	556	JAK	515
24	CO	485	CO	496	S	486	BY	499
25	TŘEBA	483	I	478	TEDA	411	TEN	486
26	ASI	479	SOU	475	TAKY	409	S	430
27	I	469	PROTOŽE	470	NEBO	401	TAKY	411
28	PROTOŽE	459	ASI	446	DO	400	EŠTĚ	390
29	KDYŽ	453	TA	382	BYLO	393	NEBO	382
30	S	435	NEVIM	377	PROTOŽE	375	PROTOŽE	369
31	UŽ	404	ABY	376	TOHO	370	BYLO	368
32	TEN	400	K	371	Z	360	TAKŽE	347
33	DO	361	TAKŽE	367	NE	352	NEVIM	333
34	JAK	357	S	363	DYŽ	351	TADY	331
35	TOHO	354	DYŽ	359	BY	342	NE	329
36	TOM	353	TOM	347	VON	323	TA	323
37	TA	348	UŽ	344	TADY	308	Z	318
38	K	344	DO	344	NÓ	297	NÓ	315
39	TAKŽE	325	KDYŽ	343	KDYŽ	288	ZA	314
40	ABY	320	DĚTI	326	NEVIM	286	TOHO	308
41	LIDI	294	Z	320	HM	285	DYŽ	295
42	DYŽ	293	JAK	317	BYL	282	VON	293
43	Z	293	TAKY	314	PAK	278	TŘEBA	285
44	NEVIM	284	TEN	313	TOM	275	ASI	280
45	BYCH	260	TOHO	309	EŠTĚ	274	KDYŽ	274
46	ČLOVĚK	257	BYCH	286	TŘEBA	273	PROSTĚ	268
47	JENOM	256	ZASE	280	TAKŽE	267	VONA	268
48	TĚCH	252	SME	260	SOU	262	SOU	268
49	TADY	251	TIM	260	ASI	261	ABY	264
50	VE	250	VE	260	VO	248	TEĎ	263