UNIVERZITA PALACKÉHO V OLOMOUCI
FILOZOFICKÁ FAKULTA
KATEDRA BOHEMISTIKY
Česká filologie

KVANTITATIVNÍ ANALÝZA FUNKČNÍCH STYLŮ

QUANTITATIVE ANALYSIS OF FUNCTIONAL STYLES

Magisterská diplomová práce

Autor: Lenka Horutová
Vedoucí práce: PhDr. Petr Pořízka, Ph.D.
Olomouc 2016

Prohlášení

Prohlašuji, že jsem magisterskou diplomovou práci vypracovala samostatně za použití literatury uvedené v seznamu na konci práce.

V Olomouci dne 17. 8. 2016 …….

Poděkování

Děkuji PhDr. Petru Pořízkovi, Ph.D., za velkou vstřícnost a podnětné připomínky při vzniku mé magisterské diplomové práce.

OBSAH

1. ÚVOD

Přestože kvantitativní lingvistika se již několik desítek let řadí mezi tradiční respektované lingvistické disciplíny s vlastním předmětem zkoumání i s vlastními metodami (srov. Karlík – Nekula – Pleskalová 2002, s. 248), v současné době ještě stále matematické ani statistické metody nejsou velkým počtem badatelů využívány. Čech, Popescu a Altmann (2014, s. 5) podle nás správně poukazují na příčinu prozatím poměrně sporadického uplatnění daného přístupu, když tvrdí, že „[…] největší potíž při aplikaci […] kvantitativnělingvistických metod spočívá hlavně v určitém ‚strachu' z modelování jazyka prostřednictvím matematických a statistických nástrojů, který panuje mezi lingvisty a studenty lingvistických oborů, přičemž tento ‚strach' je v naprosté většině případů důsledkem neznalosti či předsudků. Svou roli samozřejmě hraje i neochota překonat uzavřený metodologický rámec oboru." Naše práce by se tak chtěla zařadit mezi méně početné monografie a studie, jež ke svému lingvistickému výzkumu využívají právě exaktní matematické metody, a demonstrací snadné aplikace těchto postupů, popř. poukázáním na jejich výhody, bychom chtěli alespoň trochu přispět k jejich dalšímu rozšíření mezi lingvistickými badateli.

Předkládaná práce si klade za cíl podat charakteristiku jednotlivých funkčních stylů z hlediska pohledu lexikální statistiky i z hlediska parametrů morfologickostatistických či syntaktickostatistických. Smyslem ovšem není odhalit další příznačné rysy funkčních stylů či formulovat jejich novou klasifikaci, ale jde nám především o kvantifikaci a empirické ověření různých tezí, které se k jednotlivým funkčním stylům vztahují.

Při své analýze navazujeme na průkopnické práce z českého kvantitativnělingvistického odvětví od autorky Marie Těšitelové (vzhledem k našemu zaměření považujeme za nejdůležitější zejména její *Kvantitativní lingvistiku* (1987)), jejíž poznatky a metody byly postupem časů lingvisty dále aktualizovány. V současné době se z českých jazykovědců na kvantitativní analýzu textů zaměřuje Radek Čech, jenž také navázal úzkou spolupráci s Gabrielem Altmannem a s Ioanem Iovitzem Popescem, zřejmě nejvýznamnějšími představiteli soudobého kvantitativnělingvistického bádání. Jejich společná publikace *Metody kvantitativní analýzy (nejen) básnických textů* (Čech – Popescu – Altmann 2014) nám poskytla důležité informace týkající se problematiky kvantitativní analýzy textu, na základě této knihy jsme také ve

své práci využili některé z jejich nově představených moderních kvantitativnělingvistických metod.

Inspirací při výběru vhodných matematických postupů nám byl i článek amerických autorů Covingtona a McFalla (2010) *Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR*), publikovaný v renomovaném časopise *Journal of Quantitative Linguistics*. Právě zde byl totiž poprvé představen zdařilý výsledek dlouhodobého úsilí kvantitativních jazykovědců o eliminaci závislosti délky projevu při měření slovního bohatství textu. Poměrně významná pro nás byla i studie Jiřího Miličky a Miroslava Kubáta (2013) *Vocabulary Richness Measure in Genres*, kteří na metodu amerických lingvistů navázali a dále ji propracovali do podoby *Moving Window Type-Token Ratio Distribution*. Za velice podnětnou považujeme nakonec i dizertační práci již zmíněného Miroslava Kubáta (2015), nazvanou *Kvantitativní analýza žánrů*, v níž autor aplikuje moderní kvantitativnělingvistické indexy (na jejichž vzniku se i sám podílel) na různorodý materiál textů Karla Čapka a ověřuje jejich relevantnost pro diferenciaci žánrů.

Během vlastního výzkumu budeme uplatňovat jedny z nejnověji představených metod, jež byly prozatím testovány na poměrně úzce zaměřeném materiálu. Výše zmínění lingvisté dosud v naprosté většině případů aplikovali své indexy jen na umělecké texty (mj. Čech – Popescu – Altmann 2014; Čech – David – Davidová Glogarová 2013; Kubát – Milička 2013), jen výjimečně se zaměření jejich bádání dotýkalo také jiných typů projevů (např. Čech 2013). Proto nás bude v naší práci zajímat, zda vybrané kvantitativnělingvistické postupy budou vykazovat průkazné výsledky nejen pro texty beletristické, ale také pro specifické komunikáty z administrativní oblasti, nebo dokonce pro spontánní projevy mluvené.

Vzhledem k tomu, že náš výběrový soubor tvoří texty různého rozsahu, museli jsme pro náš výzkum zvolit jen takové metody, jež nejsou na délce textu závislé. Konkrétně v naší práci využijeme měření slovního bohatství pomocí *MATTR*, dále pak *tematickou koncentraci textu*, výpočet *vzdálenosti sloves*, index *aktivity* a *deskriptivity* textu či *distribuci slovních druhů*.

Analyzovaný materiál, jak už název naší práce vypovídá, se bude skládat z textů různých funkčních stylů, které budou vyexcerpovány z databáze *Českého národního korpusu* (ČNK) a z *Olomouckého korpusu mluvené češtiny* (OMK). Sestavený korpus budou tvořit projevy odlišných autorů i různých témat, abychom ve zkoumaném vzorku postihli rozmanitost textů spadajících do jednotlivých funkčních stylů.

Pro výzkum má nemalý význam rovněž volba konkrétního softwaru, který badatel užije pro strojové zpracování přirozených komunikátů. Pro naši kvantitativní analýzu textu jsme jako nejvhodnější vyhodnotili multifunkční nástroj *QUITA* (*Quantitative Index/Indicator Text Analyzer*), který umožňuje na velkém množství dat provádět automatizované výpočty mnoha kvantitativ-nělingvistických indexů, aniž bychom museli ovládat skriptovací jazyky či znát přesné matematické operace, na jejichž základě jsou jednotlivé kvantitativní metody postaveny. Navíc tento nástroj dokáže bez nutnosti užití dalších programů provést lemmatizaci textu či detekci slovních druhů, umí generovat frekvenční slovníky nebo mezi sebou analyzovaná data porovnávat pomocí statistických testů; v případě potřeby lze i ze zpracovaných dat v rozhraní softwaru sestavit vhodné grafy. Bližší informace o práci s nástrojem *QUITA* či o jednotlivých implementovaných indexech jsou dostupné v magisterské diplomové práci Vladimíra Matlacha (jednoho z autorů daného programu) *Kvantitativně lingvistický software* (2014) či v oficiálním manuálu vydaném k tomuto softwaru *QUITA – Quantitative Index Text Analyzer* (Kubát – Matlach – Čech 2014).

Pro metodu měření slovního bohatství textu budeme využívat také speciální program nazvaný *MATTR*, jenž byl autorem Michaelem A. Covingtonem sestaven právě za účelem výpočtu indexu *Moving-Average Type-Token Ratio*, k doplnění využijeme i software *MaWaTaTaRaD*, jejž Jiří Milička vytvořil pro možnost vykreslení grafů *Moving Window Type-Token Ratio* a jeho modifikací. Hodnoty slovního bohatství nakonec statisticky ověříme i pomocí komerčního analytického doplňku *XLSTAT*, určeného pro uživatelsky rozšířený tabelační nástroj *Microsoft Excel*.

2. KVANTITATIVNÍ LINGVISTIKA

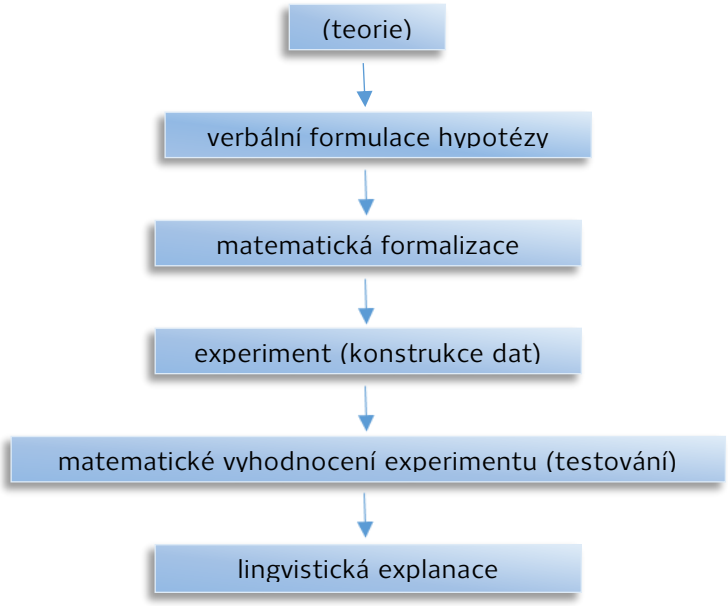
Spojení matematiky a lingvistiky není v českém lingvistickém prostředí novou záležitostí. Základy kvantitativní lingvistiky jako „složky matematické lingvistiky, která kvantifikuje (zjišťuje kvantitativní data) jevy různých jazykových rovin a modeluje jejich vztahy realizující se ve větě, v textu, abychom lépe poznali jejich příčinný mechanismus, jejich fungování, jejich stránku formální, ale i sémantickou" (Těšitelová 1987, s. 8), se u nás formovaly již ve 30. letech 20. století za výrazného přispění lingvistů, jako byli Vilém Mathesius, Bohumil Trnka nebo Josef Vachek. Na počátky mezioborové disciplíny aplikující při studiu jazykových jevů kvantitativní metody měli však kromě lingvistů velký vliv také pedagogové, mj. pedagog Václav Příhoda, který spolu s bohemistou Vladimírem Šmilauerem iniciovali vznik prvního českého frekvenčního slovníku, jenž byl původně zamýšlen pro pedagogické účely (srov. Těšitelová 1987, s. 9). Největším přínosem pro rozvoj české kvantitativní lingvistiky se však staly práce ze 70. a 80. let 20. století,[1] jejichž autorka Marie Těšitelová bývá právem považována za průkopnici kvantitativního přístupu v oblasti české jazykovědy.

Přestože byly matematické metody implementovány do lingvistického bádání již na počátku minulého století, většina provedených analýz měla často jen deskriptivní charakter. Cílem takovýchto výzkumů se stával pouhý popis jazyka a klasifikace jednotlivých jazykových jevů, doplněné o ilustrativní údaje týkající se kvantifikace (informace o frekvenci, procentuální vyjádření atd.). Radek Čech ovšem upozorňuje, že tento přístup není ničím víc než pouhým popisem s čísly, který neumožňuje vyvodit obecnější závěry ohledně fungování jazyka. „Zjištěné absolutní hodnoty se v praxi většinou normalizují do formy přehledného procentuálního vyjádření. Jenže pouhé konstatování, že pozorované jevy se vyskytují s různou frekvencí (byť normalizovanou), neříká o moc víc než prosté označení jevu číslem. Kvantifikace začíná mít smysl v případě, kdy se snažíme předpokládané (a posléze zjištěné či nezjištěné) rozdíly interpretovat, tj. dáváme do souvislosti změřený rozdíl s vlastností jinou, např. formou jazyka, žánrem, sémantickými vlastnostmi, délkou […] aj." (Čech 2014, s. 177)

Teprve od 70. let 20. století se pomalu začíná uplatňovat jiný pohled na kvantifikaci přirozeného jazyka – pohled, který se snaží popisně-klasifikační povahu dosavadních lingvistických bádání překonat a který si klade za cíl

[1] Jedná se zejména o publikace *Otázky lexikální statistiky* (1974), *Kvantitativní charakteristiky současné češtiny* (1985), *O češtině v číslech* (1987), *Kvantitativní lingvistika* (1987) aj.

odhalit mechanismy řídící naše jazykové chování pomocí empiricky testovatelných hypotéz (Čech – Popescu – Altmann 2014, s. 7). Můžeme si vzít např. dvě slova: „lev" a „krokodýl". To, že zjistíme, že se ve stomilionovém korpusu SYN2015 vyskytuje slovo „lev" ve všech svých tvarech s frekvencí 2 626 a slovo „krokodýl" pouze s frekvencí 882, nám samo o sobě neumožňuje vyslovit o fungování jazyka žádné závěry. Pokud ovšem počet výskytů vztáhneme k jiné vlastnosti pozorovaných slov, konkrétně k jejich délce, můžeme již objevit jistou zákonitost existující mezi délkou slov a jejich frekvencí a formulovat hypotézu „Čím je slovo frekventovanější, tím je kratší.", jejíž platnost je snadné, avšak nutné ověřit na rozsáhlejším materiálu. Samotné konstatování, že mezi dvěma či více jazykovými vlastnostmi byla zjištěna vzájemná korelace, ovšem nestačí, ani když tento poznatek doložíme konkrétními číselnými hodnotami. Celé naše kvantitativní šetření dostane význam až na základě relevantní lingvistické interpretace získaných dat, jež je nejdůležitější částí výzkumu. Postup, který se obvykle uplatňuje při kvantitativnělingvistickém výzkumu, Čech, Popescu a Altmann (2014, s. 8) vyjádřili také graficky:



Lingvistické bádání využívající matematické metody vychází obvykle z jisté jazykové teorie, z níž také bývá odvozena hypotéza, jejíž pravdivost je potřeba výzkumem otestovat. Protože však ne každá hypotéza je skutečně na jazykovou teorii vázána, ale její formulace může vzniknout i jiným způsobem, prvotní krok je uveden v závorce. Ověřovanou hypotézu je poté potřeba vyjádřit matematicky, aby mohl být uskutečněn experiment ve formě měření nebo

výpočtu. Výsledné hodnoty je následně nutné podrobit statistickému testu, který dokáže s přesně definovanou pravděpodobností chyby (obvykle stanovenou na 5 %) vyhodnotit, zda můžeme, nebo nemůžeme testovanou hypotézu zamítnout (formulovanou hypotézu nikdy nepřijímáme jako 100% pravdivou, vždy jen zjišťujeme, zda existuje, či neexistuje evidence pro zamítnutí této hypotézy). Zcela zásadním krokem je však závěrečná lingvistická interpretace poznatků, k nimž jsme na základě experimentu dospěli.

Důležité je si uvědomit, že přestože kvantitativní lingvistika využívá ke svému výzkumu matematické metody, její pozornost je stále soustředěna na studium přirozeného jazyka. To, čím se liší od jiných lingvistických disciplín, tak není předmět zkoumání, ale čistě jen způsob, jakým lingvisté dospívají ke svým závěrům. Metody kvantitativních lingvistů jsou založeny na experimentálním přístupu, který umožňuje mnohé do značné míry intuitivní teze týkající se přirozeného jazyka empiricky testovat. Tím se také kvantitativní lingvistika zařadila po bok exaktních věd, jako jsou fyzika či chemie.

Kvantifikace tedy není cílem kvantitativní lingvistiky, ale je jen prostředkem, který badatelům umožňuje získat intersubjektivní pohled na zkoumanou problematiku a který jim dokáže zaručit, že za dodržení stejných podmínek je kdokoli schopný daný experiment zopakovat vždy se stejnými výsledky.

3. FUNKČNÍ STYLY

S ohledem na to, že je práce věnována analýze funkčních stylů, považujeme za užitečné vymezit základní termín „funkční styl" a objasnit, které z mnoha klasifikačních kritérií, jež se v českém stylistickém prostředí uplatňují, budeme v naší práci respektovat.

Stejně jako jazyk sám, také styly podléhají neustálému vývoji a proměňují se pod vlivem společenských, kulturních a historických událostí. Kolektiv autorů *Současné stylistiky* potvrzuje, že „[j]azyk vždy reagoval na nové jevy v životě společnosti, s novými situacemi plnil nové komunikační funkce a různou měrou sloužil nově se objevujícím dobovým komunikačním potřebám. Docházelo k proměnám jazykové komunikace, což se projevuje ve vzniku a rozvoji různých stylů." (Čechová – Krčmová – Minářová 2008, s. 93)

Styl Milan Jelínek v *Příruční mluvnici češtiny* definuje jako „výsledek výběru jazykových prostředků z množin prostředků konkurenčních" (Karlík – Nekula – Rusínová (eds.) 1995, s. 699). Faktory, jež tento výběr ovlivňují, se zpravidla označují jako slohotvorné činitele, které je možné dále rozlišit v závislosti na spjatosti s osobou autora komunikátu na faktory objektivní (mimopersonální) a subjektivní (personální). Právě na stylotvorných faktorech lze založit jednu z mnoha klasifikací stylů. Milan Jelínek tak např. rozlišuje obecně styly objektivní a subjektivní, v rámci objektivních stylů pak navrhuje třídění na styly projevů psaných a mluvených, připravených a nepřipravených, styly monologické a dialogické, oficiální a familiární aj. (Karlík – Nekula – Rusínová (eds.) 1995, s. 705–706). Nejvýznamnějšími však z množiny objektivních stylů jsou styly funkční.

Současné české stylistické teorie vycházejí z přístupu, který ve 20. a 30. letech 20. století představili ve svých pracích zejména Bohuslav Havránek a Vilém Mathesius. Základním termínem jejich pojetí, jež ovlivnilo celé české stylistické poznání, je právě termín funkce, který stylistika chápe jako záměr autora komunikátu nebo účel, který daný projev v komunikaci plní. Převládající funkce textu se tak stala spolu s výběrem výrazových prostředků hlavním kritériem při rozlišování jednotlivých funkčních stylů. (Čechová – Krčmová – Minářová 2008, s. 28)

Počet a charakteristika stylů nezůstaly v průběhu času neměnné, na Havránka navázali např. K. Hausenblas, M. Jelínek nebo A. Jedlička, kteří v reakci na nově rozpoznané funkce jazyka postupně zformulovali nové názory na stylové diferenciace (více informací o vývoji představ ohledně funkce jazyka

a jeho funkční stylové diferenciaci viz Čechová – Krčmová – Minářová 2008, s. 93–105).

Ani v současné době nejsme schopni podat úplný přehled možných klasifikací funkčních stylů, neboť při jejich vymezení hraje kromě komunikační funkce značnou roli i míra zobecnění – při nižším stupni abstrakce je tak možné namísto stylu publicistického konkrétněji odlišit styl zpravodajský, úvahový, interviewový, přesvědčovací atd. (Karlík – Nekula – Rusínová (eds.) 1995, s. 725). Pojetí funkčních stylů uplatněné v naší práci se přidržuje tradičnější české stylové diferenciace, která respektuje šest funkčních stylů:

- prostěsdělovací,
- odborný,
- umělecký,
- publicistický,
- administrativní,
- řečnický.

Jak bylo nastíněno, přístupů ke stylové diferenciaci češtiny existuje celá řada, žádná klasifikace ovšem není a nemůže být univerzálně platná; stejně jako se postupem času mění a vyvíjí funkce jazyka, mění se a vyvíjí se stále i teorie funkčních stylů.

4. VOLBA JEDNOTKY

Dříve než přistoupíme k samotné analýze funkčních stylů, je potřeba si ujasnit, s jakými jednotkami budeme při zpracování kvantitativních dat pracovat. Poměrně problematické se jeví stanovení základní jednotky při aplikaci lexikálněstatistických metod. Za základní jednotku v lexikální statistice bývá zpravidla považováno slovo – to ovšem není v různých jazycích chápáno jednotně. V českém lingvistickém prostředí se tradičně problematika slova vnímá jako opozice slovního tvaru versus lemmatu. „V jazycích s bohatou morfologií, jako jsou jazyky slovanské, se tyto dvě ‚podoby slova' v lexikální statistice celkem jasně diferencují, v jazycích s morfologií chudou, jako jsou angličtina, němčina apod., nejsou tyto dvě podoby vždy dostatečně diferencovány." (Těšitelová 1974, s. 7) U češtiny jako silně flektivního jazyka tak většina jazykovědců dlouhou dobu preferovala práci s lemmatizovanými texty.

V současné době již ovšem řada lingvistů odmítá volbu lemmatu jako základní jednotky souboru takto jednoznačně přijmout. Např. Čech, Popescu a Altmann (2014, s. 10, 19–21) zdůrazňují, že jazykové jednotky jsou pouhé lingvistické nástroje, jejichž prostřednictvím se jen snažíme uvažovat o jazyce; znamená to tedy, že žádná jednotka není jazyku vlastní, přirozená, tudíž je i nemožné některou označit za lepší ve smyslu „lépe odpovídající skutečnosti" než jinou. Konkrétní volba jednotky by tak měla vždy vycházet z badatelského cíle.

Miroslav Kubát (2012) v rámci své magisterské diplomové práce provedl výzkum, v němž se snažil ověřit oprávněnost tradiční volby lemmatu jako základní jednotky lexikostatistiky, a to tak, že veškeré výpočty pro měření bohatství slovníku provedl paralelně s lemmaty i se slovoformami. Závěry, k nimž došel na základě rozsáhlého materiálu ze všech funkčních stylů, naznačují, že ačkoli se výsledky získané při práci s lemmaty liší absolutními hodnotami od výpočtů prováděných se slovními tvary, vzájemné vztahy mezi lexikostatistickými parametry zůstávají vcelku zachovány, ať už kalkulujeme s jakoukoli z těchto jednotek.

Proto se domníváme, že je výhodné při volbě jednotek vzít v úvahu i jistá omezení, která vyplývají z práce s lemmaty: pro výzkum je nutné zvolit pouze texty, jež prošly procesem lemmatizace, popř. je zapotřebí opatřit si vhodný software, který nám ke slovům jednotlivá lemmata přiřadí. Je však důležité zjistit, jak byla lemmatizace textů provedena, neboť neexistuje jednoznačné pravidlo, jež by nám říkalo, které jednotky bychom měli přiřadit k jednomu

lemmatu (jisté problémy způsobují např. komparace adjektiv, přechylování substantiv, negace aj.), tudíž texty z různých zdrojů nemusejí být lemmatizovány stejným způsobem. Další nedostatek potom souvisí s mírou nepřesnosti procesu lemmatizace, protože ať už převádíme slova do jejich základního slovníkového tvaru ručně, nebo strojově, vždy při tomto procesu existuje jisté procento chybovosti. S ohledem na tyto obtíže technického charakteru i s ohledem na srovnatelné výsledky, ke kterým je možné dojít při práci s lemmaty i slovoformami, domníváme se, že je výhodné přihlédnout k mnohem jednodušší práci se slovními tvary a není potřeba za každou cenu zůstávat u lemmatu jako tradiční základní lingvistické jednotky.

Přestože nelze žádné jednotky označit výhradně za správné, nebo špatné, existují jednotky, které jsou pro určitý výzkum vhodnější a méně vhodné. Nevyplývá z toho ovšem, že by bylo možné přiřazovat konkrétní jednotky k jednotlivým analýzám jako jediné možné řešení. Každý badatel si dle našeho názoru může zvolit jednotku, která nejlépe vyhovuje potřebám jeho výzkumu, a domníváme se, že je možné diskutovat spíše o užitečnosti volby příslušné jednotky než o jejím absolutním přijetí, či zavržení. V této práci budeme i my v souladu s výše uvedeným v závislosti na kvantitativní metodě používat různé jednotky, které dle našeho přesvědčení nejlépe vyhovují jednotlivým analýzám. Při výpočtu *tematické koncentrace* textu přihlédneme k lepším výsledkům lemmat, jež v této kvantitativní charakteristice obvykle dosahují, pro zbylé metody již shodně zvolíme práci se slovními tvary.

5. VÝBĚR ANALYZOVANÉHO MATERIÁLU

Zásadní rozhodnutí, které může značně ovlivnit výsledky každého výzkumu, spočívá i ve výběru vhodného materiálu, jenž bude podroben zkoumání. Protože cílem našeho šetření je analýza funkčních stylů, předmětem našeho bádání by se v ideálním případě měly stát všechny projevy, jež byly realizovány v současné češtině. Protože však tento požadavek není z logických důvodů reálný, musíme se v našem výzkumu omezit na pouhý vzorek jazyka. Výběrový soubor by však měl vždy maximálně vyhovovat cíli bádání, a to jak po stránce kvantitativní, tak po stránce kvalitativní.

5.1 Zdroje výběrového souboru

V dnešní době již není žádnou výjimkou, že badatel využívá ke studiu jazyka rozsáhlé jazykové databáze, které shromažďují a zpracovávají texty nejrůznějších druhů. Možnosti, jež nám jazykové korpusy nabízejí, jsou značné, vzhledem ale k tomu, že pro naše zvolené matematické a statistické výpočty používáme zejména k tomuto účelu vytvořený software *QUITA*, jenž vyžaduje materiál určený k analýze v čistém textovém formátu, naším primárním cílem v souvislosti s jazykovými korpusy je jen excerpce úplných, souvislých, reálných jazykových projevů všech funkčních stylů; rozšířené funkce těchto zdrojů jazykových dat tudíž plně nevyužijeme.

Výběrový soubor jsme nakonec sestavili na základě textů vyexcerpovaných ze dvou jazykových korpusů – k získání psaných textů a oficiálních mluvených projevů jsme využili největší databázi *Český národní korpus*, neformální mluvené projevy jsme pak vybrali z *Olomouckého korpusu mluvené češtiny*.

5.1.1 ČNK

Jak již bylo zmíněno, nejvýznamnějším zdrojem jazykových dat je v současnosti *Český národní korpus* (dále jen ČNK), jenž laické i odborné veřejnosti zpřístupňuje jazykový materiál o více než dvou miliardách slov. ČNK sestává z více než třiceti různých korpusů, z nichž jsme pro naše účely zvolili obecný korpus SYN, který obsahuje několik desítek tisíc textů synchronní psané češtiny, tj. textů, které byly publikovány od roku 1990. Z této velké množiny textových projevů jsme postupně vyfiltrovali dostupné komunikáty z jednotlivých funkčních stylů (tj. stylu publicistického, odborného, administrativního a uměleckého), z každé podskupiny jsme pak sami vybrali 20 textů tak, aby

byla ve výběrovém souboru zastoupena rozličná témata, v rámci jednotlivých stylů různé žánry a také aby zkoumaný materiál tvořily projevy odlišných produktorů, přičemž jsme zejména v uměleckém stylu preferovali texty známých autorů, které jsou opakovaně vydávány, a současnou češtinu tak stále ovlivňují.

Tímto jsme získali materiál pro analýzu prvních čtyř funkčních stylů, zbylo nám tak ještě zajistit podklady k výzkumu stylu řečnického a prostěsdělovacího. Protože oba dva styly se vyznačují primárně mluveností, bylo zapotřebí k jejich zkoumání využít korpusy mluveného jazyka. Pro kvantitativní analýzu řečnického stylu jsme zvolili specializovaný korpus SPEECHES z databáze ČNK, jenž je sestaven z oficiálních projevů prezidentů České republiky, příp. představitelů předchozích státních útvarů. Jedná se o přepisy připravených prezidentských promluv proslovených u příležitosti významných státních i jiných svátků, přičemž za oficiální podobu projevů, se kterou se v korpusu pracuje, je považována připravená psaná verze, která nereflektuje případná přeřeknutí či jiné drobné změny v realizovaném verbálním projevu (více viz charakteristika korpusu SPEECHES na stránkách ČNK).

Abychom zachovali jednotné kritérium vzniku analyzovaných textů nejpozději v roce 1990, zařadili jsme do našeho souboru jen projevy posledních tří českých prezidentů, tedy Václava Havla, Václava Klause a Miloše Zemana. Všichni zmínění své proslovy shodně přednesli při výročí vzniku samostatného československého státu, Václav Havel a Václav Klaus poté s projevy vystoupili také na Nový rok, zatímco Miloš Zeman svou další promluvu realizoval v období Vánoc.

Jsme si vědomi toho, že promluvy nejvyšších reprezentantů státu tvoří jen zlomek projevů náležejících do řečnického stylu. Protože jsme však nechtěli tento styl z naší analýzy vyřadit jen kvůli nedostatku různorodého materiálu, rozhodli jsme se šetření provést i na takto úzce zaměřené skupině textů. Jsme přesvědčeni, že i navzdory nevelké rozmanitosti textového materiálu jsou prezidentské projevy vhodným zástupcem řečnického stylu, proto i závěry, k nimž dojdeme na základě kvantitativních metod, budeme moci považovat za obecně platné.

5.1.2 OMK

Zatímco podklady pro výzkum psaných komunikátů lze poměrně bez problémů opatřit v dostatečném množství, možnosti týkající se analýzy mluveného jazyka jsou značně omezenější. Soukromá spontánní komunikace, která by nebyla vázána jen na úzce vymezenou teritoriální oblast, ale reflektovala by

projevy běžného denního styku na celém území České republiky, je z důvodu technických nároků zachycena jen ve velmi omezeném počtu jazykových databází. Protože korpus ORAL2013 z velké databáze ČNK neumožňuje uživatelům získat přístup k celým souvislým komunikátům, i když jinak výše uvedené požadavky splňuje, pro analýzu prostěsdělovacího stylu jsme zvolili přepisy nahrávek z *Olomouckého korpusu mluvené češtiny* (dále jen OMK), které jsme získali z archivu autora daného korpusu Petra Pořízky.

Velkou výhodou OMK je vždy existence dvou typů transkripcí pořízených nahrávek – badatelé tak mají k dispozici jednak přepisy ortografické, jež jsou vhodnější pro automatické lingvistické nástroje, jednak přepisy fonetické, které reflektují autentickou výslovnost i nespisovně užité gramatické tvary jednotlivých mluvčích. Vzhledem ke kvantitativnělingvistickým nástrojům užitým v naší práci vyhodnotili jsme jako vhodnější pro náš výzkum verzi ortografickou. Považujeme však za vhodné upozornit na skutečnost, že i ortografické transkripce obsahují nejrůznější strukturní značky a metatextové informace, díky kterým je možné v korpusu vyhledávat jazykové jevy podle zadaných parametrů. Aby však nebyly výsledné hodnoty našeho výzkumu zkresleny, bylo zapotřebí metatextové údaje z vybraného vzorku odstranit.

Pro všechny užité transkribované projevy platí, že se jedná o neformální spontánní promluvy dialogického charakteru, které v reálném čase trvaly zhruba 20 minut. Zúčastnění mluvčí si během hovoru nebyli vzájemně neznámí a fyzicky se vyskytovali na stejném místě, tzn. že dialogy nebyly vedeny prostřednictvím telefonu nebo jiného komunikačního přístroje. Při výběru jednotlivých komunikátů jsme stejně jako v případě psaných textů dbali na rozmanitost užitého materiálu. Původci mluvených projevů pocházejí z různých míst České republiky (zastoupena je oblast Čech, Moravy i Slezska), dosáhli různých stupňů vzdělání, věkově se pohybují v rozmezí 17–86 let a pracují na rozličných pozicích. Ani tématem hovoru nebyli mluvčí nijak limitováni, vzhledem ale k tomu, že se jedná o rozhovory realizované v každodenních komunikačních situacích, není překvapivé, že se předmětem dialogu často stala rodina, škola, práce, domácnost, současné společenské dění apod.

5.2 Kvantitativní hledisko

Jak už bylo uvedeno, při výběru zkoumaného materiálu je nutné vycházet primárně z cíle výzkumu, přičemž v potaz by se měla vzít otázka kvantity i kvality výběrového souboru. Pokud budeme pracovat s příliš malým vzorkem, mohou být naše poznatky do značné míry zkresleny vůči výsledkům, kterých

bychom dosáhli při práci s celkovou populací. Ani velký výběrový soubor nám ovšem nezaručuje přesnější výsledky. Radek Čech (2014, s. 176) k této problematice poznamenává: „[…] jakýkoliv obecný korpus je vždy souborem různých textů, jejichž jazyk je ovlivněn autorstvím, žánrem atd. Z toho plyne, že zkoumáme-li nějaký jazykový jev, pravděpodobnost jeho výskytu je pro každý text, autora, dílo atd. jiná, což má následující důsledek: zvětšujeme-li počet pozorování *pomocí* a *uvnitř* korpusu, obecně se nepohybujeme v homogenním prostoru. Extrémní růst počtu sledovaných jevů proto nemusí nutně znamenat extrémní zpřesnění poznatků." Čech (2014, s. 177) dále upozorňuje, že vzhledem k charakteru jazyka jako takového je ale v podstatě nemožné přesně určit, jaký vzorek je dostatečně reprezentativní vůči celkové populaci, již tvoří právě jazyk samotný.

Pokud však nechceme zůstat u pouhé deskripce zkoumaných jazykových jevů, ale budeme usilovat o experimentální analýzu odhalující principy, které ovlivňují uživatele jazyka v jejich jazykovém chování, můžeme vyjít rovněž z Čechova přesvědčení, že pro testování hypotéz nepředstavuje problematika reprezentativnosti zkoumaných vzorků žádnou velkou překážku. Čech (2014, s. 181) je totiž přesvědčen, že pokud jednotlivé mechanismy, které jsou předmětem ověřování našich hypotéz, skutečně platí, řídí se těmito mechanismy chování jednotlivých jazykových uživatelů, a proto se musejí tyto mechanismy projevit i v jakémkoli jednotlivém mluveném nebo psaném textu.

Vzhledem k tomu, že naši práci pojímáme jako experimentální sondu, budeme během našeho výzkumu pracovat s databází menšího rozsahu. Z každého funkčního stylu podrobíme analýze 20 textů, celkově tak budou vybrané kvantitativní metody aplikovány na 120 jazykových projevech. S ohledem na výše uvedené se přesto domníváme, že i na takto nevelkém počtu analyzovaných textů budeme schopni odhalit jisté tendence, které se v jazyce projevují a které jednotlivé funkční styly od sebe navzájem diferencují. Počítáme však s tím, že do budoucna bude potřeba naše závěry ověřit na rozsáhlejším materiálu, popř. je doplnit dalšími studiemi.

5.3 Kvalitativní hledisko

Při stanovení výběrového souboru je potřeba vzít v úvahu nejen otázku kvantity, ale je nutné zohlednit také stránku kvalitativní. Jak již bylo naznačeno, někteří lingvisté (mj. Kubát 2015; Čech 2014) důrazně upozorňují, že jednotlivé texty v sobě nesou nejen charakteristiky konkrétních funkčních stylů, ale jsou znatelně ovlivněny také stylem samotného autora, volbou tématu, příslušným žánrem a mnohými dalšími faktory. Pokud tedy do

analyzovaného materiálu zařadíme projevy různých autorů, všechny tyto vlivy budou promíchány a my pak nebudeme schopni jednoznačně rozhodnout, zda jsme k výsledným datům dospěli na základě působení konkrétního funkčního stylu, nebo individuálního stylu autorského.

Přestože s uvedeným tvrzením souhlasíme a věříme, že by korpus sestavený z textů jednoho autora vykazoval snáze akceptovatelné výsledky, do našeho analyzovaného souboru jsme zahrnuli komunikáty nejen různých funkčních stylů, ale také různých autorů. Naše rozhodnutí bylo ovlivněno jednak skutečností, že bychom jen stěží mohli najít českého autora, který by v dostatečném množství produkoval jazykové projevy všech funkčních stylů, jednak by interpretace takto získaných výsledků musela být omezena čistě jen na komunikáty příslušného autora. Navíc jsme přesvědčeni, že i navzdory působení různorodého autorství a různorodých témat budou námi užité statistické metody v textech schopny rozpoznat signifikantní rysy diferencující jednotlivé funkční styly.

Náš výběrový korpus tak tvoří komunikáty šesti funkčních stylů, každý funkční styl pak obsahuje 20 jazykových projevů od různých autorů a pokud možno také co nejvíce tematicky odlišných. Při sestavování analyzovaného vzorku jsme respektovali kritérium ucelenosti publikovaných děl, tj. náš materiál netvoří např. jednotlivé povídky a básně, ale celé soubory povídek, stejně jako samostatné básnické sbírky. Vzhledem k tomu, že se náš výzkum primárně soustředí na analýzu funkčních stylů, a ne na charakteristiky žánrů, nepovažovali jsme za nutné narušovat autorův záměr a místo izolovaných povídek či básní jsme v naší práci zachovali reálně publikované texty sebrané do logického celku. Navíc se domníváme, že bychom k případnému rozdělení jednotlivých textů museli přistoupit i v případě komunikátů publicistického stylu, neboť noviny jsou tvořeny zcela různorodými textovými útvary – od zpráv, komentářů, reportáží přes fejetony, rozhovory až např. po reklamní sdělení. Takové rozčlenění by však nekorespondovalo se záměrem našeho výzkumu, proto jsme do sestaveného korpusu jednotně zahrnuli texty v takové podobě, v jaké byly uveřejněny.

Celkový přehled analyzovaného materiálu uvádíme v tabulkách č. 1–6:

Tabulka č. 1: Přehled jazykových projevů zařazených do publicistického stylu

	typ periodika	název periodika	rok vydání
publicistický styl	bulvární noviny	*Blesk*, 30. 3. 2009	2009
		Blesk magazín, č. 39/2007	2007
		Bulvár, č. 6/2008	2008
		Nedělní Blesk, č. 13/2009	2009
		ŠÍP, 11. 1. 2007	2007
	seriózní noviny	*Hospodářské noviny*, 5. 3. 2009	2009
		Lidové noviny, 15. 1. 2009	2009
		Metro, 15. 1. 2008	2008
		Mladá fronta DNES, 9. 1. 2004	2004
		Právo, 19. 1. 2004	2004
	regionální noviny	*Haló Brno*, č. 9/2003	2003
		Karvinský zpravodaj, č. 6/2003	2003
		Kopřivnické noviny, č. 18/2005	2005
		Náchodský zpravodaj, č. 11/2004	2004
		Radniční zpravodaj – Pardubice, č. 9/2001	2001
	časopisy	*Nedělní svět*, č. 28/2004	2004
		Reflex, č. 48/2007	2007
		Respekt, č. 46/2007	2007
		Týden, č. 21/2004	2004
		Živel, č. 3/2002	2002

Tabulka č. 2: Přehled jazykových projevů zařazených do uměleckého stylu

	žánr	název textu	rok vydání
umělecký styl	romány, novely	Ladislav Fuks: *Vévodkyně a kuchařka*	2006
		Bohumil Hrabal: *Pábení*	1993
		Petra Hůlová: *Cirkus Les Mémoires*	2005
		Josef Škvorecký: *Tankový prapor*	1990
		Michal Viewegh: *Román pro ženy*	2001
	povídky	Jan Balabán: *Možná že odcházíme*	2007
		Alexandra Berková: *Knížka s červeným obalem*	2003
		Lev Blatný: *Servus, Ser-vá-ci*	2003
		Květa Legátová: *Želary*	2001
		Filip Topol: *Tři novely*	2004
	imaginativní texty	Jan Burian: *Na shledanou zítra*	1995
		Eberhardt Hauptbahnhof: *Nedokončený kalendář na tento rok a všechny roky příští*	1994
		Jiří Haussmann: *Haussmanovy méně známé texty*	1999
		Mikuláš Medek: *Texty*	1995
		Zdeněk Rotrekl: *Skryté tváře*	2005
	básně	Vladimír Holan: *Nokturnál*	2003
		František Hrubín: *Zpívám*	2002
		Josef Kainar: *Synkopy*	2003
		Bohuslav Reynek: *Ostny v závoji*	2002
		Jaroslav Seifert: *Slavík zpívá špatně*	2002

Tabulka č. 3: Přehled jazykových projevů zařazených do odborného stylu

	typ odborných textů	název textu	rok vydání
odborný styl	vědecko-naučná literatura	Lubomír Nátr: *Země jako skleník*	2006
		Pavel Přibyl – Aleš Janota – Juraj Spalek: *Analýza a řízení rizik v dopravě*	2008
		Jarmila Riegerová: *Rekondiční a sportovní masáže*	2007
		Petra Řepová: *Dítě a koktavost*	2007
		Jaroslav Zlámal – Jana Bellová: *Podniková ekonomie a management*	2007
	populárně naučná literatura	Jan Kolář: *Biologické hodiny rostlin*	2006
		Dieta, č. 10/2009	2009
		Ekonom, č. 39/2009	2009
		Benjamin Kuras: *Evropa snů a skutečností*	2007
		Stavitel, č. 3/2009	2009
	učebnice	Eva Francová: *Cestovní ruch*	2003
		Jan Ponec – Milič Jiráček: *Digitální fotografie*	2002
		Zdeněk Puchinger: *Daně; Daňová soustava ČR*	2003
		Eva Reiterová: *Základy statistiky pro studenty psychologie*	2003
		Jozef Rosina – Hana Kolářová – Jiří Stanek: *Biofyzika pro studenty zdravotnických oborů*	2006
	uspořádaná díla	Jan Císař: *Světoví dramatici*	1997
		Markéta Mikysková – Miriam Prokešová – Marcela Stehlíková: *Babiččina kuchařka*	2006
		Lydia Petráňová: *Domovní znamení staré Prahy*	2008
		Jarmila Teplíková: *Houby známé a exotické*	2004
		Václav Zelený: *Rostliny Středozemí*	2004

Tabulka č. 4: Přehled jazykových projevů zařazených do administrativního stylu

	název textu	rok vydání
administrativní styl	*České centrum čistší produkce* – výroční zpráva 1997	1998
	Českoskalický zpravodaj, č. 4/1997	1997
	Efeméra – smlouvy	1996
	Fites – stanovy	1995
	Kolektivní smlouva ČD, s. o., na rok 1999	1999
	Lesní zákon – výňatek	1995
	Město Dačice informuje, ročník 1998	1998
	Návrhy zákonů a vyhlášek Ministerstva práce a sociálních věcí ČR	2002
	Návrhy zákonů a vyhlášek Ministerstva pro místní rozvoj ČR	2002
	Návrhy zákonů a vyhlášek Ministerstva vnitra ČR	2002
	Návrhy zákonů a vyhlášek Ministerstva zemědělství ČR	2002
	Návrhy zákonů a vyhlášek Ministerstva životního prostředí ČR	2002
	Smlouva o vytvoření a užití díla	2003
	Stanovy bytového družstva	1998
	Střední zdravotnická škola Nymburk	1998
	Školní administrativa	1997
	Texty z obecního úřadu Lázně Toušeň 1	1995
	Vyhlášky MěÚ Dačice	1995
	Výroční zpráva svazu PRO-BIO za rok 2002	2003
	Zpravodaj SETUZA, č. 1/1997	1997

Tabulka č. 5: Přehled jazykových projevů zařazených do řečnického stylu

	typ projevu	název projevu	rok proslovení
řečnický styl	projev ke dni vzniku samostatného československého státu – Václav Havel	Projev prezidenta republiky Václava Havla ke státnímu svátku České republiky	1999
		Projev prezidenta republiky Václava Havla ke státnímu svátku České republiky	2000
		Projev prezidenta republiky Václava Havla ke státnímu svátku České republiky	2001
		Projev prezidenta republiky Václava Havla ke státnímu svátku České republiky	2002
	novoroční projev – Václav Havel	Novoroční projev prezidenta republiky Václava Havla	2000
		Novoroční projev prezidenta republiky Václava Havla	2001
		Novoroční projev prezidenta republiky Václava Havla	2002
		Novoroční projev prezidenta republiky Václava Havla	2003
	projev ke dni vzniku samostatného československého státu – Václav Klaus	*Překonejme minulost přítomností* – Projev prezidenta republiky u příležitosti státního svátku 28. 10. 2004	2004
		Projev prezidenta republiky u příležitosti státního svátku dne 28. října 2007	2007
		Projev prezidenta republiky k 28. říjnu 2010 o smyslu našeho státu	2010
		Projev prezidenta republiky k 28. říjnu 2011	2011
	novoroční projev – Václav Klaus	Novoroční projev prezidenta republiky	2006
		Novoroční projev prezidenta republiky	2008
		Novoroční projev prezidenta republiky	2012
		Novoroční projev prezidenta republiky 2013	2013
	projev ke dni vzniku samostatného československého státu – Miloš Zeman	Projev prezidenta republiky při slavnostním ceremoniálu udílení státních vyznamenání	2013
		Projev prezidenta republiky během slavnostního ceremoniálu udělení státního vyznamenání České republiky ve Vladislavském sále	2014
	vánoční poselství – Miloš Zeman	Vánoční poselství prezidenta republiky Miloše Zemana	2013
		Vánoční poselství prezidenta republiky	2014

Tabulka č. 6: Přehled jazykových projevů zařazených do hovorového stylu

	lokalita, z níž pocházejí mluvčí	počet mluvčích	věk mluvčích	téma rozhovoru	rok uskutečnění rozhovoru
hovorový styl	Vlkoš, Kyjov	2	51, 25	rodina, domácnost	2010
	Přerov	2	48, 44	počasí, peněžní výdaje, domácí práce	2010
	Ostrava, Dvůr Králové	2	44, 40	škola, aktuální společenské dění, domácnost	2010
	Ostrava, Nová Paka	3	41, 20, 26	nakupování, vaření, pečení, práce	2010
	Brno, Frýdek-Místek	3	22, 24, 20	volný čas, Expo Shangai, Olympijské hry, škola	2010
	Chodovice, Vrchlabí	2	75, 43	diskuze nad časopisem, studánky	2010
	Krnov	2	28, 17	volejbal, oslava narozenin	2010
	Olomouc	2	36, 23	diskuze při snídani, nemoci, zásnubní prsten	2010
	Vrbno pod Pradědem	3	30, 52, 19	ceny a kvalita zboží, film, facebook, kamarádi	2010
	Nové Město na Moravě	2	47, 20	fotbal, sport, škola	2010
	Zlín	2	22, 19	škola, cestování, oblečení	2010
	Hronov, Jablonec nad Nisou	2	54, 55	práce, počasí, rodina, kouření, zahrada	2010
	Moravské Budějovice, Znojmo	3	19, 57, 86	příbuzní, výchova dětí, vzpomínky z dětství, historka z talk show	2010
	Karviná	2	19, 20	doprava, studium, spolužáci, Olomouc	2010
	Vsetín, Uherské Hradiště	2	46, 22	dovolená, cestování, práce v lese	2010
	Rožnov pod Radhoštěm	2	21, 21	bakalářská práce, pohádky, seriály, partner	2010
	Pardubice, Hradec Králové	2	21, 23	výlet, oslava narozenin, Silvestr, vztahy na facebooku, TV pořad	2010
	Litomyšl	2	20, 21	místní zábava, brigáda, praxe, festivaly	2010
	Havířov	2	20, 22	násilí, pojištění, peníze	2010
	Valašská Bystřice, Hutisko-Solanec	2	21, 21	jména, fotografie, výlet, přátelé, ubytování	2010

6. KVANTITATIVNÍ METODY

6.1 Bohatství slovníku

Výpočet slovního bohatství textu je problém, kterému lexikální statistika věnuje dlouhou dobu svou pozornost. Tato kvantitativní charakteristika je poměrně snadno interpretovatelná: čím méně se v textu opakují jednotlivá slova, tím lze daný projev považovat za lexikálně bohatší. Dlouhodobý a stále trvající zájem o tento fenomén není však způsoben jednoduchým porozuměním danému konceptu, ale převážně tím, že dosavadní indexy vytvořené pro výpočet bohatství slovníku nedokázaly zatím eliminovat vliv délky textu. Vzhledem k tomu, že slovní zásoba žádného mluvčího či pisatele není neomezená, je logické, že delší texty budou obsahovat větší množství lexikálních jednotek, které byly v komunikátu již dříve použity, a budou tedy mít i nižší slovní bohatství než texty kratšího rozsahu. Neúměrný vztah mezi délkou textu a nárůstem počtu nových slov je však problematický, neboť neumožňuje vzájemné srovnání textů odlišného rozsahu.

Možností, jak vypočítat slovní bohatství textu, byla již navržena celá řada. Mezi nejvýznamnější badatele v této oblasti se řadí mj. George Udny Yule, Pierre Guiraud, Ioan-Iovitz Popescu, Gabriel Altmann, ze slovenských lingvistů např. Jozef Mistrík, Gejza Wimmer, z českých lingvistů pak můžeme jmenovat zejména Marii Těšitelovou či Radka Čecha (k vývoji konceptu slovního bohatství a k rozličným indexům pro výpočet bohatství slovníku viz např. Těšitelová 1987; Kubát 2012, 2015; Čech – Popescu – Altman 2014).

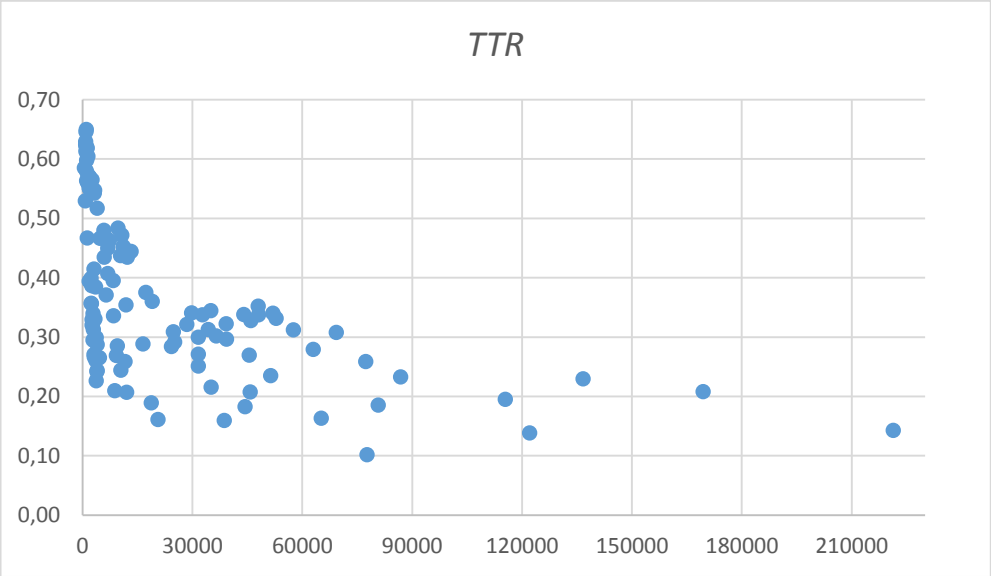
Zcela základní formulí pro výpočet této kvantitativní charakteristiky je *type-token ratio*, tedy poměr počtu typů k počtu tokenů v textu:

$$TTR = \frac{V}{N},$$

V = počet typů
N = počet tokenů

přičemž *type* (česky někdy také *typ*) chápeme jako abstraktní jednotku (obvykle slovní tvar nebo lemma), která je nezávislá na kontextu a jejíž konkrétní realizace (*token*) se v analyzovaném materiálu vyskytuje alespoň jedenkrát. Součet všech typů tak udává počet *různých* jednotek v textu, zatímco celkový počet tokenů značí množství *všech* jednotek, kterými je komunikát tvořen, tj. celkovou délku textu (Cvrček – Richterová (eds.) 2014a, b). *Type-token poměr* však výrazně ilustruje závislost míry opakování slov na délce textu. S přibývající délkou projevu se slova častěji opakují, tudíž hodnota *TTR*

postupně klesá. Pro názornost také uvádíme graf, v němž jsou zaznačeny hodnoty *TTR* pro všechny námi analyzované texty z různých funkčních stylů:



Graf č. 1: Závislost *TTR* na délce komunikátu ve 120 textech různých funkčních stylů

Problematice délky textu a slovního bohatství se věnovala celá řada lingvistů (viz např. níže), kteří se snažili s tímto nedostatkem vyrovnat. K tomu, aby se vliv délky textu redukoval, nebo dokonce zcela eliminoval, prováděli obvykle badatelé dva druhy úprav:

a) transformaci rovnice pro výpočet bohatství slovníku;
b) mechanické zkrácení analyzovaných textů do stejné délky, popř. rozdělení daných textů do několika částí vždy o stejné délce.

6.1.1 Modifikace rovnice pro výpočet slovního bohatství

Možností, jak modifikovat rovnice pro výpočet indexu měřícího slovní bohatství, existuje celá řada. Některé navržené postupy byly úspěšnější, jiné už méně, přesto se má v současné době za to, že stále nebyla představena taková transformace, která by vykazovala nulovou závislost na délce textu.

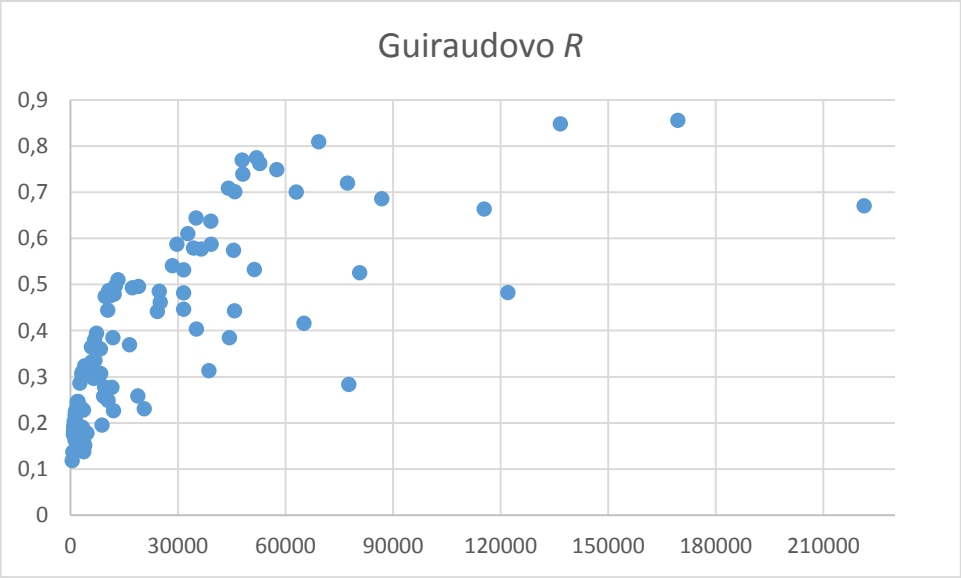
Zřejmě nejjednodušší modifikaci základního výpočtu *TTR* provedl Pierre Guiraud, na něhož navázala i Marie Těšitelová, který do jmenovatele k délce textu přidal odmocninu, jež měla závislost na délce projevu omezit (Těšitelová 1987, s. 67–69). Tato úprava rovnice:

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

V = počet typů

N = počet tokenů

však nebyla dostačující, jak lze vidět i na grafu č. 2 znázorňujícím výsledky této rovnice pro všechny texty, jež byly zařazeny do naší analýzy funkčních stylů. Tentokrát však platí, že s přibývající délkou textu hodnota indexu R postupně narůstá.



Graf č. 2: Závislost Guiraudova R na délce komunikátu ve 120 textech různých funkčních stylů

Jsou známé ovšem i případy, kdy některý pokus o úpravu rovnice, jež by počítala slovní bohatství textu bez jakéhokoli omezení, jistou dobu platil za zcela nezávislý na délce textu, později se ale zjistilo, že mezi rozsahem komunikátu a vypočítanými hodnotami indexu jistý vztah přece jen existuje. Takovým příkladem může být index lambda, jehož nezávislost byla prověřena na tisících textů z různých jazyků, posléze však sami autoři upozornili na jistou zákonitost, jež byla mezi délkou textu a indexem objevena (tento vztah nebyl snadno postřehnutelný, neboť hodnota lambdy s rostoucí délkou textu zpočátku stoupá, poté je v určitém intervalu na délce nezávislá, ovšem následně její hodnota opět klesá) (Čech 2016, s. 59).

Podobným způsobem pracuje i jiný index, jenž eliminuje vliv délky textu jen v omezeném intervalu a jehož transformaci Čech (2016, s. 91–93) považuje

za dosud nejméně závislou na délce komunikátu. Jedná se o McIntoshovu transformaci indexu opakování slov RR_{MC}:

$$RR_{MC} = \frac{1 - \sqrt{RR}}{1 - 1/\sqrt{V}} ,$$

přičemž

$$RR = \sum_{r=1}^{V} p_r^2 ,$$

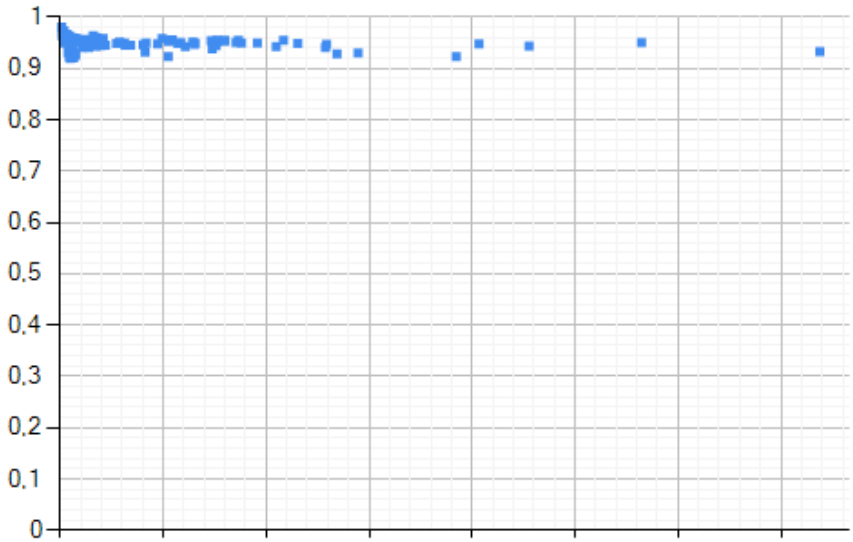
V = počet typů
p_r = pravděpodobnost výskytu slova r

a

$$p_r = \frac{f_r}{N}$$

f_r = absolutní frekvence daného slova
N = počet tokenů

Přestože se z grafu č. 3 může zdát, že tato rovnice nevykazuje závislost na velkém rozptylu délek textů, Čech empiricky ověřil, že nulový vliv na rozsah textu má jen v omezeném intervalu N ∈ <1 300; 5 000> tokenů.



Graf č. 3: Závislost RR_{MC} na délce komunikátu ve 120 textech různých funkčních stylů

Ačkoli tedy dosud nebyla sestavena rovnice, která by dokázala zcela eliminovat vliv délky textu, pro některé textové analýzy mohou být vyhovující i indexy, jejichž nezávislost na délce je omezena určitým intervalem. V takových případech je však nutné pracovat výhradně s texty o délce přísluš-

ného intervalu. Protože však námi vyexcerpované komunikáty z různých funkčních stylů nesplňují toto kritérium, není možné žádný z těchto indexů v naší práci aplikovat.

6.1.2 Mechanické krácení a rozdělení textů do menších subtextů

Pravděpodobně nejjednodušší strategií, jak lze komparovat texty odlišných délek bez toho, aby došlo k jakémukoli zkreslení, je provést analýzu jen části textů, tzn. lze např. srovnávat jen prvních tisíc slov. Tento způsob ovšem není zcela přijatelný, protože nerespektuje text jako homogenní celek: „[…] představme si například detektivní povídku, ve které autor úmyslně použije velké množství nových slov v závěrečné části (např. v souvislosti s odhalením okolností zločinu); použít pro stanovení slovního bohatství počátečních sto či tisíc slov je v takovém případě jistě zavádějící." (Čech – Popescu – Altmann 2014, s. 30)

Kubát (2015, s. 40) pro eliminaci délky textu zmiňuje další oblíbený přístup spočívající v rozčlenění každého textu na menší části o shodném rozsahu, přičemž hodnoty slovního bohatství vypočítané pro jednotlivé subtexty se nakonec zprůměrují. Výhodou této metody známé jako *standardized type-token ratio* (*STTR*) je skutečnost, že analýza je validní pro texty v celém rozsahu, problematickým se ale jeví fakt, že k rozdělení jednotlivých subtextů dochází zcela mechanicky, aniž by se přihlíželo k přirozeným strukturám v textu.

Rozdělení komunikátu na více částí o stejné délce je i základem metody, již autoři Covington a McFall představili v roce 2010 a jež se na základě výsledků nejrůznějších studií dosud jeví jako strategie, která dokázala vliv délky textu úspěšně omezit. Celý přístup pojmenovaný jako *Moving Average Type-Token Ratio* (*MATTR*) spočívá v aplikaci základního výpočtu *TTR*, jenž vykazuje značnou závislost na délce textu, novým způsobem. „We choose a window length (say 500 words) and then compute the TTR for words 1–500, then for words 2–501, then 3–502, and so on to the end of the text. The mean of all these TTRs is a measure of the lexical diversity of the entire text and is not affected by text length nor by any statistical assumptions." (Covington – McFall 2010, s. 96) Výpočet *MATTR* lze definovat takto (Kubát 2015, s. 41):

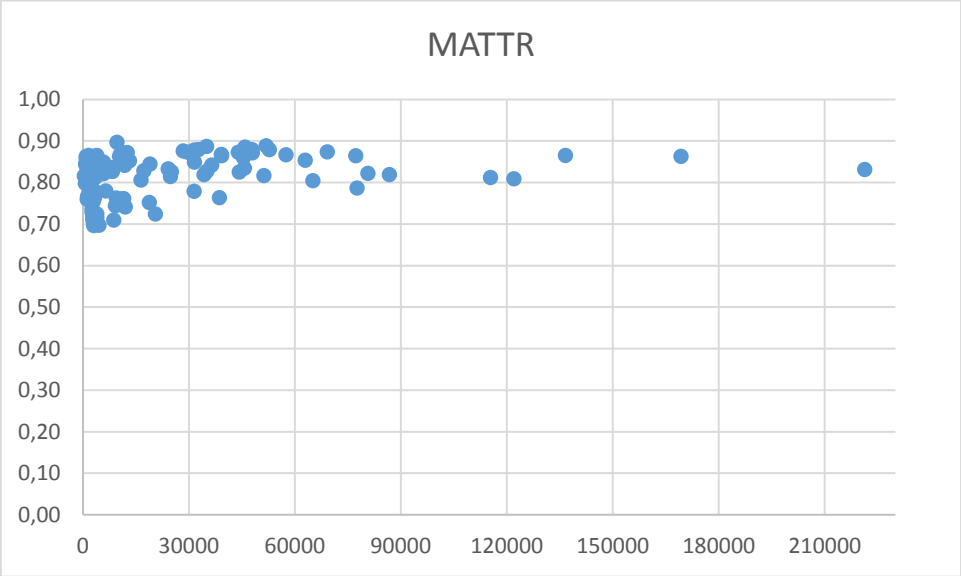
$$MATTR = \frac{\sum_{i=1}^{N-L} V_i}{L(N-L+1)},$$

L = arbitrárně zvolená velikost okna, přičemž L < N

N = délka textu uvedená v počtu tokenů

V_i = počet typů v jednotlivém okně

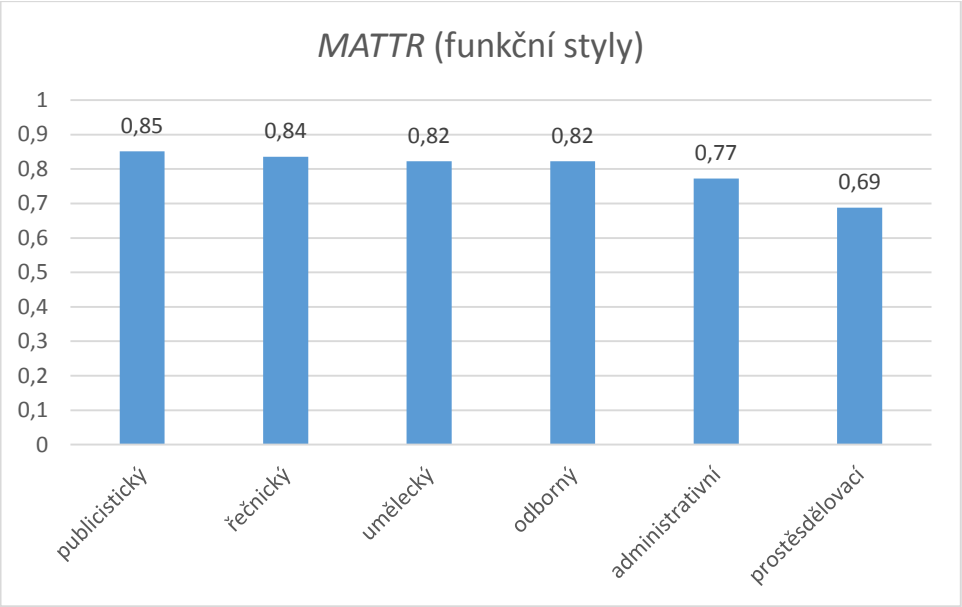
Nezávislost hodnot *MATTR* na délce textu ověřili mj. Kubát a Milička (2013) nebo Kubát (2015), výsledné hodnoty *MATTR* vypočítané pro náš analyzovaný materiál jsou zaznačeny v grafu č. 4:



Graf č. 4: Závislost *MATTR* na délce komunikátu ve 120 textech různých funkčních stylů

6.1.3 Analýza funkčních stylů pomocí *MATTR*

K výpočtu hodnot *MATTR* pro texty různých funkčních stylů jsme využili software *MATTR*, jehož autorem je Michael A. Covington. Délku tzv. okna jsme nastavili na 100 tokenů, tím, že se pak okno o dané délce vždy posouvá o jeden token, nedochází k rozpadu textu jako homogenní jednotky. Výsledné hodnoty *MATTR* vyjadřující bohatost slovníku jednotlivých funkčních stylů jsou zachyceny v grafu č. 5:



Graf č. 5: Výsledné hodnoty *MATTR* vypočítané pro jednotlivé funkční styly

Přestože již z grafu jsou patrné jisté diference mezi jednotlivými funkčními styly, komparovat takto získaná data mezi sebou není dostačující. Abychom nehodnotili různou velikost pozorovaného rozdílu vágními frázemi typu: výsledky se „velmi liší", „jsou téměř stejné", zjištěný rozdíl je „malý", „velký" apod., je potřeba výstupní hodnoty ověřit statistickým testem (srov. Čech 2014, s. 177–178). Užití statistického testu nám umožní jednoznačně rozhodnout, zda lze změřenou diferenci považovat za signifikantní, či nikoli, a rovněž nám vyloučí možnost vlivu náhody na sledovaný rozdíl. Pro srovnání vypočítaných hodnot *MATTR* z jednotlivých funkčních stylů použijeme asymptotický *u-test*[2], k jehož výpočtu využijeme program *Microsoft Excel*, konkrétně pak jeho komerční statistický doplněk *XLSTAT*:

$$u = \frac{|\overline{X_1} - \overline{X_2}|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

$\overline{X_1}$, $\overline{X_2}$ = aritmetický průměr výsledných hodnot *MATTR* každého souboru

s_1^2, s_2^2 = rozptyl *MATTR* každého souboru

n_1, n_2 = počet textů v každém souboru

[2] Označení *u-test* přebíráme od Čecha, Popesca a Altmanna (2014), Kubáta (2015), Kubáta, Matlacha a Čecha (2014), Matlacha (2014) aj. Ve statistice je tento vzorec znám jako *z-test*, který testuje odchylku průměrných hodnot od nulové hypotézy při normálním rozdělení dat (více např. viz Hendl 2004).

Signifikantní i nesignifikantní rozdíly mezi funkčními styly zjištěné na základě asymptotického *u-testu* jsou znázorněny v tabulce č. 7:

	adm	uměl	prostěsděl	odbor	řeč	pub
adm	×					
uměl	4,94	×				
prostěsděl	7,90	13,64	×			
odbor	4,41	0,02	12,16	×		
řeč	7,52	1,77	18,36	1,49	×	
pub	9,28	3,89	19,97	3,27	3,37	×

Tabulka č. 7: *u-test* hodnot *MATTR* (α = 0,05; u ≥ 1,96 vyjadřuje signifikantní diferenci)

Pro větší přehlednost můžeme z údajů v tabulce č. 7 vypočítat ještě váženou hodnotu u_v, která nám signalizuje míru vzájemné blízkosti či rozdílnosti skupiny textů určitého stylu vůči ostatním funkčním stylům. Hodnoty váženého rozdílu vypočítáme na základě rovnice (Čech 2016, s. 124):

$$u_v = \frac{\sum |u_i|}{\sqrt{k}},$$

u = testová hodnota u,
k = počet možných srovnání pro každou skupinu.

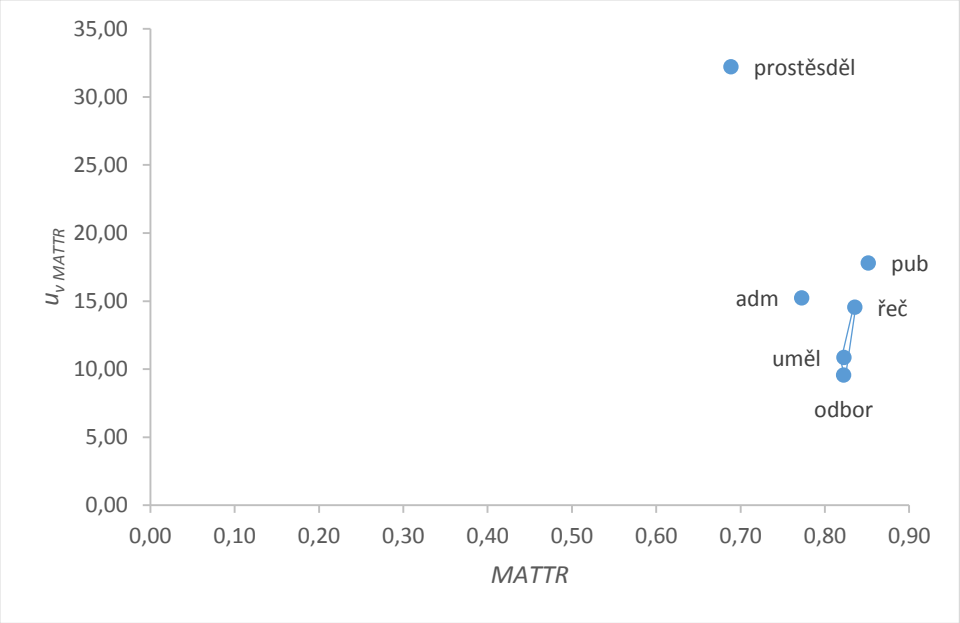
Porovnáváme-li tedy mezi sebou šest funkčních stylů, pak pro každou skupinu textů připadá v úvahu pět vzájemných srovnání, tedy vážená hodnota u_v např. pro administrativní styl by vypadala následovně:

$$u_v = \frac{4{,}94 + 7{,}90 + 4{,}41 + 7{,}52 + 9{,}28}{\sqrt{5}} = \frac{34{,}05}{\sqrt{5}} = 15{,}23.$$

V tabulce č. 8 jsou zaznamenány hodnoty váženého rozdílu u jednotlivých stylů, seřazené podle velikosti, graf č. 6 pak graficky vyjadřuje míru odlišnosti funkčních stylů navzájem:

styl	$u_{v\ MATTR}$
prostěsděl	32,22
pub	17,79
adm	15,23
řeč	14,54
uměl	10,85
odbor	9,55

Tabulka č. 8: Vážené rozdíly $u_{v\ MATTR}$ mezi jednotlivými funkčními styly



Graf č. 6: Podobnost/rozdílnost jednotlivých funkčních stylů na základě hodnot *MATTR* a $u_{v\ MATTR}$ (čára značí nesignifikantní rozdíl *MATTR* mezi styly)

6.1.3.1 Prostěsdělovací styl

Z výše uvedených tabulek č. 7 a č. 8 a z grafů č. 5 a č. 6 je z pohledu slovního bohatství zřetelné zcela specifické postavení stylu prostěsdělovacího. Mluvené nepřipravené projevy dosahují jednoznačně nejnižších hodnot *MATTR*, zároveň se od projevů ostatních stylů velice nápadně odlišují, vůči ostatním funkčním stylům dokonce vykazují největší diferenci. Tato tendence zcela koresponduje s obecnou představou tradiční stylistiky o užší slovní zásobě komunikátů mluvených oproti komunikátům psaným, pro lexikum takovýchto projevů je dokonce příznačná „preference slov s velkým rozsahem a malým obsahem" (Čechová – Krčmová – Minářová 2008, s. 204). Ve spontánních soukromých promluvách poměrně často nad funkcí informativní převládá funkce fatická, mluvčí tak v hojném počtu užívají opakující se kontaktové výrazy, zájmena či

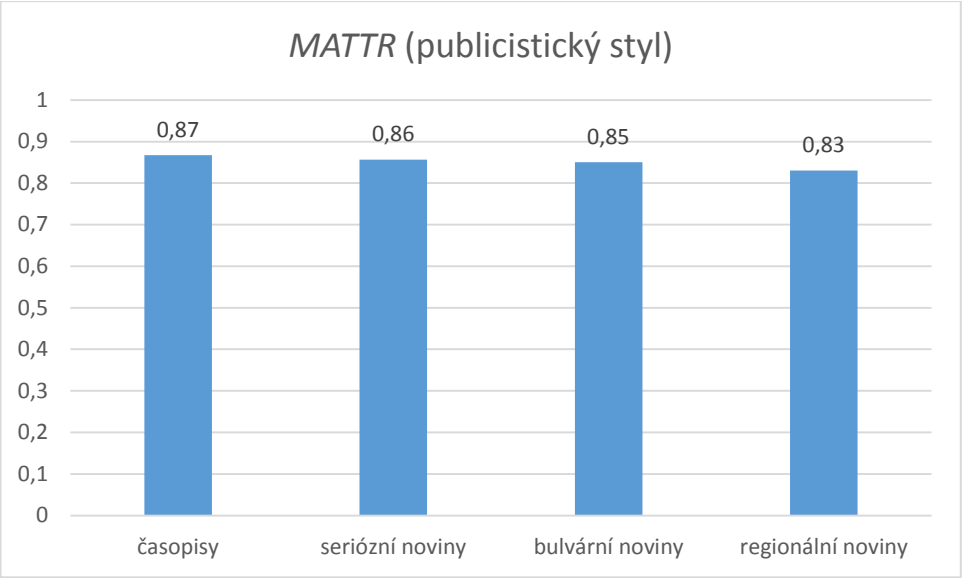
citoslovce, navíc při tvorbě promluvy neusilují o neotřelé, kreativní vyjádření, ale často se uchylují k opětovnému užití již vyřčených slov ve snaze o formulaci jednoznačné a srozumitelné výpovědi. Omezený inventář výrazových prostředků, zdá se, se v neformálních dialozích přitom uplatňuje i navzdory rozmanitým tématům projevů zařazených do analyzovaného vzorku.

6.1.3.2 Administrativní styl

Ani druhá pozice z hlediska nejužší slovní zásoby není pro nás nijak překvapivá. Texty administrativní povahy, podobně jako komunikáty odborné preferují vyjadřování přesné, citově neutrální, proto se v nich nezřídka uplatňují specifické termíny. S ohledem na omezenou synonymii odborné terminologie i s ohledem na stereotypní fráze a konstrukce užívané v administrativní komunikaci je slovní zásoba textů tohoto typu málo rozsáhlá a přispívá k jejich celkově úspornému charakteru vyjadřování.

6.1.3.3 Publicistický styl

Zcela opačnou pozici, tedy pozici s nejvyšším slovním bohatstvím, zaujímá styl publicistický. Vysoké hodnoty *MATTR* publicistických textů můžeme patrně přičítat jejich značné tematické různorodosti i odlišným funkcím jednotlivých publicistických žánrů. V textech publicistického charakteru se tak prolínají výrazy odborné i prostředky stylu uměleckého, jako např. obrazná vyjádření, frazémy či různě modifikovaná přísloví, využívány bývají i prvky z různých útvarů národního jazyka, čímž také přirozeně dochází k rozšíření lexika publicistických projevů. Vzhledem k velké mnohotvárnosti celého publicistického stylu zkusíme metodu *MATTR* aplikovat rovněž na jednotlivé subkategorie daného funkčního stylu s cílem odhalit mezi jednotlivými typy periodik jisté diference z pohledu slovního bohatství. Výsledné hodnoty indexu *MATTR*, asymptotického *u-testu* i vážených rozdílů spolu s grafickým znázorněním vzájemné blízkosti jednotlivých podkategorií publicistického stylu jsou zachyceny v následujících grafech a tabulkách:



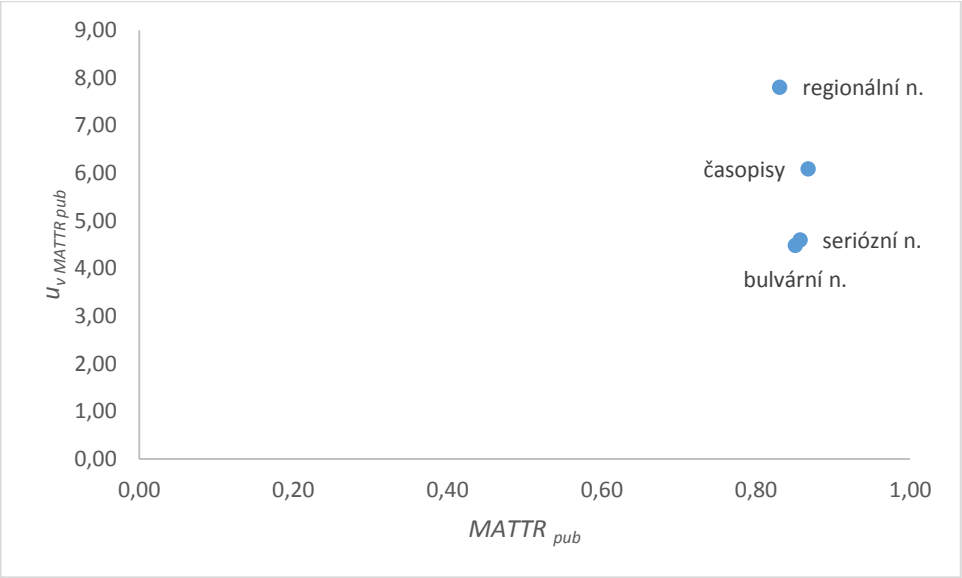
Graf č. 7: Výsledné hodnoty *MATTR* $_{pub}$ vypočítané pro jednotlivé subkategorie publicistického funkčního stylu

	bulvární noviny	časopisy	seriózní noviny	regionální noviny
bulvární noviny	×			
časopisy	**3,00**	×		
seriózní noviny	1,37	**2,01**	×	
regionální noviny	**3,40**	**5,54**	**4,58**	×

Tabulka č. 9: *u-test* hodnot *MATTR* $_{pub}$ ($\alpha = 0{,}05$; $u \geq 1{,}96$ vyjadřuje signifikantní diferenci)

žánr	$u_{v\ MATTR\ pub}$
regionální noviny	7,81
časopisy	6,09
seriózní noviny	4,60
bulvární noviny	4,49

Tabulka č. 10: Vážené rozdíly $u_{v\ MATTR\ pub}$ mezi jednotlivými subkategoriemi publicistického funkčního stylu



Graf č. 8: Podobnost/rozdílnost jednotlivých subkategorií publicistického funkčního stylu na základě hodnot *MATTR* $_{pub}$ a $u_{v\ MATTR\ pub}$ (čára značí nesignifikantní rozdíl *MATTR* $_{pub}$ mezi skupinami textů)

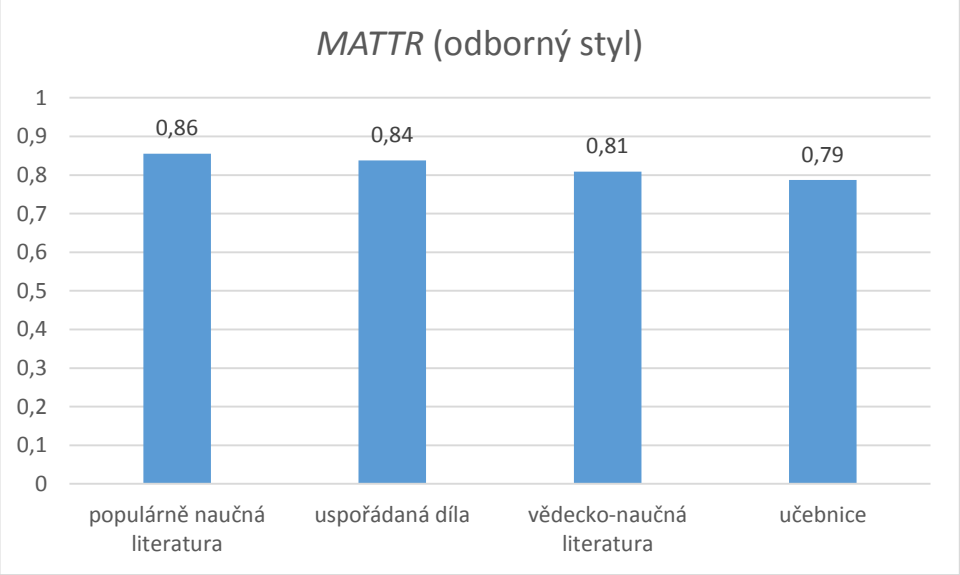
Přestože graf. č. 7 nezaznamenává příliš rozdílné hodnoty slovního bohatství v jednotlivých typech periodik, po statistickém ověření zjišťujeme, že lze i takto mírně odlišné výsledky považovat za signifikantní. Nejvíce se vymykající subkategorií publicistického stylu a zároveň subkategorií s nejchudší slovní zásobou se staly noviny regionální. Příčinu můžeme hledat v poněkud stereotypních tématech vztahujících se výhradně ke konkrétnímu regionu, a tedy i v nutnosti hojně opakovat označení příslušné lokality či jména místních organizací, obvykle v sobě rovněž nesoucích název daného města. Městské zpravodaje, které v našem vzorku reprezentují regionální periodika, se navíc obvykle své čtenáře nesnaží upoutat příliš originálními formulacemi, ale dbají spíše na naplnění informační funkce pomocí osvědčených automatizovaných vyjádření.

Naopak časopisy jsou více zaměřeny analyticky, publikované informace dále rozebírají, komentují, odhadují či odhalují příčiny a důsledky různých jevů, popř. se snaží předkládané události nějak hodnotit. Právě v publicistických projevech tohoto typu je proto dostatek prostoru pro aktualizované jazykové prostředky, figury či tropy jsou navíc pro recipienta nezřídka působivější a snáze v něm zanechají dojem, který jej přesvědčí o správnosti prezentovaného názoru. Úsilí o neobvyklou formu sdělení a o porušení modelového vyjádření se potom dle našeho názoru promítá i ve větší slovní zásobě analytické publicistiky.

Poněkud překvapivý je pro nás zjištěný nesignifikantní rozdíl mezi novinami seriózního a bulvárního typu. Očekávali bychom přitom, že právě tato periodika se budou vůči sobě nejvíce vymezovat. Nižší hodnoty slovního bohatství seriózních novin vzhledem k časopisům je možné odůvodnit zastoupením nejen analytických útvarů, ale také žánrů zpravodajského typu, jejichž snaha o přesné, jednoznačné a neutrálně zabarvené vyjádření má za následek limitovanou možnost užití vhodných výrazů, a tedy i užší slovní zásobu. Chudší slovník bulvárních textů je pak pravděpodobně způsoben použitím omezenějšího inventáře výrazových prostředků s ohledem na usnadnění recepce méně vzdělaným recipientům. Do jisté míry mohou být ovšem podobné hodnoty slovního bohatství novin seriózního a bulvárního typu také důsledkem stírání rozdílů mezi oběma typy periodik, menší slovní zásoba pak může poukazovat právě na tendenci k bulvarizaci seriózního tisku. Ačkoli tedy za obdobnými výslednými hodnotami slovního bohatství bulvárních a seriózních novin mohou stát různé příčiny, z hlediska kvantitativní metody *MATTR* lze v našem výzkumu oba typy periodik považovat za rovnocenné.

6.1.3.4 Odborný styl

Styly odborný, umělecký a řečnický se co do bohatosti slovníku nacházejí zhruba na střední pozici, mezi sebou se však již nijak signifikantně neliší. Protože jsou ovšem příslušné styly tvořeny texty poměrně rozmanitého charakteru, podíváme se blíže, zda se jistá diferenciace projevuje alespoň mezi jejich jednotlivými subkategoriemi:



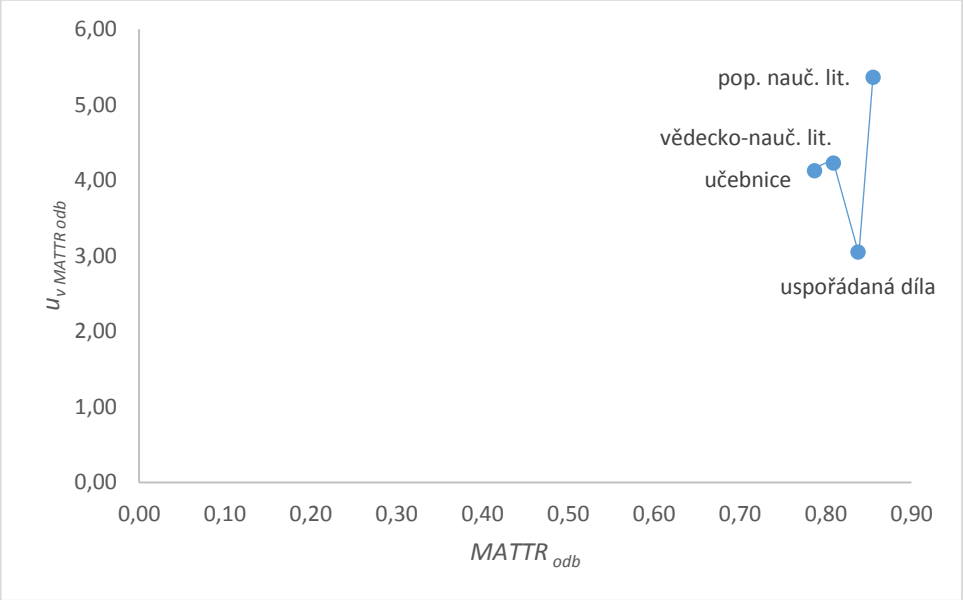
Graf č. 9: Výsledné hodnoty *MATTR* $_{odb}$ vypočítané pro jednotlivé subkategorie odborného funkčního stylu

	uspořádaná díla	pop. nauč. lit.	vědecko-nauč. lit.	učebnice
uspořádaná díla	×			
pop. nauč. lit.	1,18	×		
vědecko-nauč. lit.	1,80	**4,39**	×	
učebnice	**2,30**	**3,72**	1,13	×

Tabulka č. 11: *u-test* hodnot *MATTR* $_{odb}$ (α = 0,05; u ≥ 1,96 vyjadřuje signifikantní diferenci)

žánr	$u_{v\ MATTR\ odb}$
pop. nauč. lit.	5,36
vědecko-nauč. lit.	4,23
učebnice	4,13
uspořádaná díla	3,05

Tabulka č. 12: Vážené rozdíly $u_{v\ MATTR\ odb}$ mezi jednotlivými subkategoriemi odborného funkčního stylu



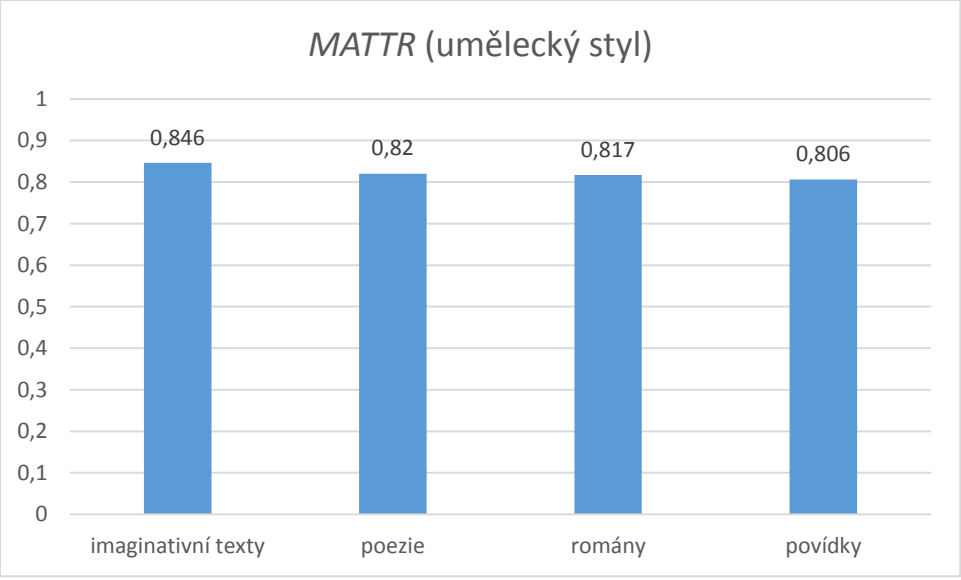
Graf č. 10: Podobnost/rozdílnost jednotlivých subkategorií odborného funkčního stylu na základě hodnot *MATTR* $_{odb}$ a $u_{v\ MATTR\ odb}$ (čára značí nesignifikantní rozdíl *MATTR* $_{odb}$ mezi skupinami textů)

Z grafu č. 9 můžeme vypozorovat jistou tendenci k rozdělení textů odborného charakteru do dvou skupin. První skupinu s menším rozsahem slovní zásoby tvoří komunikáty vědecko-naučné a učebnice. Na první pohled bychom mohli tuto skupinu označit za poněkud nesourodou, neboť pod učebnicemi si obvykle představujeme zjednodušené texty určené pro žáky základních, popř. středních škol. Pokud se ovšem podíváme na konkrétní texty

reprezentující v našem výzkumu příslušnou kategorii, zjistíme, že se jedná hlavně o skripta určená vysokoškolským studentům. Projevy vědecko-naučné i učební lze tedy považovat za více exaktní, s tím potom také souvisejí přísnější normy pro vyjadřování. Texty dané povahy jsou omezeny jen na spisovný jazyk, navíc jsou svázány odbornou terminologií, která příliš variantních výrazů nepřipouští. Nižší hodnoty *MATTR* jsou proto pro tyto skupiny textů očekávatelné. Oproti tomu uspořádaná díla a díla populárně naučná berou na zřetel méně poučeného čtenáře, přizpůsobují se mu výběrem i formou prezentovaných faktů, přičemž kladou důraz na snadnou srozumitelnost. Při vyjadřování se tak striktně neřídí pravidly odborného stylu a využívají širších možností slovní zásoby. Nesignifikantní diference mezi texty vědecko-naučnými a uspořádanými díly ovšem i tuto mírnou vzdálenost mezi oběma skupinami odborných textů stírá, z prezentovaných výsledků kvantitativní charakteristiky *MATTR* tak zřejmě nelze odhalit větší diferenciaci odborného stylu.

6.1.3.5 Umělecký styl

Výsledné hodnoty indexu *MATTR*, asymptotického *u-testu* a vzájemnou odlišnost jednotlivých uměleckých žánrů prezentují následující grafy a tabulky:



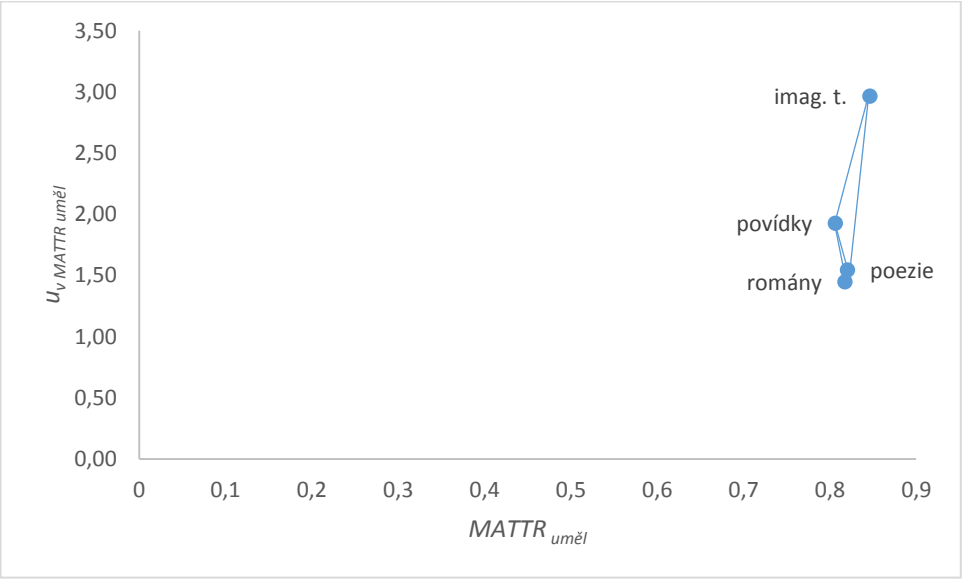
Graf č. 11: Výsledné hodnoty *MATTR* $_{uměl}$ vypočítané pro jednotlivé subkategorie uměleckého funkčního stylu

	povídky	imaginativní texty	romány	poezie
povídky	×			
imaginativní texty	1,93	×		
romány	0,57	1,65	×	
poezie	0,84	1,55	0,29	×

Tabulka č. 13: *u-test* hodnot *MATTR* $_{uměl}$ (α = 0,05; u ≥ 1,96 vyjadřuje signifikantní diferenci)

žánr	$u_{v\ MATTR\ uměl}$
imaginativní texty	2,97
povídky	1,93
poezie	1,55
romány	1,45

Tabulka č. 14: Vážené rozdíly $u_{v\ MATTR\ uměl}$ mezi jednotlivými subkategoriemi uměleckého funkčního stylu



Graf č. 12: Podobnost/rozdílnost jednotlivých subkategorií uměleckého funkčního stylu na základě hodnot *MATTR* $_{uměl}$ a $u_{v\ MATTR\ uměl}$ (čára značí nesignifikantní rozdíl *MATTR* $_{uměl}$ mezi skupinami textů)

Ačkoli bychom očekávali, že texty různých uměleckých žánrů se budou vzájemně značně odlišovat, asymptotický *u-test* neprokázal z hlediska slovního bohatství mezi jednotlivými subkategoriemi uměleckého stylu žádný signifikantní rozdíl. Překvapivé jsou zejména blízké hodnoty *MATTR* básní a románů, neboť právě v poezii mají autoři pravděpodobně největší možnost

uplatnit ve svém textu výrazy z různých stylových vrstev či užít pojmenování s různým příznakem, od nespisovného přes dobový až po citově zabarvený, navíc nemusejí být básnické texty ani tolik svázány gramatickými pravidly, a tudíž ani gramatická slova není potřeba užívat v tak velké frekvenci jako v jiných komunikátech. Také malý rozsah tohoto typu textů a obvyklá snaha nestavět blízko sebe tytéž pojmenovávací jednotky podporují předpoklad většího slovního bohatství básnických děl, zejména např. oproti dlouhým prozaickým románům.

Pokud se ovšem podíváme na výsledky, k nimž došel Kubát při srovnávání uměleckých žánrů Karla Čapka z hlediska *MATTR*:

	román	povídka	cestopis	studie	sloupek	pohádka	dopis
román	x						
povídka	1,35	x					
cestopis	**3,31**	1,57	x				
studie	**5,04**	**5,78**	**7,64**	x			
sloupek	1,63	0,24	1,33	**6,05**	x		
pohádka	**7,14**	**7,68**	**9,04**	**3,13**	**7,88**	x	
dopis	0,07	1,10	**2,47**	**4,01**	1,31	**6,25**	x
báseň	1,00	1,44	**1,99**	0,94	1,54	**2,57**	0,94

Tabulka č. 15: Výsledky *u-testu* mezi žánry Karla Čapka (Kubát 2015, s. 43)

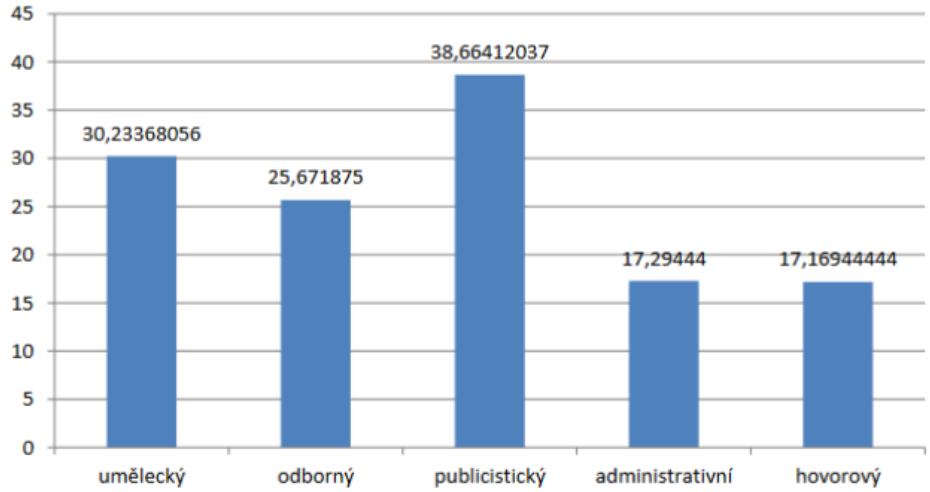
zjistíme, že zvolená metoda sice dokázala detekovat rozdíly mezi některými skupinami textů, ovšem žánry, jež byly zahrnuty do naší analýzy, se ani v Kubátově práci nejeví signifikantně odlišné. Rozdílně se Kubátovi vyjevila jen studie, již můžeme zařadit do stylu odborného, pohádka, která je silně přizpůsobena recepci malých dětí, a cestopis, jenž svou výraznou popisností stojí také spíše na okraji uměleckého stylu. Tradiční umělecké texty určené dospělým čtenářům se tedy z hlediska *MATTR* ani v Kubátově analýze nijak výrazně od sebe neodlišují. Dá se sice předpokládat, že jistý vliv na naše zjištěné výsledky může mít malý počet textů, jež jsme použili pro dané srovnání, vzhledem ale k nevelkým rozdílům odhaleným uvnitř dalších stylů i vzhledem k obdobným zjištěním v Kubátově práci zůstává otázkou, zda je kvantitativní metoda *MATTR* vhodným nástrojem pro diferenciaci tak úzce vymezených skupin textů, jako jsou žánry.

6.1.3.6 Řečnický styl

Ani styl řečnický, zdá se, není z pohledu slovního bohatství nijak výrazně vyhraněný, v grafu č. 6 vidíme, že se jeho pozice nachází mezi stylem uměleckým a publicistickým, nesignifikantní odlišnost však vykazuje i vůči stylu odbornému. Podobně jako komunikáty publicistické i slavnostní rétorické projevy podávají do určité míry věcné informace o událostech, při jejichž příležitosti je projev prosloven, zároveň však usilují také o persvazi širokého okruhu posluchačů, kterou se snaží podpořit přesnou terminologií a věcnými argumenty blízkými komunikátům odborným. V neposlední řadě navozují příležitostné řeči slavnostní atmosféru a snaží se recipienta zaujmout, k čemuž nezřídka využívají i obrazná pojmenování, tropy či figury, jež se často uplatňují rovněž v textech uměleckých. Slavnostní promluvy tedy využívají prvky z různých funkčních stylů, s čímž nakonec koresponduje i jejich výsledné postavení v našem výzkumu.

6.1.4 Srovnání výsledků *MATTR* a *rozsahu lexika*

Zjištěné tendence můžeme porovnat i s výsledky výzkumu Miroslava Kubáta (2012), který se soustředil na výpočet bohatství slovníku v jednotlivých funkčních stylech mj. za pomoci *rozsahu lexika*, tj. základní charakteristiky slovního bohatství od Marie Těšitelové:



Graf č. 13: Bohatství slovníku ve funkčních stylech (slovoforma) (Kubát 2012, s. 54)

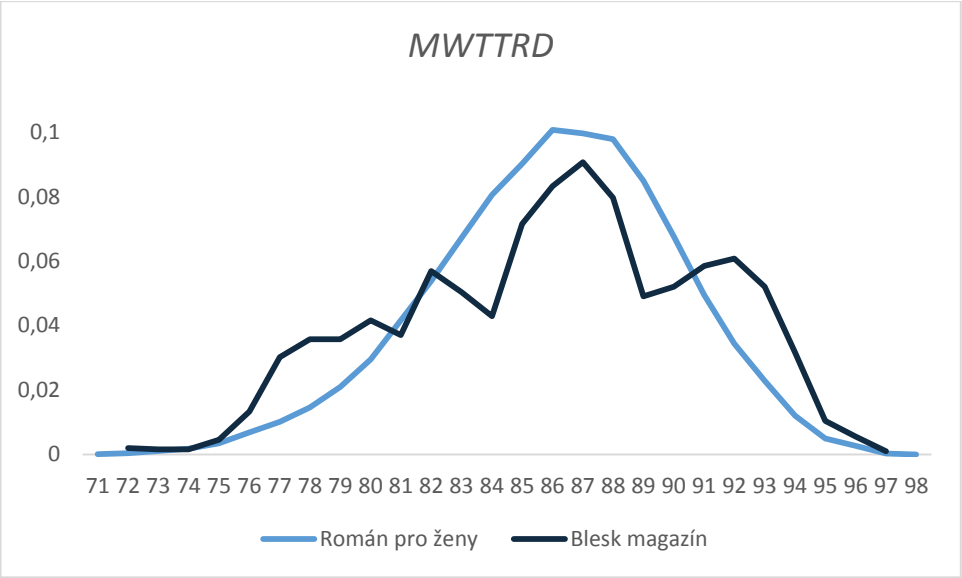
Ačkoli je naše metoda od výpočtu tzv. *rozsahu slovníku* značně odlišná, výsledné hodnoty statisticky ověřuje a pracuje s celými texty (na rozdíl od zmíněného indexu Těšitelové, který na statistické testování rezignuje a je závislý na délce projevu, tudíž jej lze počítat jen pro komunikáty upravené na

stejnou délku), výsledné tendence v Kubátově šetření se do značné míry shodují i s našimi zjištěními. Vzhledem ke skutečnosti, že Kubát do svého výzkumu nezařadil promluvy řečnického stylu, nemůžeme bohužel postavení tohoto typu textů nijak porovnat, s výjimkou téměř shodné pozice stylu administrativního a prostěsdělovacího v Kubátově výzkumu však výsledné vztahy mezi ostatními funkčními styly odpovídají i vztahům odhaleným v naší analýze. Velice malý rozdíl mezi projevy z administrativní a z neoficiální soukromé oblasti sám Kubát ve své práci přičítá poněkud odlišnému způsobu zpracování korpusů SYN2010 a PMK, z nichž komunikáty daných stylů excerpoval, naše výsledky potvrzující nejmenší slovní bohatství prostěsdělovacího stylu tak zřejmě lépe odpovídají obecné představě o užší slovní zásobě mluvených projevů oproti komunikátům psaným.

6.1.5 *Moving Window Type-Token Ratio Distribution* (*MWTTRD*)

Ačkoliv se *MATTR* jako první metoda výpočtu slovního bohatství bez závislosti na délce komunikátu jeví jako vhodný a užitečný nástroj, Kubát s Miličkou (2013) upozorňují na její potenciální nevýhodu, jíž je výsledek ve formě jediné číselné hodnoty, na jejímž základě se poté interpretuje celý text. Hodnoty *TTR* v průběhu textu mohou totiž značně kolísat, jediná průměrná hodnota ovšem taková specifika textu nedokáže postihnout, proto pak může dojít k jistému zkreslení. Z tohoto důvodu zmíněná dvojice lingvistů navrhla modifikaci původní metody, nazvanou *Moving Window Type-Token Ratio Distribution* (*MWTTRD*), která postupuje stejně při výpočtu typů v jednotlivých oknech, která se ale liší ve formě prezentace konečného výsledku. *MWTTRD* totiž postihuje text jako celek, když ve formě křivky znázorňuje poměrné zastoupení počtu typů ve všech oknech.

Pro snazší představu jsme metodu aplikovali na dva komunikáty odlišných funkčních stylů, vybrané z našeho korpusu:



Graf č. 14: Výsledné hodnoty *MWTTRD* vypočítané pro umělecký a publicistický text

Kubát s Miličkou jsou přesvědčeni, že jimi navržená metoda vykazuje přesnější výsledky měření slovního bohatství textu, sami však uznávají, že zjištěné rozdíly po statistickém testování srovnávaných komunikátů lze poměrně obtížně interpretovat. Proto nebudeme tuto metodu na náš analyzovaný soubor textů aplikovat, i když Kubát a Milička ve své studii prokázali, že pro některé typy výzkumů, např. týkající se diferenciace jednotlivých jazyků, může být tento nástroj užitečný a může vykazovat dobré výsledky.

6.2 Tematická koncentrace textu

Asi u každého textu v první řadě zjišťujeme, o čem vlastně je, jakému tématu, popř. jakým tématům se věnuje. Z odbornějšího hlediska nás však téma může zajímat např. i v souvislosti s otázkou, zda autor během projevu dodržel zadané téma nebo který z posuzovaných textů (či skupin textů) je více tematicky vyhraněný. Pokud bychom chtěli mezi sebou jednotlivé komunikáty z hlediska tematické charakteristiky porovnávat, museli bychom se obvykle spolehnout pouze na naše intuitivní hodnocení. U jednotlivých krátkých textů bychom zřejmě dokázali určit, který projev se více zaměřuje na dané téma a který je tematicky méně vyhraněný, o dlouhých textech, jako jsou např. romány, nebo o celých souborech textů, jež např. spadají do různých funkčních stylů, však zřejmě ani takto subjektivně rozhodnout nedokážeme. Aby se překonalo intuitivní hodnocení a odhadování tematické zaměřenosti textu, vytvořil Popescu, později i s pomocí Altmanna a Čecha, metodu měřící tzv. *tematickou koncentraci textu* (*TK*), kterou později doplnili indexem *sekundární tematické koncentrace* (*STK*) a *proporcionální tematické koncentrace* (*PTK*) (bližší informace o dané metodě a o její aplikaci na texty různého charakteru viz např. Čech – Popescu – Altmann 2014; Čech 2016; Čech – David – Davidová Glogarová 2013; Kubát 2015; všechny prezentované vzorce přebíráme z Čech – Popescu – Altmann 2014).

Celá metoda je založená na tom, že jsou v textu identifikována tematická slova, u nichž je vypočítána tematická váha a nakonec tematická koncentrace celého textu. Aby mohla být tematická slova v komunikátu detekována, využívá se vlastnosti tzv. *h-bodu* a kvantitativních charakteristik jazykových jednotek ve frekvenční distribuci. Pokud totiž slova určitého textu uspořádáme podle klesající frekvence, jako *h-bod* se označuje místo, kde pořadí slova i jeho frekvence mají stejnou hodnotu, tzn.:

$$h = r, \quad \text{pokud } r = f(r)$$

r = pořadí slova ve frekvenční distribuci
$f(r)$ = absolutní frekvence slova v daném pořadí

Nenachází-li se ovšem v distribuci žádný takový bod, je nutné jeho hodnotu vypočítat podle vzorce:

$$h = \frac{f(i)j - f(j)i}{j - i + f(i) - f(j)}, \qquad \text{pokud } r \neq f(r)$$

i = pořadí slova, u něhož platí, že se jedná o největší číslo, u něhož $i < f(i)$ (a zároveň $i < j$)
j = pořadí slova, u něhož platí, že se jedná o nejmenší číslo, u něhož $j > f(j)$ (a zároveň $i < j$)
$f(i)$, $f(j)$ = frekvence slov v daném pořadí

Tuto obecnou a možná poněkud obtížně srozumitelnou rovnici pro ilustraci aplikujeme na konkrétní text administrativního stylu. Následující tabulka zachycuje prvních 15 nejfrekventovanějších slov daného textu:

pořadí	frekvence	lemma
1	28	dílo
2	26	autor
3	24	objednatel
4	22	a
5	20	v
6	16	smlouva
7	16	být
8	14	tento
9	13	nebo
10	**11**	li
11	**10**	právo
12	10	s
13	9	část
14	8	ten
15	7	na

Tabulka č. 16: Patnáct nejfrekventovanějších slov v textu administrativního charakteru

Kdyby se substantivum „právo“, jež se v tabulce nachází na 11. pozici, vyskytovalo v textu 11×, také hodnota *h-bodu* by se rovnala 11. Protože však v uvedené frekvenční distribuci nenajdeme slovo, jehož pořadí i frekvence by se shodovaly, vypočítáme hodnotu *h-bodu* dle výše uvedené rovnice:

$$h = \frac{f(i)j - f(j)i}{j - i + f(i) - f(j)} = \frac{11 \cdot 11 - 10 \cdot 10}{11 - 10 + 11 - 10} = 10{,}5\ .$$

H-bod je pro výpočet tematické koncentrace důležitý, protože ve frekvenční distribuci textu vytváří neostrou hranici mezi slovy autosémantickými

a synsémantickými. Oblast nad *h-bodem* bývá obvykle určena synsémantikům, která jsou v textech pravidelně přítomna s velkou frekvencí; pokud se tedy nad hranici *h-bodu* dostane i některé z užitých autosémantik, můžeme tuto skutečnost interpretovat jako doklad větší koncentrovanosti mluvčího či pisatele na téma reprezentované *tematickým slovem*, tj. autosémantickým slovem v oblasti nad *h-bodem*. V našem příkladovém textu tedy můžeme za *tematická slova* označit výrazy „dílo", „autor", „objednatel" a „smlouva" (pomocné sloveso „být" se za tematické slovo nepovažuje), z nichž si již snadno odvodíme, že analyzovaným textem je *Smlouva o vytvoření a užití díla*.

Detekce tematických slov je ovšem jen dílčím krokem v měření tematické koncentrace celého textu, i když pro badatele s vhodným zaměřením práce může být i tento krok užitečný. Pro stanovení celkové tematické zaměřenosti textu je však potřeba dále znát *tematickou váhu* jednotlivých *tematických slov*, která se vypočítá jako vzdálenost mezi *h-bodem* a příslušným *tematickým slovem*. Aby bylo možné porovnávat také texty odlišné délky, je nutné *tematickou váhu* následně normalizovat. *Normalizovanou tematickou váhu* vypočítáme jako:

$$TV_{slovo} = 2\frac{(h-r\prime)f(r')}{h(h-1)f(1)},$$

r' = pořadí autosémantického slova nad *h-bodem*
h = *h-bod*
$f(r')$ = frekvence autosémantického slova nad *h-bodem*

Tematickou koncentraci textu potom tvoří součet tematických vah všech tematických slov:

$$TK = \sum_{r'=1}^{T} 2\frac{(h-r\prime)f(r')}{h(h-1)f(1)},$$

T = počet tematických slov
r' = pořadí autosémantického slova nad *h-bodem*
h = *h-bod*
$f(r')$ = frekvence autosémantického slova nad *h-bodem*

V našem příkladu s administrativním textem se tedy *TK* vypočítá jako:

$$TK_{adm} = TV_{dílo} + TV_{autor} + TV_{objednatel} + TV_{smlouva} =$$
$$= 2\frac{(10{,}5-1)28}{10{,}5(10{,}5-1)28} + 2\frac{(10{,}5-2)26}{10{,}5(10{,}5-1)28} + 2\frac{(10{,}5-3)24}{10{,}5(10{,}5-1)28} + 2\frac{(10{,}5-6)16}{10{,}5(10{,}5-1)28} =$$
$$= 0{,}19048 + 0{,}15825 + 0{,}12889 + 0{,}05156 = 0{,}52918.$$

6.2.1 Analýza funkčních stylů pomocí *TK*

Výsledné hodnoty tematické koncentrace mohou být ovšem ovlivněny několika faktory. První spočívá v rozhodnutí, s jakými jednotkami budeme při kvantifikaci *TK* pracovat. Přestože jsme již ve 4. kapitole věnující se volbě jazykové jednotky poukázali na výhody práce se slovními tvary, na základě srovnání výsledků dosažených při analýze tematických charakteristik shodných textů lemmatizovaných i nelemmatizovaných (srov. např. Čech 2016, s. 38–43) bylo prokázáno, že při kvantifikaci s lemmaty se obecně v menším počtu případů vyskytuje nulová *TK* textu, což pro lingvistickou interpretaci znamená lepší výsledek, neboť nulová hodnota o tematické charakteristice textu téměř nic neprozrazuje. V případě této kvantitativní metody tedy považujeme za příhodnější pracovat s lemmaty, která bývají u flektivních jazyků obecně považována za nejvhodnější volbu mezi jazykovými jednotkami.

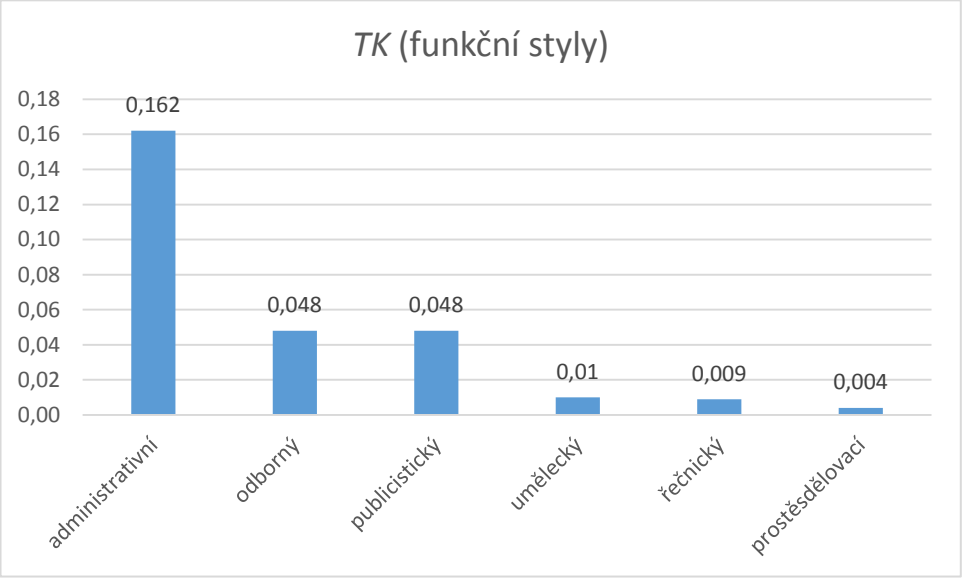
Druhým úskalím spojeným s výpočtem *TK* může být otázka, jaké slovní druhy pokládat za tematická slova. Přestože se domníváme, že za významné reprezentanty tematických charakteristik textu lze považovat substantiva, adjektiva, verba i adverbia, Čech, Popescu a Altmann (2014) i Kubát (2015) volí za tematická slova jen substantiva a jejich predikáty prvního řádu, tedy adjektiva a verba. Vzhledem k tomu, že k výpočtu *TK* všech 120 textů různých funkčních stylů budeme používat software *QUITA*, který jako tematická slova identifikuje jen podstatná jména, přídavná jména a slovesa, nezbývá nám, než se k tomuto způsobu nazírání na výrazy reprezentující hlavní téma textu připojit.

Poslední problém, jejž je potřeba vzít do úvahy před aplikací indexu *TK* na náš zkoumaný vzorek, souvisí s ne/závislostí výsledných hodnot na rozsahu textu. Třebaže dosavadní studie měly za to, že metoda měření tematické koncentrace textu dokázala vliv délky projevu zcela eliminovat, nejnovější publikace *Tematická koncentrace textu v češtině*, která vyšla na konci března 2016, tento poznatek mírně koriguje. Čech (2016, s. 55–67) v monografii uvádí, že zcela nezávislá na délce textu je metoda v intervalu $N \in \langle 200; 6\,500 \rangle$ slov, u dlouhých komunikátů podle něj poté dochází k postupnému ustálení hodnot indexu *TK* v rámci malého intervalu. Čech tuto skutečnost vysvětluje tím, že u velmi dlouhých textů autor již zpravidla nebývá schopen záměrně kontrolo-

vat frekvenční charakteristiky komunikátu, v textu o velkém rozsahu se proto postupně začnou prosazovat přirozené mechanismy vlastní každému textu.

Aby se badatel vešel do optimálního rozsahu textu v rozmezí 200 až 6 500 tokenů, doporučuje Čech (2016, s. 136) rozčlenit delší komunikáty na menší části, např. román na jednotlivé kapitoly. Takové subtexty, které ovšem ani po logickém rozdělení nespadají do vhodného intervalu, Čech jednoduše z analýzy vyřadí. Protože nám tento způsob nepřipadá vědecky úplně čistý (v analyzovaném materiálu část textu po takovém zásahu chybí), smíříme se raději s nepatrnou tendencí, která se u delších textů objevuje, kterou ovšem ani Čech nakonec nepovažuje za výrazně zřetelnou. Metodu měření tematické koncentrace tedy můžeme s vědomím této skutečnosti považovat za nezávislou na rozsahu komunikátu.

V následujícím grafu uvádíme výsledné hodnoty tematické koncentrace vypočítané pro skupiny textů z jednotlivých funkčních stylů:



Graf č. 14: Výsledné hodnoty *TK* vypočítané pro jednotlivé funkční styly

Stejně jako v případě *MATTR* musíme i nyní ověřit zjištěné hodnoty statistickým *u-testem*:

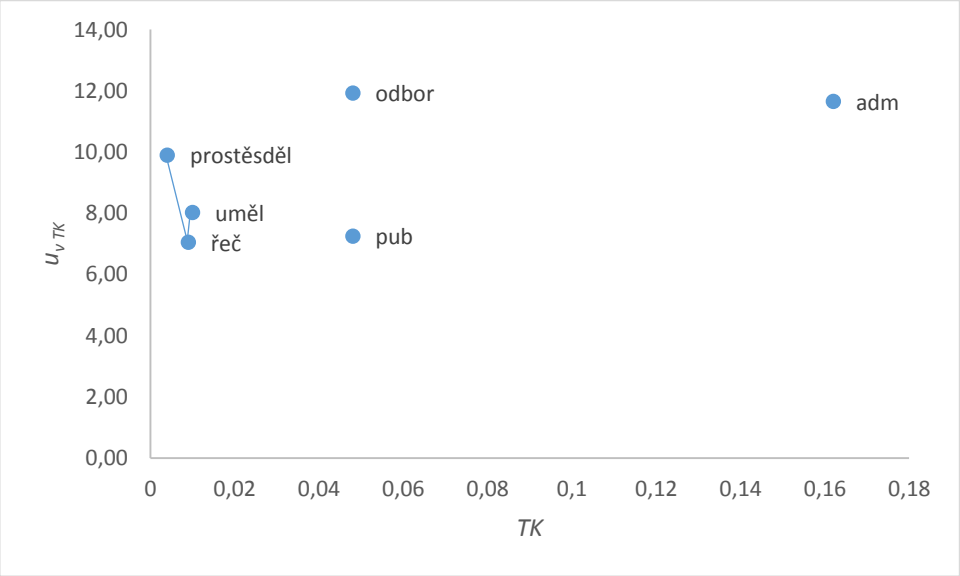
	adm	uměl	prostěsděl	odbor	řeč	pub
adm	×					
uměl	5,56	×				
prostěsděl	5,80	3,81	×			
odbor	4,06	6,07	7,28	×		
řeč	5,58	0,37	1,78	5,86	×	
pub	5,06	2,13	3,46	3,40	2,16	×

Tabulka č. 17: *u-test* hodnot *TK* (α = 0,05; u ≥ 1,96 vyjadřuje signifikantní diferenci)

Tabulka č. 18 a graf č. 15 dále zaznamenávají hodnoty váženého rozdílu mezi jednotlivými styly a míru odlišnosti funkčních stylů navzájem:

styl	$u_{v\,TK}$
odbor	11,93
adm	11,65
prostěsděl	9,90
uměl	8,02
pub	7,25
řeč	7,04

Tabulka č. 18: Vážené rozdíly $u_{v\,TK}$ mezi jednotlivými funkčními styly



Graf č. 15: Podobnost/rozdílnost jednotlivých funkčních stylů na základě hodnot *TK* a $u_{v\,TK}$ (čára značí nesignifikantní rozdíl *TK* mezi styly)

Jak lze názorně vidět na grafu č. 15, nejnižších hodnot *TK* dosahují styly prostěsdělovací, řečnický a umělecký, mezi nimiž dokonce v rámci této kvantitativní charakteristiky není signifikantní rozdíl. Těsná blízkost daných funkčních stylů je nicméně do jisté míry dána tím, že velké množství analyzovaných textů v těchto stylech vykazuje nulové hodnoty indexu *TK*, tím pádem jsou menší rozdíly mezi styly nivelizovány (prostěsdělovací a řečnický styl dosáhly nenulové hodnoty *TK* v méně než polovině případů zkoumaných textů, v uměleckém stylu nebyl nalezen žádný reprezentant tematického slova v 25 % uměleckých projevů; viz tabulky č. 19 a č. 20). Z toho důvodu se pokusíme aplikovat na náš výběrový materiál variantní metodu analýzy tematické koncentrace textu – *sekundární tematickou koncentraci*.

text prostěsděl	*TK*
1	0,00
2	0,00
3	0,00
4	0,00
5	0,00
6	0,00
7	0,01
8	0,00
9	0,01
10	0,01
11	0,00
12	0,00
13	0,01
14	0,01
15	0,00
16	0,00
17	0,00
18	0,01
19	0,00
20	0,00

text *uměl*	*TK*
1	0,02
2	0,01
3	0,01
4	0,01
5	0,00
6	0,03
7	0,00
8	0,01
9	0,00
10	0,02
11	0,01
12	0,01
13	0,01
14	0,02
15	0,02
16	0,01
17	0,01
18	0,00
19	0,00
20	0,01

Tabulky č. 19 a č. 20: Hodnoty *TK* v jednotlivých textech prostěsdělovacího a uměleckého stylu

6.2.2 Analýza funkčních stylů pomocí *STK*

Jak již bylo naznačeno výše, v některých případech (ne výjimečných) texty vykazují nulovou hodnotu *TK*, protože do oblasti nad *h-bodem* nepronikl žádný

reprezentant tematického slova. Takové komunikáty potom zpravidla chápeme jako texty z pohledu *TK* neutrální. Nestačí-li nám však např. při komparaci celých skupin textů toto tvrzení, můžeme pro detailnější vyjádření míry tematické vyhraněnosti využít doplňující index *sekundární tematické koncentrace* (*STK*). Tato metoda je založená na operaci, při níž se hodnota *h-bodu* vynásobí dvěma, tím pádem se také hranice mezi synsémantiky a autosémantiky posune, a nastane tak větší pravděpodobnost, že se nad *h-bodem* vyskytne nové tematické slovo. Výpočet *STK* je definován takto (Kubát – Matlach – Čech 2014, s. 53):

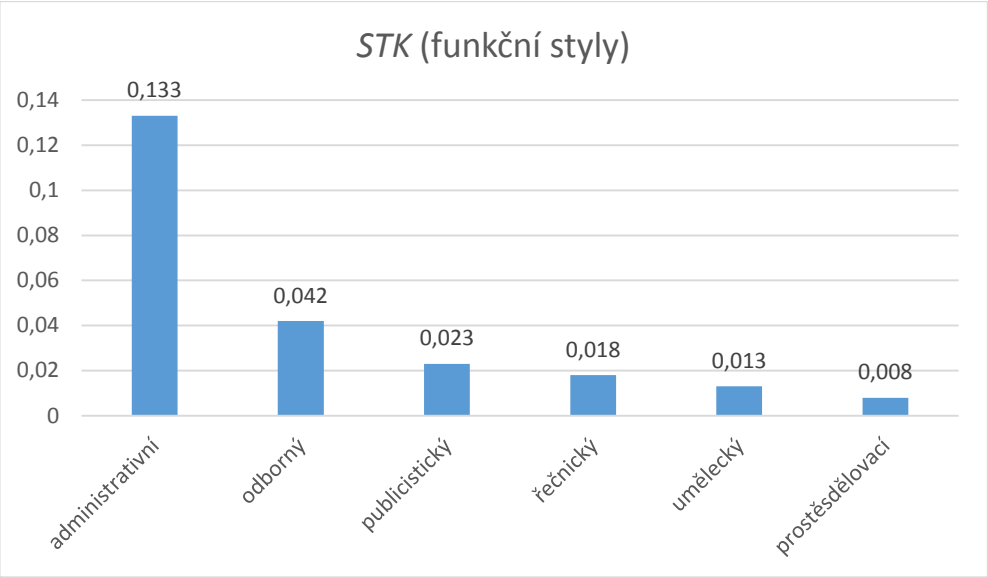
$$STK = \sum_{r'=1}^{2h} \frac{(2h - r\prime) f(r')}{h(2h-1) f(1)} .$$

r' = pořadí autosémantického slova nad $2h$
h = *h-bod*
$f(r')$ = frekvence autosémantického slova nad $2h$

Ani takovýto postup ovšem není zcela bezproblémový. Kubát (2015, s. 61) ve své dizertační práci upozorňuje, že „[…] *h*-bod byl v podstatě uměle ad hoc zdvojnásoben čistě z pragmatických důvodů konkrétních analýz. Je tedy třeba konstatovat, že tato metoda sice odstraňuje určité nevýhody *TK*, ale zároveň do měření tematické koncentrace přináší jiný metodologický problém. Máme totiž za to, že *h*-bod vyjadřuje určitou vlastnost textu, v našem případě rozděluje frekvenční distribuci na synsémantika a autosémantika […], proto, alespoň z našeho pohledu, jakékoliv dodatečné zasahování je značně problematické."

Ačkoliv je otázka, jak moc lze úpravu hodnoty *h-bodu* ospravedlnit, pravdou zůstává, že tento index vykazuje poměrně dobré výsledky. Je tedy otázkou času a dalšího ověřování, zda se potvrdí vhodnost užití této metody při výpočtu tematické koncentrace textu.

V grafu č. 16 jsou vizualizovány výsledky *STK* aplikované na analyzované texty jednotlivých funkčních stylů, tabulky č. 21 a č. 22 pak zachycují signifikantní rozdíly hodnot *STK* a vážené rozdíly mezi danými styly. V grafu č. 17 je opět zachycena vzájemná vzdálenost jednotlivých stylů na základě hodnot *STK* a $u_{v\,STK}$.



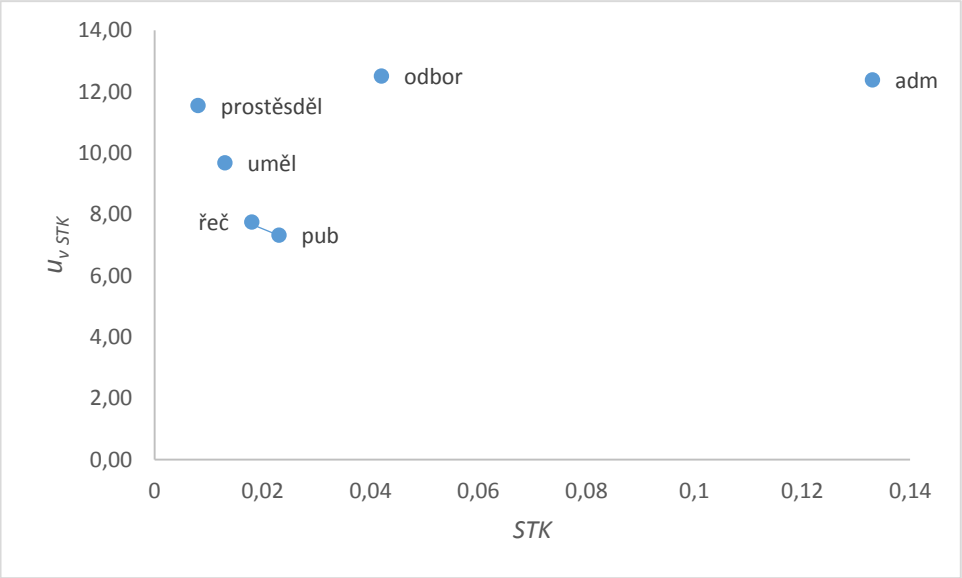
Graf č. 16: Výsledné hodnoty *STK* vypočítané pro jednotlivé funkční styly

	adm	uměl	prostěsděl	odbor	řeč	pub
adm	×					
uměl	6,00	×				
prostěsděl	6,23	3,74	×			
odbor	4,44	7,11	8,23	×		
řeč	5,66	2,09	3,72	4,84	×	
pub	5,37	2,72	3,90	3,36	1,02	×

Tabulka č. 21: *u-test* hodnot *STK* (α = 0,05; u ≥ 1,96 vyjadřuje signifikantní diferenci)

styl	$u_{v\ STK}$
odbor	12,51
adm	12,39
prostěsděl	11,55
uměl	9,69
řeč	7,75
pub	7,32

Tabulka č. 22: Vážené rozdíly $u_{v\ STK}$ mezi jednotlivými funkčními styly



Graf č. 17: Podobnost/rozdílnost jednotlivých funkčních stylů na základě hodnot *STK* a $u_{v\ STK}$ (čára značí nesignifikantní rozdíl *STK* mezi styly)

6.2.2.1 Administrativní styl

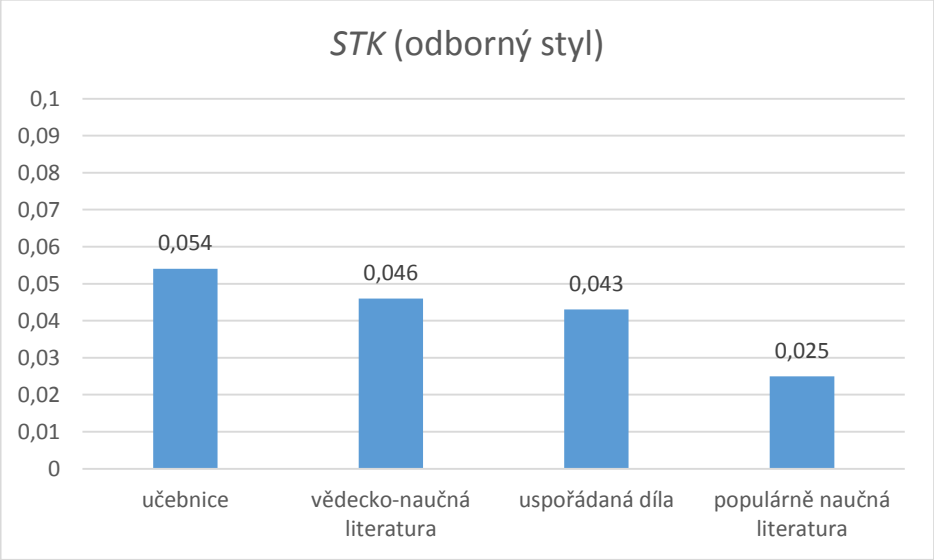
Aplikace indexu *sekundární tematické koncentrace* na náš analyzovaný materiál se ukázalo být dobrým rozhodnutím, neboť po uplatnění této kvantitativní metody dosáhl signifikantních rozdílů větší počet funkčních stylů oproti indexu *TK*. Jako nejvíce tematicky vyhraněný se ukázal styl administrativní, přičemž nejvyšší hodnoty tohoto stylu korespondují i s našim intuitivním očekáváním. Charakteristickým rysem administrativních projevů je značná schematičnost; texty administrativního charakteru usilují o přesné a jednoznačné vyjádření, užívají ustálené formulace, přitom variabilita výrazů ani konstrukcí nebývá příliš žádoucí. Texty administrativní sféry jsou nanejvýš stručné a výstižné, jejich koncentrace na téma je tedy maximální.

6.2.2.2 Prostěsdělovací styl

Naopak nejmenší hodnoty *STK* byly zjištěny u stylu prostěsdělovacího. Vzhledem k tomu, že do našeho vzorku byly zařazeny jen projevy mluvené, které jsou v souladu s vymezením tohoto stylu neformální, spontánní a pro něž spíše než obsah je důležitý kontakt samotný, je zřejmé, že se mluvčí po celou dobu konverzace nesoustředí jen na jedno téma, ale jejich dialog je značně polytematický. Nejnižší hodnoty *STK* v rámci funkčních stylů jsou tedy pro skupinu nepřipravených soukromých projevů předpokládatelné.

6.2.2.3 Odborný styl

Druhou pozici co do nejvyšších hodnot *tematické koncentrace* obsadil styl odborný. Podobně jako ve stylu administrativním projevuje se i v odborných komunikátech snaha o explicitnost, objektivnost a silná zaměřenost na sledovanou problematiku. Na rozdíl od textů administrativní povahy ovšem projevy odborné nejsou tolik stručné, ale zpravidla pojednávané téma zpracovávají šířeji. Vnitřní diferenciaci odborného stylu lze dále sledovat na následujících grafech a tabulkách:



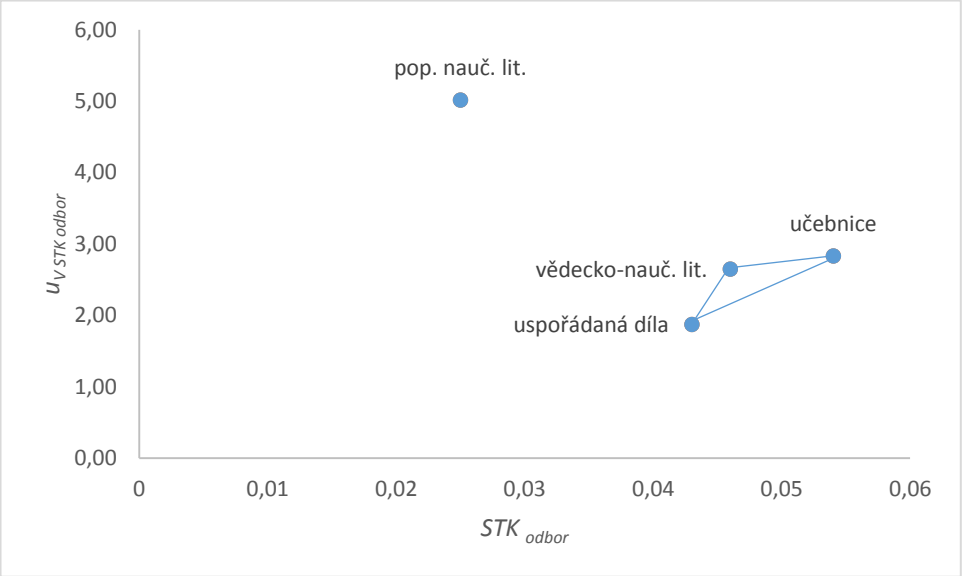
Graf č. 18: Výsledné hodnoty *STK* $_{odbor}$ vypočítané pro jednotlivé subkategorie odborného funkčního stylu

	uspořádaná díla	pop. nauč. díla	vědecko-nauč. díla	učebnice
uspořádaná díla	×			
pop. nauč. lit.	**2,03**	×		
vědecko-nauč. lit.	0,31	**3,47**	×	
učebnice	0,91	**3,19**	0,81	×

Tabulka č. 23: *u-test* hodnot *STK* $_{odbor}$ (α = 0,05; u ≥ 1,96 vyjadřuje signifikantní diferenci)

žánr	$u_{v\ STK\ odbor}$
pop. nauč. lit.	5,02
učebnice	2,83
vědecko-nauč. lit.	2,65
uspořádaná díla	1,88

Tabulka č. 24: Vážené rozdíly $u_{v\ STK\ odbor}$ mezi jednotlivými subkategoriemi odborného funkčního stylu



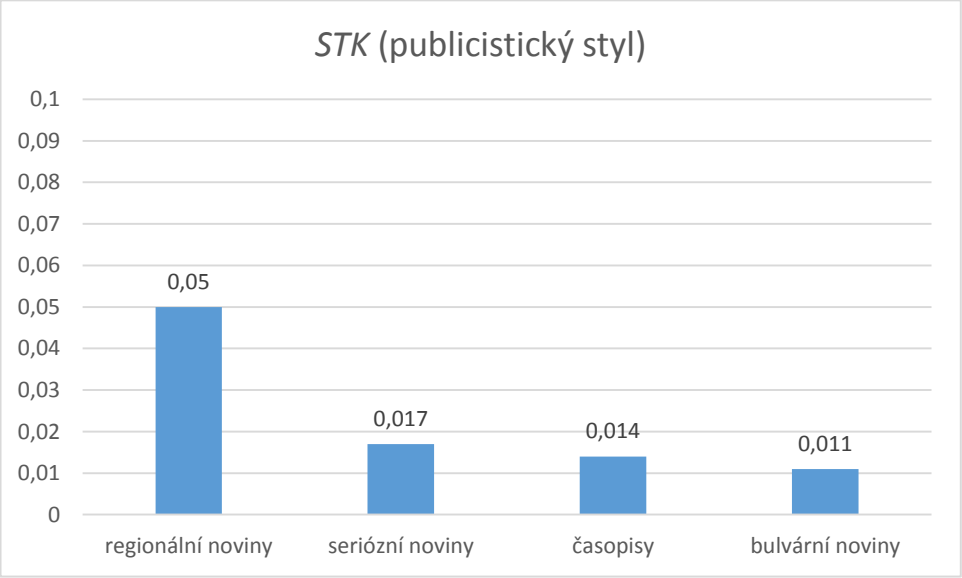
Graf č. 19: Podobnost/rozdílnost jednotlivých subkategorií odborného funkčního stylu na základě hodnot STK_{odbor} a $u_{v\,STK\,odbor}$ (čára značí nesignifikantní rozdíl STK_{odbor} mezi skupinami textů)

Jak vyplývá z grafu č. 19, poslední pozici z hlediska hodnot *STK* zaujala literatura populárně naučná, která se také jako jediná signifikantně liší od zbylých subkategorií odborného stylu. Důvod takové odlišnosti lze hledat patrně v délce a v rozmanitosti populárně naučných komunikátů. Do kategorie s nejnižší tematickou koncentrací textu spadají i odborné časopisy určené pro širší okruh čtenářů, v nichž se logicky publikují texty kratšího rozsahu, zato více tematicky rozrůzněné. Jedno hypertéma určující zaměření daného periodika, zdá se, je oproti ostatním monotematičtějším odborným textům přitom méně silným prvkem tematické vyhraněnosti. Neperiodické komunikáty populárně naučné povahy nejsou potom sice v rámci jedné publikace tolik tematicky různorodé, sledovanou problematiku ale doplňují i méně podstatnými informacemi, množstvím příkladů a ani při vyjadřování nejsou tolik omezeny přísnou odbornou terminologií. Celkově tak u populárně naučné literatury dochází k beletrizaci, čemuž odpovídá i menší koncentrace textu na jedno úzké téma. Ostatní subkategorie odborného stylu nevykazují mezi sebou žádné signifikantní rozdíly a svými vyššími hodnotami indexu *STK* odpovídají běžné stylové charakteristice odborných projevů.

6.2.2.4 Řečnický, publicistický styl

Zbývající funkční styly již dosahují spíše nižších hodnot *STK*, styl publicistický a řečnický se co do tematické vyhraněnosti od sebe dokonce ani signifikantně

neliší. Čechová, Krčmová a Minářová (2008, s. 285) vzájemnou blízkost dvou posledně zmíněných stylů potvrzují: „Rétorický styl je nejblíže k (mluvené) publicistice, s ní jej spojuje jak veřejnost a kolektivní adresát, tak i [...] persvaze, kterou pokládáme pro formování stylu za podstatnou. Rétorika má však diferencovanější typ působení na vnímatele než publicistika a zasahuje více komunikačních témat, než je pro publicistku obvyklé." Publicistický i řečnický styl také dosahují nejmenších hodnot vážených rozdílů $u_{v\ STK}$, což ukazuje na jejich malou odlišnost od ostatních funkčních stylů. Přestože jsou tedy řečnické projevy specifické svou formou realizace, v níž hraje podstatnou úlohu osobnost řečníka a jeho pečlivá artikulace, míra zaměření na téma zřejmě v těchto typech projevů není vůči ostatním funkčním stylům dostatečně diferenční. Podobně je na tom také styl publicistický, vzhledem ale k jeho velké vnitřní rozrůzněnosti zkusíme porovnat *tematickou koncentraci textu* alespoň v rámci jeho jednotlivých subkategorií:



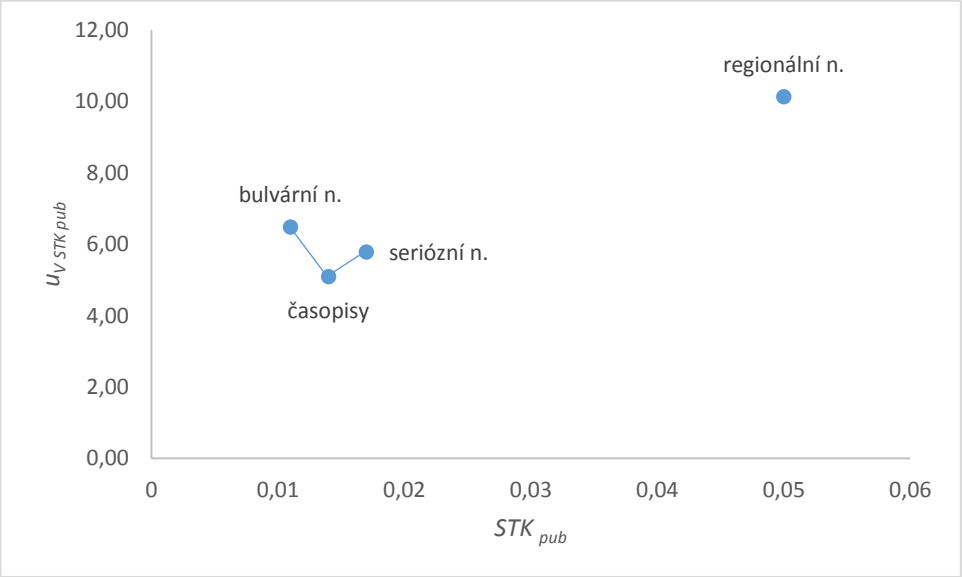
Graf č. 20: Výsledné hodnoty *STK* $_{pub}$ vypočítané pro jednotlivé subkategorie publicistického funkčního stylu

	bulvární noviny	časopisy	seriózní noviny	regionální noviny
bulvární noviny	×			
časopisy	1,57	×		
seriózní noviny	**3,27**	1,43	×	
regionální noviny	**6,39**	**5,83**	**5,33**	×

Tabulka č. 25: *u-test* hodnot *STK* $_{pub}$ (α = 0,05; u ≥ 1,96 vyjadřuje signifikantní diferenci)

žánr	$u_{v\ STK\ pub}$
regionální noviny	10,13
bulvární noviny	6,48
seriózní noviny	5,79
časopisy	5,10

Tabulka č. 26: Vážené rozdíly $u_{v\ STK\ pub}$ mezi jednotlivými subkategoriemi publicistického funkčního stylu



Graf č. 21: Podobnost/rozdílnost jednotlivých subkategorií publicistického funkčního stylu na základě hodnot STK_{pub} a $u_{v\ STK\ pub}$ (čára značí nesignifikantní rozdíl STK_{pub} mezi skupinami textů)

Na grafu č. 21 je patrné zcela specifické postavení novin regionálních. Příčinu vysokých hodnot *STK* i značně velkou míru odlišnosti vůči ostatním publicistickým subkategoriím je možné odůvodnit prostou skutečností, že regionální periodika ve svých textech užívají s velkou frekvencí označení příslušného regionu, popř. z něj utvořená adjektiva (*Magistrát města Karviné*, *historie karvinské radnice* atd.). Neobvykle vysokou četnost takovýchto výrazů dokazuje i jejich umístění na předních pozicích frekvenčního slovníku daných komunikátů. Bývá zvykem, že prvních deset nejfrekventovanějších slov v textu patří mezi slova gramatická, zpravidla se jedná o spojky, předložky a pomocné sloveso „být". Ve třech z pěti případů analyzovaných regionálních novin však název příslušné lokality obsadil pátou nejčetnější pozici ve frekvenční distribuci slov, v jednom periodiku se mezi gramatickými slovy označení daného místa prosadilo na pozici šesté a jen v jednom případě skončilo těsně

pod hranicí pásma slov s nejvyšší frekvencí, tedy na pozici jedenácté (viz tabulky č. 27–31). Také samotné slovo „město“ obsadilo ve čtyřech případech některou z deseti nejfrekventovanějších pozic ve frekvenční distribuci slov. Máme za to, že takto častý výskyt daných výrazů se tudíž nutně musel projevit na větší míře *tematické koncentrace* analyzovaných komunikátů.

pořadí	lemma	frekvence
1	být	320
2	v	308
3	se	274
4	a	261
5	na	178
6	z	106
7	ten	105
8	o	92
9	který	90
10	do	77
11	**Kopřivnice**	74
12	rok	69
13	po	66
14	s	65
15	hodina	54

pořadí	lemma	frekvence
1	v	616
2	a	571
3	být	429
4	na	328
5	**Brno**	286
6	se	274
7	**město**	203
8	z	168
9	s	157
10	o	132
11	do	128
12	pro	127
13	rok	123
14	i	120
15	ten	109

pořadí	lemma	frekvence
1	a	236
2	v	226
3	být	208
4	na	179
5	se	139
6	**Pardubice**	73
7	rok	69
8	**město**	56
9	do	56
10	s	55
11	pro	54
12	z	54
13	který	53
14	i	48
15	ten	46

pořadí	lemma	frekvence
1	v	169
2	a	149
3	být	126
4	na	99
5	**Karviná**	81
6	se	77
7	s	55
8	**město**	51
9	který	40
10	karvinský	40
11	do	37
12	rok	37
13	pro	36
14	o	32
15	dítě	31

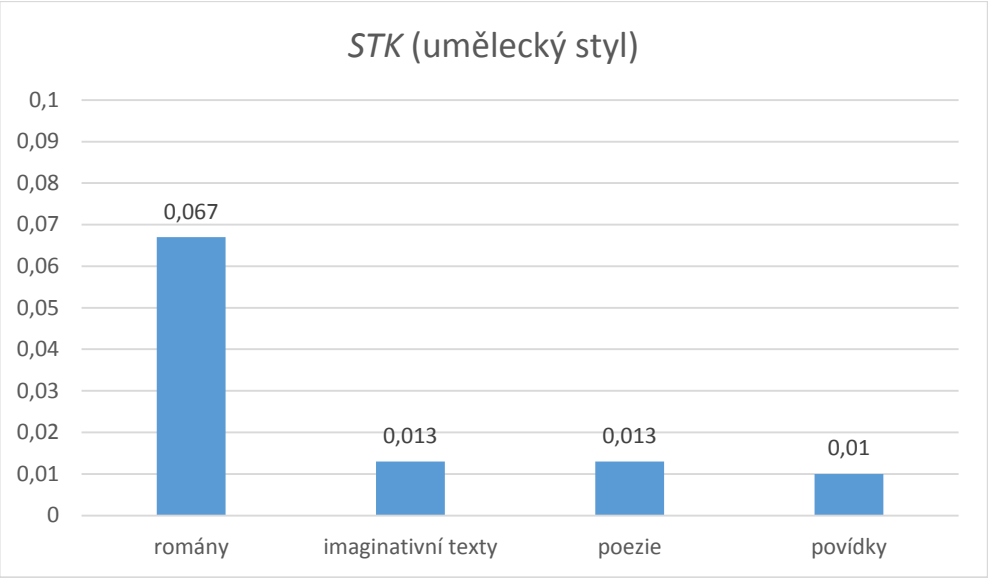
pořadí	lemma	frekvence
1	v	257
2	a	232
3	na	148
4	se	130
5	**Náchod**	128
6	být	124
7	s	82
8	z	68
9	o	61
10	**město**	58
11	do	54
12	rok	46
13	pro	45
14	který	45
15	i	41

Tabulky č. 27–31: Patnáct nejfrekventovanějších slov v regionálních periodikách

Ostatní skupiny publicistických projevů se mezi sebou již tolik neliší, neboť všechny obsahují texty kratšího rozsahu z různých tematických oblastí. Nejnižší hodnoty *STK* u bulvárních periodik jsou pak pravděpodobně způsobeny největším zájmem o zpracování pestrých témat, v nichž figurují co možná nejrůznější osoby, neboť sdělení bulvárního charakteru bývají krátká, avšak usilují o co nejširší záběr s ohledem na velký okruh recipientů.

6.2.2.5 Umělecký styl

Stylem s druhou nejmenší mírou tematické zaměřeností je nakonec styl umělecký. Možnost využití neomezeného množství prostředků z různých stylových vrstev pro vyjádření estetického záměru se podle našeho mínění odráží i v nižších hodnotách *STK*. Kvůli mnohotvárnosti uměleckých komunikátů se ale pokusíme kvantifikovat tematickou zaměřenost i v rámci jednotlivých uměleckých žánrů:



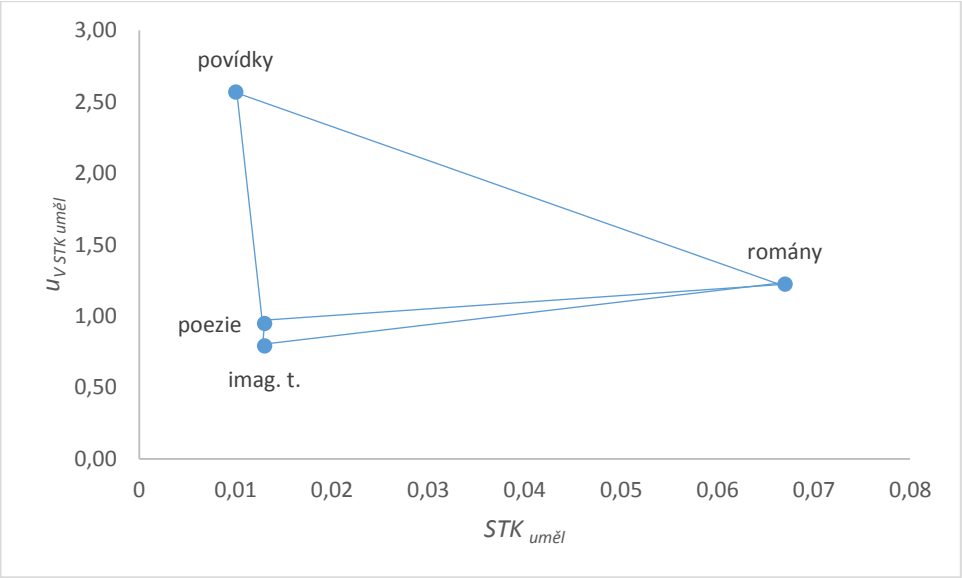
Graf č. 22: Výsledné hodnoty *STK* $_{uměl}$ vypočítané pro jednotlivé subkategorie uměleckého funkčního stylu

	povídky	imaginativní texty	romány	poezie
povídky	×			
imaginativní texty	1,16	×		
romány	1,83	0,16	×	
poezie	1,46	0,05	0,13	×

Tabulka č. 32: *u-test* hodnot *STK* $_{uměl}$ ($\alpha = 0{,}05$; $u \geq 1{,}96$ vyjadřuje signifikantní diferenci)

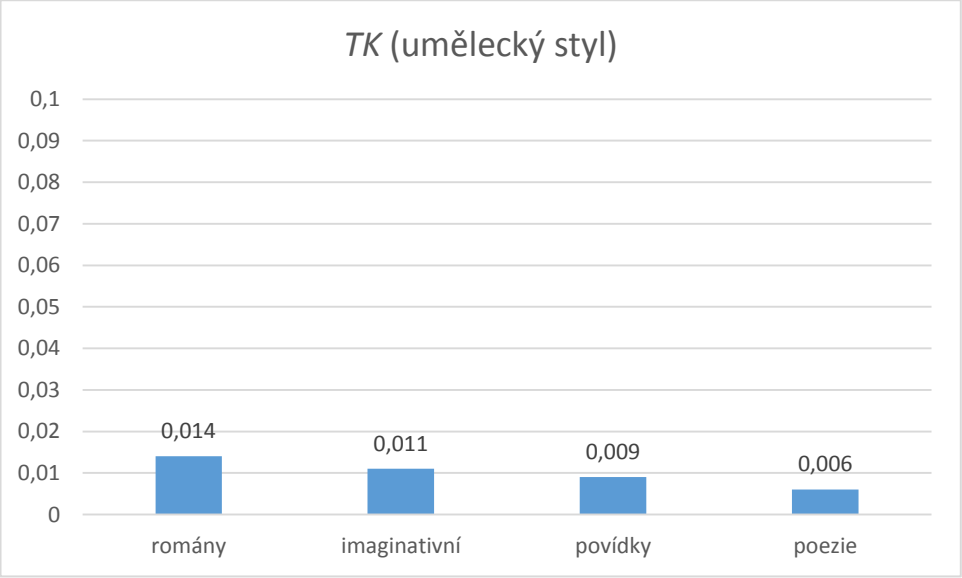
žánr	$u_{v\ STK\ uměl}$
povídky	2,57
romány	1,22
poezie	0,95
imaginativní texty	0,79

Tabulka č. 33: Vážené rozdíly $u_{v\ STK\ uměl}$ mezi jednotlivými subkategoriemi uměleckého funkčního stylu



Graf č. 23: Podobnost/rozdílnost jednotlivých subkategorií uměleckého funkčního stylu na základě hodnot *STK uměl* a *u v STK uměl* (čára značí nesignifikantní rozdíl *STK uměl* mezi skupinami textů)

Jak vyplývá z výše vypočítaných hodnot *STK* a statistického *u-testu*, mezi jednotlivými žánry uměleckého stylu není z hlediska zaměřenosti na téma žádný signifikantní rozdíl. Na základě kvantitativní metody *sekundární tematické koncentrace* můžeme považovat takové výsledky za validní, protože však „žádný rozdíl" není z lingvistického hlediska příliš zajímavý, můžeme pro jednotlivé podkategorie uměleckého stylu zkusit vypočítat také základní index *TK*, jenž nám může podat mírně odlišné výstupní hodnoty:



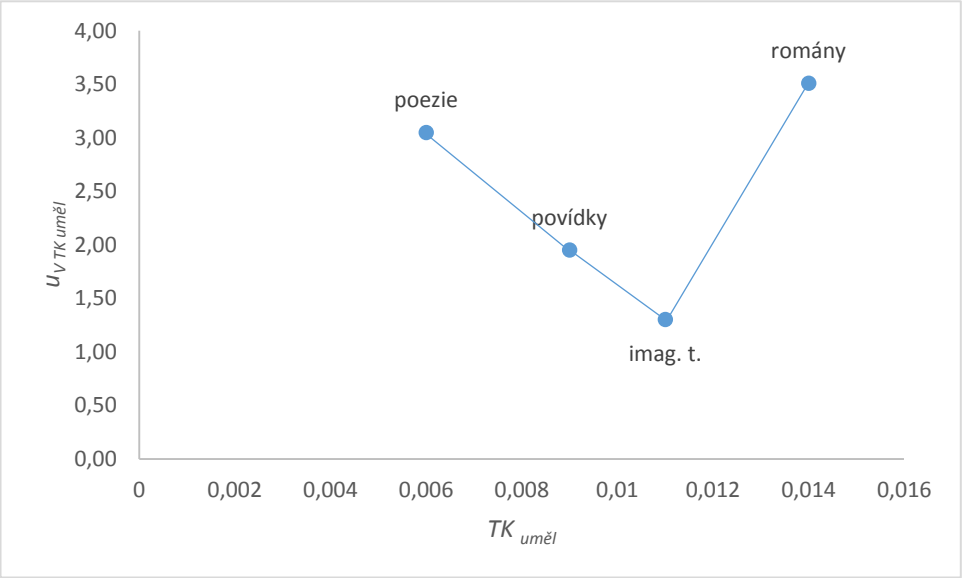
Graf č. 24: Výsledné hodnoty *TK uměl* vypočítané pro jednotlivé subkategorie uměleckého funkčního stylu

	povídky	imaginativní texty	romány	poezie
povídky	×			
imaginativní texty	0,43	×		
romány	**2,01**	0,78	×	
poezie	0,94	1,05	**3,29**	×

Tabulka č. 34: *u-test* hodnot $TK_{uměl}$ ($\alpha = 0{,}05$; $u \geq 1{,}96$ vyjadřuje signifikantní diferenci)

žánr	$u_{v\ TK\ uměl}$
romány	3,51
poezie	3,05
povídky	1,95
Imaginativní texty	1,30

Tabulka č. 35: Vážené rozdíly $u_{v\ TK\ uměl}$ mezi jednotlivými subkategoriemi uměleckého funkčního stylu



Graf č. 25: Podobnost/rozdílnost jednotlivých subkategorií uměleckého funkčního stylu na základě hodnot $TK_{uměl}$ a $u_{v\ TK\ uměl}$ (čára značí nesignifikantní rozdíl $TK_{uměl}$ mezi skupinami textů)

Z grafu č. 24 lze zjistit, že výsledné tendence pro tematickou koncentraci textu zůstaly v jednotlivých žánrech téměř zachovány jak v případě aplikace *TK*, tak v případě *STK*, i když po užití *TK* dosahují výsledky menších hodnot. Tabulka č. 34 však již prozrazuje, že metoda *TK* dokázala na rozdíl od své modifikované verze mezi některými uměleckými žánry detekovat signifikantní

rozdíl. Z toho důvodu budeme v tomto případě index *TK* považovat s ohledem na náš cíl za vhodnější.

Tabulka č. 35 i graf č. 25 ukazují, že nejvíce odlišným žánrem vůči ostatním jsou romány. V naší analýze nejvyšších hodnot *TK* romány dosahují pravděpodobně kvůli svému většímu rozsahu, i když na první pohled bychom v souvislosti s touto charakteristikou očekávali opačnou tendenci. Delší prozaické útvary rozpracovávají více dějových linií, vystupuje v nich více postav a obsahují různé popisné a úvahové pasáže, které koncentraci na hlavní téma textu rozptylují. Oproti tomu povídka jako zástupce kratší prózy se soustředí jen na jednu dějovou linku, jež má rychlý spád, a od ústředního tématu příliš neodbíhá. Z toho vyplývá, že by žánr povídky měl vykazovat vyšší hodnoty tematické zaměřenosti, než je tomu u románů. K podobným závěrům došel dokonce i Kubát (2015, s. 59–70), který z hlediska tematické koncentrace porovnával různé žánry textů Karla Čapka – v jeho analýze tento index dosáhl v povídkách o něco vyšších hodnot, diference mezi oběma žánry ovšem nakonec nebyla signifikantní. Rozdíl mezi výsledky v Kubátově výzkumu a v naši práci ovšem spočívá v analyzovaném vzorku. Kubát totiž mezi srovnávané texty zařadil jednotlivé kapitoly románů a samostatné povídky, které si rozsahem více odpovídají, zatímco my jsme při komparaci zachovali ucelenost publikovaných děl, tzn. do analýzy jsme zahrnuli celé romány, stejně jako povídkové soubory. Protože nám šlo v první řadě více o komparaci funkčních stylů než jednotlivých žánrů, považovali jsme za podstatné nepozměňovat záměr autora a nenarušovat žádná specifika reálných publikovaných textů různých stylů, např. dodatečným rozdělením jednotlivých povídek, navíc vzhledem k tomu, že užité metody nevykazují závislost na délce textu, nepokládali jsme ani za nutné rozčleňovat romány na kratší úseky.

Z našeho materiálu potom vyplývá, že menší tematická koncentrace povídek je pravděpodobně způsobena větší rozmanitostí témat v rámci kratších textů jedné povídkové sbírky, stejně jako tomu je u sbírky básnické nebo u souboru imaginativních textů. Dlouhé, souvislé romány s více či méně jasně vymezeným dějem tak z tohoto srovnání vycházejí jako umělecký žánr s nejvyššími hodnotami *TK*.

6.2.3 Proporcionální tematická koncentrace (*PTK*)

K základnímu způsobu měření *TK* byl kromě *STK* vyvinut i další doplňující index, a to *proporcionální tematická koncentrace textu* (*PTK*). Tato metoda, měřící proporci mezi frekvencemi *tematických slov* a frekvencemi všech slov, která se vyskytují nad *h-bodem*, již tolik neřeší problém s nulovou hodnotu *TK*,

jako je tomu u výše představené *STK*, přínosem tohoto indexu je ale možnost pomocí statistického testu ověřovat i texty, v nichž byl detekován pouze jeden reprezentant *tematického slova*, což základní výpočet *TK* neumožňuje (Čech 2016, s. 30–32). Vzorec pro výpočet *PTK* vypadá následovně:

$$PTK = \frac{1}{N_h} \sum_{r' < h} f(r') \ ,$$

N_h = frekvence všech slov nad *h-bodem*,
f(r') = frekvence tematického slova.

Použití této kvantitativní metody je však mírně problematické. Čech (2016, s. 56–57) ve své nejnovější publikaci přišel s poznatkem, že výpočet *PTK* není vhodné aplikovat na texty krátkého rozsahu, konkrétně na komunikáty kratší než 2 000 slov, protože se u nich projevuje závislost výsledných hodnot *PTK* na délce projevu. U textů s délkou nad 2 000 tokenů již tato závislost není patrná, přesto nám tato nově objevená skutečnost znemožnila uplatnit daný index při výpočtu tematické koncentrace textů různých funkčních stylů, neboť bychom byli nuceni ze své analýzy vyřadit celý řečnický styl, který reprezentují projevy českých prezidentů, jejichž rozsah je (až na jednu výjimku) menší než 2 000 slov.

6.3 Vzdálenosti sloves (*VD*)

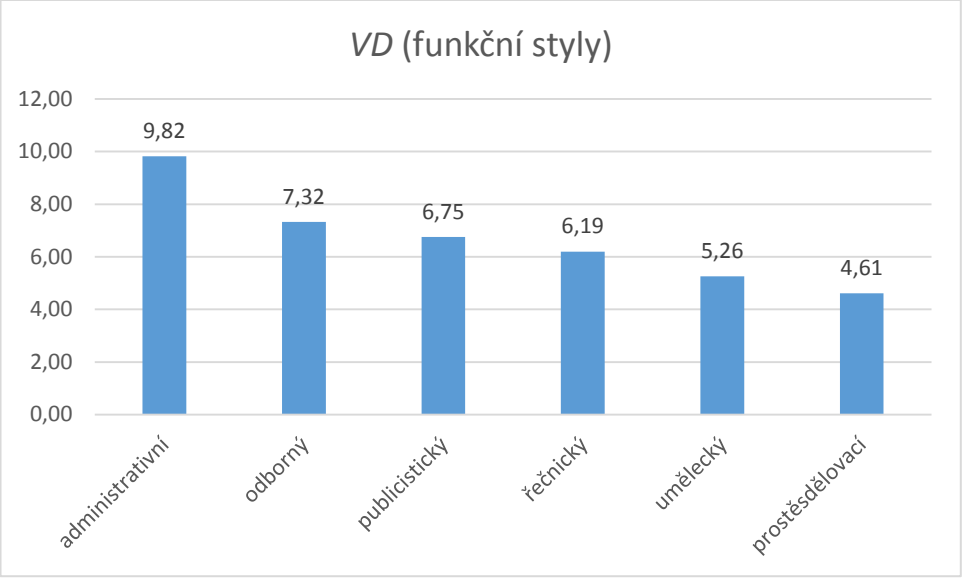
Velká část kvantitativních lingvistů při výzkumu soustředí svou pozornost zejména na lexikální stránku textů. Měření vzdálenosti sloves ovšem patří mezi méně početné metody, jež nám umožňují zkoumat texty i z pohledu syntaktického. Index *vzdálenosti sloves* Kubát (2015, s. 70–74) přibližuje jako indikátor průměrného počtu tokenů mezi dvěma nejbližšími slovesy, přičemž se dá předpokládat, že čím vyšší hodnotu index vyjadřuje, tím komunikát obsahuje delší nominální fráze a jiné konstrukce poukazující na obtížnější syntaktickou strukturu textu. „Vycházíme z prosté úvahy, kdy extrémně jednoduchý text složený pouze z tzv. holých vět má nejkratší vzdálenosti mezi slovesy. Čím je syntaktická struktura složitější, tím by se měla i prodlužovat vzdálenost mezi slovesy. S tím souvisí samozřejmě i obtížnost textu (readability)." (Kubát 2015, s. 73–74) Velká výhoda tohoto kvantitativního přístupu pak spočívá v jednoduchosti, a zejména v možnosti pracovat i s texty bez syntaktické anotace.

Přestože výsledné tendence (viz níže) odpovídají i intuitivním charakteristikám jednotlivých funkčních stylů, chtěli bychom explicitně upozornit na jisté limity dané metody, o nichž se autoři zatím nezmiňují. Vzhledem k tomu, že při výpočtu indexu se nepřihlíží k větněčlenské funkci sloves, ale jsou detekována všechna verba bez rozdílu, myslíme si, že je nutné považovat výsledné hodnoty minimálně za zkreslené. Slovesa se v textu totiž nevyskytují jen v určitém tvaru ve funkci predikátu, ale běžně v infinitivu zastávají funkci subjektu, atributu aj., a bývají tedy také součástí nominalizovaných konstrukcí, zejména jejichž délka se pomocí indexu *vzdálenosti sloves* měří. Infinitivizace podle nás patří mezi běžné transformační procesy, jež prodlužují vzdálenost mezi predikáty a ztěžují syntaktickou strukturu věty, podobně jako je tomu u konstrukcí s deverbativními adjektivy, s deverbálními substantivy aj. Současné nastavení softwaru *QUITA* však počítání tokenů zastaví kdykoli, když detekuje sloveso, a to ve finitním i infinitním tvaru, tedy např. i v případě infinitivní nominalizace, přitom na měření podobných struktur se má index *vzdálenosti sloves* zaměřovat.

6.3.1 Analýza funkčních stylů pomocí *VD*

Přestože máme k současné podobě praktického měření této jednoduché kvantitativní charakteristiky výhrady, s ohledem na výsledky, jež tento index podává, považujeme i přes zmíněné nedostatky danou metodu za vhodnou pro diferenciaci funkčních stylů. V grafu č. 26 jsou zaznamenány hodnoty *VD*

vypočtené pro jednotlivé styly, dále pak následují výsledky *u-testu* s vyznačenými signifikantními rozdíly a znázornění vzájemné blízkosti stylů z hlediska *VD* a $u_{v\,VD}$.



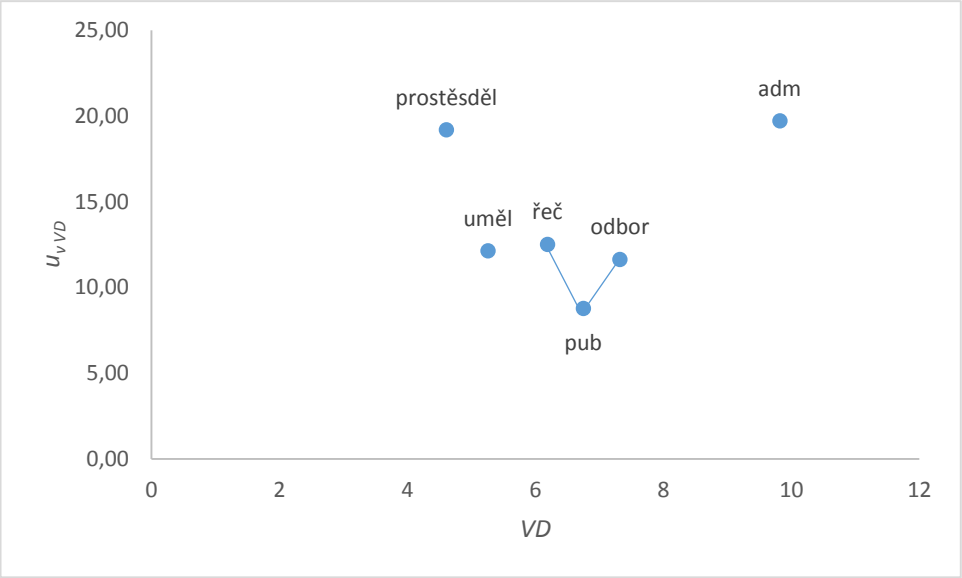
Graf č. 26: Výsledné hodnoty *VD* vypočítané pro jednotlivé funkční styly

	adm	uměl	prostěsděl	odbor	řeč	pub
adm	×					
uměl	10,45	×				
prostěsděl	13,38	3,07	×			
odbor	5,29	6,02	9,68	×		
řeč	8,84	3,70	10,18	3,66	×	
pub	6,15	3,92	6,61	1,36	1,6	×

Tabulka č. 36: *u-test* hodnot *VD* (α = 0,05; u ≥ 1,96 vyjadřuje signifikantní diferenci)

styl	$u_{v\,VD}$
adm	19,73
prostěsděl	19,19
řeč	12,51
uměl	12,15
odbor	11,63
pub	8,78

Tabulka č. 37: Vážené rozdíly $u_{v\,VD}$ mezi jednotlivými funkčními styly



Graf č. 27: Podobnost/rozdílnost jednotlivých funkčních stylů na základě hodnot *VD* a $u_{v\,VD}$ (čára značí nesignifikantní rozdíl *VD* mezi styly)

6.3.1.1 Administrativní, prostěsdělovací styl

Jak z grafu č. 26 vyplývá, nejvyšších hodnot *VD* dosáhl styl administrativní, jehož specifické postavení mezi ostatními styly názorně ilustruje i graf. č. 27. Výsledek není překvapivý, neboť administrativní texty usilují zejména o věcné, fakticky maximálně nasycené sdělení. Projevy jsou často tvořeny dlouhými větami s malým počtem určitých slovesných tvarů, zato s větší frekvencí jmenných konstrukcí, vysvětlujících přístavků, a to vše často i na úkor snadného porozumění. Výjimkou v administrativních textech nejsou ani opakované výčty, které maximální zhuštěnost vyjádření podporují.

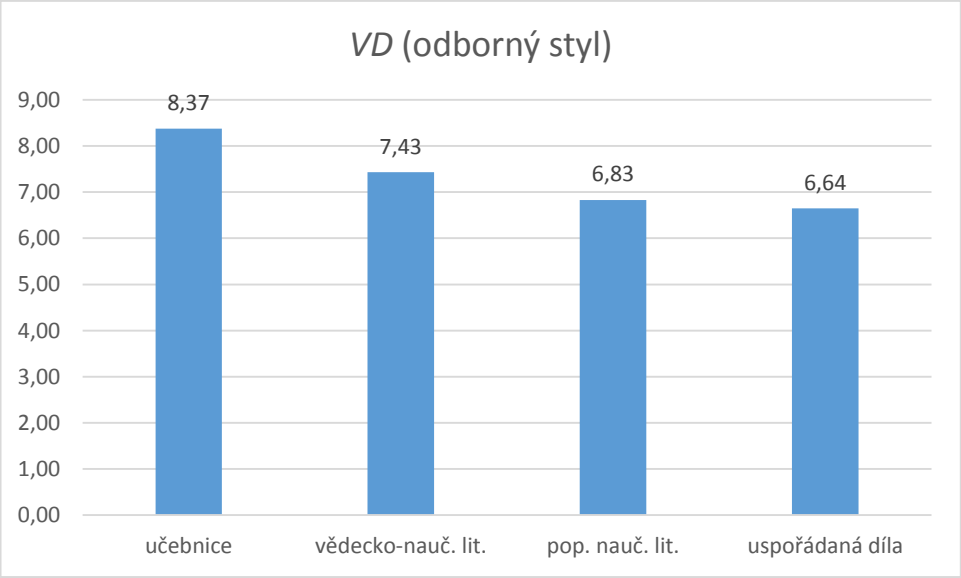
Na opačném konci pak stojí styl prostěsdělovací, jenž se spolu s administrativním stylem od ostatních funkčních stylů statisticky nejvíce odlišuje. Pro prostěsdělovací styl jsou charakteristické projevy mluvené a spontánní, nepřipravené výpovědi přitom nemívají příliš propracovanou syntaktickou strukturu; některé větné členy se připojují za výpověď až dodatečně, často však ani nedojde k ukončení výpovědi. Navíc jsou projevy soukromé oblasti (obvykle dialogického charakteru) zakotvené v konkrétní komunikační situaci, což umožňuje komunikantům vzájemné porozumění, aniž by museli explicitně verbalizovat některé části výpovědi. *Ukaž mi to.* nebo *Polož to tam.*, doplněné např. ještě gestem, je pro zúčastněné dostatečně srozumitelným sdělením, i když takovéto výpovědi jsou nevelkého rozsahu.

Mezi těmito póly se nacházejí ostatní funkční styly, které k sobě již mají z hlediska *vzdálenosti sloves* blíže. Publicistický styl se dokonce z hlediska

zkoumaného indexu ani signifikantně neliší od stylu řečnického a stylu odborného, což je v souladu s představou tradiční stylistiky, která tvrdí, že se v publicistice projevují prvky řečnických projevů (snaha o persvazi recipientů) i prvky odborných textů (kondenzované formulace) a že se jedná o nejméně vyhraněný funkční styl (Čechová – Krčmová – Minářová 2008, s. 248). Protože styly se středními hodnotami *VD* jsou vnitřně značně rozmanité, pokusíme se analyzovat jejich syntaktickou stránku i z hlediska jednotlivých subkategorií spadajících do příslušného funkčního stylu.

6.3.1.2 Odborný styl

Hned za administrativním stylem druhou nejvyšší hodnotu *VD* vykazuje styl odborný. V odborných projevech bývají prezentovány nové myšlenky, objasňují se komplikované vztahy, vysvětlují se složité mechanismy, čemuž také odpovídají dlouhé věty se složitou větnou stavbou. Podobně jako u administrativního stylu i odborné texty kumulují maximální množství informací do jednoduchých vět za pomoci nominalizovaných konstrukcí. Skutečnost, jak moc je odborný styl dále diferencován z hlediska *VD*, zobrazují následující grafy a tabulky:



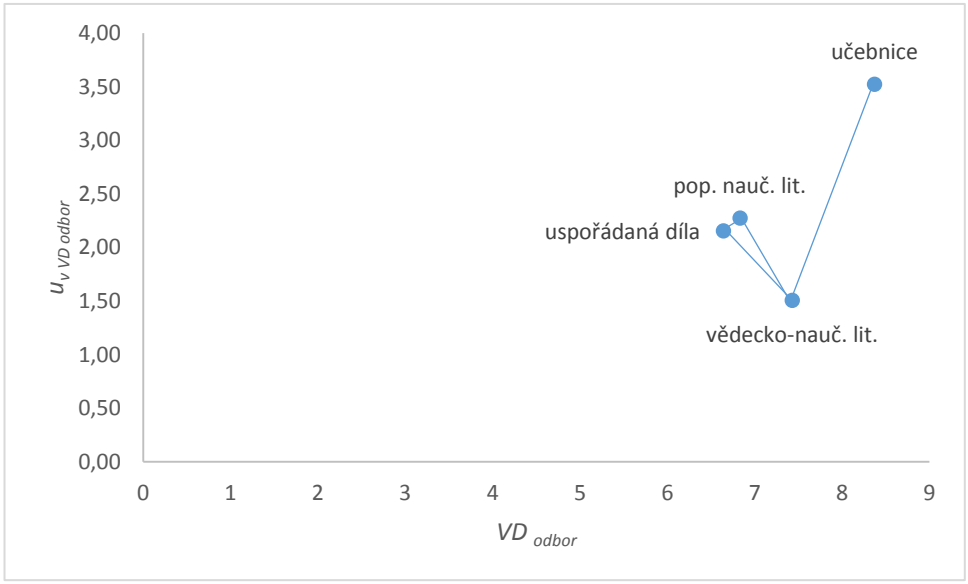
Graf č. 28: Výsledné hodnoty *VD* $_{odbor}$ vypočítané pro jednotlivé subkategorie odborného funkčního stylu

	uspořádaná díla	pop. nauč. lit.	vědecko-nauč. lit.	učebnice
uspořádaná díla	×			
pop. nauč. lit.	0,30	×		
vědecko-nauč. lit.	1,18	1,20	×	
učebnice	**2,25**	**2,44**	1,41	×

Tabulka č. 38: *u-test* hodnot *VD* $_{odbor}$ (α = 0,05; u ≥ 1,96 vyjadřuje signifikantní diferenci)

žánr	$u_{v\ VD\ odbor}$
učebnice	3,52
pop. nauč. lit.	2,27
uspořádaná díla	2,15
vědecko-nauč. lit.	1,51

Tabulka č. 39: Vážené rozdíly $u_{v\ VD\ odbor}$ mezi jednotlivými subkategoriemi odborného funkčního stylu

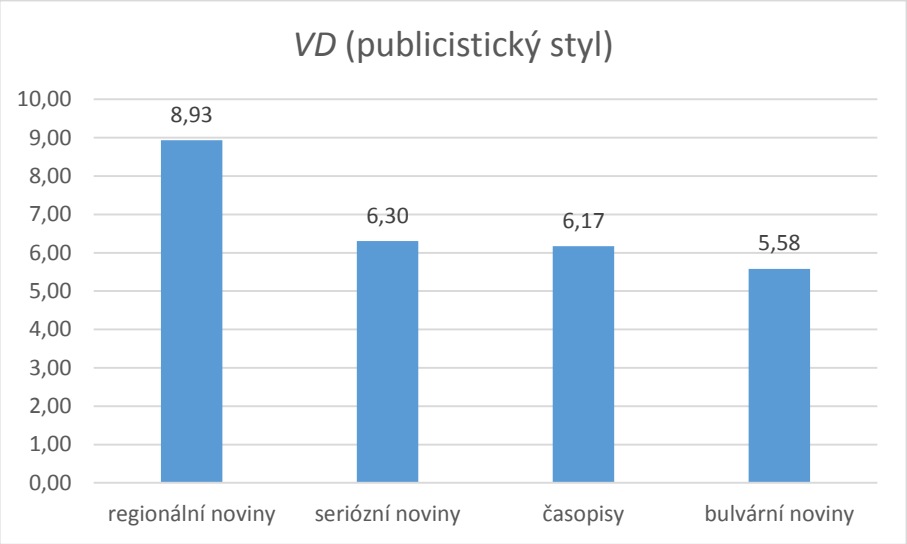


Graf č. 29: Podobnost/rozdílnost jednotlivých subkategorií odborného funkčního stylu na základě hodnot *VD* $_{odbor}$ a $u_{v\ VD\ odbor}$ (čára značí nesignifikantní rozdíl *VD* $_{odbor}$ mezi skupinami textů)

Hodnoty *VD* a vztahy znázorněné v grafu č. 29 ukazují, že jednotlivé subkategorie odborného stylu se z pohledu syntaktického mezi sebou signifikantně neliší. Jedinou výjimku tvoří texty učební, které také dosahují nejvyšších hodnot *vzdálenosti sloves*. Tento výsledek může být poněkud překvapivý, protože se má obecně za to, že učebnice jako texty určené žákům

a studentům jsou po obsahové i formální stránce přizpůsobené jazykovým kompetencím mladších recipientů, a tedy ani syntax by v tomto typu textů neměla být tak složitá jako v tradičních odborných projevech. Konkrétní texty reprezentující v našem výzkumu kategorii učebnic jsou ovšem hlavně skripta určená vysokoškolským studentům. Takové učební texty již příliš nepočítají se zprostředkující rolí učitele, ale jsou určeny čtenářům poučeným v daném oboru nebo alespoň čtenářům, kteří by měli být schopni složitým odborným formulacím rozumět. Na rozdíl od teoretických vědeckých projevů, jež prezentují aktuální výsledky výzkumů nebo nové interpretace dosavadních poznatků v příslušné odborné sféře, vysokoškolská skripta spíše jen předkládají nejdůležitější informace vyplývající z dosud provedených bádání a nastiňují základní problémy daného odvětví. I když se tedy tyto typy odborných textů liší obsahově, forma jejich vyjadřování bývá obdobná, hodně exaktní, což nakonec potvrzují i výsledné hodnoty *VD* a statistického *u-testu*, které ukazují na nesignifikantní rozdíl mezi texty učebními a literaturou vědecko-naučnou.

6.3.1.3 Publicistický styl



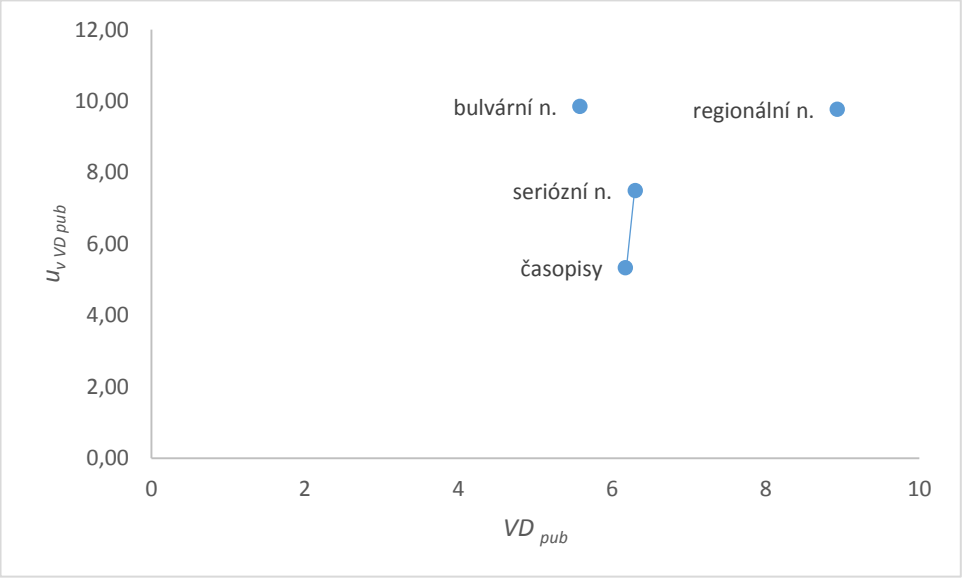
Graf č. 30: Výsledné hodnoty *VD* $_{pub}$ vypočítané pro jednotlivé subkategorie publicistického funkčního stylu

	bulvární noviny	časopisy	seriózní noviny	regionální noviny
bulvární noviny	×			
časopisy	3,35	×		
seriózní noviny	7,13	0,70	×	
regionální noviny	6,58	5,19	5,15	×

Tabulka č. 40: *u-test* hodnot *VD* $_{pub}$ ($\alpha = 0{,}05$; $u \geq 1{,}96$ vyjadřuje signifikantní diferenci)

žánr	$u_{v\ VD\ pub}$
bulvární noviny	9,85
regionální noviny	9,77
seriózní noviny	7,49
časopisy	5,33

Tabulka č. 41: Vážené rozdíly $u_{v\ VD\ pub}$ mezi jednotlivými subkategoriemi publicistického funkčního stylu



Graf č. 31: Podobnost/rozdílnost jednotlivých subkategorií publicistického funkčního stylu na základě hodnot VD_{pub} a $u_{v\ VD\ pub}$ (čára značí nesignifikantní rozdíl VD_{pub} mezi skupinami textů)

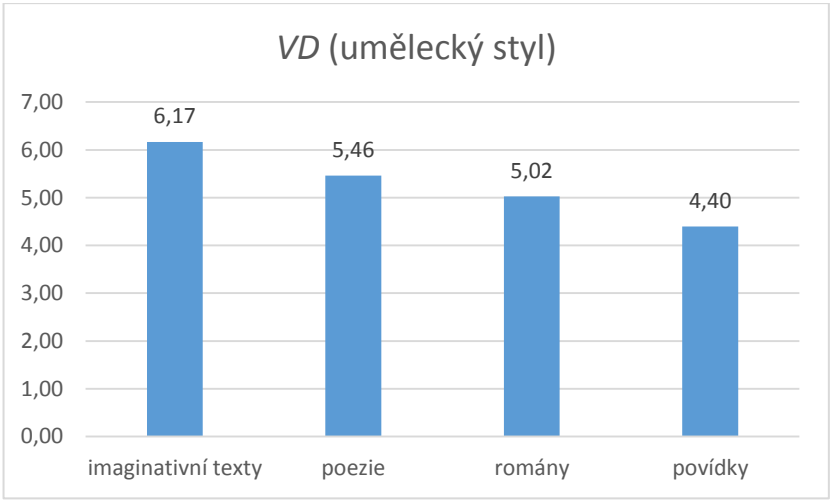
Jak jsme již naznačili výše, z hlediska syntaktického mají texty publicistického stylu často blízko ke stylu odbornému, zejména to pak platí o textech zpravodajských, v nichž se redaktoři snaží na malém rozsahu objektivně prezentovat co nejvíce dostupných informací, což vede ke kondenzovanému vyjadřování, k většímu užívání rozvitých přístavků, několikanásobných větných členů a k menšímu výskytu sloves ve finitních tvarech. Obsah seriózních novin a časopisů typu *Reflex*, *Týden* apod. tvoří ovšem kromě zpravodajských sdělení také texty analytické, které již více usilují o aktualizaci vyjádření a jejich syntaktická stavba není tolik standardizovaná. Proto zřejmě časopisy i seriózní noviny dosahují v naší analýze středních hodnot indexu *vzdálenosti sloves* a vzájemně se nijak signifikantně neliší.

Nicméně důvod, proč noviny regionální vykazují nejvyšší hodnoty *VD*, není na první pohled vůbec zřejmý. Částečně to může být tím, že městské

zpravodaje zastupující zde regionální publicistiku publikují většinou texty zpravodajského typu, mnohem závažněji se zde však pravděpodobně projevuje fakt, že v těchto novinách jsou ve velké míře zastoupena i sdělení formálně blízká administrativnímu stylu, v nichž např. město informuje své občany o rozhodnutích, která přijalo na svém zasedání zastupitelstva; tato oznámení bývají zveřejňována ve formě výčtů, často bez užití jediného slovesného tvaru, zato však s velkou frekvencí vlastních jmen (*ZMB schválilo: rozpočtová opatření týkající se: finančního vypořádání za rok 2002, přesunu provozních výdajů z důvodu poskytnutí příspěvku na obnovu kulturních památek, změny schváleného rozpočtu roku 2003,* […] *uzavření smlouvy o poskytnutí dotace SK Královo Pole na provoz sportovního areálu na ul. Vodova 108 v Brně,* […] *zrušení usnesení: ZMB Z4/003, bod 28, týkající se prodeje pozemku v k. ú. Trnitá* […]). Navíc bývají součástí těchto periodik i strany, až dvoustrany zveřejňující program jednotlivých kulturních či sportovních zařízení – takový přehled možných volnočasových aktivit je ovšem rovněž tvořen převážně jen údaji o čase a místě konání spolu s názvem či popisem příslušné činnosti. Dlouhé výčty a přehledy bez užitého slovesa poté nutně musí prodloužit průměrnou *vzdálenost sloves* v celém periodiku.

Opačných, tedy nejmenších hodnot indexu *VD* dosahují noviny bulvární. Vzhledem k tomu, že tento typ periodik je určen široké veřejnosti, zejména pak recipientům nižšího vzdělání, musí s touto skutečností korespondovat také jednodušší výstavba textů. Kratší, nepříliš složité větné celky přitom zajišťují snadnou recepci nezávažných sdělení bulvárních novin právě i pro čtenáře méně vzdělané.

6.3.1.4 Umělecký styl



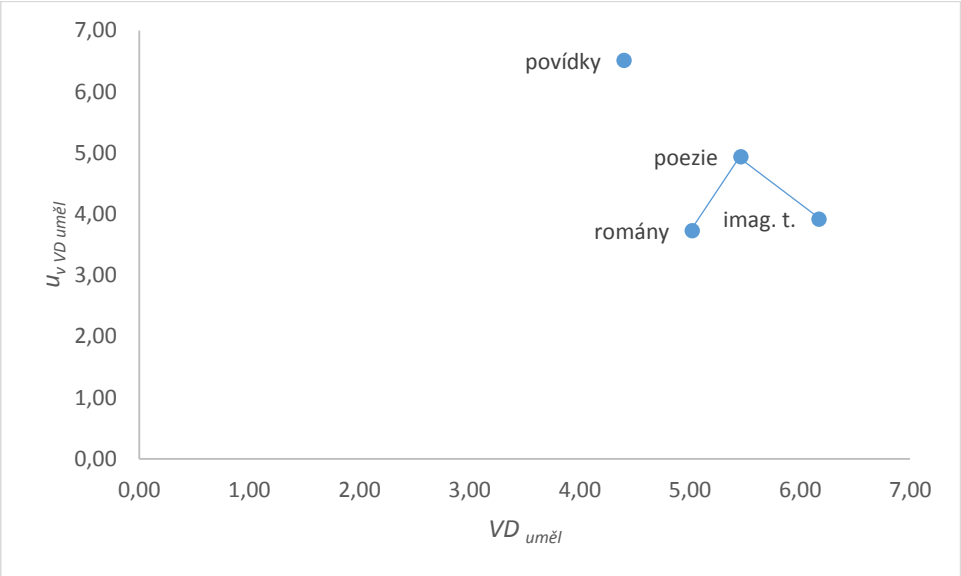
Graf č. 32: Výsledné hodnoty *VD* $_{uměl}$ vypočítané pro jednotlivé subkategorie uměleckého funkčního stylu

	romány	povídky	imaginativní texty	poezie
romány	×			
povídky	**2,55**	×		
imaginativní texty	**2,09**	**3,35**	×	
poezie	1,82	**5,38**	1,35	×

Tabulka č. 42: *u-test* hodnot *VD* $_{uměl}$ (α = 0,05; u ≥ 1,96 vyjadřuje signifikantní diferenci)

žánr	$u_{v\ VD\ uměl}$
povídky	6,51
poezie	4,94
imaginativní texty	3,92
romány	3,73

Tabulka č. 43: Vážené rozdíly $u_{v\ VD\ uměl}$ mezi jednotlivými subkategoriemi uměleckého funkčního stylu



Graf č. 33: Podobnost/rozdílnost jednotlivých subkategorií uměleckého funkčního stylu na základě hodnot *VD* $_{uměl}$ a $u_{v\ VD\ uměl}$ (čára značí nesignifikantní rozdíl *VD* $_{uměl}$ mezi skupinami textů)

Umělecký styl je z pohledu syntaktického nesmírně variabilní a není nesnadné v textech tohoto stylu nalézt příklady zcela protichůdných tendencí. Umělecké texty mohou obsahovat dlouhá souvětí, stejně jako krátké jednoduché věty či větné ekvivalenty, v řeči postav lze např. simulovat syntax běžně mluvených projevů, zatímco v pásmu vypravěče se mohou uplatnit i extrémně dlouhé

popisné pasáže. Z grafu č. 33 vyplývá, že texty imaginativní a poetické s nejvyššími hodnotami *VD* směřují k rozvolněnější syntaktické stavbě, což je v souladu i s našim intuitivním poznáním. V jejich blízkosti se pak vyskytují romány, jejichž výsledné hodnoty interpretujeme spíše jako důsledek dlouhých úvahových a popisných úseků typických pro velkou epiku než jako podobnost s veršovým principem členění textu.

Statisticky nejvíce odlišným žánrem v uměleckém stylu je však z hlediska měření *vzdálenosti sloves* žánr povídky. V tomto kratším epickém útvaru je obvykle pozornost zaměřena jen na jednu dějovou linii, která má rychlý spád a jsou v ní minimalizovány statické popisné pasáže. Větší důraz na dějovou složku má pak za následek menší vzdálenosti mezi slovesy, což získaná data také potvrzují.

Na základě uvedených výsledků lze konstatovat, že index *vzdálenosti sloves* splnil naše očekávání a dokázal rozpoznat rozdíly v syntaktické stavbě jednotlivých funkčních stylů. Domníváme se tedy, že nejen slovní bohatství, ale i tato kvantitativní charakteristika může být užitečná pro nejrůznější stylometrické analýzy.

6.4 Aktivita (*Q*) a deskriptivita (*D*) textu

Kromě indexů lexikálních nebo syntaktických můžeme u analýzy funkčních stylů uplatnit také matematické metody založené na kvantifikaci slovních druhů. Pozornost vzájemným vztahům mezi slovními druhy věnovala už v 60. a 70. letech 20. století Marie Těšitelová (1987, s. 89–91), která se snažila odhalit poměry mezi jistými slovními druhy např. pomocí *koeficientů rozvíjení v nominální* a *ve verbální* skupině, popř. také pomocí *koeficientu nominálnosti*.[3] Podobných poměrů by bylo možné vytvořit celou řadu, pravděpodobně nejvíce však lingvisty upoutal Busemannův koeficient, jenž počítá proporce mezi slovníkem adjektiv a slovníkem sloves (Těšitelová, s. 91). Právě poměrné zastoupení slovních druhů, jimiž se spíše popisuje, vůči slovním druhům vyjadřujícím zejména děj některé badatele zaujalo natolik, že na původní metodu navázali a dále ji modifikovali.

Čech, Popescu a Altmann (2014, s. 52–73) představili upravený Busemannův postup jako index měřící *aktivitu* a *deskriptivitu textu*. Za nositele aktivity autoři považují v první řadě slovesa (s výjimkou sloves stavových a modálních), popř. k nim přiřazují také deverbativní substantiva; deskriptivitu pak podle nich vyjadřují zejména adjektiva, někdy je ovšem možné za jejího reprezentanta považovat i adverbia odpovídající na otázku „jak?“. Software *QUITA*, který pro výpočet obou zmíněných indexů opět použijeme, kalkulaci *aktivity* i *deskriptivity* provádí jen na základě kvantifikace adjektiv a sloves. Celkovou aktivitu textu (*Q*) pak lze jednoduše vypočítat jako poměr:

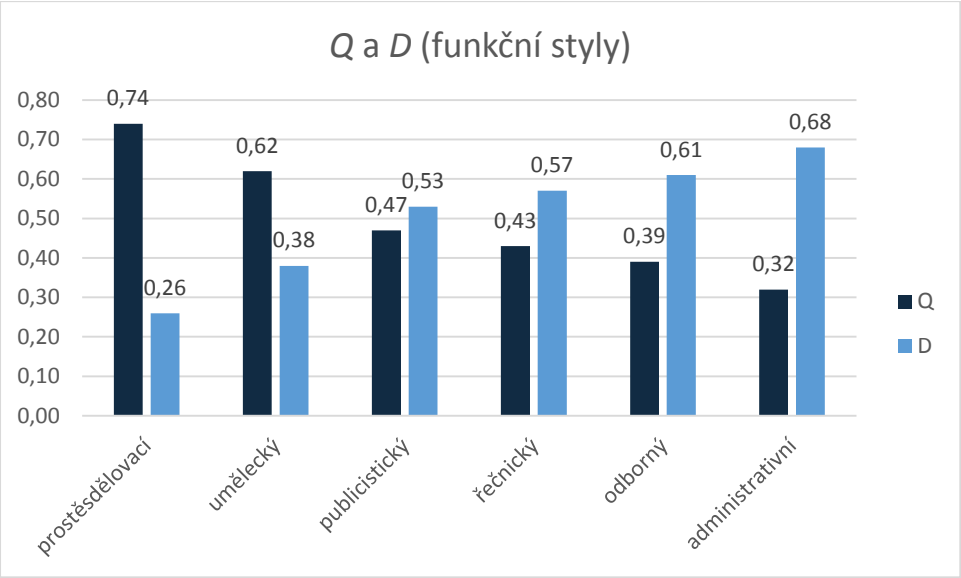
$$Q = \frac{V}{V+A}$$

V = počet sloves
A = počet adjektiv

6.4.1 Analýza funkčních stylů pomocí *Q*

V následujícím grafu zachycujeme vypočítané průměrné hodnoty *aktivity* jednotlivých funkčních stylů, *deskriptivita*, jak je z grafu patrné, poté odpovídá rozdílu hodnoty *aktivity* od 1. Čím větší tedy má určitý funkční styl *aktivitu*, tím má logicky menší hodnotu *deskriptivity* a naopak.

[3] *Koeficient rozvíjení v nominální skupině* lze vypočítat jako poměr počtu různých adjektiv k počtu různých substantiv, *koeficient rozvíjení ve skupině slovesné* jako poměr slovníku adjektiv ke slovníku sloves, *koeficient nominálnosti* jako poměr počtu různých substantiv ke slovníku sloves.



Graf č. 34: Výsledné hodnoty *aktivity* (*Q*) a *deskriptivity* (*D*) vypočítané pro jednotlivé funkční styly

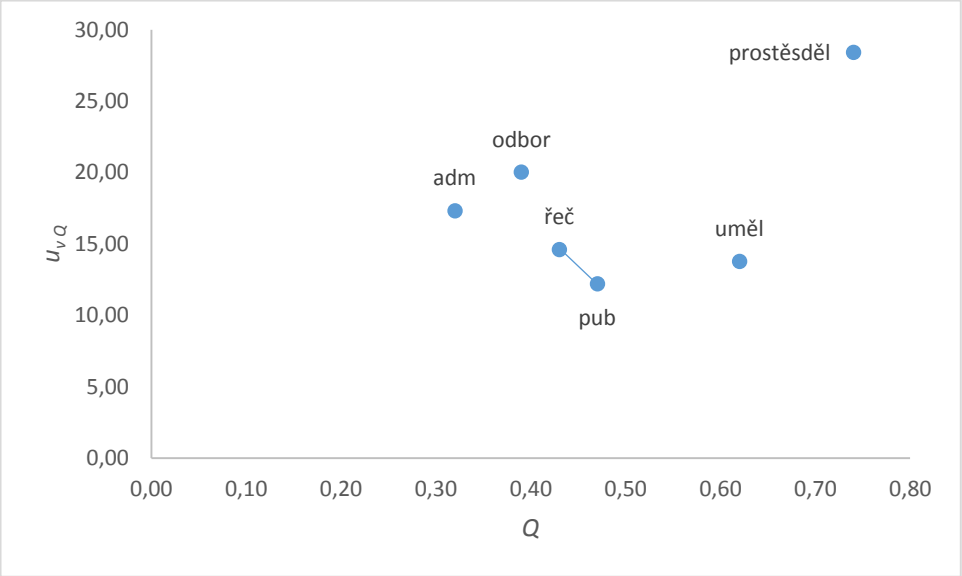
I tentokrát podrobíme výsledné hodnoty statistickému *u-testu* a naznačíme vzájemnou blízkost, či rozdílnost funkčních stylů z hlediska míry *aktivity*:

	adm	uměl	prostěsděl	odbor	řeč	pub
adm	×					
uměl	11,79	×				
prostěsděl	18,16	4,60	×			
odbor	2,92	8,59	14,23	×		
řeč	5,88	8,10	15,12	2,12	×	
pub	6,09	5,69	11,05	3,01	1,49	×

Tabulka č. 44: *u-test* hodnot *Q* (α = 0,05; u ≥ 1,96 vyjadřuje signifikantní diferenci)

styl	$u_{v\,Q}$
prostěsděl	28,44
odbor	20,05
adm	17,34
řeč	14,63
uměl	13,81
pub	12,22

Tabulka č. 45: Vážené rozdíly $u_{v\,Q}$ mezi jednotlivými funkčními styly



Graf č. 35: Podobnost/rozdílnost jednotlivých funkčních stylů na základě hodnot Q a $u_{v\,Q}$ (čára značí nesignifikantní rozdíl Q mezi styly)

Graf č. 35 rozdělil jednotlivé funkční styly na ty, v nichž hrají hlavní úlohu děj, dynamické prvky, a na ty, které jsou více statické, popisné. Protože předchozí index *vzdálenosti sloves* byl podobně jako index *aktivity* založen na kvantifikaci sloves, jejich výsledné hodnoty spolu do značné míry korelují. Z toho důvodu nebudeme již z hlediska *aktivity* zobrazovat vnitřní diferenciaci jednotlivých funkčních stylů, neboť zjištěné tendence i jejich interpretace již byly ozřejmeny výše.

Administrativní styl a styl odborný vykazují nejmenší hodnoty *aktivity* textu, což zcela odpovídá jejich nedějovému charakteru a dlouhým kondenzovaným formulacím s velkou frekvencí adjektiv a substantiv. Styl umělecký a hovorový poté tvoří protiklad těchto silně popisných funkčních stylů. Umělecké projevy epické obsahují větší množství sloves, aby mohly vyjadřovat děj, lyrická díla jsou zase více orientována na popis, tudíž v těchto typech komunikátu stoupá četnost adjektiv. Přestože tedy umělecký styl obsahuje jak texty dějové, tak texty deskriptivní, jako celek, zdá se, se řadí mezi styly s větší *aktivitou* textů.

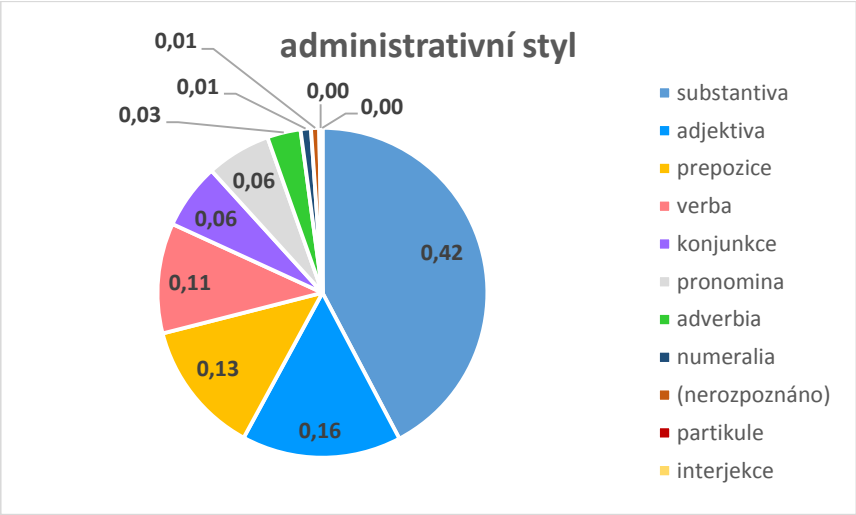
Nejvýraznější odlišnosti však z hlediska *aktivity* vykazuje styl prostěsdělovací. Protože se jedná o projevy mluvené, obvykle produktoři tvoří jen krátké výpovědi, které si jejich komunikační partner je schopný zapamatovat a vnímat. Cílem takové komunikace bývá zejména komunikace samotná, neformální rozhovory tedy neusilují o maximální věcnost; případné předání informace či jednoduché vyprávění nejsou předem připravené ani nemají

složitou strukturu, naopak při spontánní promluvě bývají výpovědi tvořeny asociativně, nebývají dokončeny a některé výrazy, popř. i větší motivy se několikrát opakují či opravují. Mluvčí také nezřídka uplatňují vycpávkové fráze typu *víš co*, *víš jak*, *já nevím*... To vše potom zvyšuje frekvenci použitých sloves reprezentujících dějovost ve spontánním dialogu.

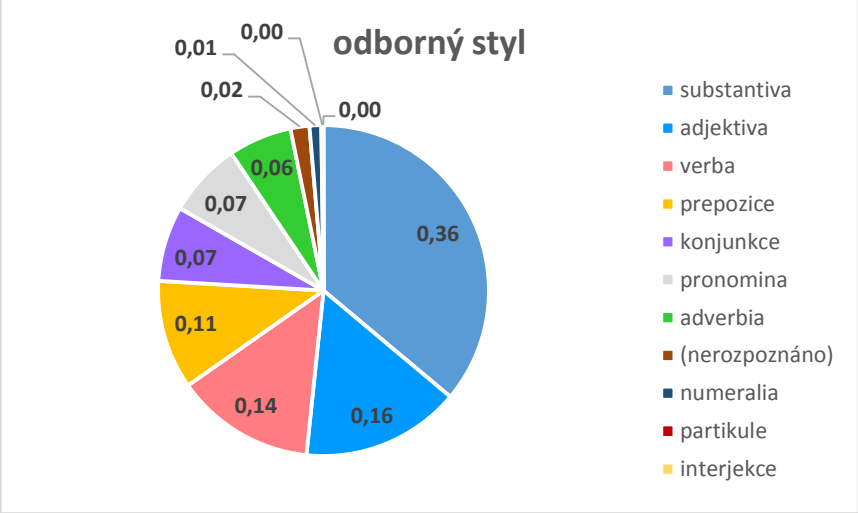
Hraniční pásmo mezi styly dějovými a popisnými tvoří nakonec funkční styl publicistický a řečnický. Publicistický styl, jak jsme již ukázali výše, je značně diferencován; zpravodajské útvary se recipientovi snaží objektivně předložit nezbytná fakta ve značně omezeném rozsahu, což vede k větší kumulaci substantiv a adjektiv, analytické projevy inklinují naopak k větší četnosti sloves, kterými objasňují a interpretují prezentované skutečnosti. Nejvíce dějové jsou ovšem z publicistického stylu texty bulvární, jež nezřídka prezentují i obyčejné skutečnosti jako vzrušující příběhy. Publicistický styl celkově tak není z hlediska *aktivity* nijak jednoznačně vyhraněný. Podobně je na tom také styl řečnický. Stejně jako publicistika usilují řečnické projevy promyšlenou argumentací a předneseným závažnějších skutečností o persvazi svého posluchače, čímž posilují věcnost daného sdělení. Přestože se ale jedná o projev připravený (často s psanou předlohou), neméně důležitý vliv má i zvuková realizace daného komunikátu, jež se musí přizpůsobit recepčním schopnostem posluchačů. Proto stejně jako neformální hovorové projevy mají i rétorické komunikáty jednodušší větnou stavbu s větším množstvím určitých slovesných tvarů a pro snazší pochopení doplňují závažné informace názornými příklady. Množství substantiv a adjektiv, jimiž se prezentují relevantní fakta, tak zhruba odpovídá četnosti sloves v jednodušších výpovědích, proto *aktivita* řečnického stylu vykazuje přibližně stejné hodnoty jako doplňující *deskriptivita*.

6.4.2 Distribuce slovních druhů

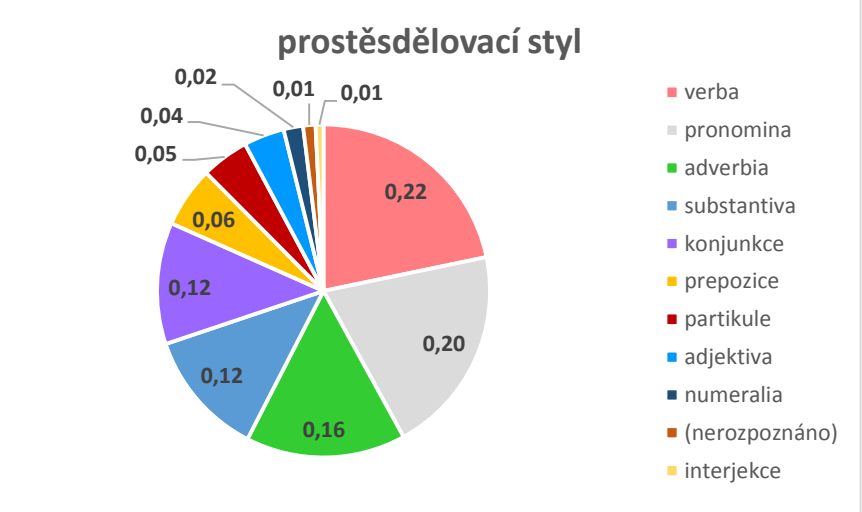
O přibližném zastoupení sloves a adjektiv v projevech různých funkčních stylů jsme si udělali představu již na základě indexu *aktivity* a *deskriptivity* textu. Abychom si potvrdili zjištěné tendence a abychom získali také větší přehled o celkové distribuci slovních druhů, identifikovali jsme pomocí softwaru *QUITA* v jednotlivých stylech všechny slovní druhy a na základě údajů o jejich absolutní frekvenci jsme dále vypočetli jejich frekvenci relativní. Výsledné proporce jednotlivých slovních druhů ve všech funkčních stylech jsou vizualizovány v grafech č. 36 až č. 41:



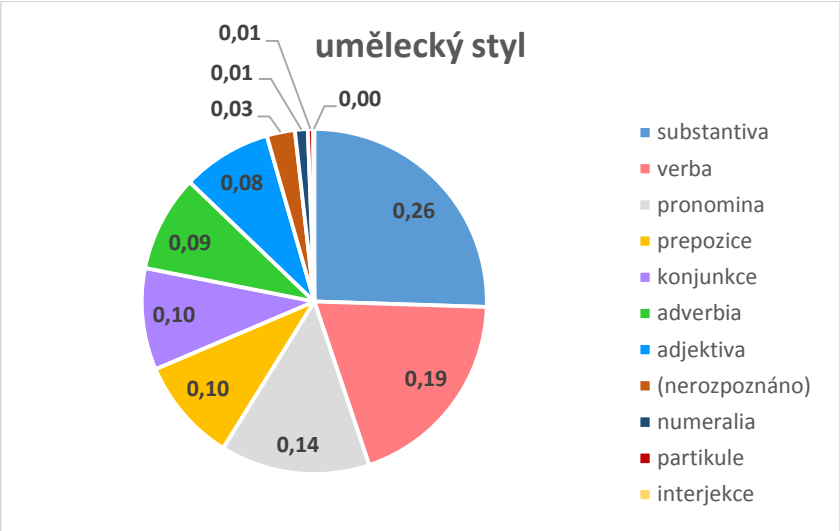
Graf č. 36: Relativní frekvence slovních druhů v administrativním stylu



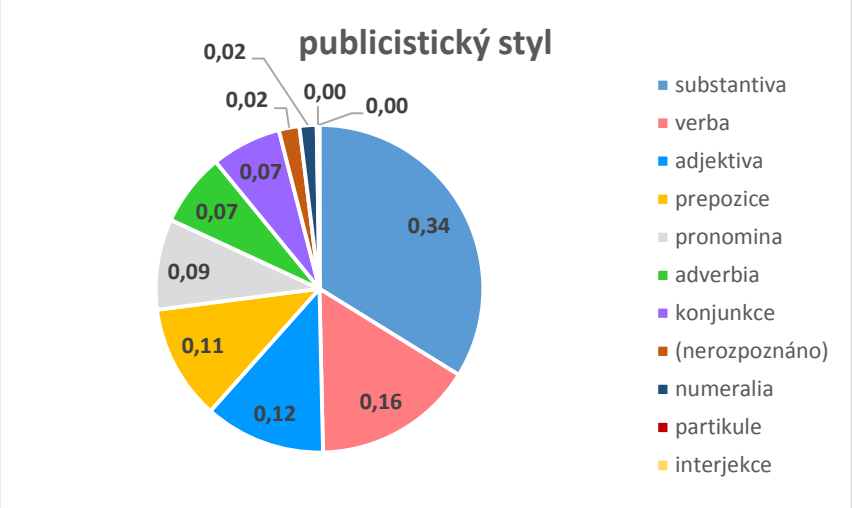
Graf č. 37: Relativní frekvence slovních druhů v odborném stylu



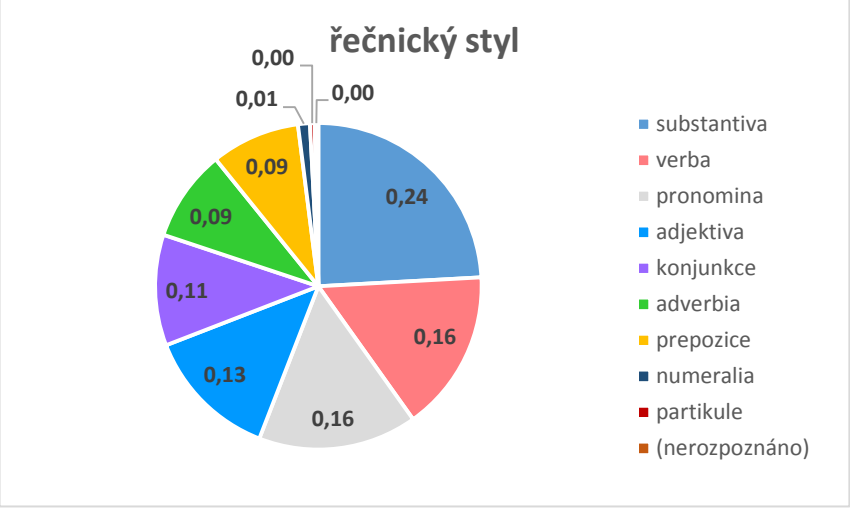
Graf č. 38: Relativní frekvence slovních druhů v prostěsdělovacím stylu



Graf č. 39: Relativní frekvence slovních druhů v uměleckém stylu



Graf č. 40: Relativní frekvence slovních druhů v publicistickém stylu



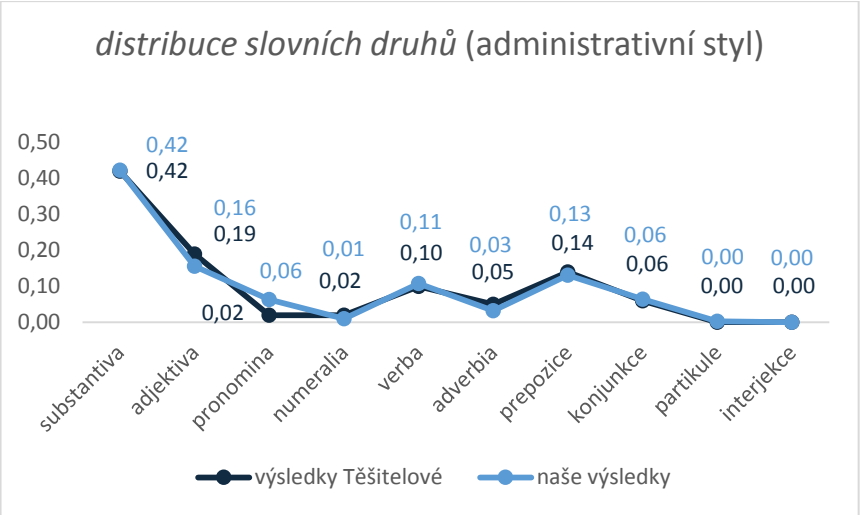
Graf č. 41: Relativní frekvence slovních druhů v řečnickém stylu

Uvedené proporce ve výše zobrazených grafech korespondují s výsledky získanými na základě výpočtu *aktivity* a *deskriptivity* textu. Také distribuce slovních druhů potvrdila výrazně věcný a popisný charakter stylu administrativního a odborného, součet relativní frekvence jejich substantiv a adjektiv tvoří dokonce více než 50 % celkové slovnědruhové distribuce. S vysokým počtem substantiv poté logicky souvisí i větší výskyt předložek. S ohledem na neosobní vyjadřování a snahu o co nejjasnější formulaci je pro dané skupiny textů charakteristický i menší podíl zájmen, tato skutečnost se pak ale zřejmě opět promítá do většího počtu podstatných jmen. Nižší výskyt sloves v odborném a v administrativním stylu má následně vliv na menší zastoupení adverbií, specifická je i poměrně malá četnost spojek, jež může signalizovat časté užití výčtů a asyndetických spojení.

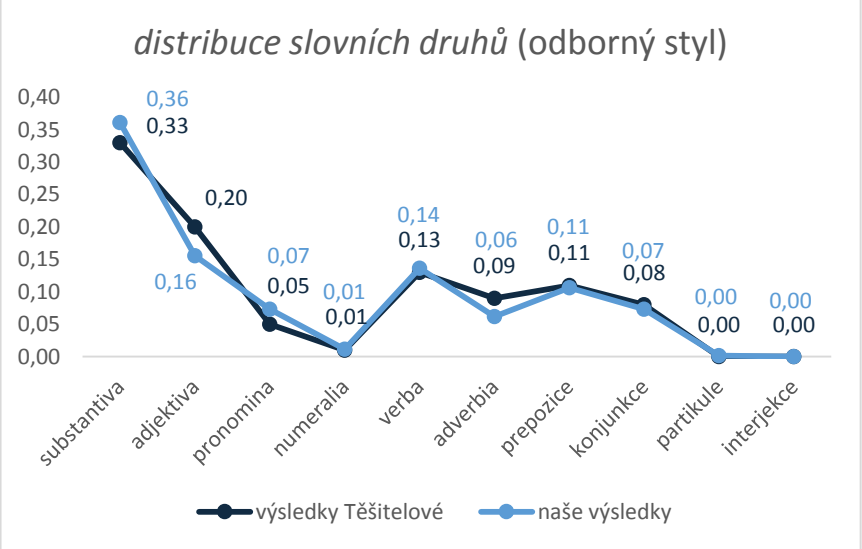
Naopak větší zastoupení sloves oproti adjektivům u stylů uměleckého a prostěsdělovacího podpořilo náš předpoklad o jejich více dějovém charakteru, v nepřipravených neoficiálních promluvách dokonce verba dosáhla největší relativní frekvence ze všech slovních druhů. Pro oba styly je rovněž typický větší podíl zájmen, vzhledem ale k tomu, že v osobní neformální komunikaci se předmětem hovoru stávají častěji konkrétní lidé než abstraktní představy či neživé věci a rovněž umělecké texty velice často zobrazují postavy, na něž se obvykle odkazuje zájmeny, není tento výsledek nijak překvapivý. V prostěsdělovacím stylu je ve srovnání s ostatními styly také patrný nárůst frekvence částic a citoslovcí, které vyjadřují vztah mluvčího k výpovědi, jeho náladu či pocity, v ostatních textech bývá výskyt těchto výrazů a jiných vycpávkových slov víceméně zanedbatelný.

Publicistické texty a projevy řečnické mají dle našeho šetření téměř totožné zastoupení sloves i adjektiv, toto zjištění pak odpovídá i přibližně shodným hodnotám vypočtené *aktivity* a *deskriptivity* v komunikátech daných stylů. Distribucí zbylých slovních druhů, zdá se, je pak publicistický styl blíže k projevům administrativním a odborným, řečnický styl svým větším zastoupením zájmen, a tedy i menší četností substantiv a prepozic se pak více přibližuje stylu prostěsdělovacímu a uměleckému.

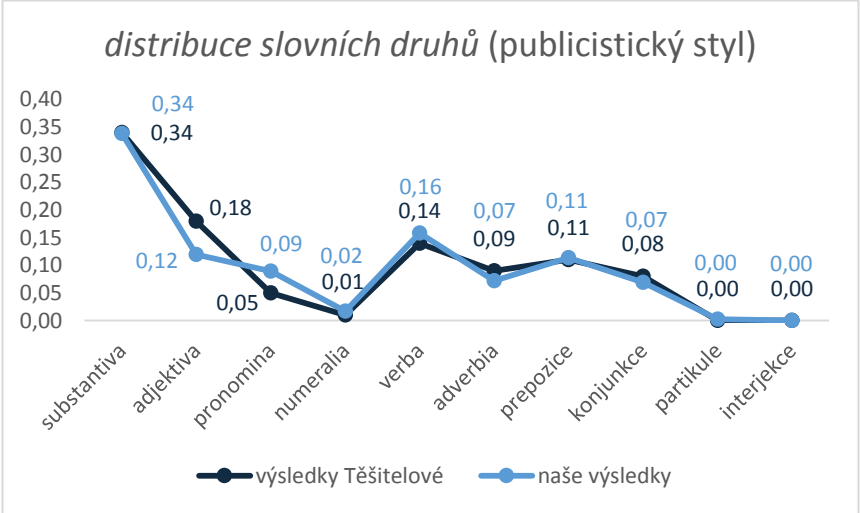
Prezentovanou slovnědruhovou distribuci můžeme pro ilustraci srovnat také s výsledky, k nimž dospěla Marie Těšitelová ve svých publikacích zaměřených na kvantitativní charakteristiky věcného stylu (Těšitelová 1982, 1983, 1985):



Graf č. 42: Distribuce slovních druhů v administrativním stylu



Graf č. 43: Distribuce slovních druhů v odborném stylu



Graf č. 44: Distribuce slovních druhů v publicistickém stylu

Srovnáním slovnědruhové distribuce vypočtené pro projevy věcného stylu v 1. polovině 80. let 20. století s rozložením slovních druhů identifikovaným v našem vzorku jsme zjistili, že z pohledu frekvence zůstává pořadí jednotlivých slovních druhů v obou distribucích v podstatě zachováno. Jistý rozdíl se projevil jen mezi adjektivy a pronominy, konkrétně v našem vzorku dosáhla přídavná jména oproti zjištěním Těšitelové obecně nižších relativních frekvencí ve všech třech funkčních stylech, naopak zájmena opět v celém věcném stylu vykazují v našich textech vyšší četnosti. Vzhledem k tomu, že hodnoty relativních frekvencí u ostatních slovních druhů se liší max. jen o 1–2 %, může se zdát, že v novějších datech došlo k jistému posunu ve vyjadřování. Rozdíl je však pravděpodobně způsoben jen odlišným chápáním adjektiv a zájmen, neboť pronomina, která se formálně shodují s adjektivy, jsou u Těšitelové zařazena mezi adjektiva – tím pak dochází k poklesu frekvence zájmen právě ve prospěch jmen přídavných.

Dá se tedy předpokládat, že program *QUITA*, jejž jsme pro určení slovnědruhové distribuce zvolili, dokázal jednotlivé slovní druhy detekovat úspěšně, a tedy i výsledky *aktivity* a *deskriptivity* textů, které je s rozložením slovních druhů v souladu, lze považovat za obecně platné.

7. ZÁVĚR

V naší práci jsme se zaměřili na analýzu funkčních stylů za využití několika moderních, avšak méně často užívaných kvantitativnělingvistických metod. Zjištěné stylové charakteristiky zřejmě nelze označit za zcela nové, za přínosný však považujeme způsob, jímž jsme k prezentovaným poznatkům dospěli, neboť jsme dokázali dosud intuitivní teze týkající se jednotlivých funkčních stylů empiricky ověřit. Intersubjektivní náhled na jednotlivé styly tak přináší exaktní výsledky, jež můžeme s vědomím limitů konkrétních metod považovat za obecně platné.

Na základě dílčích kvantitativních analýz a jejich statistického vyhodnocení lze konstatovat, že největší diference vůči ostatním stylům vykazuje styl prostěsdělovací. Vzhledem ke skutečnosti, že řečnické projevy jsou sice ve výsledku promlouvány, ale jejich psaná předloha má značný vliv na konečnou podobu komunikátu, můžeme považovat prostěsdělovací styl za jediného reprezentanta spontánních mluvených projevů. Získaná data ukazují, že neformální nepřipravené mluvené dialogy mají nejchudší lexikum, jejich výpovědi jsou krátké, asociativně tvořené a mají jednoduchou syntaktickou strukturu s množstvím odchylek od pravidelné větné stavby. Vzhledem k polytematickému charakteru vykazují také nejmenší tematickou koncentraci. Co se týče možností a limitů užitých metod, projevila se zde rovněž skutečnost, že index *TK* je u mluvených komunikátů vhodný jen pro komparaci celkové tematické zaměřenosti jednotlivých textů či skupin textů, vzhledem k častým výsledným nulovým hodnotám se však již nehodí pro detekci tematických slov.

Mezi psanými texty poté vykazují největší odlišnosti texty administrativní. Výrazná stylová specifika tohoto stylu opět podporují oprávněnost rozhodnutí tradiční stylistiky (např. Čechová – Krčmová – Minářová 2008) nezahrnovat projevy administrativní povahy do stylu odborného, ale vydělit jej jako samostatný funkční styl. Pro komunikáty z administrativní oblasti jsou charakteristické nízké hodnoty slovního bohatství, přesně dodržované, až stereotypní formulace a maximální koncentrace na téma bez jakékoli redundance. Věcné, nanejvýš zhuštěné sdělení s převahou substantiv a adjektiv je rovněž příznačným rysem pro texty z administrativní oblasti, podobné charakteristiky ovšem najdeme i ve stylu odborném.

Právě odborný styl lze rovněž zařadit do trojice nejvíce se odlišujících stylů. Svými výslednými hodnotami mají texty odborného charakteru blízko ke komunikátům administrativním, jejich koncentrace na sledovanou problema-

tiku, ekonomičnost vyjadřování a schematičnost je však přece jen méně striktní než u textů administrativních; zejména pak projevy populárně naučné jsou svou formou zpracování více uvolněné ve snaze přizpůsobit se méně poučenému recipientovi.

Zbývající trojice funkčních stylů již mezi sebou tolik výrazné rozdíly nemá. Trochu překvapivé je zjištění, že styl řečnický se na základě většiny našich vybraných kvantitativních charakteristik od stylu publicistického dokonce ani signifikantně neliší. Přestože si jsou oba styly kolektivním adresátem a snahou o persvazi vzájemně blízké, očekávali bychom, že jako samostatně vyčleněné styly budou jistou míru diference vykazovat. Ačkoli bychom tedy v praxi pravděpodobně dokázali rétorický projev na základě zvukové podoby intuitivně rozpoznat, zdá se, že pro stylometrické metody je slavnostní promluva příliš málo specifická.

V rámci stylu publicistického se jako nejvíce odlišné jeví noviny regionální a bulvární. U lokálně zaměřených periodik byly poněkud překvapivě zjištěny vysoké hodnoty indexu *VD*. Tato skutečnost může být interpretována jako důsledek častých oznámení o schválených rozhodnutích města, která jsou dlouhými výčty a uváděním přesných názvů formálně blízká textům administrativním, jistý vliv může mít i několik stran publikovaného kulturního přehledu bez užitého slovesa, jenž bývá běžnou součástí regionálních novin. Zcela opačných výsledků pak dosahují bulvární sdělení; jednodušší syntax a malé zaměření na jedno téma ovšem koresponduje s jejich povrchním zpracováním velkého počtu nezávažných událostí s cílem oslovit co největší počet čtenářů.

Ani umělecký styl se nakonec v naší analýze neprojevil jako výrazně odlišný, v jednotlivých kvantitativních charakteristikách dosahoval spíše středních hodnot, jen při měření aktivity textu se zařadil těsně za styl prostěsdělovací, tj. na pozici signalizující v textech větší převahu prvků dějových nad prvky deskriptivními. Tato tendence se nám poté potvrdila i na základě distribuce slovních druhů, neboť umělecké texty dosáhly v průměru 19% zastoupení sloves, jimiž se právě dějovost vyjadřuje. Většího podílu verb dosáhly již jen zmíněné mluvené projevy neformálního charakteru, z psaných komunikátů však lze považovat právě umělecké texty za nejvíce aktivní, resp. nejvíce zaměřené na příběh.

Na závěr můžeme konstatovat, že zvoleným kvantitativním metodám se podařilo poměrně úspěšně rozlišit jednotlivé funkční styly, přičemž na základě různých sledovaných parametrů dokázaly stanovit i některé jejich společné, či naopak zcela rozdílné charakteristiky. Protože jsme však v naší práci

aplikovali uvedené matematické a statistické metody jen na omezený vzorek, bude do budoucna potřeba výsledky, k nimž jsme dospěli, ověřit i na rozsáhlejším materiálu.

8. ANOTACE

Autor diplomové práce: Lenka Horutová

Název katedry a fakulty: Filozofická fakulta Univerzity Palackého v Olomouci, Katedra bohemistiky

Název diplomové práce: Kvantitativní analýza funkčních stylů

Vedoucí diplomové práce: PhDr. Petr Pořízka, Ph.D.

Počet znaků: 124 189

Počet příloh: 0

Počet titulů použité literatury: 23

Klíčová slova: kvantitativní lingvistika, funkční styly, slovní bohatství textu, tematická koncentrace textu, aktivita textu, vzdálenosti sloves, statistické testy

Charakteristika diplomové práce:

Práce se zaměřuje na analýzu funkčních stylů za využití vhodných kvantitativních metod. Výzkum a komparace jednotlivých stylů se dotýkají nejen oblasti lexikální statistiky (např. při výpočtu slovního bohatství textů), ale uplatňován je i pohled morfologickostatistický (např. prostřednictvím měření aktivity a deskriptivity textu) nebo syntaktickostatistický (např. zkoumání složitosti syntaktické struktury na základě vzdálenosti sloves). Analýza je založena na výběrovém materiálu excerpovaném z databází Českého národního korpusu a Olomouckého korpusu mluvené češtiny. Za hlavní cíl si práce klade empirické ověření obecných tezí vztahujících se k jednotlivým funkčním stylům.

9. RESUMÉ

Our work focuses on the analysis of functional styles using selected methods of quantitative linguistics. The research and comparison of various styles concern not only the lexical statistics, the methods of quantitative linguistics which are most frequently used in the context of the quantitative linguistics, but we also take in consideration the morphological statistics and syntactic statistics. However, the aim of the work is not to reveal other characteristic features of functional styles or to formulate their new classification. We focus mainly on the quantification and empirical verification of the various intuitive propositions that are related to each individual style.

The analysis is based on a selected material excerpted from the databases of the Czech National Corpus and from Olomouc Corpus of Spoken Czech. The assembled body then consists of a total of 120 speeches of six functional styles, different authors and different themes so that the examined sample covers the variety of texts which belong to the various functional styles.

We used methods that do not show any dependence on the length of the text, namely: *Moving Average Type-Token Ratio* (*MATTR*), *thematic concentration of text* (*TK*), *secondary thematic concentration of text* (*STK*), *verb distances* (*VD*), *activity* (*Q*) and *descriptivity* (*D*). All resulting values obtained by mentioned quantitative methods are statistically tested and we tried to relevantly interpret them in terms of linguistics.

On the basis of partial quantitative analysis and the statistical evaluation we can say that the neutral style has the biggest difference from the other styles. The neutral style has the poorest vocabulary, simple syntactic structure with a number of deviations from regular sentence structure, and also shows the slightest thematic concentration. This result confirms the significant distinction between the spoken and written texts. From the written texts, the most distinguished are the technical and administrative forms which are characterized by low levels of vocabulary richness, precisely adhered and stereotypical formulations and the maximal concentration on the subject without any redundancy. The remaining trio of functional styles shows no significant differences among themselves, the rhetorical style and journalistic style are not even significantly different on the basis of most of our selected quantitative characteristics of both styles.

The selected quantitative methods quite successfully distinguished individual functional styles, and were able to specify some of the common or completely different characteristics based on various observed parameters. In addition, the intersubjective view brought the exact results which can be regarded as generally valid, knowing the limits of specific methods. Because we have applied these mathematical and statistical methods to a limited sample, the results will need to be verified using larger material in the future.

10. LITERATURA

COVINGTON, Michael A. – MCFALL, Joe D. Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR). *Journal of Quantitative Linguistics*, 17 (2), 2010, s. 94–100.

CVRČEK, Václav – RICHTEROVÁ, Olga (eds). Pojmy: typ (type). *Příručka ČNK*, 2014 [online]. Dostupné z: <https://wiki.korpus.cz/doku.php/pojmy:typ> [cit. 20. 6. 2016].

CVRČEK, Václav – RICHTEROVÁ, Olga (eds). Pojmy: token. *Příručka ČNK*, 2014 [online]. Dostupné z: <https://wiki.korpus.cz/doku.php/pojmy:typ> [cit. 20. 6. 2016].

ČECH, Radek. Tematická analýza novoročních projevů československých a českých prezidentů z let 1949–2013. In: *Slovo a text v historickém kontextu*. Brno: Host – Ostrava: Ostravská univerzita v Ostravě, 2013, s. 42–61.

ČECH, Radek. Jen popis s čísly? Perspektivy korpusové lingvistiky. *Naše řeč*, 97 (4–5), 2014, s. 171–184.

ČECH, Radek. *Tematická koncentrace textu v češtině*. Praha: ÚFAL, 2016.

ČECH, Radek – DAVID, Jaroslav – DAVIDOVÁ GLOGAROVÁ, Jana. Analýza tematické koncentrace textu: komparace publicistiky Ladislava Jehličky a Karla Čapka. *Slovo a slovesnost*, 74 (1), 2013, s. 41–54.

ČECH, Radek – POPESCU, Ioan Iovitz – ALTMANN, Gabriel. *Metody kvantitativní analýzy (nejen) básnických textů*. Olomouc: UPOL, 2014.

ČECHOVÁ, Marie – KRČMOVÁ, Marie – MINÁŘOVÁ, Eva. *Současná stylistika*. Praha: Nakladatelství Lidové noviny, 2008.

HENDL, Jan. *Přehled statistických metod zpracování dat*. Praha: Portál, s. r. o., 2004.

KARLÍK, Petr – NEKULA, Marek – PLESKALOVÁ, Jana (eds.). *Encyklopedický slovník češtiny*. Praha: Nakladatelství Lidové noviny, 2002.

KARLÍK, Petr – NEKULA, Marek – RUSÍNOVÁ, Zdenka (eds.). *Příruční mluvnice češtiny*. Praha: Nakladatelství Lidové noviny, 1995.

KUBÁT, Miroslav. *Funkční styly z hlediska lexikální statistiky* [magisterská diplomová práce]. Olomouc: UPOL, 2012.

KUBÁT, Miroslav. *Kvantitativní analýza žánrů* [dizertační práce]. Olomouc: UPOL, 2015.

KUBÁT, Miroslav – MATLACH, Vladimír – ČECH, Radek. *QUITA – Quantitative Index Text Analyzer*. Lüdenscheid: RAM, 2014.

KUBÁT, Miroslav – MILIČKA, Jiří. Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics*, 20 (4), 2013, s. 339–349. Dostupné z: <https://www.researchgate.net/publication/258518594_Vocabulary_Richness_Measure_in_Genres>.

MATLACH, Vladimír. *Kvantitativně lingvistický software* [magisterská diplomová práce]. Olomouc, 2014.

TĚŠITELOVÁ, Marie. *Otázky lexikální statistiky*. Praha: Academia, 1974.

TĚŠITELOVÁ, Marie. *Kvantitativní charakteristiky současné české publicistiky*. Praha: Československá akademie věd, Ústav pro jazyk český, 1982.

TĚŠITELOVÁ, Marie. *Kvantitativní charakteristiky současné odborné češtiny (v rámci věcného stylu)*. Praha: Československá akademie věd, Ústav pro jazyk český, 1983.

TĚŠITELOVÁ, Marie. *Současná česká administrativa z hlediska kvantitativního*. Praha: Československá akademie věd, Ústav pro jazyk český, 1985.

TĚŠITELOVÁ, Marie. *Kvantitativní charakteristiky současné češtiny*. Praha: Nakladatelství Československé akademie věd, 1985.

TĚŠITELOVÁ, Marie. *Kvantitativní lingvistika*. Praha, 1987.

Korpusové databáze:

Český národní korpus – korpusy psané i mluvené češtiny. Dostupné z: <http://www.korpus.cz>.

POŘÍZKA, Petr a kol. *Olomoucky korpus mluvené češtiny*. Interní materiály KB FF UPOL, archiv autora.

Softwarové nástroje:

COVINGTON, Michael A. *MATTR (Moving-Average Type-Token Ratio)* (software). Athens–Georgia: Artificial Intelligence Center, The University of Georgia, 2007 (dostupné z: <http://ai1.ai.uga.edu/caspr/>).

MATLACH, Vladimír – KUBÁT, Miroslav – ČECH, Radek. *QUITA – Quantitative Text Analyzer* (software). Olomouc: UPOL, 2014 (dostupné z: <https://code.google.com/p/oltk/>).

MILIČKA, Jiří. *MaWaTaTaRaD* (software). Praha, 2013 (dostupné z: <http://milicka.cz/mawatatarad/>).

XLSTAT (statistický doplněk pro *Microsoft Excel*) (dostupné z: <https://www.xlstat.com/en/download>).