UNIVERZITA PALACKÉHO

FILOZOFICKÁ FAKULTA

Katedra obecné lingvistiky

Miroslav Kubát

Kvantitativní analýza žánrů

Disertační práce

Školitel: Mgr. Radek Čech, Ph.D.

Olomouc 2015

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

V Olomouci 25. dubna 2015 ……

Rád bych poděkoval především Radku Čechovi za cenné rady, vstřícné vedení a všestrannou podporu při psaní této práce.

Obsah

1. Úvod

Tato práce stojí na pomezí tradiční lingvistiky a mezioborové disciplíny, jež se označuje jako matematická či kvantitativní lingvistika. Pokud bychom měli vymezit co nejpřesněji obor, do nějž spadá náš výzkum, pak jej můžeme bez větších obtíží zařadit do aplikované kvantitativní jazykovědy, konkrétně pak do stylometrie. Základní cíle této práce jsou dva, prvním je na základě analýzy konkrétních textů za použití kvantitativních a statistických metod prozkoumat některé textové vlastnosti, přičemž nám nejde o předkládání nových teorií či rozdělení stylů nebo žánrů, ale o kvantifikaci zkoumaných jevů založenou na experimentálních metodách. Druhým cílem je zjistit, do jaké míry jsou jednotlivé metody relevantní pro diferenciaci žánrů.

První pokusy o kvantitativní žánrovou analýzu můžeme v českém jazykovědném prostředí sledovat v díle Marie Těšitelové,[1] která byla průkopníkem kvantitativního přístupu ke zkoumání jazyka u nás. Vedle Těšitelové můžeme zmínit také jejího vrstevníka, slovenského jazykovědce Jozefa Mistríka.[2] Oba zmínění lingvisté sehráli zásadní roli v rozvoji kvantitativní lingvistiky v československé jazykovědě, zejména pak v oblasti lexikální statistiky. Tato práce do značné míry navazuje na výše uvedené badatele a aktualizuje stylometrické poznatky o českém jazyce za použití metod současné kvantitativní lingvistiky. Přehledu dnešního stavu stylometrického bádání ve světě je pak v textu věnována samostatná kapitola.[3]

Klíčovou roli jakéhokoliv kvantitativního výzkumu hraje výběr vhodného materiálu, v našem případě půjde o texty Karla Čapka. Dílo jednoho z nejslavnějších českých spisovatelů totiž nabízí neobyčejně pestrou paletu textů různých žánrů (romány, povídky, pohádky, cestopisy, básně, novinové sloupky, dramata, odborné studie či dopisy). Omezením korpusu analyzovaných textů na jediného autora jsme dosáhli eliminace nežádoucího (z hlediska stylometrie) vlivu různých autorských stylů, jenž nutně devalvuje vypovídací hodnotu podobných výzkumů. Každý autor

[1] Např.:
Těšitelová, M. (1972).
Těšitelová, M. (1974).
Těšitelová, M. (1987).
Těšitelová, M. (1983).
Těšitelová, M. a kol. (1987).
[2] Např.:
Mistrík, J. (1969).
Mistrík, J. (1985)..
[3] Kap. 2.2. *Současná stylometrie*.

má totiž svůj jedinečný způsob psaní, který prostupuje napříč žánry. V případě zařazení více autorů do korpusu nutně znemožníme relevantní vyhodnocení výsledků, neboť nebudeme s to zjistit, zda získané hodnoty vypovídají spíše o autorovi nebo o žánru. Na druhou stranu je třeba poznamenat, že omezením korpusu na jediného autora nelze závěry zobecnit, protože nevíme, jak by se sledované ukazatele chovaly v případě jiných autorů. Materiál Čapkových textů je však dle našeho názoru dostatečně velký a rozmanitý na to, aby mohl sloužit jako výchozí bod v této dosud jen málo prozkoumané oblasti. Problematice vytváření korpusu je v textu věnována samostatná kapitola.[4]

Použité metody lze rozdělit do dvou skupin. První tvoří stylometrické ukazatele, které umožňují přímou lingvistickou interpretaci. Důraz je kladen zejména na slovní bohatství, které patří k základním stylometrickým indexům již od počátků kvantitativní lingvistiky.[5] Protože většina metod měření slovního bohatství je závislá na délce textu, kvantitativní lingvisté se této problematice věnují již desítky let. Teprve v roce 2010 byla navržena metoda *MATTR* (moving average type-token ratio)[6], která je prvním indexem měření slovního bohatství bez vlivu délky textu. Z *MATTR* byla v roce 2013 odvozena přesnější metoda *MWTTRD* (moving type-token ratio distribution)[7], která pracuje s celou distribucí hodnot.

Právě nezávislost na délce textu byla pro výběr použitých metod zásadní. Proto jsme vedle slovního bohatství, jež má v této práci dominantní postavení, zvolili následující metody, které taktéž nejsou ovlivněny délkou textu: tematická koncentrace textu, vzdálenosti sloves, průměrná délka tokenu, aktivita a deskriptivita textu, distribuce slovních druhů. Druhou skupinu použitých metod tvoří ty, které jsou primárně určeny pro automatickou klasifikaci textů a určování autorství. Konkrétně jsme vybrali víceúrovňový autorský profil *AMNP* (Author's Multilevel N-gram Profile), který kombinuje bigramy a trigramy grafémů a slov. Další metodou je potom *MFW* analýza, která porovnává nejfrekventovanější slova. Obě metody se vyznačují vysokou přesností klasifikace textů, zejména pokud jde o určování

[4] Kap. 3.3 *Korpus*.
[5] Viz např. Yule, G.U. (1944).
[6] Covington, M. A., McFall J. D. (2010).
[7] Kubát, M., Milička, J. (2013).

autorství. Naším cílem bude ověřit, zda a do jaké míry jsou použitelné pro žánrovou analýzu.

Důležitým aspektem každého kvantitativnělingvistického výzkumu je výběr vhodného nástroje pro zpracování dat. Dnes je již nepředstavitelné a vzhledem k množství analyzovaných textů často i nemožné provádět měření a výpočty ručně. Proto v případě našeho korpusu, který sestává ze 760 textů, bylo nezbytné použít adekvátní software. V tomto výzkumu byl pro zpracování dat jako základní nástroj použit multifunkční software *QUITA*[8] (Quantitative Index Text Analyzer). Kromě *QUITA* byly v dílčích analýzách použity také další programy, přičemž k nejdůležitějším patří nástroj k měření slovního bohatství *MaWaTaTaRaD*[9], statitistický software *R*[10] nebo stylometrický program *Stylo*[11].

Nedílnou součástí předkládaného textu je také příloha, jež mimo jiné obsahuje frekvenční slovníky použitého korpusu, a to jak slovních tvarů, tak lemmat.

[8] Matlach, V., Kubát, M., Čech, R. (2014).
[9] Milička, J. (2013).
[10] R Core Team (2013).
[11] Eder, M., Kestemont, M., Rybicki, J. (2013).

2. Stav oboru

2.1. Kvantitativní lingvistika

Ačkoliv mají kvantitativní metody v jazykovědném bádání již pevné místo,[12] nelze přehlédnout fakt, že pro mnohé lingvisty je využití matematiky stále čímsi neznámým a cizím. Někteří jen zatím neměli příležitost nahlédnout do tajů kvantitativní lingvistiky, někteří ji odmítají z principu. Gabriel Altmann, považovaný za zakladatele současné kvantitativní lingvistiky, k tomu poznamenává: „The conventional university education in linguistics does not stimulate a linguist to deal with language in any way other than the orthodox one. There are unwritten norms for the extent of knowledge, some things are important, and some others will not even be taken into consideration."[13] Čech a kol. pak hovoří o strachu lingvistů z matematických metod, který znemožňuje širší uplatnění kvantitativního přístupu v jazykovědě: „Praxe však ukazuje, že největší potíž při aplikaci těchto a jim podobných 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."[14] Proto považujeme za vhodné alespoň stručně tento obor vymezit a nastínit jeho základní východiska.

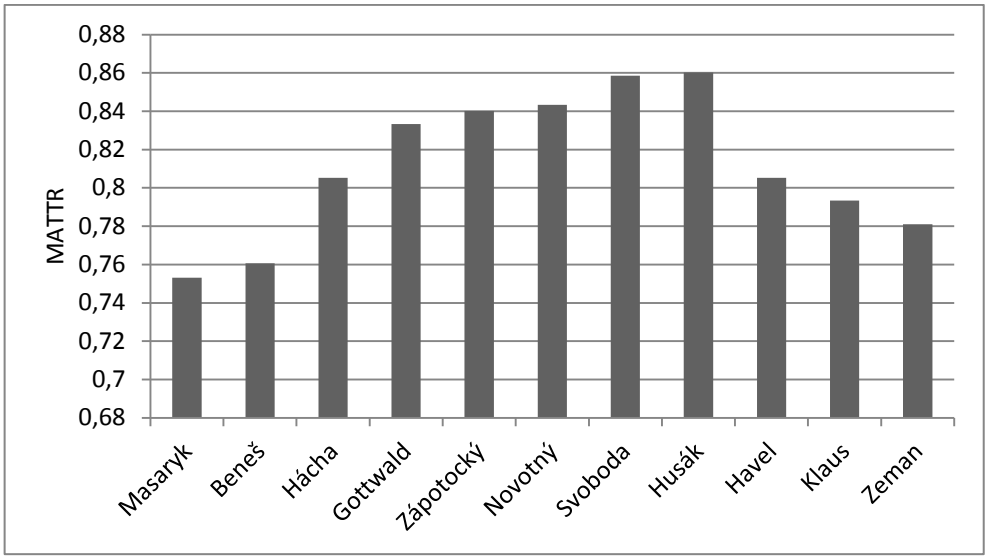
Základním cílem kvantitativní lingvistiky je poznání přirozeného jazyka z nejrůznějších hledisek. Užití matematických metod tedy v žádném případě neznamená odvrácení pozornosti od jazyka jiným směrem, rozdíl spočívá v cestě, kterou se k poznatkům badatelé dostávají. Kvantitativní lingvisté docházejí k závěrům pouze na základě experimentu, což umožnuje intersubjektivní pohled na zkoumanou problematiku. Základní výhodou takového přístupu je to, že kdokoliv daný experiment zopakuje, získá stejné výsledky. Pokud bychom však zadali kvalitativní analýzu jednoho textu několika lidem, jen stěží budou výsledky totožné.

[12] Srov. Uhlířová (2005).
[13] Altmann, G. (1997), s. 13.
[14] Čech, R., Popescu, I. I., Altmann, G. (2014), s. 5.

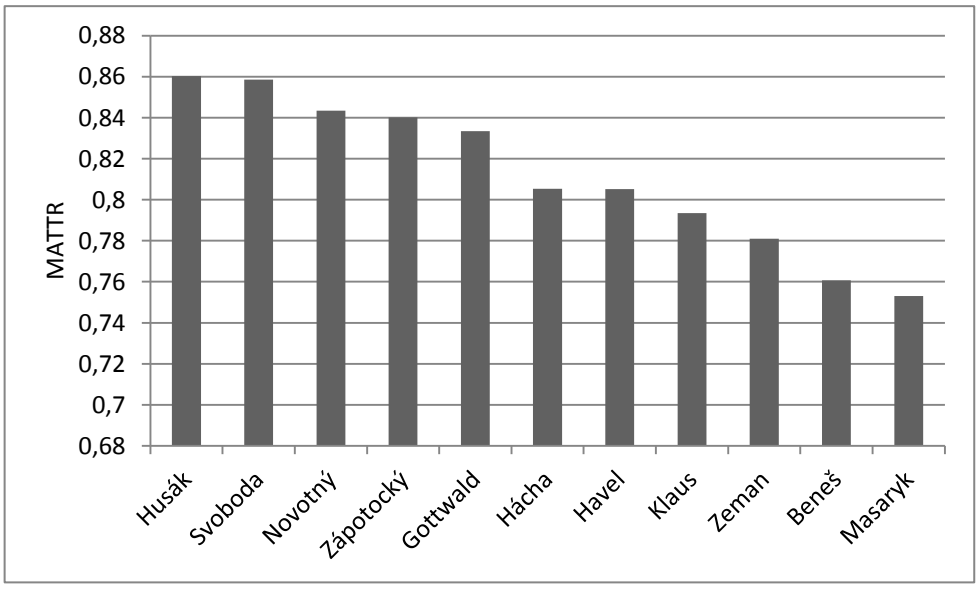
Kvantitativní přístup nám umožňuje ověřovat naše intuitivní a do značné míry subjektivní předpoklady. Problematiku shrnuje Wimmer a kol. takto: „Či už používame verbálne alebo matematické modely, *všetko* je v nich zjednodušené. Každý model zachytáva len isté aspekty javu, a aj tie musí relatívne izolovať, t. j. musí ignorovať mnohé jeho vnútorné i vonkajšie vzťahy, ktoré sú zvyčajne implicitne obsiahnuté *ceteris paribus*. Nie je jasné, ako by sa dala komplexita objektov zachytiť presnejšie alebo detailnejšie prirodzeným jazykom, ktorý je sám plný vágnosti, nepresnosti, mnohoznačnosti atď. Nazhromaždenie opisných údajov je len predpokladom pre analýzu, nie samotná analýza. Kvantifikácia a matematizácia nie sú teda ničím iným než spresnením a prehĺbením výskumu exaktnejšími metódami, ktoré umožňujú dedukciu a výstavbu teórie.“[15]

Jako příklad může sloužit jazyk totalitních režimů, který asi většina lidí označí za maximálně zjednodušený a vyprázdněný s chudým slovníkem. Jedním z nejznámějších popisů takového jazyka je Orwellův newspeak v románu *1984*.[16] Pokud bychom tedy v našem prostředí použili novoroční a vánoční projevy československých a českých prezidentů, můžeme předpokládat v případě projevů komunistických prezidentů tendenci k nižšímu slovnímu bohatství. Naopak u demokratických prezidentů, zvláště pak v případě dramatika Václava Havla, vyšší slovní bohatství. Bez opory kvantitativních metod si však jen těžko vytvoříme představu o tom, jaké bude pořadí jednotlivých prezidentů a jak velké rozdíly mezi nimi budou. Experiment provedený v podobě aplikace metody měření slovního bohatství *MATTR*[17] nám umožní nahlédnout na sledovanou problematiku intersubjektivně s přesnou kvantifikací, výsledky najdeme na Obr. 1 a Obr. 2.

[15] Wimmer a kol. (2003), s. 15.
[16] Viz Orwell, G. (2009).
[17] Této metodě se detailně věnuje kapitola 4.1 *Slovní bohatství*.



Obr. 1. Chronologicky seřazené výsledky *MATTR* u prezidentů



Obr. 2. Sestupně seřazené výsledky *MATTR* u prezidentů

Naměřené hodnoty slovního bohatství umožňují naše intuitivní předpoklady a kvalitativní analýzy kvantifikovat a korigovat. V případě prezidentů je evidentní, že slovní bohatství se dokonce chová přesně opačně, než jak bychom očekávali. Právě intersubjektivní náhled na problematiku je klíčovou výhodou kvantitativního přístupu. To však neznamená, že jsou kvalitativní analýzy nepotřebné, aplikaci kvantitativních metod lze považovat za vhodné doplnění tradičně pojatých výzkumů. Samotná kvantifikace však není konečným stádiem analýzy, protože například sice

víme, že Husák má v novoročních projevech vyšší slovní bohatství (0,86) oproti Havlovi (0,81), ale nevíme, zda je tento rozdíl natolik velký, abychom z něj mohli vyvozovat nějaké závěry. K tomu slouží další nezbytný krok každé kvantitativní analýzy, a to statistický test, který za daných podmínek určí, zda je rozdíl mezi hodnotami statisticky významný neboli signifikantní. Výsledkem pak může být tabulka, která zobrazuje výsledky testu, v našem případě hodnoty nad 1,96 znamenají, že na hladině významnosti 0,05 jsou rozdíly signifikantní. V Tab. 1 tak vidíme, že Husák a Havel se statisticky významně liší.

Tab. 1. Výsledky *u*-testu mezi prezidenty (signifikantní $u \geq 1{,}96$, $\alpha = 0{,}05$)

	Masaryk	Beneš	Hácha	Gottwald	Zápotocký	Novotný	Svoboda	Husák	Havel	Klaus
Masaryk	x									
Beneš	1.76	x								
Hácha	**4.11**	**3.32**	x							
Gottwald	**6.80**	**5.78**	1.62	x						
Zápotocký	**16.28**	**11.55**	**2.53**	0.52	x					
Novotný	**42.60**	**17.17**	**2.96**	0.83	0.56	x				
Svoboda	**19.39**	**14.09**	**3.86**	1.94	**2.42**	**2.61**	x			
Husák	**35.29**	**18.84**	**4.21**	**2.20**	**3.27**	**4.55**	0.26	x		
Havel	**7.60**	**5.49**	0.00	**2.06**	**4.01**	**5.31**	**6.09**	**7.33**	x	
Klaus	**9.13**	**5.29**	0.88	**3.17**	**6.73**	**10.19**	**9.30**	**12.46**	1.72	x
Zeman	1.36	0.97	1.00	**2.21**	**2.78**	**3.02**	**3.65**	**3.81**	1.12	0.59

Matematika a statistika se ukazuje jako efektivní nástroj napříč různými obory, proto považujeme za vhodné aplikovat kvantitativní přístup i v lingvistice. Altmann to komentuje takto „The history of all sciences shows that the best conceptual instruments or analytical means are just the mathematical methods. If something is good enough for 'harder' sciences, it cannot be bad for linguistics, even if language can turn out to have its own mathematics."[18] Důležité je však mít vždy na vědomí, že jakkoliv složité matematické výpočty jsou jen nástrojem k poznání jazyka. Teprve správná lingvistická interpretace získaných dat je konečným cílem každého výzkumu.

Kvantitativní výzkum zpravidla sestává z několika kroků, který můžeme znázornit jako uzavřený cyklus, viz Obr. 3. Na počátku stojí nějaká teorie, na základě které

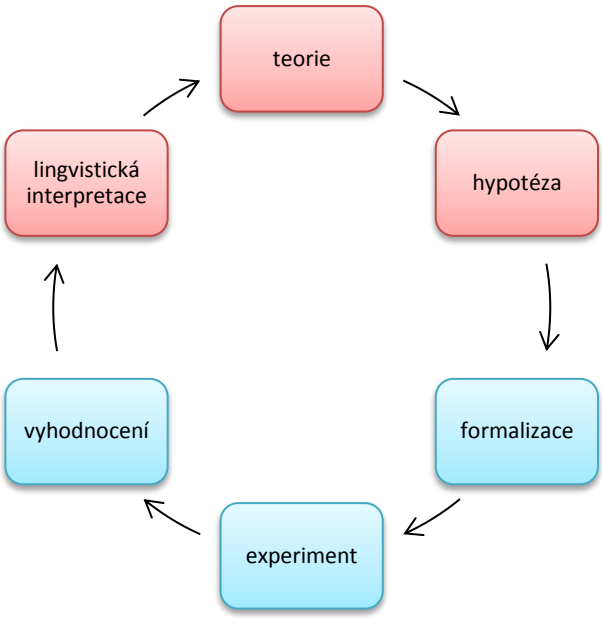
[18] Altmann, G. (1997), s. 13.

dospíváme k určitému předpokladu, který formulujeme do hypotézy.[19] Tuto hypotézu následně formalizujeme, tj. převedeme do řeči čísel. Poté provedeme experiment, jenž má zpravidla podobu výpočtu či měření. Získané výsledky následně podrobíme statistickému vyhodnocení, které za daných podmínek potvrdí, nebo vyvrátí stanovenou hypotézu. Samotným přijetím či zamítnutím hypotézy však výzkum nekončí, je třeba získané výsledky lingvisticky interpretovat a vrátit se zpět k teorii, která stála na počátku. Tato teorie může být na základě výzkumu potvrzena, vyvrácena nebo nějak modifikována, přičemž je třeba si uvědomit, že „No hypothesis should be definitively rejected or definitively accepted. Corroboration is a matter of degree.“[20] Důležité je, že lingvistika stojí jak na začátku, tak na konci výzkumu, experiment je jen nástrojem, který nám umožňuje objektivně nebo spíše intersubjektivně zkoumat jazyk. Pro detailnější informace odkazujeme na bohatou literaturu.[21]

[19] Je třeba však poznamenat, že ne vždy stojí na počátku nějaká teorie, hypotézy bývají inspirovány i jinými podněty, což nelze považovat za chybu. Formulace hypotézy na základě určité teorie je jen „ideální“ způsob, ostatně jako celý uvedený cyklus.

[20] Stauss, U. a kol. (2008), s. III.

[21] Např.:
Altmann, G. (1997).
Altmann, G. (2006).
Altmann, G. (2012).
Köhler, R., Altmann, G. (2005).
Köhler, R., Altmann, G., (2011).
Benešová, M. (2011).
Čech, R., Popescu, I. I., Altmann, G. (2014).



Obr. 3. Zjednodušený cyklus zobrazující jednotlivé kroky výzkumu

2.2. Současná stylometrie

Stylometrie měla vždy své pevné místo v kvantitativní lingvistice a nejinak je tomu i dnes.[22] Nicméně je třeba poznamenat, že většina badatelů soustředí svůj zájem primárně na problematiku určování autorství a teprve sekundárně na žánrovou klasifikaci. Spíše než detailními žánrovými analýzami, jejichž hlavním cílem jsou lingvistické interpretace dat, se častěji setkáváme s předložením výsledků v podobě číselných hodnot, které ukazují míru efektivity dané statistické metody pro automatickou diferenciaci textů. Obecně lze konstatovat, že prudký vývoj počítačových technologií umožnil v posledních letech značný posun stylometrických analýz, které vynikají propracovanými statistickými metodami a precizními vizualizacemi v podobě různých grafů, dendrogramů a sítí. Na druhou stranu lze ale také sledovat jisté odchýlení těchto výzkumů od původního cíle, ke kterému by měly směřovat, a to k závěrečné jazykovědné interpretaci získaných dat, jež ústí k vědecké explanaci.

V této kapitole chceme představit nejdůležitější osobnosti současné stylometrie. Protože je tento obor značně závislý na dostupnosti potřebných nástrojů, tj. softwaru,

[22] Srov. Juola, P. (2006).

jsou někdy badatelé nuceni před jednotlivými výzkumy vytvořit vhodný program. Vzhledem k tomu, jak důležité postavení některé programy mají, pokládáme za vhodné je alespoň stručně představit, protože kvalitní software nezřídka znamená pro stylometrii minimálně stejný význam jako samotný výzkum daného lingvisty.

Důležitým jménem v současném kvantitativním stylometrickém bádání zaujímá řecký lingvista George Mikros, jehož zájem se soustředí zejména na automatické určování autorství,[23] kde dosahuje vysoké přesnosti predikce přesahující i 90% hranici.[24] Mikros s obdobnou úspěšností však také aplikoval své metody na klasifikaci textů na základě pohlaví.[25] Základním kamenem analýz je kombinace několika stylových charakteristik, mezi něž patří například bohatství slovníku, poměr synsémantik k autosémantikům, entropie, průměrná délka slova a věty nebo frekvence slovních druhů. Vedle těchto tradičních ukazatelů však Mikros nejvíce používá grafémových a slovních n-gramů (konktrétně bigramů a trigramů), jejichž kombinací ve víceúrovňovém autorském profilu *AMNP* (Author's Multilevel N-gram Profile) dosahuje nejlepších výsledků. Z hlediska statistiky jsou data zpracována zejména pomocí metod Random forest (*RF*) a Suport Vector Machine (*SVM*).[26] Mikros zkoumá nejen beletrii, ale také texty z internetového prostředí, kam patří například e-maily nebo tzv. tweety. Mikrosovy práce tak nabízejí cenná data, která pokrývají širokou škálu nejrůznějších stylů převážně řeckých textů.

Neméně důležité postavení v současné stylometrii zaujímá trojice Maciej Eder, Jan Rybicki a Mike Kestemont. Zmínění autoři se zaměřují zejména na textové analýzy literárních textů z období středověku, renesance či baroka,[27] dále také aplikují své metody na analýzu překladů různých děl.[28] Stejně jako v případě

[23] Viz např.:
Mikros, G. K., Perifanos, K. (2011).
Mikros, G. K., Perifanos, K.. (2013).
Mikros, G. K. (2006).
Mikros, G. K. (2007a).
Mikros, G. K. (2007b).
Mikros, G. K. (2009).
Mikros, G. K., Argiri, E. K. (2007).
[24] Viz Mikros, G. K., Perifanos, K. (2013).
[25] Viz Mikros, G. K. (2013).
[26] Pro více informací o těchto metodách viz např. Berka, P. (2003).
[27] Viz např.:
Eder, M. (2014).
Eder, M., Rybicki, J. (2009).
[28] Viz např.:

Mikrose i zde je ústředním zájmem určování autorství. Z metodologického hlediska používají zejména obsahovou analýzu, přičemž velký důraz je kladen na vizualizaci získaných dat, a to zejména pomocí sítí.

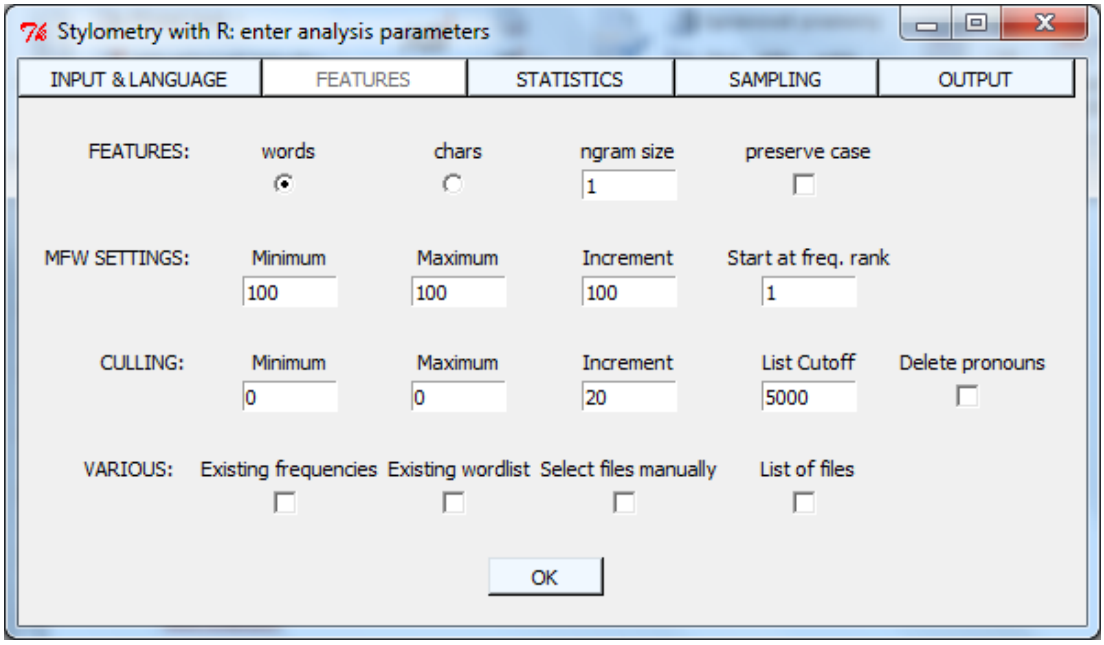
Zřejmě nejvýznamnějším počinem této trojice je však software pro stylometrickou analýzu textů *Stylo*.[29] Ve skutečnosti jde o balíček pro statistický software *R*[30]. Autoři umožňují pracovat jak se standardními příkazovými řádky, tak v grafickém uživatelském rozhraní (GUI). Každý si tak může zvolit prostředí, které mu lépe vyhovuje. Díky implementaci GUI stačí pro práci s programem *Stylo* skutečně jen elementární znalost softwaru *R*. Základními daty pro statistické zpracování jsou slova, grafémy nebo jejich n-gramy. Zpravidla se nastaví jen určité množství jednotek, se kterými se pracuje, typicky 100 nejfrekventovanějších slov. Pro statistické vyhodnocení si může uživatel vybrat z klastrové analýzy, mnohorozměrného škálování (*MDS*), analýzy hlavních komponent (*PCA*) a dalších. Pro detailnější statistické vyhodnocení jsou dostupné další nadstavbové balíčky. Výsledky jsou exportovány jako .txt soubory a grafy jako .pdf, .png, .jpg nebo .svg soubory. Pokud chce uživatel vytvořit sítě, je třeba použít další nástroj (např. *Gephi*[31]), který zpracuje získaná data do požadované podoby. Screenshot grafického uživatelského rozhraní je k nahlédnutí na Obr. 4, ukázka výstupu ve formě dendrogramu pak na Obr. 5.

Rybicki, J., Heydel, M. (2013).
Heydel, M., Rybicki J. (2012).
Rybicki, J. (2012).

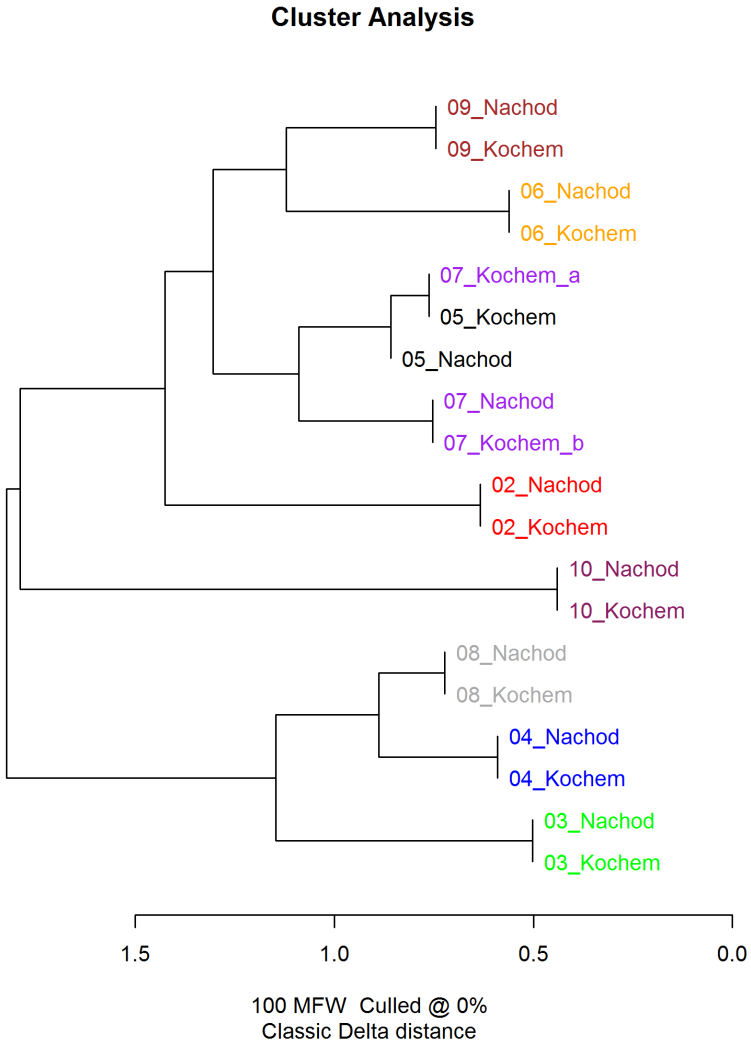
[29] Eder, M., Kestemont, M., Rybicki, J. (2013).

[30] R Core Team (2013).

[31] *Gephi* je software s licencí GNU volně stažitelný na https://gephi.github.io/ [cit. 25. 3. 2015]



Obr. 4. Grafické uživatelské rozhraní programu *Stylo*

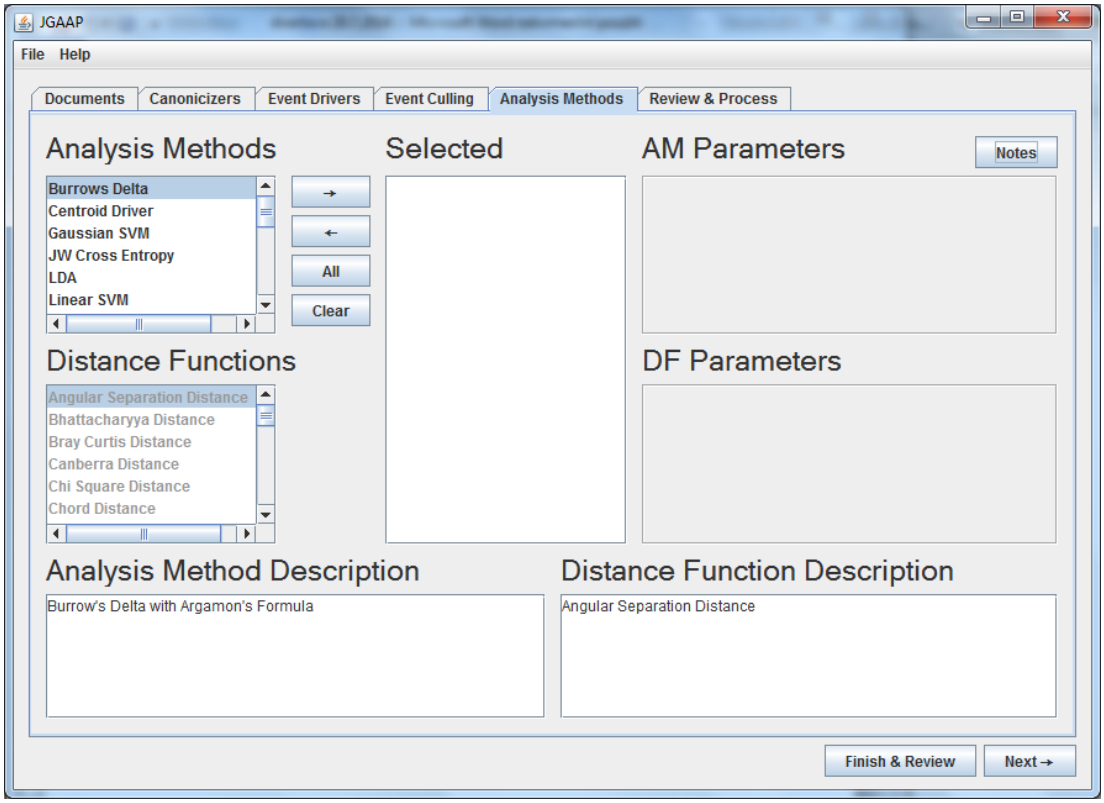


Obr. 5. Ukázka grafického výstupu programu *Stylo* ve formě dendrogramu

Mezi nejvýznamnější kvantitativní lingvisty v současném stylometrickém bádání patří Patrick Juola. Tento americký badatel se zaměřuje zejména na určování autorství. Jeho základním dílem v daném oboru je publikace *Authorship Attribution.*[32] Juola se svým týmem na univerzitě v Pittsburghu vytvořil program pro určování autorství a textovou klasifikaci Java Graphical Authorship Attribution Program (*JGAAP*), který poskytuje jednoduché uživatelské prostředí (viz Obr. 6) pro všechny zájemce bez hlubších znalostí dané problematiky. *JGAAP* vyhodnocuje rozdíly mezi texty zejména na základě tradičních ukazatelů, jako jsou frekvence grafému, slov, n-gramů, délek vět apod. Následné statistické vyhodnocení nabízí pestrou škálu možností (Support Vector Machine, lineární diskriminační analýzu a

[32] Joula, P. (2008).

další). Joula klade značný důraz na aplikaci stylometrie v praxi, tedy na řešení konkrétních reálných případů.[33]



Obr. 6. Grafické uživatelské rozhraní programu *JGAAP*

V současném českém jazykovědném prostředí se kvantitativní analýze textů systematicky věnuje Radek Čech, jehož výzkum se soustředí zejména na oblast indexů slovního bohatství, tematickou koncentraci textu a syntax. Z hlediska našeho zájmu patří k nejdůležitějším autorovým publikačním počinům *Metody kvantitativní analýzy (nejen) básnických textů*[34]. Tato kniha poskytuje českému čtenáři nejen teoretický úvod do světa kvantitativní analýzy textu, ale také představení několika indexů s konkrétními příklady jejich užití. Posledním obdobným titulem publikovaným v češtině byla *Kvantitativní lingvistika*[35] již zmíněné Těšitelové vydaná před více než 25 lety. Pod autorovým vedením vznikl také projekt

[33] Joula je zakladatelem a ředitelem společnosti Juola & Associates, která komerčně poskytuje služby spojené se stylomerií (určování autorství pro forenzní účely, plagiátorství, vypracování autorského profilu apod.).

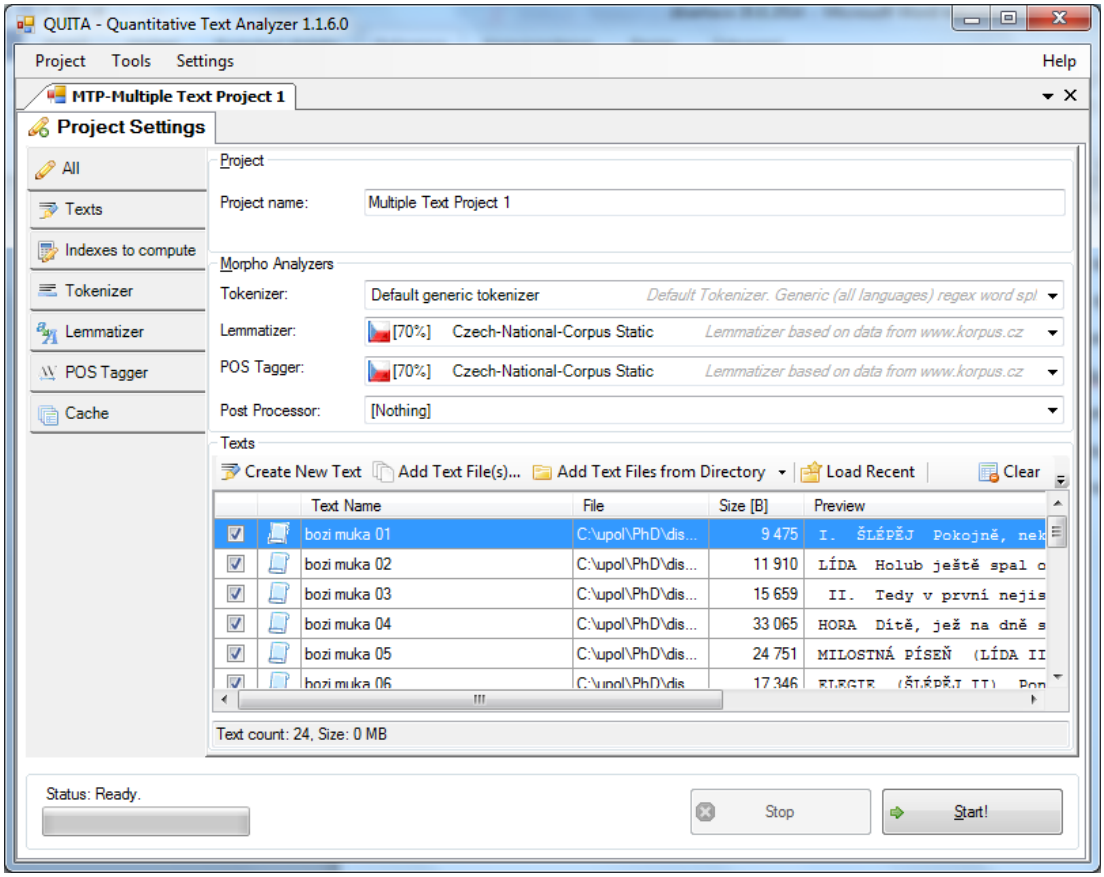
[34] Čech, R., Popescu, I. I., Altmann, G. (2014).

[35] Těšitelová, M. (1987).

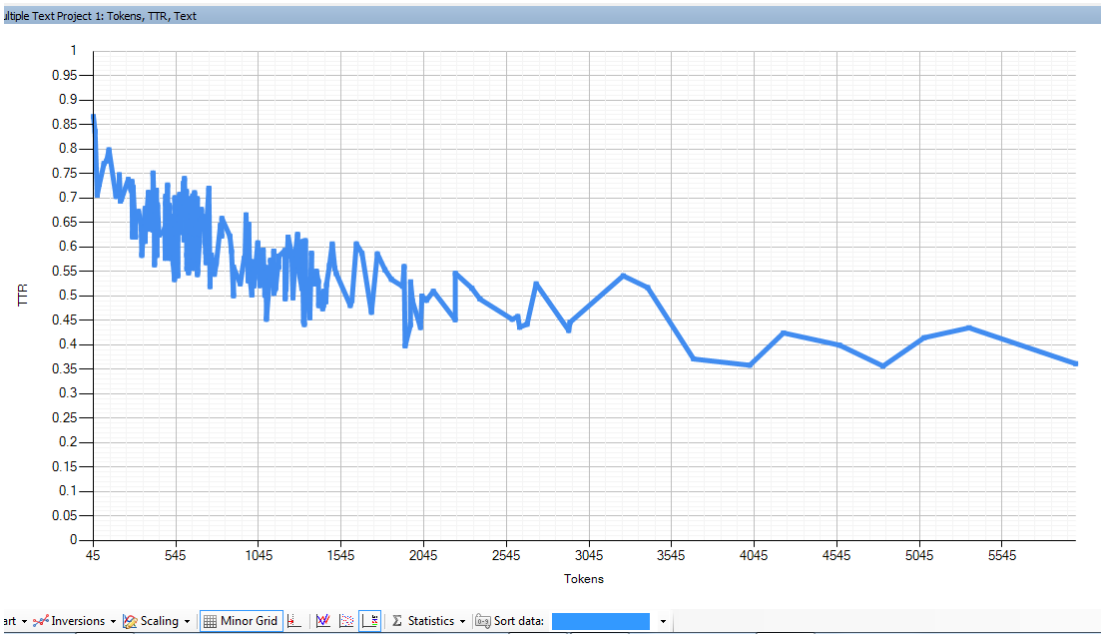
Quantitative Index Text Analyzer (*QUITA*), jehož primárním výstupem je software pro kvantitativní analýzu textů *QUITA*.[36]

Quantitative Index Text analyzer (*QUITA*) je univerzální nástroj s přehledným intuitivním uživatelským prostředím, který je určen pro nejširší spektrum uživatelů. Tomu odpovídá také výběr základních kvantitativních indikátorů, jako jsou různé indexy slovního bohatství, tematická koncentrace, aktivita a deskriptivita, průměrná délka tokenu apod. *QUITA* také umožňuje provádět základní operace s textem, kam patří zejména tvorba frekvenčních slovníků, lemmatizace textu nebo rozlišování slovních druhů. Získaná data lze jednoduše statisticky testovat a také vytvořit požadované grafy či tabulky, které lze exportovat. Podrobné informace o programu, včetně popisů všech indexů s konkrétními příklady, lze najít v knize *QUITA – Quantitative Index Text Analyzer*[37] a v diplomové práci *Kvantitativně lingvistický software.*[38]

[36] Matlach, V., Kubát, M., Čech, R. (2014).
[37] Kubát, M., Matlach, V., Čech, R. (2014).
[38] Matlach, V. (2014).



Obr. 7. Grafické uživatelské rozhraní programu *QUITA*



Obr. 8. Ukázka vizualizace výsledků ve formě grafu v programu *QUITA*

V našem přehledu nejdůležitějších jmen současné stylometrie nelze opomenout Gabriela Altmanna, původem orientalistu, a Ioana Iovitze Popesca, jednoho z nejvýznamnějších rumunských fyziků plazmatu, který se v současnosti zabývá také matematickou lingvistikou. Zatímco Altmann, považovaný za zakladatele moderní kvantitativní lingvistiky, přinesl rozsáhlé dílo, jež zasahuje snad do všech odvětví jazykovědy, Popescu je autorem několika indexů, jimiž lze charakterizovat různé vlastnosti textu.[39] Mezi nejvýznamnější indexy patří například index slovního bohatství R_1, index frekvenční struktury textu lambda (Λ) nebo tematická koncentrace textu (*TC*). Důležitým nástrojem pro kvantitativní analýzu textů je také Altmanův software Altmann-Fitter[40], který umožňuje aplikovat distribuce na teoretická rozdělení.

Detailně se kvantitativní analýzou textu dlouhodobě zabývá také italská badatelka Arjuna Tuzzi, jež působí na univerzitě v Padově. Pokud se podíváme detailněji na její specializaci, zjistíme, že se věnuje zejména těmto disciplínám: obsahová analýza, klastrování textů, text mining, určování autorství, statistické vyhodnocovací metody. Z mnoha autorčiných publikací je z hlediska naší práce patrně nejdůležitější kniha *Quantitative Analysis of Italian texts*[41], kde analyzuje italské texty pomocí několika stylometrických metod včetně slovního bohatství či tematické koncentrace. Tuzzi je také zakládající členkou italského sdružení *GIAT* (Gruppo Interdisciplinare di Analisi Textuale), jež se zabývá textovou analýzou z různých interdisciplinárních pohledů, kterému dominuje kvantitativní přístup. Výsledkem práce tohoto sdružení je kromě množství publikací také vytváření softwaru pro automatickou analýzu textu. Jednotlivé programy včetně manuálů lze najít na webu *GIAT*.[42]

[39] Např.:
Popescu, I. I., Altmann, G. (2011).
Popescu, I. I., Altmann, G. (2007).
Popescu, I. I., Altmann, G. (2011).
Popescu, I. I. a kol. (2009).
Popescu, I. I. a kol. (2010).
Popescu, I. I., Čech, R., Altmann, G. (2011).
Popescu, I. I., Čech, R., Altmann, G. (2012).
Popescu, I. I., Mačutek, J., Altmann, G. (2009).
[40] Altmann, G. Altmann–Fitter (software).
[41] Tuzzi, A., Popescu, I. I., Altmann, G. (2010b).
[42] http://www.giat.org/

3. Metodologické aspekty

3.1. Metodologická východiska

Než přejdeme k samotné analýze a představení jednotlivých metod, považujeme za nezbytné vymezit elementární metodologická východiska této práce.

V českém jazykovědném prostředí navazujeme zejména na průkopnici lexikální statistiky Marii Těšitelovou, jejíž dodnes cenné publikace spadají do 70. a 80. let 20. stol.[43] Tato disertace tak do značné míry aktualizuje metody Těšitelové a přináší do českého kontextu nové poznatky kvantitativní analýzy textu, jež zaznamenala v posledních letech zejména díky rozvoji počítačových technologií značný pokrok.

Vzhledem k tomu, že práce je věnována textové analýze žánrů, je nezbytné vymezit zdánlivě jednoduché pojmy „text" a „žánr". V mnoha případech se setkáváme s pojetím, jež za text považuje např. román, sbírku povídek, cestopis, sbírku pohádek či básní.[44] Vůči tomuto pojetí bychom se chtěli jasně vymezit a uvést naše chápání výrazu „text", přičemž vycházíme z tradičních výkladů, které definují text jako „obsahově i formálně relativně celistvý, uzavřený, spojitý útvar znakové povahy […]," [45] nebo „Jazykový projev, komplexní, uspořádaná promluva, psaná i mluvená. Různí autoři text definují různě. Vedle komplexnosti, vnitřní uspořádanosti a organizovanosti se jako význačné rysy uvádějí koherence (spojitost), tematická a funkční jednota (informativnost, intence), relativní uzavřenost, ohraničenost apod. […]." [46] Pokud má být text celistvou homogenní jednotkou, považujeme za takřka nemožné pracovat např. s celým románem jako s jedním textem. Schopnost autora psát texty o desítkách či stovkách tisíc slov jako jednu koherentní jednotku shledáváme jako nepravděpodobnou. Pokud navíc přihlédneme k faktu, že mnozí spisovatelé píší svá díla i roky, je zřejmé, že takové knihy jen stěží mohou sloužit k textové analýze našeho druhu. Také celé sbírky povídek či pohádek

[43] Např.:
Tešitelová, M. (1972).
Těšitelová, M. (1974).
Těšitelová, M. (1987).
Těšitelová, M. (1983).
Těšitelová, M. a kol. (1987).
[44] Viz např. ČNK.
[45] Lotko, E. (2005), s. 117.
[46] Nekula, M. (2002b), s. 489.

nelze z našeho pohledu vnímat jako jeden text, což je ještě zřejmější než u románu. Wimmer a kol. k tomu poznamenávají: „Ak chceme overovať nejaký zákon, tak nesmieme miešať texty, lebo v každom texte sú iné tzv. počiatočné podmienky. Dokonca je niekedy potrebné analyzovať oddelene aj jednotlivé kapitoly románu nebo symfónie.“[47]

Je vhodné také zmínit problematiku autorství textu, protože předpoklad jediného podavatele je klíčový. Směs různých textů, které byly následně někým spojeny do jednoho celku, nám jen stěží může sloužit jako vhodný materiál pro jazykovědný výzkum. Bohužel je velmi problematické, ba nemožné, získat texty, které by splňovaly všechny požadavky lingvistů, tj. zejména: jediný autor, minimum pozdějších zásahů, napsání textu najednou bez delších přerušení. Abychom demonstrovali, jak složité je získání textů, které by splňovaly všechny uvedené požadavky, můžeme blíže nahlédnout do struktury novinových článků, které se z různých důvodů považují za vhodný materiál pro zkoumání jazyka.[48] Pokud se však podíváme z hlediska autorství na novinové články blíže, zjistíme, že i na tvorbě těch nejkratších textů se podílí několik lidí. Výsledné články tak jsou zpravidla vždy mozaikou několika různých textů, které se postupně slepují dohromady. Spíše než o novináři jakožto autorovi článku je vhodnější mluvit o institucionálním podavateli. Důležitým specifikem publicistických textů je také vkládání různých vyjádření v podobě přímých citací, kdy do textu vstupuje mnohdy i několik dalších subjektů v rámci jednoho článku. V Tab. 2 je znázorněn obecný rámec vytváření novinového článku. Toto schéma je obecné, konkrétní realizace může být různě modifikována, ve většině případů však těmito procesy prochází vytváření většiny článků, zvláště pak v případě velkých vydavatelských domů.

[47] Wimmer, G. a kol. (2003), s. 21.

[48] To dokládají i tzv. reprezentativní korpusy ČNK, kde publicistické texty například tvoří 33 % korpusu SYN2010 a 60 % korpusu SYN2000. Z hlediska celého ČNK pak i vůbec největší korpus SYN2013PUB (935 000 000 slov) je tvořen právě publicistickými texty.

Tab. 2. Obecný rámec vytváření novinového článku

událost	→	**podpůrné zdroje informací**	→	**redaktor**	→	**korektor**	→	**editor**
tisková zpráva		jiná média						
komerční sdělení		vyjádření účastníků události						
tisková konference		vyjádření odborníků						
skutečná událost		vyjádření tiskových mluvčích						
fiktivní událost		literatura						
událost v jiném médiu		internetové zdroje						
redaktorova intence		rešerše						

Novinový článek nám posloužil jako ukázka toho, jak komplikované je zvolit vhodný materiál pro jazykovědný výzkum. Z tohoto pohledu se zdají být v rámci publicistického stylu Čapkovy sloupky téměř ideální, neboť je zřejmé, že oproti většině dnešních novinových článku tyto texty psal jen samotný Čapek bez větších zásahů dalších osob.

Jednotlivá díla Karla Čapka jsme tak dle výše uvedeného rozdělili na texty následovně. Za jeden text v této práci považujeme:

- kapitolu románu, cestopisu nebo odborné studie,
- jednotlivou povídku, pohádku, báseň, dopis či novinový sloupek.

Ačkoliv považujeme výše uvedené pojetí textu pro naši analýzu za vhodné, je třeba říct, že jsme si vědomi jistých problémů při komparaci žánrů, kde v jedné rovině stojí kapitola románu, povídka, dopis či novinový sloupek. Jakkoliv takové rozdělení nemusí být ideální,[49] jsme přesvědčeni, že v kontextu naší práce je nejméně zavádějící, což je evidentní, postavíme-li vedle sebe například celý román a dopis.

Pokud máme definovat druhý klíčový termín této práce – žánr, musíme se vyrovnat s jistou nejednotností jeho chápání. Na tuto skutečnost upozorňuje také literární teoretik Eduard Petrů: „Ačkoliv by se mohlo předpokládat, že genologická studia přispěla alespoň k upřesnění terminologie užívané v této oblasti, literárněvědná literatura nás přesvědčuje o tom, že zde panuje značná různorodost. Dokonce i základní pojmy toho stupně obecnosti jako *druh* a *žánr* jsou užívány promiskue

[49] Otázka je, zda vůbec nějaké ideální pojetí textu může existovat.

[…].“[50] Naše pojetí žánrů je v souladu s Petrů, který tento termín definuje takto: „Jako literární žánr označujeme ty literární útvary, které se realizují uvnitř literárních druhů (epos, komedie, hymnus apod.), s vědomím, že tyto literární žánry jsou dále vnitřně diferencovány na žánrové varianty (například milostný román, historický román, sociální román atd.) a využívají různých žánrových forem.“[51]

V této práci budeme konkrétně analyzovat osm žánrů: román, povídku, cestopis, pohádku, studii, sloupek, báseň a dopis.

3.2. Jazykové jednotky

Protože v jednotlivých analýzách této práce budeme pracovat s různými jednotkami, považujeme za vhodné vysvětlit náš přístup k této problematice. V první řadě je nezbytné uvést, že principálně odmítáme dělení jednotlivých jednotek na správné nebo špatné, popř. přirozené či umělé. Vycházíme z faktu, že všechny známé jednotky jsou z principu pouze lingvistické konstrukce, které nám umožňují nějakým způsobem uvažovat o jazyce. Je třeba jasné říct, že stanovení konkrétní jednotky pro každou analýzu vychází výhradně z cíle daného výzkumu. Proto může být např. lemma v případě některých jazyků a některých charakteristik (např. tematická koncentrace textu) vhodnější než slovní tvar. Neznamená to však, že takový výběr je absolutní a neměnný, vždy záleží na rozhodnutí daného badatele, který musí zvolit jednotku pro svůj výzkum.

Právě uvedený příklad lemmatu a slovního tvaru je asi nejčastějším problémem, na který lingvisté narážejí. Pokud zůstaneme u češtiny, nelze přehlédnout fakt, že jde o výrazně flexivní jazyk, a tudíž se zdá logické pracovat pouze s lemmatizovanými texty. Jakkoliv je takový předpoklad pravdivý, stejně tak pravdivá je skutečnost, že i volba konkrétního slovního tvaru vykazuje určité charakteristiky textu a není náhodná.[52] Dogmatické přijímání či odmítání jednotlivých jednotek v konkrétních analýzách tak považujeme za nesprávné. Domníváme se, že debata může být vedena pouze o míře vhodnosti té či oné jednotky, nikoliv však o jednoznačném odmítnutí.

[50] Petrů, E. (2006), s. 71.
[51] Tamtéž.
[52] Srov. např. Čech, R., Kelih, E., Mačutek, J. (2014).

V souvislosti s výběrem jednotek v textové analýze nelze opomenout fakt, že z hlediska tradiční lingvistiky se v kvantitativní lingvistice někdy používají jednotky (např. n-gramy, délkové motivy, hreby), které nemají oporu v klasických jazykovědných popisech. Jak vyplývá z výše uvedeného, náš postoj je takový, že tyto nové jednotky nelze odmítat jen proto, že nemají oporu v tradiční jazykovědě, a to ze dvou důvodů: jednak nelze vyloučit, že taková teorie se časem v lingvistice uplatní, jednak nelze přehlížet výsledky, které ukazují funkčnost těchto jednotek v daných oblastech.

Ať už se rozhodneme použít v analýze jakoukoliv jednotku, platí, že nikdy daná volba nebude ideální. Každý výběr bude vždy zákonitě z nějakého hlediska špatný. S tímto poznatkem také přistupujeme ke všem analýzám v této práci.

3.3. Korpus

Základem jakéhokoliv seriózního výzkumu je výběr vhodného vzorku zkoumaného materiálu, a to jak z hlediska kvantitativního, tak i kvalitativního. Nejdůležitějším kritériem pro sestavení výběrového souboru je zaměření a cíl celé práce. V případě jazyka zkoumáme texty, které jsou snadno dosažitelné zejména prostřednictvím různých databází, zejména pak velkých národních korpusů. Český národní korpus (ČNK) v současnosti umožňuje pracovat s materiálem o velikosti stovek milionů slov. Z kvantitativního hlediska jsou tedy možnosti značné, mnohem problematičtější proto bývá hledisko kvalitativní.

Použití ČNK či podobné databáze má pro žánrovou analýzu značná úskalí. Je třeba si uvědomit, že jazykový styl konkrétního textu odráží nejen vliv žánru, ale zejména také vliv samotného autora a dalších faktorů. Právě různé autorské styly mohou při žánrové analýze způsobit značné problémy. Pokud zahrneme do výběrového souboru texty různých autorů, je nemožné následně seriózně interpretovat výsledná data, neboť nemůžeme vědět, zda výsledky vypovídají více o autorství, nebo o žánrových charakteristikách.[53] Vypovídací hodnota takových výzkumů je přinejmenším diskutabilní. Jan Chromý k tomu poznamenává: „[...]

[53] Srov. Králík, J. (2013).

takzvané reprezentativní korpusy určitého jazyka jsou z technického hlediska nereprezentativní, a tedy nevyužitelné pro některé funkce, které lingvisté od korpusu očekávají. O reprezentativnosti lze mluvit pouze u specializovaných korpusů, které mají jasně ohraničenou populaci. Právě tyto korpusy mohou poskytnout lingvistům solidní oporu pro poznávání toho, jak jazyk ve svých různých formách skutečně funguje.“[54] Máme za to, že pokud skutečně chceme pozorovat vlastnosti textů pouze z hlediska žánru, je nezbytné různorodé autorství zcela eliminovat. Z tohoto důvodu jsme se rozhodli vytvořit korpus textů složený z textů jediného autora, a to konkrétně Karla Čapka. Tento spisovatel byl zvolen čistě z pragmatického hlediska, neboť jen stěží najdeme jiného českého autora, který publikoval tak velké množství textů různých žánrů.

Jakkoliv považujeme za správné omezit výzkum na jediného autora, byť s poměrně velkým korpusem, je třeba zmínit, že stejně jako všechna podobná rozhodnutí i toto přináší určité problémy. V tomto případě musíme veškeré závěry omezit pouze na korpus Karla Čapka, protože nevíme, jaké výsledky by přinesly analýzy textů jiných autorů. Náš korpus však považujeme z hlediska kvantitativního i kvalitativního za relevantní výchozí materiál, na který mohou navázat v budoucnu další studie.

Korpus je rozdělen do osmi žánrů (román, povídka, cestopis, studie, sloupek, pohádka, dopis, báseň). Celkový přehled děl zařazených do našeho výběrového souboru je uveden v Tab. 3. Jak již bylo zmíněno výše, uvedené roztřídění textů odpovídá pojetí Petrů[55].

Tab. 3. Přehled textů zařazených do korpusu Karla Čapka

žánr	dílo
román	Hordubal Krakatit Obyčejný život Povětroň První parta Továrna na absolutno Válka s mloky

[54] Chromý, J. (2014), s. 192.
[55] Viz Petrů, E. (2006).

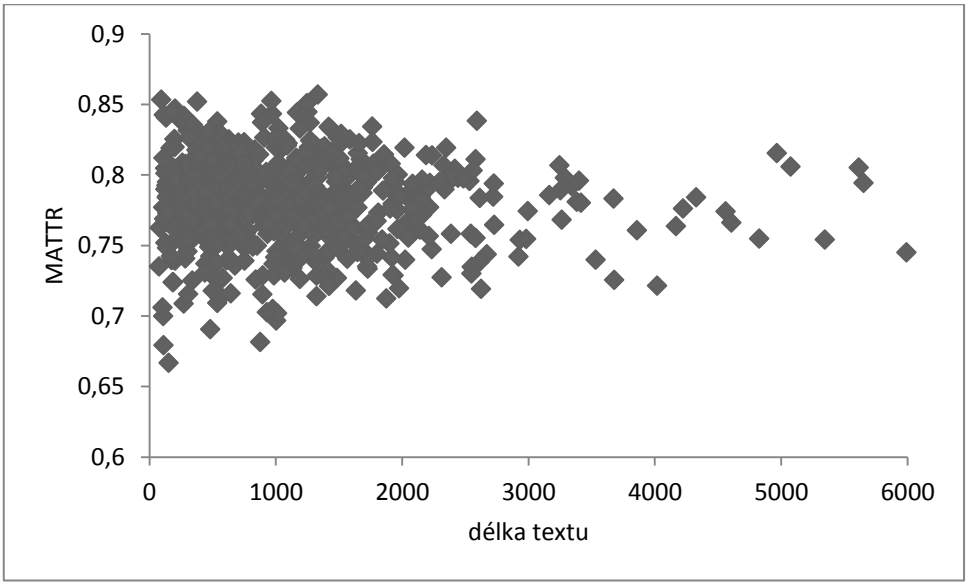
	Život a dílo skladatele Foltýna
povídka	Boží muka Povídky z druhé kapsy Povídky z jedné kapsy Trapné povídky
cestopis	Anglické listy Cesta na sever Italské listy Obrázky z Holandska Výlet do Španěl
studie	Objektivní metoda v estetice se zřením k výtvarnému umění Směry v nejnovější estetice Pragmatismus
sloupek	Jak se co dělá Zahradníkův rok výběr z Lidových novin
pohádka	Dášeňka čili Život štěněte Devatero pohádek
dopis	Anne Nešporové Heleně Čapkové S. K. Neumannovi Olze Scheinpflugové T. G. Masarykovi Věře Hrůzové
báseň	výběr z Lidových novin výběr z týdeníku Nebojsa

3.4. Výběr metod

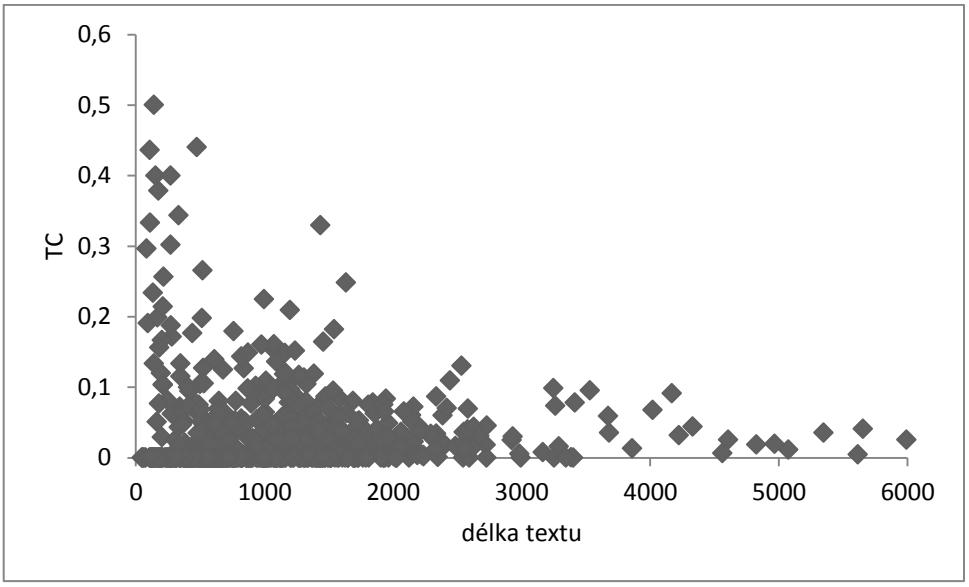
Jak již bylo zmíněno v úvodu této práce, do analýzy jsme vybrali metody dvojího druhu. První skupina představuje indexy, které lze přímo lingvisticky interpretovat; druhá pak takové metody, jejichž využití patří spíše do oblasti určování autorství a automatické klasifikace textů s poměrně komplikovanou lingvistickou interpretací. Pro výběr jednotlivých indexů byla důležitá skutečnost, že mnohé indexy, zvlášť ty,

které mají co do činění s frekvenční strukturou textů (nejčastěji slovní bohatství), jsou závislé na délce textu.

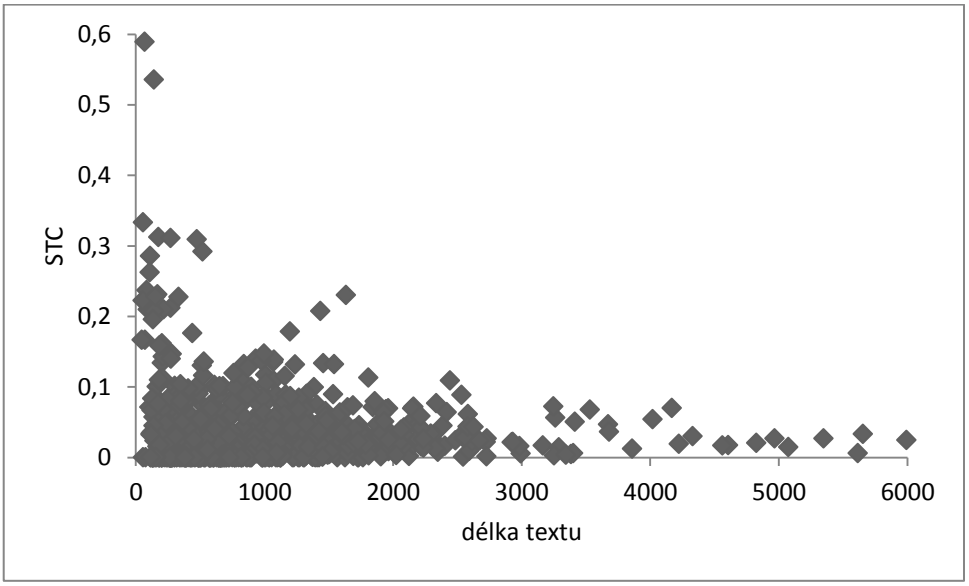
Právě nezávislost na délce textu proto byla klíčovým faktorem pro výběr stylometrických indikátorů. Následující grafy přehledně ukazují, jakým způsobem se výsledné hodnoty zvolených indexů chovají vzhledem k délce textu. Pro výpočet bylo použito vždy 760 stejných Čapkových textů. V této práci proto použijeme následující metody: slovní bohatství (*MATTR*), tematickou koncentraci textu (*TC*, *STC*, *PTC*), vzdálenosti sloves (*VD*), průměrnou délku tokenu (*ATL*), aktivitu (*Q*) a distribuci slovních druhů.



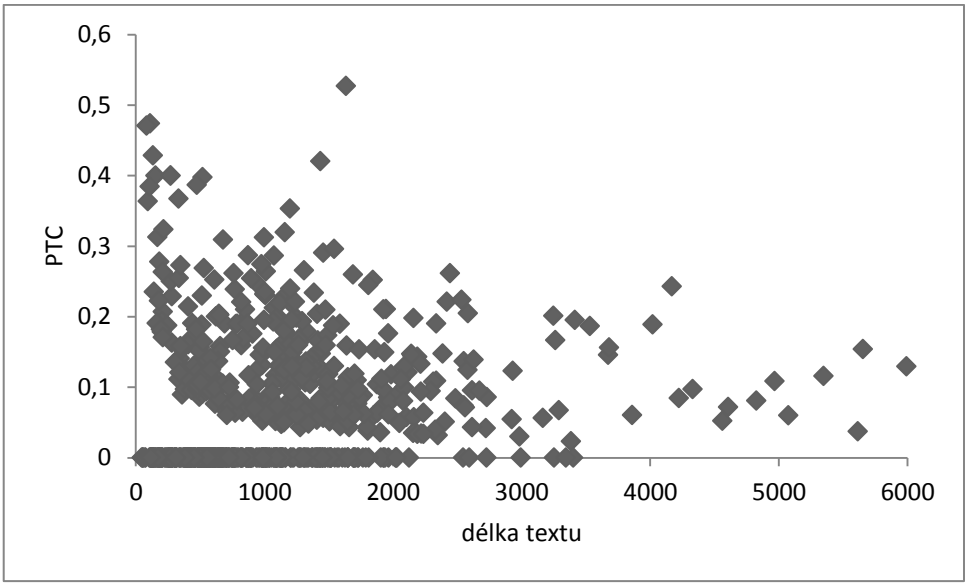
Obr. 9. Závislost *MATTR* na délce textu v 760 Čapkových textech



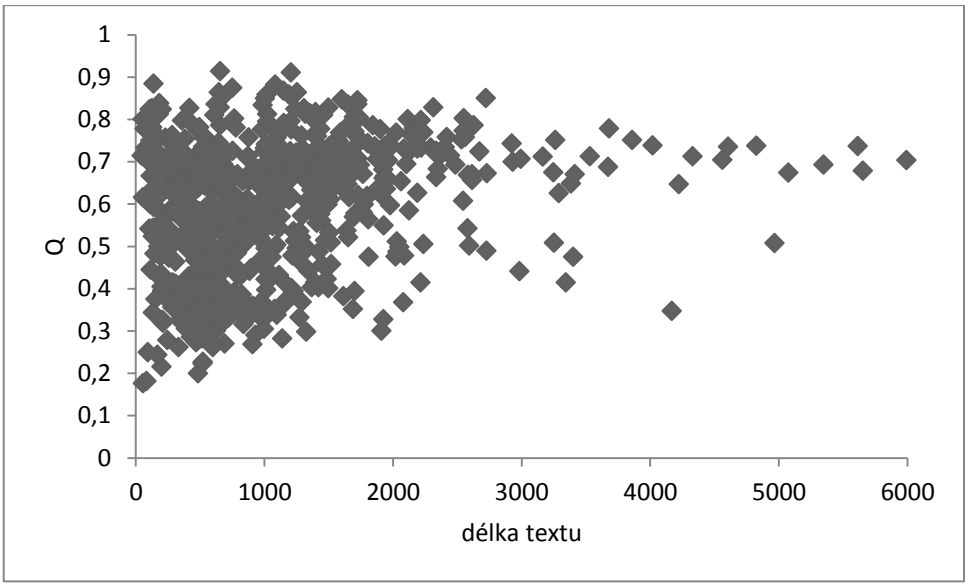
Obr. 10. Závislost *TC* na délce textu v 760 Čapkových textech



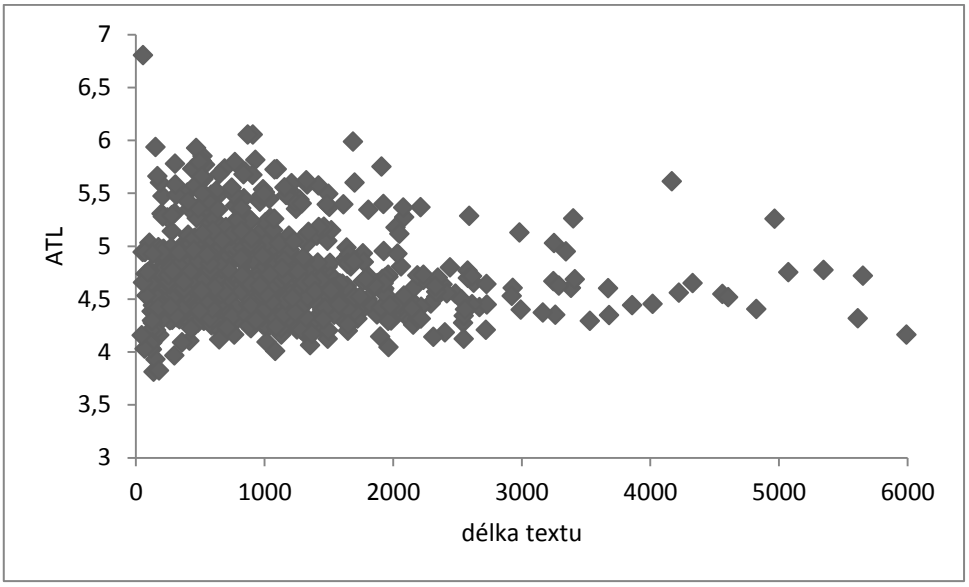
Obr. 11. Závislost *STC* na délce textu v 760 Čapkových textech



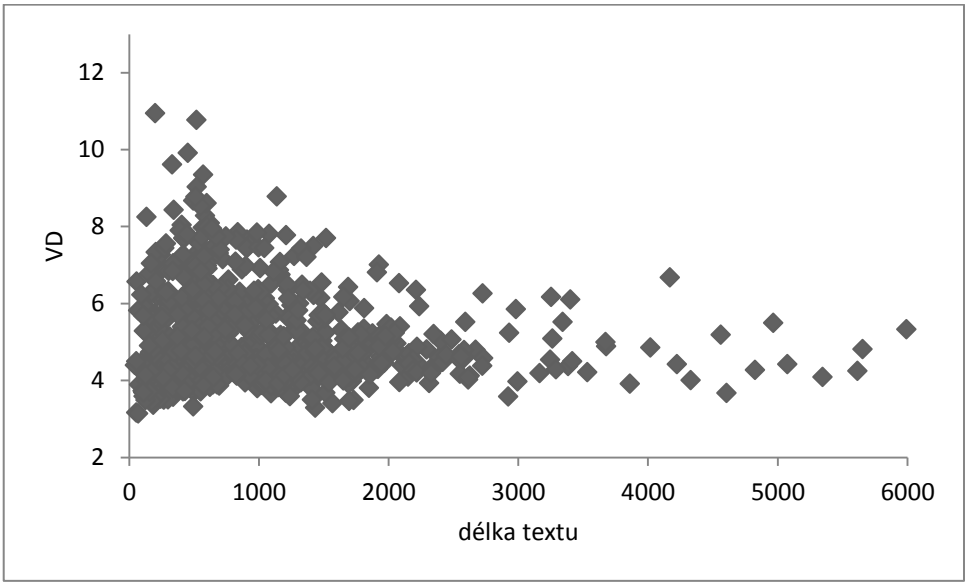
Obr. 12. Závislost *PTC* na délce textu v 760 Čapkových textech



Obr. 13. Závislost aktivity na délce textu v 760 Čapkových textech



Obr. 14. Závislost *ATL* na délce textu v 760 Čapkových textech



Obr. 15. Závislost *VD* na délce textu v 760 Čapkových textech

Druhou skupinu tvoří metody, které se primárně používají k automatické klasifikaci textů, zejména pak v oblasti určování autorství. Tyto metody na rozdíl od výše uvedených mají velmi komplikovanou lingvistickou interpretaci, na druhou stranu však dosahují poměrně dobrých výsledků v klasifikaci textů. V naší práci použijeme dvě takové metody, a to *AMNP* a *MFW* analýzu.[56]

[56] Detailní informace ke všem metodám jsou uvedeny v následujících kapitolách.

4. Žánrová analýza

4.1. Slovní bohatství

Bohatství slovníku patří mezi tradiční oblasti zájmu kvantitativní lingvistiky, první pokusy o měření této charakteristiky sahají k samotným počátkům celého oboru, za zakladatele soustavného studia této oblasti je považován George Udny Yule, jehož kniha *The statistical study of literary vocabulary*[57] patří k základním dílům zkoumání slovního bohatství jako stylometrického nástroje. Neutuchající zájem o tento koncept trvající již desítky let pramení zejména ze závažného nedostatku měření slovního bohatství, a to závislosti na délce textu (viz dále). Badatelé proto hledali a stále hledají takový způsob měření, který by dokázal tento nedostatek eliminovat.[58]

Koncept slovního bohatství vychází z předpokladu, že každý člověk má individuální slovní zásobu, která se odráží v produkovaných textech. Pro úspěšnou komunikaci je nezbytné, aby mluvčí vždy respektoval komunikační kompetenci adresáta, tedy i jeho slovník. Každý tak může produkovat texty s různou slovní zásobou, a to v závislosti na adresátovi nebo na jiných faktorech (zvlášť v případě umělecké literatury), kam patří např. téma, záměr, čas, kontext apod.

Závislost indexů slovního bohatství na délce textu je způsobena tím, že slovní zásoba je omezená, a tudíž je nemožné, aby s rostoucí délkou textu mohl úměrně také narůstat počet nových lexikálních jednotek. Krátké texty proto mají z principu vyšší slovní bohatství než dlouhé texty. Tento fakt znemožňuje relevantně porovnávat různě dlouhé texty, což vylučuje použití bohatství slovníku z většiny stylometrických analýz. Proto bylo učiněno mnoho pokusů o eliminaci vlivu délky textu na měření slovního bohatství, což shrnuje Popescu a kol. v knize *The lambda-structure of texts*: „[…] vocabulary richness – whose history is almost an epos describing the battle against the influence of text length N – can be estimated only if

[57] Yule, G.U. (1944).
[58] Za všechny navržené pokusy jmenujme alespoň několik nejvýznamnějších:
Yule, G.U. (1944).
Popescu, I. I., Čech, R., Altmann, G. (2011).
Covington, M. A., McFall J. D. (2010).
Kubát, M., Milička, J. (2013).
Popescu, I. I. a kol.. (2009).

one eliminates the detrimental factor of text length by some transformation. This has been tried in many cases but the solutions did not seem to be satisfactory. The same is the fate of all indicators of other text properties depending on text length. A typical case is the restriction to the frequently used relativized hapax legomena (*V1*/*N*) as an indicator of richness. It has several flaws: (i) hapaxes are not the exclusive indicators of richness; the same holds for dislegomena, etc. (ii) Very short texts can get maximal richness 1, while very long texts in which all words are repeated would have richness 0; texts of intermediate length, say from *N* = 100 to *N* 100000 would still depend on *N*. Hence taking a special class of words is not sufficient. Indicators like *V1/N* do not work correctly for any text length even if it is possible to set up a test for comparison for not too short and not too long texts. But what is too short and too long?"[59]

Vůbec neviditelnější je vliv délky textu na základním způsobu výpočtu slovního bohatství, kterým je type-token ratio (*TTR*), což je jednoduchý poměr počtu typů k počtu tokenů v textu. Než přejdeme k samotnému výpočtu *TTR*, je třeba alespoň stručně vysvětlit distinkci type-token. Termín „type" (v češtině také někdy označovaný jako „typ") je jakákoliv abstraktní jednotka, v kvantitativní lingvistice nejčastěji slovo (slovní tvar nebo lemma), „token" pak představuje jeho konkrétní realizaci. V následující číselné sekvenci 1, 2, 3, 4, 2, 2, 1 se tak nachází 4 typy (1, 2, 3, 4) a 7 tokenů. Pojednání o distinkci type-token v lingvistice podal již koncem 70. let Bohumil Palek[60], novější definici pak můžeme najít například na webu Českého národního korpusu (ČNK)[61] nebo v *Encyklopedickém slovníku češtiny*[62]. V zahraniční literatuře se pak distinkci type-token věnovali např. Wetzel[63], Wimmer[64], Herdan[65] nebo Peierce[66]. V této práci budeme opozici type-token chápat v rámci současné kvantitativní lingvistiky, přičemž základní jednotkou bude zpravidla slovní tvar (označovaný také jako slovoforma či slovní forma).

[59] Popescu, I. I., Čech, R., Altmann, G. (2011), s. 1.
[60] Palek B. (1969).
[61] <http://wiki.korpus.cz/doku.php/pojmy:typ> cit. 16. 10. 2014.
< http://wiki.korpus.cz/doku.php/pojmy:token> cit 16. 10. 2014.
[62] Nekula, M. (2002a).
[63] Wetzel, L. (2006).
Wetzel, L. (2014).
[64] Wimmer, G. (2005).
[65] Herdan, G. (1960)
[66] Peirce, Ch. S. (1958).

K distinkci type-token je třeba alespoň stručně poznamenat, že toto rozlišení sahá do starověké filozofie a souvisí se základním ontologickým sporem, kde proti sobě stály Platónovy ideje jakožto pravdivý svět versus svět stínů a Aristotelův realismus, který jednotlivým jsoucnům přiznával skutečné bytí. Tyto myšlenky později významně ovlivnily středověkou filozofii v tzv. sporu o univerzálie, kde proti sobě stáli realisté (např. Anselm z Cantenbury či Johannes Scotus Eriugena) hlásající, že obecné existuje nezávisle na jednotlivém, a nominalisté (např. Roscellinus, Wilhelm Ockham), podle nichž jsou tzv. univerzálie jen rozumem vytvořené pojmy sloužící k označení jednotlivostí. Nominalismus se stal základním kamenem novověkého empirismu, který umožnil rozvoj vědeckého myšlení. Více o filozofickém aspektu distinkce type-token nalezneme ve nejrůznějších přehledech dějin filozofie.[67]

K poměru type-token (*TTR*) jakožto indikátoru slovního bohatství je třeba poznamenat, že někteří badatelé jej vnímají spíše jako projev tzv. informačního toku (information flow).[68] To znamená, že každý podavatel textu vědomě distribuuje lexikální jednotky v určitém množství, a to zejména na základě předpokládané délky textu. Jinými slovy, má se za to, že ke každému sdělení má podavatel na základě tématu a dalších faktorů k dispozici určité množství slov, která může použít. Pokud tato slova užije v krátkém textu, bude lexikum nutně koncentrovanější a slovní bohatství vyšší. Naopak delší text nutí autora dané lexikum opakovat a více rozprostřít, čímž by se slovní bohatství mělo snížit. Pokud by tento předpoklad byl správný, museli bychom pravděpodobně ve stylometrii zcela rezignovat na měření slovního bohatství jako takového. Na základě měření metodou *MATTR* se však ukazuje, že tento předpoklad je chybný, neboť se neprojevuje žádná závislost bohatství slovníku na délce textu.[69]

Rovnice 1

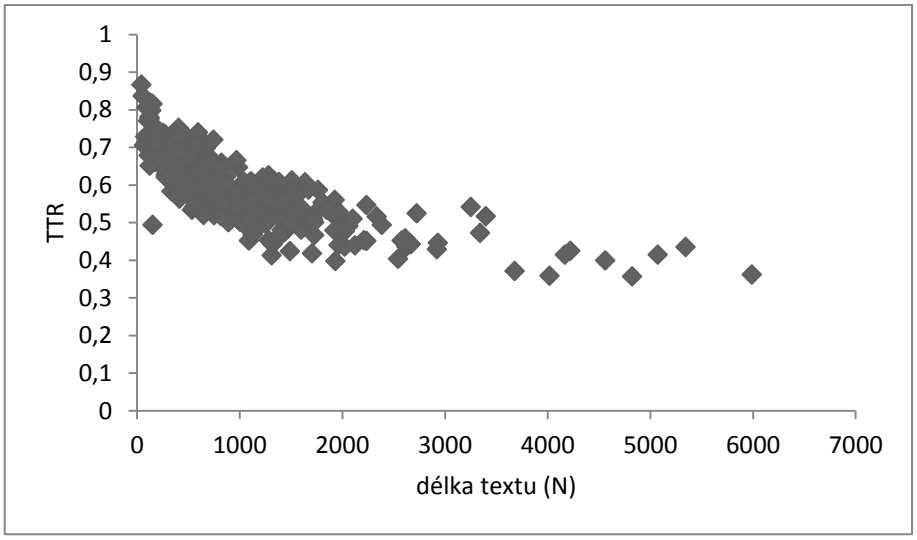
$$TTR = \frac{V}{N}$$

V…počet typů

[67] Např. Blecha, I. (2004).
[68] Viz např. Wimmer, G. (2005).
[69] Viz Obr. 9.

N…počet tokenů

Pro ilustraci závislosti *TTR* na délce textu uvádíme na Obr. 16 hodnoty *TTR* v celém korpusu 760 Čapkových textů.



Obr. 16. *TTR* v 760 textech Karla Čapka

Eliminaci závislosti indexů slovního bohatství na délce textu můžeme v zásadě rozdělit do dvou skupin:

a) Modifikace rovnice pro výpočet slovního bohatství takovým způsobem, aby vliv délky textu byl buď částečně, nebo zcela eliminován.[70]
b) Zásah do textu, a to jeho zkrácením nebo rozdělením do více subtextů o stejné délce.[71]

[70] Viz např.:
Altmann, G., Wimmer, G. (1999).
Guiraud, P. (1954).
Popescu, I. I., Čech, R., Altmann, G. (2011).
Popescu, I. I. a kol. (2009).
Těšitelová, M. (1987).
Tuzzi, A., Popescu, I. I., Altmann, G. (2010b).
Wimmer, G. a kol. (2003).
Yule, G.U. (1944).
[71] Viz např.:
Covington, M. A., McFall J. D. (2010).
Kubát, M. (2013).
Kubát, M., Matlach, V., Čech, R. (2014).
Kubát, M., Milička, J. (2013).
Scott, M. (2013).

Ad a) Je třeba poznamenat, že přestože se tento způsob jeví jako ideální, ani po více než padesáti letech nebyl představen jediný index slovního bohatství, který by byl skutečně zcela nezávislý na délce textu (z posledních pokusů stojí za zmínku zejména index frekvenční struktury lambda[72]). Ačkoliv tedy pravděpodobně musíme vzhledem k dlouhodobému úsilí mnoha badatelů rezignovat na možnost sestavení takového vzorce, který by měřil slovní bohatství bez nežádoucího zkreslení, je nutné uvést, že několik indexů výrazně vliv délky textu redukuje. Proto nelze slovní bohatství a priori vyloučit ze stylometrického výzkumu. Je však nezbytné vždy vybrat takový index, který je vhodný pro konkrétní textovou analýzu. Jednotlivé indexy tak mohou být aplikovány pouze na soubory textů, jejichž délka se pohybuje v určitém intervalu.[73] Spolehlivost jednotlivých indexů pro použití v konkrétním výzkumu je třeba vždy dostatečně předem ověřit. Z výše uvedeného je tedy zřejmé, že vzhledem k velkému rozptylu délky textu v různých žánrech (např. básně vs. povídky, novinové články vs. romány) je aplikace těchto indexů v naší práci zcela vyloučena.

Abychom věděli, jakým způsobem se vzorce pro výpočet slovního bohatství zpravidla upravují, aby se eliminoval vliv délky textu, ukážeme si několik konkrétních příkladů. Asi nejjednodušším zásahem do výpočtu je Guiraudova[74] modifikace *TTR*, kterou do českého prostředí převzala např. Těšitelová[75]. Celá úprava *TTR* spočívá v tom, že se délka textu (*N*) odmocní, a tím se sníží nežádoucí vliv textu.

Rovnice 2

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

V…počet typů

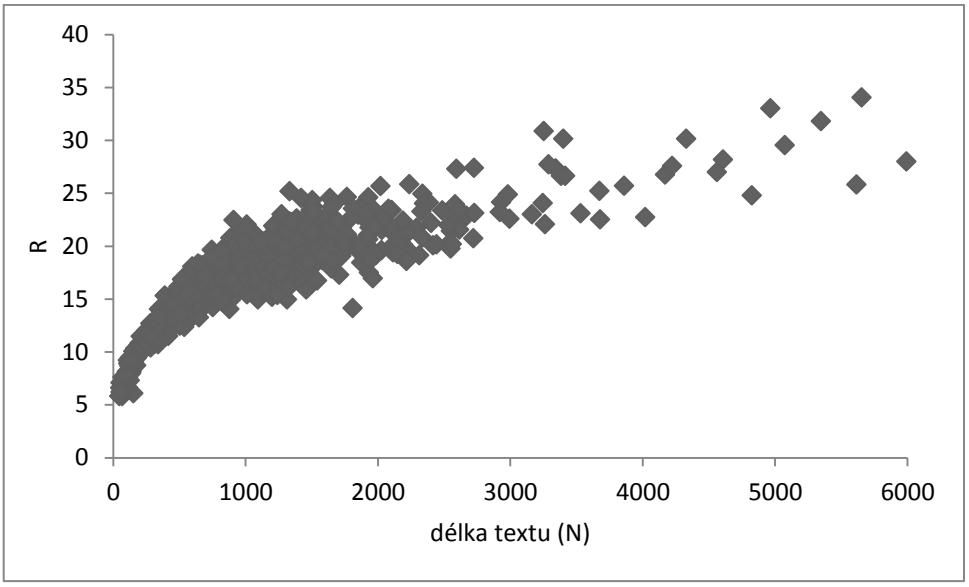
N…počet tokenů

[72] Popescu, I. I., Čech, R., Altmann, G. (2011).

[73] Např. výše zmíněný index frekvenční struktury lambda sice není zcela nezávislý na délce textu, ale v určitém intervalu je použitelný, viz Čech, R. (2015).

[74] Guiraud, P. (1954).

[75] Těšitelová, M. (1987).



Obr. 17. Závislost indexu slovního bohatství *R* na délce textu v 760 Čapkových textech

Dalším prostředkem pro omezení vlivu délky textu bývá použití logaritmu, což můžeme demonstrovat například na indexu lambda, který byl představen jako index frekvenční struktury, který je zcela nezávislý na délce textu.[76] Ani tento pokus však nebyl úspěšný, na což posléze upozornil jeden z autorů.[77]

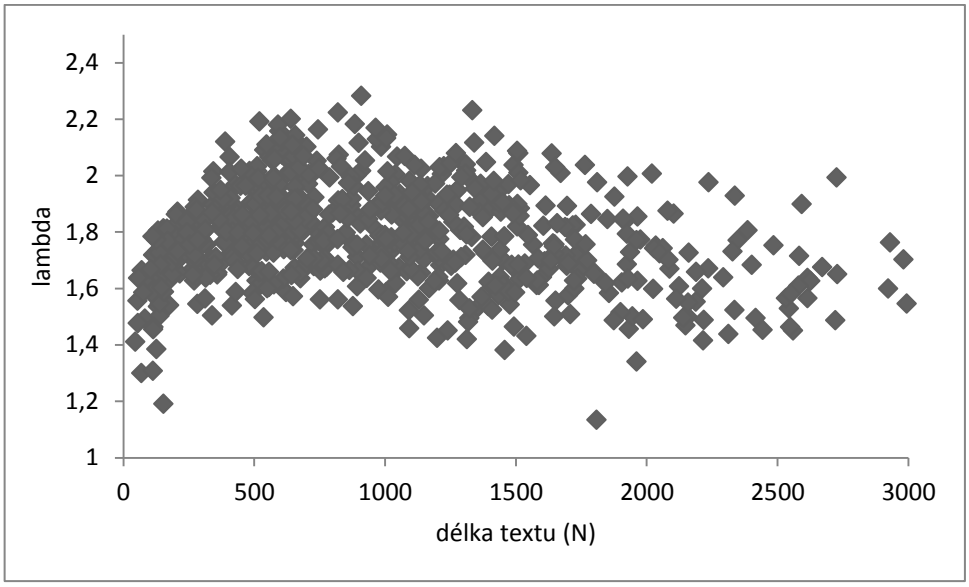
$$\Lambda = \frac{L(log_{10}\, N)}{N}$$

N...délka textu
L...délka křivky distribuce rank-frekvence

$$L = \sum_{i=1}^{V-1} [(f_i - f_{i+1})^2 + 1]^{\frac{1}{2}}$$

f_i…absolutní frekvence
V...nejvyšší pořadí

[76]Popescu, I. I., Čech, R., Altmann, G. (2011).
[77] Čech, R. (2015).



Obr. 18. Závislost lambdy na délce textu v 760 Čapkových textech

Slovní bohatství *R* a lambda mají kromě závislosti na délce textu ještě jeden podstatný nedostatek, a to nejasný interval mezních hodnot, což je v případě jednoduchého *TTR* <0;1>. Pokud tak zobrazujeme výsledky těchto indexů v grafu, interpretace je ovlivněna použitým měřítkem, které je zvoleno zcela arbitrárně.

Ad b) Na jedné straně je zkrácení textů na stejnou délku nejjednodušším opatřením, kterým zcela eliminujeme nežádoucí vliv délky textu, na druhé straně se však z lingvistického hlediska dopouštíme nepřijatelného zásahu do textu. Jestliže zájmem analýzy jsou texty jakožto homogenní uzavřené jednotky, potom zkoumání jejich částí nutně devalvuje výsledné hodnoty a tím i celý výzkum. Další možností je rozdělit texty na více menších částí o stejné délce a jednotlivé dílčí hodnoty použít pro výsledný výpočet aritmetického průměru. Tím dosáhneme pokrytí celých textů a zároveň se vyhneme problému vlivu délky textu. Tato metoda je známa jako standardized type-token ratio (*STTR*) a je použita v softwaru *WordSmith Tools*[78]. Přestože je tato metoda do značné míry průkopnická a víceméně použitelná i pro žánrovou analýzu, její hlavní nedostatek spočívá v tom, že hranice mezi jednotlivými subtexty jsou zcela umělé a nerespektují přirozené celky uvnitř textu. Dvojice amerických vědců Covington a McFall proto navrhla v roce 2010 moving average

[78] Scott, M. (2013).

type-token ratio (*MATTR*),[79] kde se jednotlivá „okna“[80] posouvají vždy jen o jediný token. *MATTR* se tak jeví jako nejspolehlivější způsob měření slovního bohatství, který je zcela nezávislý na délce textu a respektuje text jakožto homogenní celek (viz dále).

Z výše uvedeného je zřejmé, že pro naší žánrovou analýzu můžeme pro měření slovního bohatství použít pouze metody založené na průběžném *TTR*. Rovnice pro výpočet *MATTR* je následující:

Rovnice 3

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

L…arbitrárně zvolená velikost okna v tokenech, $L<N$

N…délka textu v tokenech

V_i…počet typů v jednotlivém okně

Výpočet *TTR* a *MATTR* ukážeme na úryvku básně *Kyjov* od Petra Bezruče. Tento krátký text má 10 tokenů (N) a 6 typů (V). V případě výpočtu *MATTR* zvolíme arbitrárně délku okna na 6 tokenů (L).

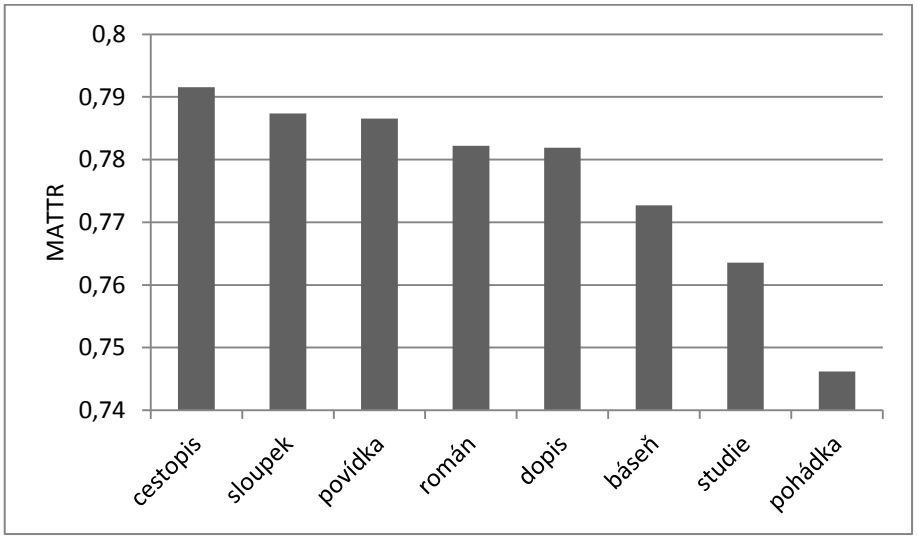
(...)
vždy veselo bývalo v Kyjově,
vždy veselo v Kyjově bude
(...)
(P. Bezruč: Kyjov)

$$TTR = \frac{V}{N} = \frac{6}{10} = 0{,}6$$

$$MATTR(L) = \frac{\sum_{i=1}^{N-L} V_i}{L(N-L+1)} = \frac{5+5+5+4+5}{6(10-6+1)} = \frac{24}{30} = 0{,}8$$

[79] Covington, M. A., McFall J. D. (2010).
[80] Též označovaná jako „windows“, tj. subtexty daného textu o arbitrárně zvolené délce.

Samotný výpočet hodnot *MATTR* v jednotlivých žánrech provedeme pomocí softwaru MaWaTaTaRaD[81], délku okna (*L*) jsme zvolili na 100 tokenů. Výsledné hodnoty uvádíme na Obr. 19.



Obr. 19. Výsledné hodnoty *MATTR* v různých žánrech v Čapkových textech

Ačkoliv lze na Obr. 19 sledovat určité rozdíly mezi jednotlivými žánry, je nutné podrobit výsledky statistickému testu, abychom zjistili, zda jsou zjištěné rozdíly signifikantní. V tomto případě použijeme asymptotický *u*-test, který je ve statistice známý také jako *z*-test. Výsledky jsou zobrazeny v Tab. 4 a pro přehlednost také pomocí sítě na Obr. 20 a Obr. 21. Vztahy mezi žánry jsou dále vizualizovány v Tab. 5 a na Obr. 22.

Rovnice 4

$$u = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{X}_1$, $\bar{X}_2$...aritmetický průměr výsledků každé skupiny

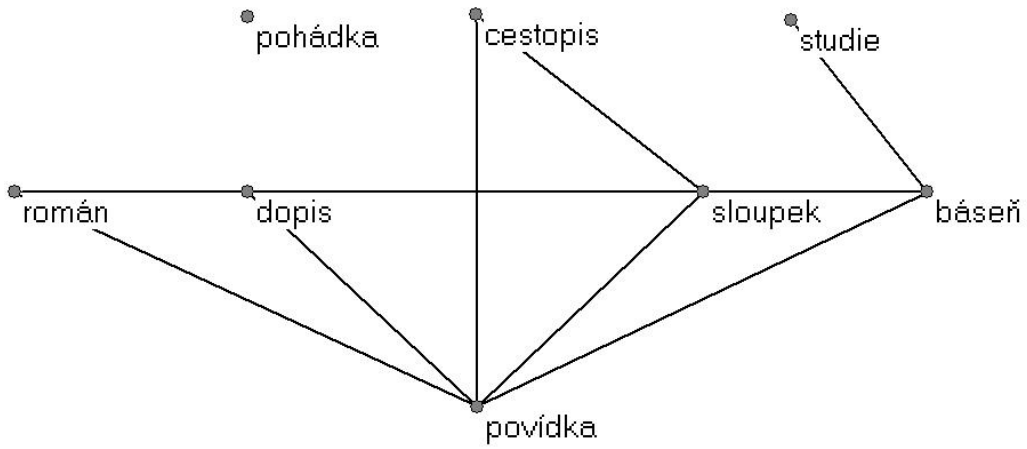
S_1, S_2…standardní odchylka

[81] Milička, J. (2013).

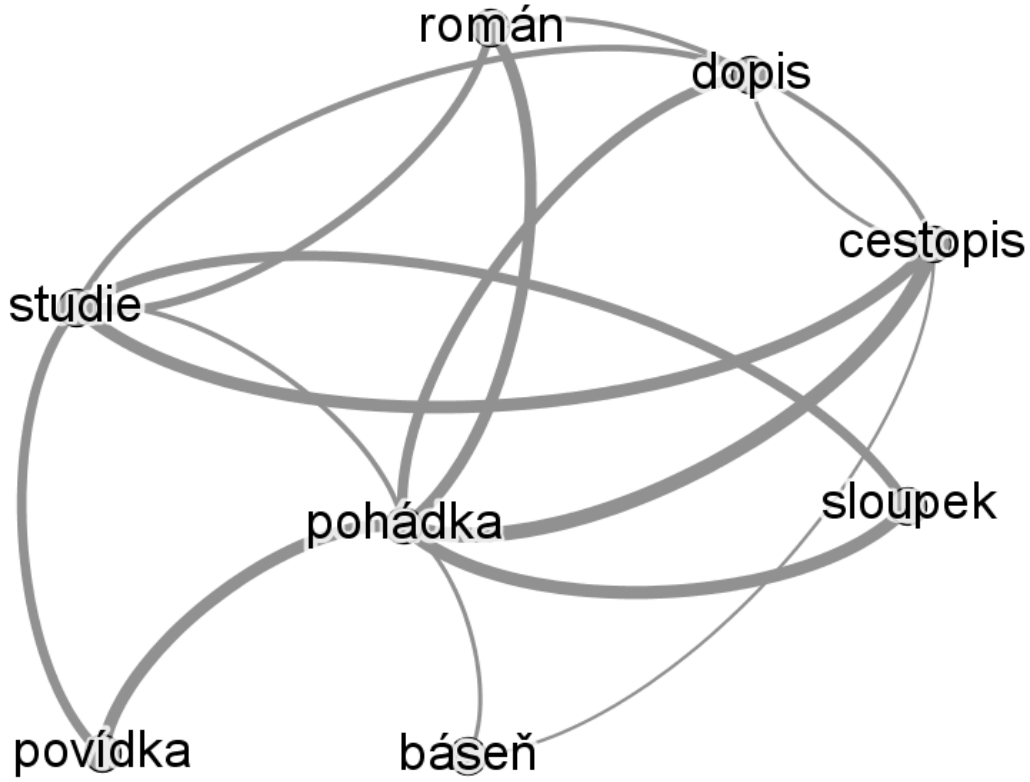
n_1, n_2…počet výsledků v každé skupině

Tab. 4. Výsledky *u*-testu mezi žánry, signifikantní rozdíly ($u \geq 1{,}96$, $\alpha = 0{,}05$) jsou vyznačeny tučně

	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



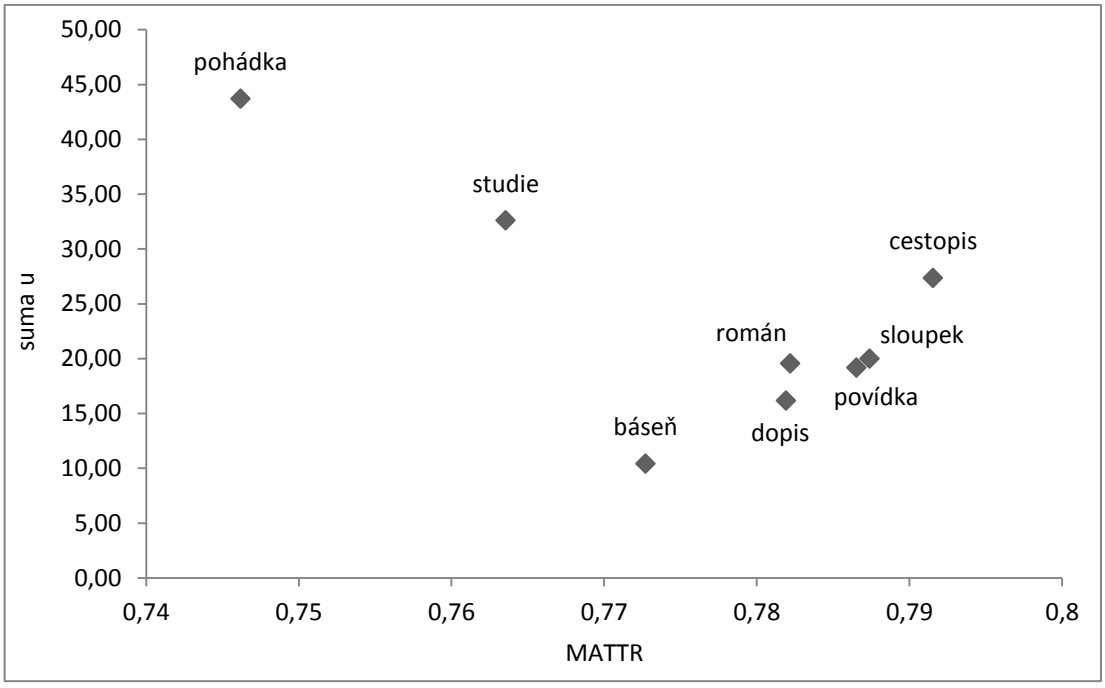
Obr. 20. Síť zobrazující rozdíly mezi žánry v *MATTR* (hrany značí nesignifikantní rozdíl).



Obr. 21. Síť zobrazující rozdíly mezi žánry v *MATTR* (čím širší hrany, tím větší rozdíl)

Tab. 5. Sumy hodnot *u*-testu

pohádka	43,69
studie	32,59
cestopis	27,35
sloupek	19,99
román	19,56
povídka	19,17
dopis	16,16
báseň	10,42



Obr. 22. Graf zobrazující vzdálenosti jednotlivých žánrů na základě hodnot *u* a *MATTR*

Na základě výše uvedených výsledků, kde více než polovina rozdílů je signifikantních, můžeme tvrdit, že slovní bohatství je z hlediska diferenciace žánrů poměrně silný nástroj. To znamená, že slovní bohatství je faktor, který hraje v žánrové klasifikaci důležitou roli.

Z hlediska stylistiky jsou získaná data důležitá nejen vzhledem ke specifickému postavení pohádky, studie a cestopisu, ale také vzhledem k ostatním žánrům. Nízké hodnoty *MATTR* u pohádky odpovídají obecnému předpokladu, že slovník literatury určené primárně dětem musí být nutně přizpůsoben jejich schopnostem. Pohádka se dokonce signifikantně liší od všech ostatních analyzovaných žánrů, čímž se jí dostává skutečně výjimečné pozice, což je nejlépe vidět na Obr. 22. Druhé nejnižší hodnoty bohatství slovníku dosáhla studie, což ukazuje na striktní pravidla odborného stylu, kde je formální kreativita potlačena ve prospěch srozumitelnosti a stručnosti. Na opačném konci škály stojí cestopis, který dosáhl vůbec nejvyššího slovního bohatství. Tento fakt lze vysvětlit zejména vyšší frekvencí proprií, zejména pak toponym. Zatímco výsledky pohádky, studie a cestopisu odpovídají obecným předpokladům, zejména hodnoty básně a dopisu mohou působit poněkud překvapivě. Ostatní žánry (román, povídka, sloupek) naopak nejsou v rozporu s očekáváním.

Přestože *MATTR* se zdá být jedinečným nástrojem, který měří slovní bohatství bez vlivu délky textu a zároveň respektuje text jakožto uzavřený homogenní celek, nabízí se další problematický aspekt všech podobných indexů používaných ve stylometrii vůbec. Jde o to, že výsledkem těchto indikátorů je jediná číselná hodnota, jež má charakterizovat celý text. Otázka je, zda se takto nedopouštíme příliš velké simplifikace, neboť hodnota průběžného *TTR* může v textu značně oscilovat, jediná průměrná hodnota tak může do určité míry specifickou charakteristiku textu zkreslovat. Z tohoto důvodu byl navržen moving type-token distribution (*MWTTRD*)[82]. Tato metoda vychází z *MATTR* a liší se pouze ve formě konečného výsledku. Nejdříve jsou získána data jednotlivých oken (windows), tj. počet typů v každém okně. Na rozdíl od *MATTR*, který tyto hodnoty zprůměruje, *MWTTRD* vypočítá poměrné zastoupení jednotlivých výskytů typů. Distribuce znázorněná v grafu jako křivka je pak konečným výsledkem *MWTTRD*.

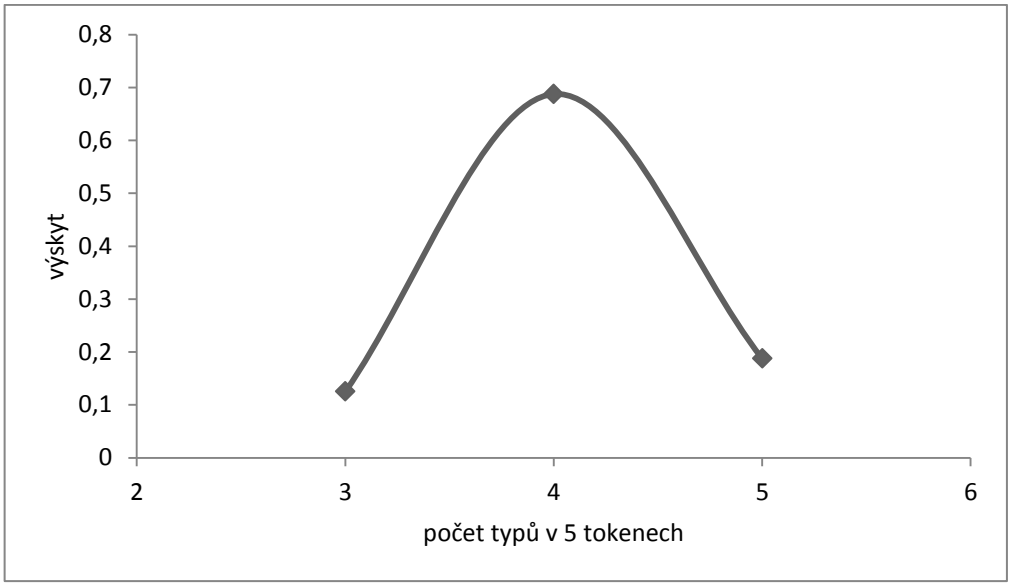
Abychom lépe porozuměli metodě *MWTTRD*, ukážeme si výpočet na jednoduchém příkladu. Máme sekvenci 20 písmen: *a, a, b, c, d, a, f, g, c, c, b, c, i, d, d, j, k, b, b, d*. Velikost okna arbitrárně zvolíme na 5 tokenů ($L = 5$) a spočítáme počet typů v každém okně. Tím získáme následující distribuci: 4, 4, 5, 5, 5, 4, 4, 3, 3, 4, 4, 4, 4, 4, 4, 4, kterou převedeme do tabulky s frekvencemi, viz Tab. 6.

Tab. 6. Distribuce počtu typů v jednotlivých oknech s frekvencemi

Počet typů	Abs.frekvence	Relativní frekvence
3	2	0,125
4	11	0,6875
5	3	0,1875

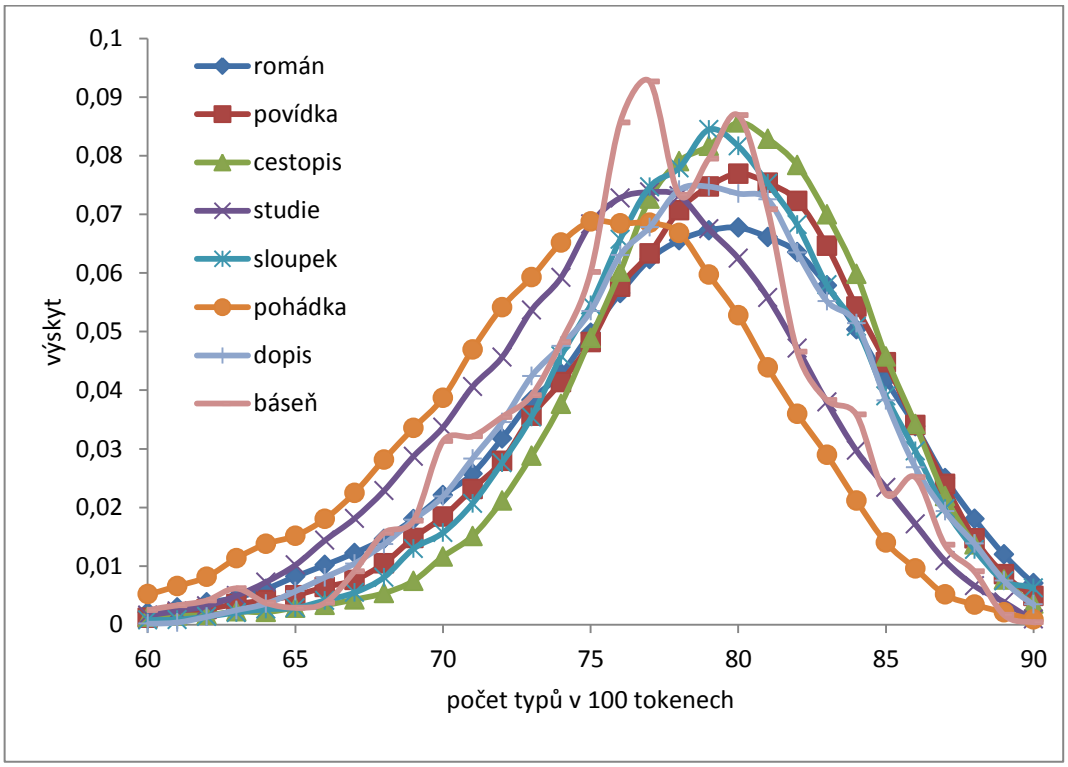
Z tabulky zjistíme, že v jednotlivých oknech o velikosti 5 tokenů se objevilo 3, 4 nebo 5 typů, přičemž 3 typy v 12,5 %, 4 typy v 68,75 % a 5 typů v 18,75 %. Tyto hodnoty následně převedeme do grafu, který nám umožní detailně sledovat celou distribuci průběžného type-token poměru, viz Obr. 23.

[82] Kubát, M., Milička, J. (2013).



Obr. 23. Ukázka zobrazení hodnot *MWTTRD* ze sekvence písmen

Výsledky *MWTTRD* v jednotlivých žánrech jsou uvedeny na Obr. 24.



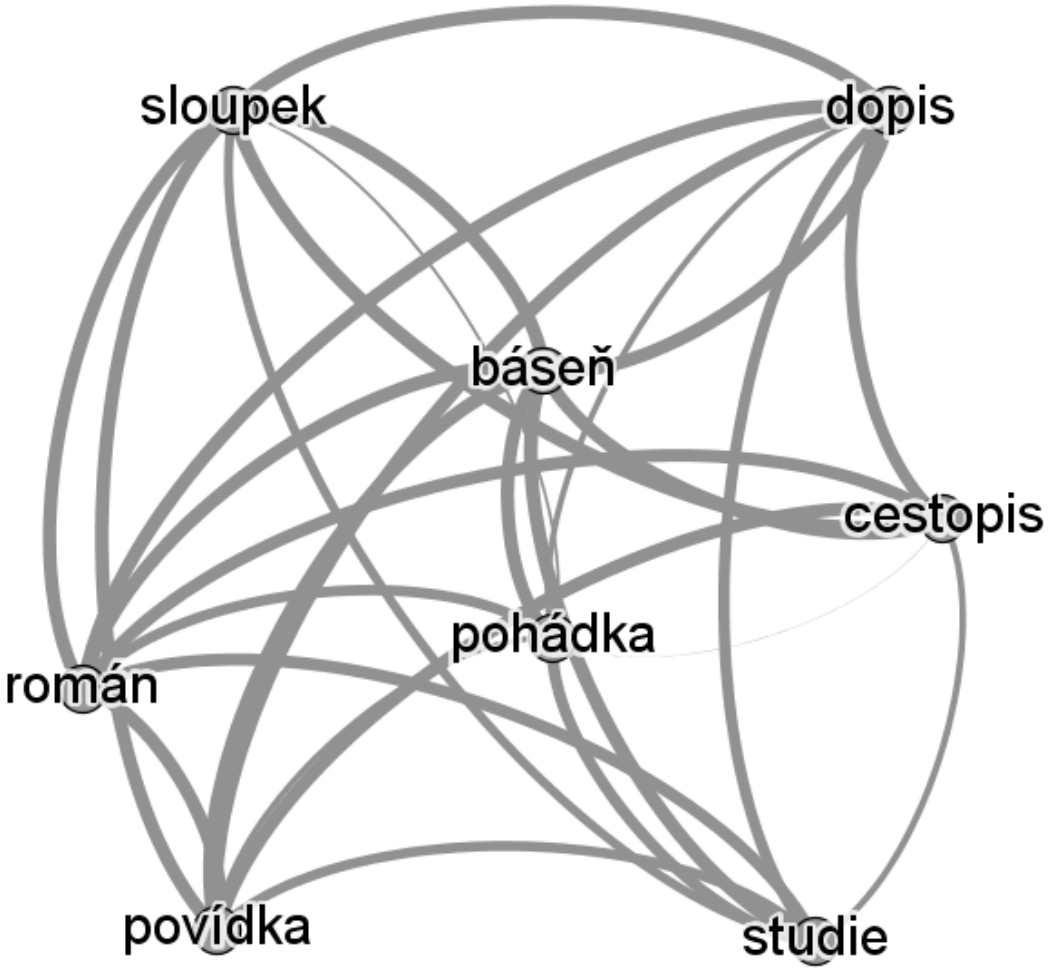
Obr. 24. Výsledné hodnoty *MWTTRD* v různých žánrech v Čapkových textech

Ze získaných distribucí na Obr. 24 jsou patrné jisté diference, nicméně pro porovnání jednotlivých žánrů je třeba použít přesnější metodu. Pro tento účel lze aplikovat x^2 diskrepanční koeficient (C), který se zpravidla užívá pro testování shody měřené distribuce s konkrétním rozdělením.[83] Za hraniční hodnotu C pro stanovení rozdílu jsme určili 0,05. Výsledky jsou uvedeny v Tab. 7. a pro větší přehlednost také na Obr. 25, kde jsou rozdíly znázorněny pomocí sítě, v Tab. 8 pak najdeme sumy hodnot C v jednotlivých žánrech.

Tab. 7. Porovnání *MWTTRD* pomocí C ($C \geq 0{,}05$ znamená, že se distribuce liší)

	román	povídka	cestopis	studie	sloupek	pohádka	dopis
román	x						
povídka	0,005	x					
cestopis	0,013	0,008	x				
studie	0,018	0,042	**0,08**	x			
sloupek	0,006	0,004	0	**0,051**	x		
pohádka	0,036	**0,087**	**0,146**	0,025	**0,117**	x	
dopis	0,003	0,004	0,019	0,028	0,007	**0,085**	x
báseň	0,001	0,002	0,004	0,002	0,004	0,008	0,005

[83] Srov. Mačutek, J., Wimmer, G. (2013).



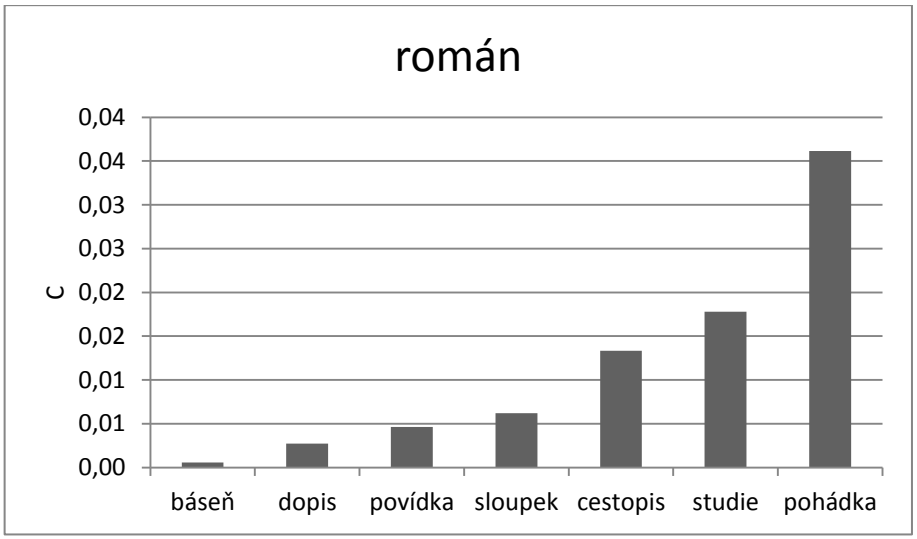
Obr. 25. Síť zobrazující rozdíly mezi žánry v *MWTTRD* (čím širší hrany, tím menší rozdíl)

Tab. 8. Sumy hodnot *C* u jednotlivých žánrů

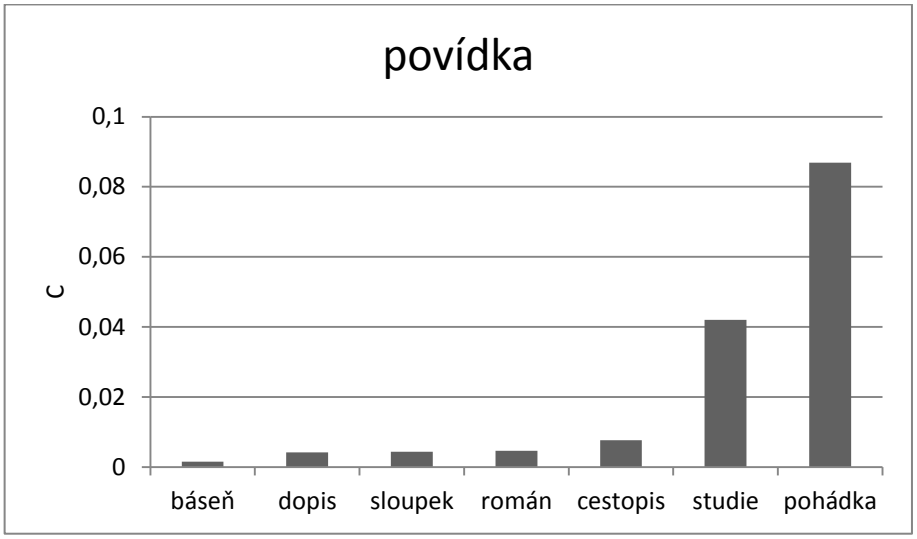
pohádka	0,504
cestopis	0,27
studie	0,246
sloupek	0,19
dopis	0,151
povídka	0,151
román	0,081
báseň	0,026

Z výše uvedených výsledků na Obr. 24, v Tab. 7 a v Tab. 8 se potvrzuje výjimečné postavení pohádky. Další rozdíly pak byly nalezeny mezi studií a cestopisem a sloupkem, což potvrzuje vysoké hodnoty *u*-testu u těchto dvojic

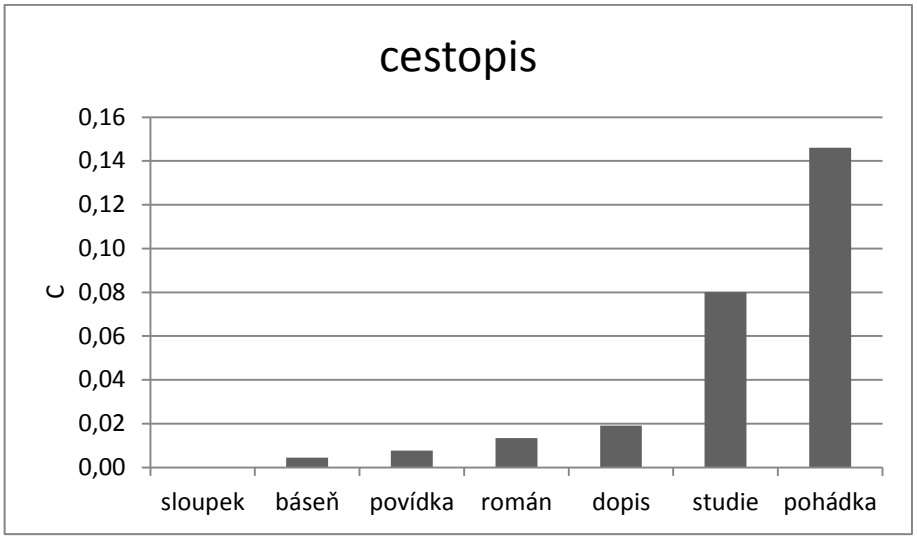
v *MATTR*. Abychom ukázali, které žánry mají k sobě naopak nejblíže, uvádíme níže přehledové grafy jednotlivých žánrů uspořádané podle vzestupných hodnot *C*, přičemž platí, že čím vyšší je *C*, tím je žánr specifičtější, tedy odlišný od ostatních žánrů.



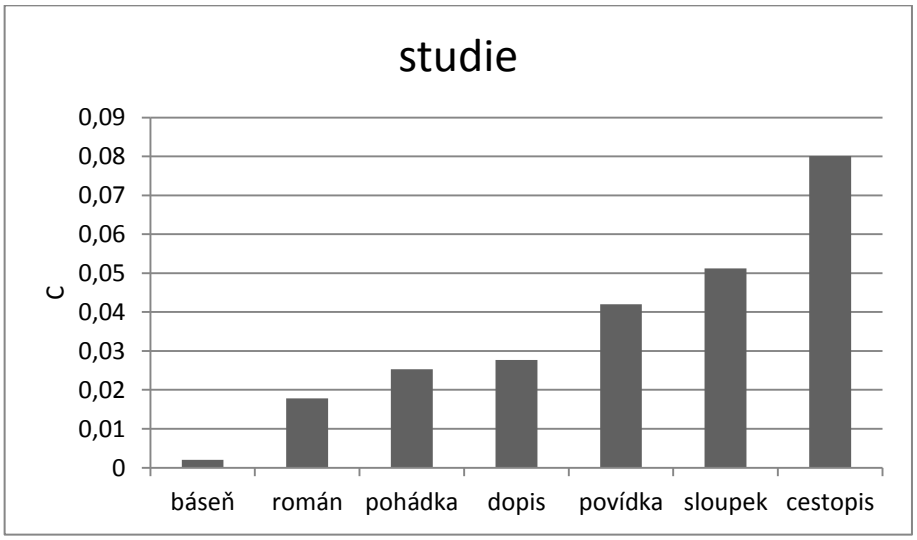
Obr. 26. Nejbližší žánry románu dle hodnot *C* (Čím nižší *C*, tím menší rozdíl)



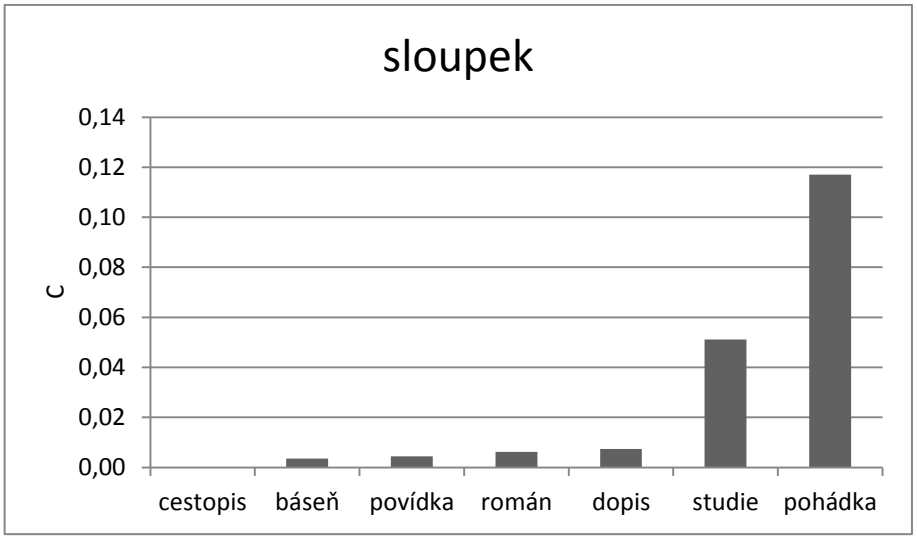
Obr. 27. Nejbližší žánry povídky dle hodnot *C* (Čím nižší *C*, tím menší rozdíl)



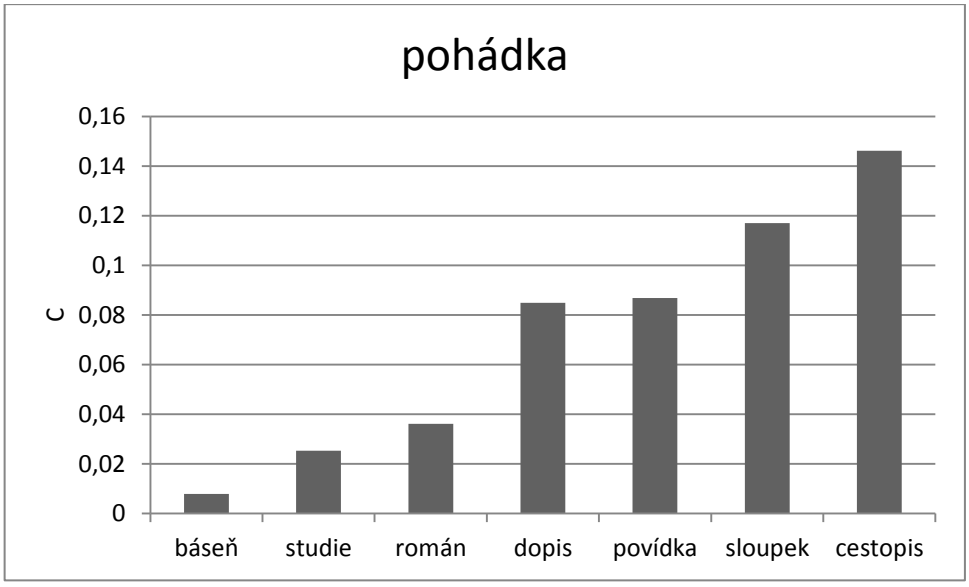
Obr. 28. Nejbližší žánry cestopisu dle hodnot *C* (Čím nižší *C*, tím menší rozdíl)



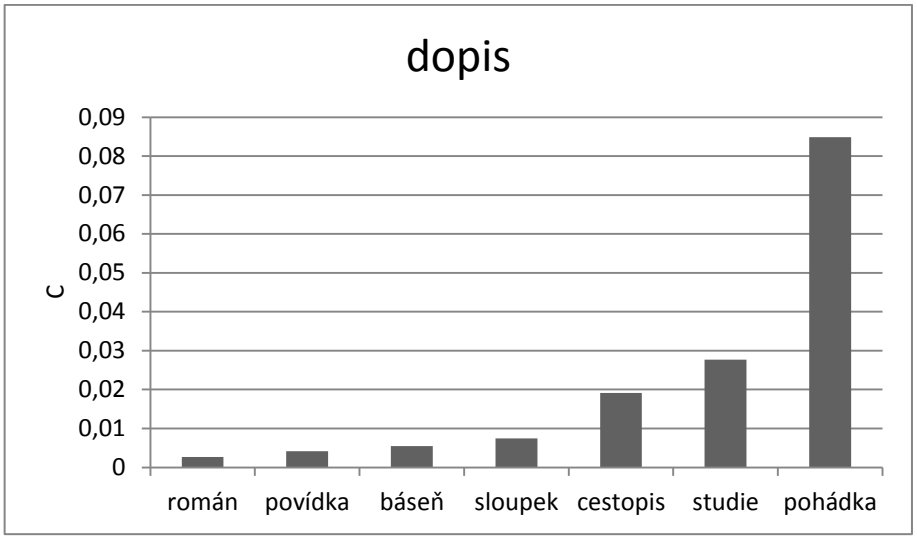
Obr. 29. Nejbližší žánry studie dle hodnot *C* (Čím nižší *C*, tím menší rozdíl)



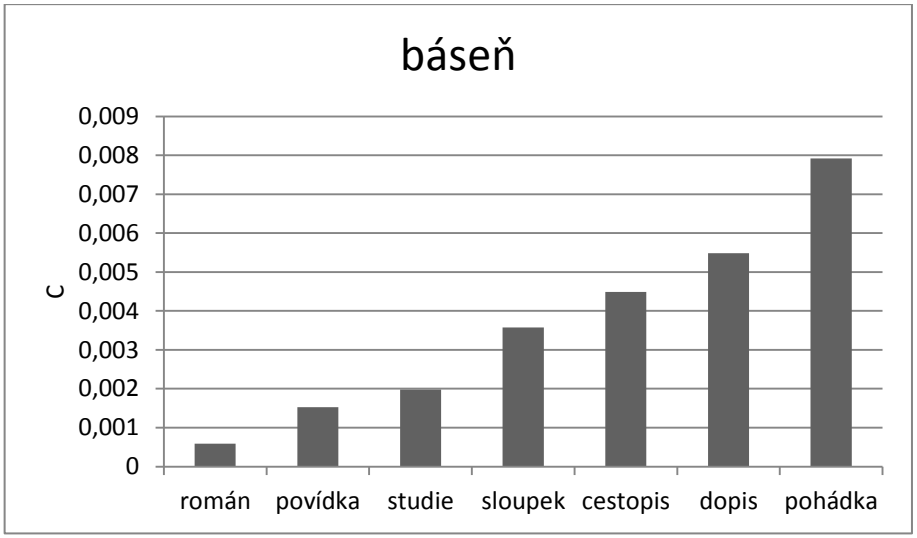
Obr. 30. Nejbližší žánry sloupku dle hodnot *C* (Čím nižší *C*, tím menší rozdíl)



Obr. 31. Nejbližší žánry pohádky dle hodnot *C* (Čím nižší *C*, tím menší rozdíl)



Obr. 32. Nejbližší žánry dopisu dle hodnot *C* (Čím nižší *C*, tím menší rozdíl)

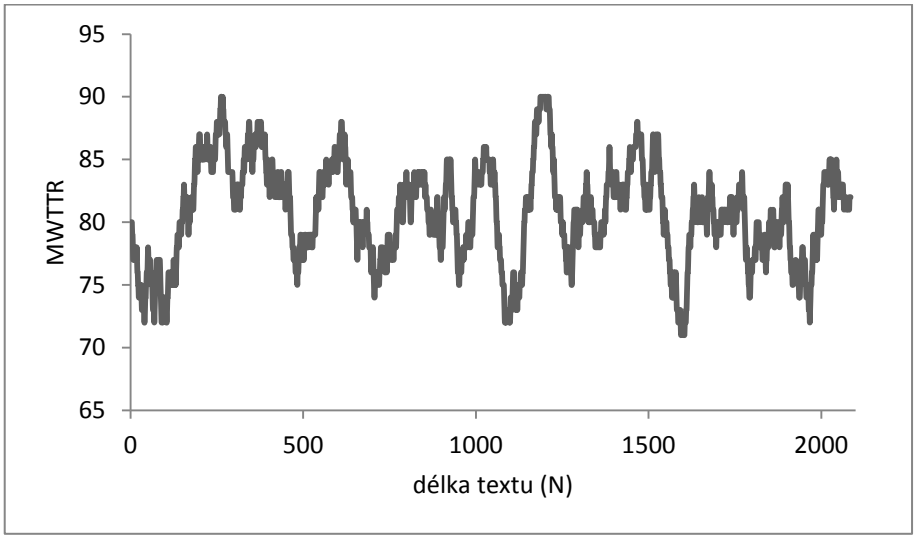


Obr. 33. Nejbližší žánry básně dle hodnot *C* (Čím nižší *C*, tím menší rozdíl)

Pro úplnost ještě zmíníme jednu metodu, jež je založena na průběžném měření *TTR* v textu, jde o moving window type-token ratio (*MWTTR*).[84] Tato metoda je vhodná zejména pro detailní analýzu jednotlivých textů, neboť zobrazuje průběh jednotlivých hodnot *TTR* v každém okně. Získáme tak informaci, jak se *TTR* v textu vyvíjí. Lze tak např. zjistit, zda se určitá pasáž textu nějak liší od ostatních apod.

[84] Köhler, R., Galle, M. (1993).
Covington, M. A., McFall J. D. (2010).

Jako příklad uvádíme na Obr. 34 *MWTTR* v Čapkově *Povídce starého kriminálníka* z *Povídek z druhé kapsy*.



Obr. 34. *MWTTR* v *Povídce starého kriminálníka* z *Povídek z druhé kapsy*

Jak již bylo zmíněno výše, moving window type-token ratio je metoda vhodná spíše pro zkoumání rozdílů mezi jednotlivými texty než pro žánrovou analýzu. *MWTTR* tak nebudeme aplikovat na náš primární výzkum, nicméně je třeba konstatovat, že výsledky této metody potvrzují předpoklad, že jediná výsledná hodnota může být vhledem k nelineárnímu průběhu *TTR* v textu do jisté míry zavádějící.

Za nejpřesnější metodu pro měření slovního bohatství v žánrech tak považujeme *MWTTRD*, ale je třeba uvést, že toto měření není z hlediska textové klasifikace tak účinné jako *MATTR* a je také mnohem komplikovanější pro statistické vyhodnocení výsledků. Pro výběr vhodné metody je tedy vždy nezbytné přihlédnout ke konkrétním požadavkům a cílům dané analýzy.

4.2. Tematická koncentrace textu

Jakýkoliv text se soustředí na určité téma či témata. Je zřejmé, že míra koncentrace témat se u různých textů liší. U některých kratších textů, např. básní, jsme zpravidla schopni rozhodnout, který text je tematicky koncentrovanější. Jestliže však chceme porovnat delší texty nebo celé skupiny textů, naše subjektivní hodnocení nutně selhává. Abychom mohli objektivně hodnotit tuto charakteristiku textu, byl Popescem[85] navržen a dalšími autory[86] rozpracován index tematické koncentrace textu (*TC*). Vzhledem k tomu, že v poslední době byl výpočet tematické koncentrace rozšířen o index sekundární tematické koncentrace *STC* a index proporcionální tematické koncentrace *PTC*, rozhodli jsme se pro naši analýzu použít všechny tři uvedené indexy.[87]

Přestože se tyto způsoby výpočtu do určité míry liší, všechny jsou založeny na *h*-bodu, což je hranice rozdělující frekvenční distribuci textu na synsémantika a autosémantika. Původně byl *h*-bod zaveden Hirschem[88] ve scientometrii, do kvantitativní lingvistiky jej uvedl Popescu.[89] V oblasti nad *h*-bodem však můžeme zpravidla najít také některá autosémantika a naopak. Tuto skutečnost chápeme jako určitou anomálii, která signalizuje zvláštní postavení těchto autosémantik. Ve frekvenční distribuci slov je *h*-bod místo, kde se pořadí rovná frekvenci ($r = f(r)$). Pokud takový bod v distribuci není, vypočítáme jej pomocí následujícího vzorce.

Rovnice 5

$$h = \frac{f(r_1)\,(r_2 - r_1) - \,[f(r_2) - \,f(r_1)]\,r_1}{r_2 - r_1 - [f(r_2) - \,f(r_1)]} = \frac{f(r_1)\,r_2 - f(r_2)r_1}{r_2 - r_1 + f(r_1) - f(r_2)}$$

r…pořadí

f(r)…frekvence daného pořadí

[85] Popescu, I. I. (2007).
[86] Popescu, I. I. a kol. (2009).
Popescu, I. I., Altmann, G. (2011), s. 110–116.
Čech, R., Popescu, I. I., Altmann, G. (2014).
[87] Srov. Čech, R., Garabík, R., Altmann, G. (2015).
[88] Hirsch, J. E. (2005).
[89] Popescu, I. I. (2007).

Výpočet *h*-bodu vysvětlíme na dvou příkladech, v prvním použijeme povídku *Historie beze slov* ze sbírky *Boží muka*, kde v Tab. 9 vidíme, že u tokenu *byl* má pořadí a frekvence stejnou hodnotu 8, tudíž i *h*-bod je 8.

Tab. 9. Deset nejfrekventovanějších slov v povídce *Historie beze slov*

pořadí	frekvence	token
1	49	a
2	31	se
3	18	na
4	13	je
5	12	ježek
6	11	v
7	10	tak
8	**8**	byl
9	7	by
10	7	že

V druhém případě použijeme povídku *Šlépěj* ze sbírky *Boží muka*, jejíž frekvenční distribuce je uvedena v Tab. 10. Na rozdíl od předchozího příkladu zde nenajdeme žádný token, který by měl stejnou hodnotu pořadí a frekvence, proto se pro výpočet *h*-bodu použije Rovnice 5.

Tab. 10. Deset nejfrekventovanějších slov v povídce *Šlépěj*

pořadí	frekvence	Token
1	77	a
2	47	se
3	30	to
4	28	je
5	22	že
6	21	ale
7	21	na
8	20	by
9	17	šlépěj
10	16	si
11	16	v
12	**13**	snad

13	**12**	ní
14	12	byla
15	10	sníh
16	10	když
17	10	tam

$$h = \frac{f(r_1)\, r_2 - f(r_2) r_1}{r_2 - r_1 + f(r_1) - f(r_2)} = \frac{13 \cdot 13 - 12 \cdot 12}{13 - 12 + 13 - 12} = 12{,}5$$

4.2.1. Tematická koncentrace (*TC*)

Jak bylo uvedeno výše, autosémantika nad *h*-bodem jsou považována za slova, která odráží specifika daného textu, v našem případě jeho téma. Tato slova označujeme jako tematická. Pokud změříme vzdálenost tematického slova od *h*-bodu, získáme jeho tematickou váhu (*TW*), tu vypočítáme pomocí Rovnice 6.

Rovnice 6

$$TW_{word} = 2\frac{(h - r')f(r')}{h(h-1)f(1)}$$

Celková tematická koncentrace textu se potom vypočítá jako součet tematických vah (Rovnice 7).

Rovnice 7

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

r'…pořadí autosémantika nad *h*-bodem
h…*h*-bod
T…počet tematických slov

Je nezbytné zmínit, že všechny tři způsoby výpočtu tematické koncentrace (*TC, STC, PTC*) narážejí na dvě úskalí. Prvním je rozhodnutí, které slovní druhy zahrneme mezi tematická slova. Zpravidla jsou použita substantiva, adjektiva a verba, nicméně

je možné zařadit i další slovní druhy, např. příslovce.[90] S tímto problémem úzce souvisí také fakt, že například pomocná slovesa by neměla patřit mezi tematická slova. Druhým problémem je stanovení základní jednotky, neboť jak již bylo zmíněno výše, právě u flexivních jazyků může být vhodné pracovat s lemmatizovanými texty, a tudíž považovat za základní jednotku lemma. Existuje však i jiný pohled na věc, ke kterému se přikláníme i my, a to skutečnost, že volba slovních tvarů je důležitou charakteristikou stylu daného textu.[91] Výhodou (spíše technického rázu) práce se slovními tvary je poměrně snadná segmentace textu. Je zřejmé, že segmentace textu na slovní tvary je nesrovnatelně jednodušší, než je tomu v případě lemmat. Problematika lemmatizace má dvě základní roviny, první spočívá v nejednotnosti způsobu lemmatizace (tedy vymezení toho, které jednotky budeme řadit k jednomu lemmatu), druhá pak v nepřesnosti samotného procesu lemmatizace (ať už ručního, nebo strojového). Jak vyplývá z metodologických východisek celé práce, naše rozhodnutí pracovat při výpočtu tematické koncentrace se slovními tvary jakožto základními jednotkami v žádném případě neznamená, že je tento způsob lepší než práce s lemmaty.

Výpočet tematické koncentrace si ukážeme opět na povídce *Historie beze slov* ze sbírky *Boží muka*. V Tab. 11 vidíme, že se nad *h*-bodem vyskytuje pouze jedno tematické slovo – *ježek*.

Tab. 11. Deset nejfrekventovanějších slov v povídce *Historie beze slov*

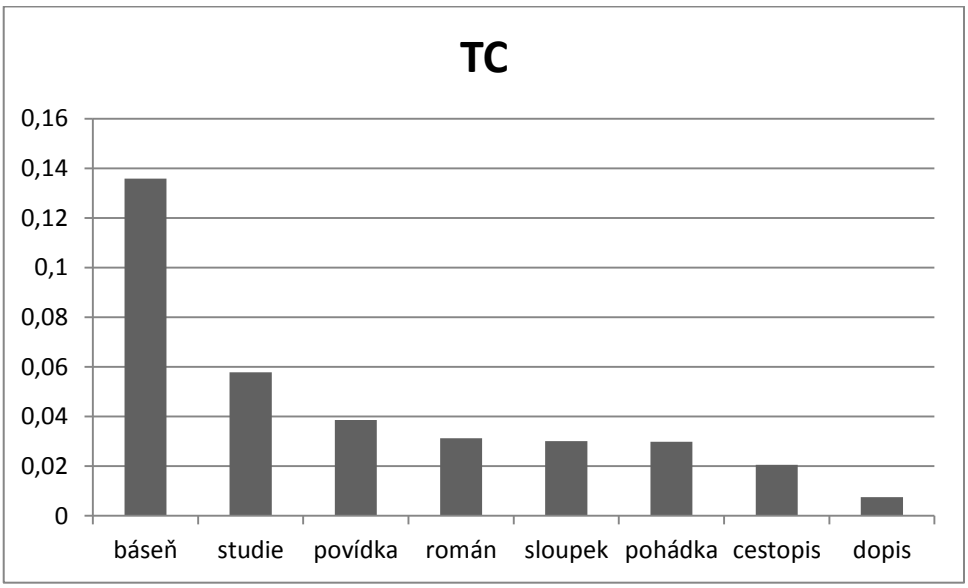
pořadí	frekvence	token
1	49	a
2	31	se
3	18	na
4	13	je
5	**12**	**ježek**
6	11	v
7	10	tak
8	8	byl

[90] Srov.:
Popescu, I. I. a kol. (2009). Word frequency studies. Berlin/New York: Mouton de Gruyter.
Popescu, I. I., Altmann, G. (2011), s. 110–116.
Čech, R., Popescu, I. I., Altmann, G. (2014).
[91] Srov. např. Čech, R., Kelih, E., Mačutek, J. (2014).

9	7	by
10	7	Že

$$TC = \sum_{r'=1}^{T} 2\frac{(h-r')f(r')}{h(h-1)f(1)} = 2\frac{(8-5)12}{8(8-1)49} = 2\frac{36}{2744} = 0{,}026239$$

Konečné výsledky *TC* jsou zobrazeny na Obr. 35 a statistické vyhodnocení pak v Tab. 12 a Tab. 13.



Obr. 35. Hodnoty tematická koncentrace v různých žánrech v Čapkových textech

Tab. 12. *u*-test hodnot *TC* ($u \geq 1{,}96$ značí signifikantní rozdíl, $\alpha = 0{,}05$)

	román	povídka	cestopis	studie	sloupek	pohádka	dopis
román	x						
povídka	1,07	x					
cestopis	1,41	1,90	x				
studie	**3,31**	1,94	**3,58**	x			
sloupek	0,14	0,87	0,93	**2,60**	x		
pohádka	0,20	0,93	0,92	**2,70**	0,03	x	
dopis	**6,30**	**4,49**	1,70	**6,22**	**2,85**	**2,93**	x
báseň	**2,22**	**2,05**	**2,43**	1,64	**2,22**	**2,23**	**2,73**

Tab. 13. Sumy hodnot *u*-testu *TC*

dopis	27,21
studie	21,98
báseň	15,53
román	14,65
povídka	13,25
cestopis	12,87
pohádka	9,94
sloupek	9,65
celkem	125,07

4.2.2. Sekundární tematická koncentrace (*STC*)

Výpočet tematické koncentrace (*TC*) naráží zejména na jeden problém, a to skutečnost, že se často nad *h*-bodem nevyskytují žádná autosémantika, tudíž je výsledkem nulová hodnota *TC*. Aby se tento nedostatek odstranil, byla navržena sekundární tematická koncentrace (*STC*), která se liší tím, že je *h*-bod vynásoben dvěma.[92] Výpočet provedeme pomocí Rovnice 8. Tímto se výrazně zvyšuje pravděpodobnost výskytu tematických slov.

Rovnice 8

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

Pro lepší ilustraci dané problematiky můžeme použít například analýzu tematické koncentrace v *Povídkách z druhé kapsy*. Z Tab. 14 můžeme zjistit, že z 24 povídek jich 12 má nulovou hodnotu *TC*, což je způsobeno tím, že se v daných textech nevyskytlo žádné tematické slovo (viz výše). Takový výsledek analýzy můžeme interpretovat v podstatě dvojím způsobem: buď texty s nulovou hodnotou *TC* budeme považovat za z hlediska tematické koncentrace neutrální či nevyhraněné, nebo takové texty vzhledem k nulovým hodnotám vyřadíme z analýzy. Pokud však použijeme pro výpočet tematické koncentrace metodu *STC,* z 24 povídek má pouze jediná nulovou

[92] Srov. Čech, R., Garabík, R., Altmann, G. (2015).

hodnotu, což nám umožňuje mnohem přesněji porovnávat jednotlivé texty, neboť se nemusíme smířit s tím, že celá polovina zkoumaných textů je tematicky neutrální, ale získáme konkrétnější vyjádření míry tematické koncentrace.

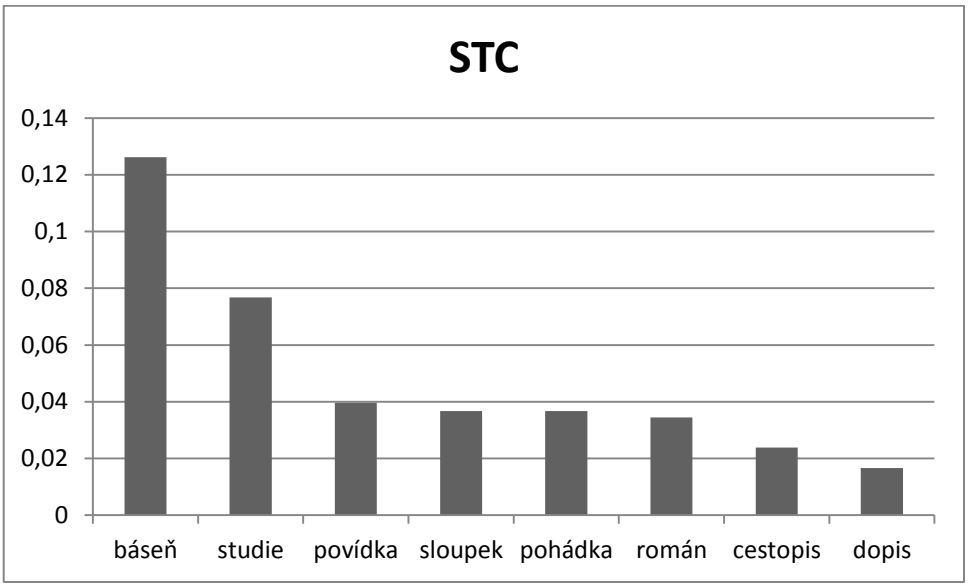
Jakkoliv se *STC* jeví jako vhodnější metoda pro analýzu textů, je třeba zmínit poměrně problematický aspekt sekundární tematické koncentrace, a to klíčový fakt, že *h*-bod byl v podstatě uměle ad hoc zdvojnásoben čistě z pragmatických důvodů konkrétních analýz. Je tedy třeba konstatovat, že tato metoda sice odstraňuje určité nevýhody *TC*, 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 (viz výše), proto, alespoň z našeho pohledu, jakékoliv dodatečné zasahování je značně problematické.

V současné době tematická koncentrace stále není důkladně prostudována, a nelze proto zde konstatovat obecnější závěry ohledně použití *STC*. Můžeme pouze zopakovat, že tato metoda má jak výhody, tak nevýhody, a tedy není zřejmé, do jaké míry je zásah do *h*-bodu je obhájitelný.

Tab. 14. Hodnoty *TC* a *STC* v *Povídkách z druhé kapsy*

	TC	STC
1	0	0,007618
2	0	0,006185
3	0,049167	0,069617
4	0,007653	0,017002
5	0,005967	0,018621
6	0,004722	0,015605
7	0,070454	0,063936
8	0	0,007523
9	0	0,001658
10	0,039393	0,05835
11	0,046213	0,045234
12	0	0,033356
13	0,005442	0,02067
14	0,083317	0,063677
15	0	0,015328
16	0	0,015256
17	0,030085	0,035547
18	0	0,005968

19	0,094644	0,089495
20	0	0,006033
21	0	0,001404
22	0	0,001319
23	0	0
24	0,019724	0,046746



Obr. 36. Hodnoty sekundární tematické koncentrace v různých žánrech v Čapkových textech

Tab. 15. *u*-test hodnot *STC* ($u \geq 1{,}96$ značí signifikantní rozdíl, $\alpha = 0{,}05$)

	román	povídka	cestopis	studie	sloupek	pohádka	dopis
román	x						
povídka	1,64	x					
cestopis	1,16	1,71	x				
studie	**6,19**	**4,07**	**5,68**	x			
sloupek	0,87	0,23	1,78	**2,23**	x		
pohádka	0,85	0,22	1,75	**2,20**	0,69	x	
dopis	**3,12**	**2,95**	0,47	**4,82**	1,60	1,62	x
báseň	**2,62**	**2,38**	**2,86**	1,85	**2,60**	**2,61**	**3,27**

4.2.3. Proporcionální tematická koncentrace (*PTC*)

Proporcionální tematická koncentrace (*PTC*) vznikla jako alternativní výpočet k *TC* a *STC*. Jak již bylo zmíněno výše, i tento index je založen na *h*-bodu, samotný výpočet je pak proporcí frekvence autosémantik nad *h*-bodem vůči frekvenci všech slov nad *h*-bodem. Výpočet provedeme pomocí Rovnice 9. Hlavní výhoda *PTC* spočívá v tom, že lze oproti *TC* a *STC* pomocí statistického testu porovnávat i takové texty, kde je pouze jedno tematické slovo.[93]

Rovnice 9

$$PTC = \frac{1}{N_h} \sum_{r' \leq h} f(r')$$

N_h…frekvence všech slov nad *h*-bodem
$f(r')$…frekvence autosémantik nad *h*-bodem

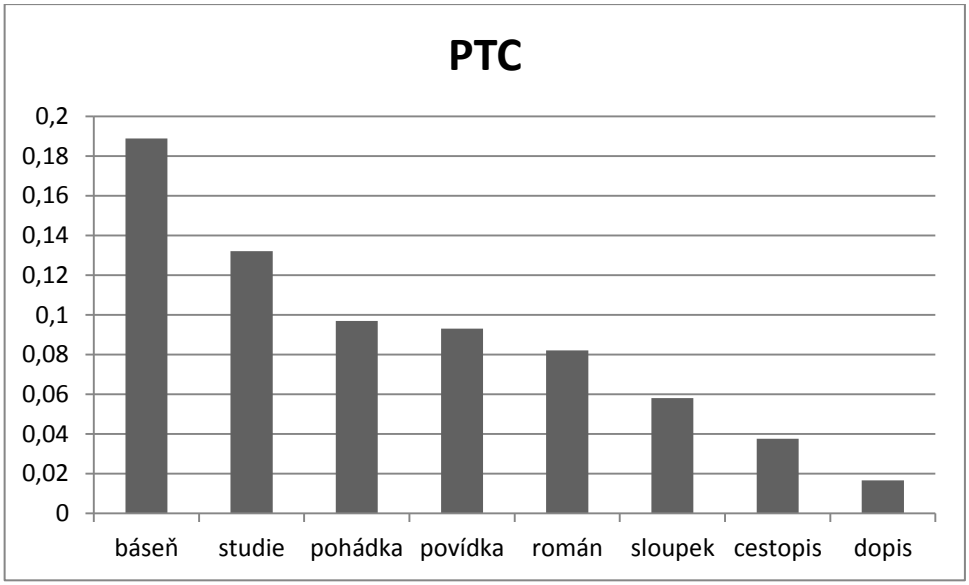
Výpočet proporcionální tematické koncentrace si ukážeme opět na povídce *Historie beze slov* ze sbírky *Boží muka*. Z tabulky vidíme, že se nad *h*-bodem vyskytuje pouze jedno tematické slovo – *ježek*.

Tab. 16. Deset nejfrekventovanějších slov v povídce *Historie beze slov*

pořadí	frekvence	token
1	49	a
2	31	se
3	18	na
4	13	je
5	**12**	**ježek**
6	11	v
7	10	tak
8	8	byl
9	7	by
10	7	Že

$$PTC = \frac{1}{N_h} \sum_{r' \leq h} f(r') = \frac{12}{144} = 0{,}08333$$

[93] Viz Čech, R., Garabík, R., Altmann, G. (2015).



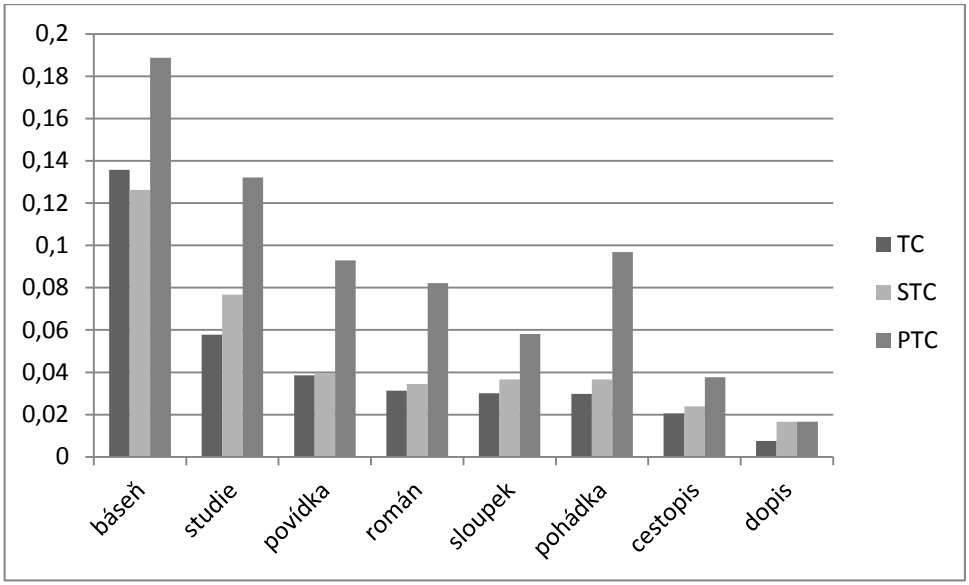
Obr. 37. Hodnoty proporcionální tematické koncentrace v různých žánrech v Čapkových textech

Tab. 17. *u*-test hodnot *PTC* ($u \geq 1{,}96$ značí signifikantní rozdíl, $\alpha = 0{,}05$)

	román	povídka	cestopis	studie	sloupek	pohádka	dopis
román	x						
povídka	**5,61**	x					
cestopis	0,68	0,09	x				
studie	**9,03**	**7,43**	**8,59**	x			
sloupek	**2,64**	1,66	**3,09**	0,02	x		
pohádka	**4,73**	**3,87**	**4,96**	**2,50**	**4,30**	x	
dopis	**2,38**	**2,60**	0,42	**4,37**	1,45	1,45	x
báseň	**2,42**	**2,30**	**2,57**	**2,00**	**2,43**	**2,43**	**2,79**

Abychom mohli přejít k lingvistické interpretaci získaných výsledků, je třeba rozhodnout, ze které metody měření tematické koncentrace (*TC, STC, PTC*) budeme vycházet. Při pohledu na Obr. 38 je patrné, že hodnoty tematické koncentrace a sekundární tematické koncentrace se nijak výrazně neliší, v případě proporcionální tematické koncentrace jen můžeme konstatovat, že výsledné hodnoty kromě pohádky víceméně odpovídají distribucím *TC* a *STC*. Pro upřesnění je třeba dodat, že ačkoliv výsledné hodnoty *PTC* jsou vždy vyšší než u *TC* a *STC*, jde jen o důsledek různého měření tematické koncentrace. Podstatné je, že pokud jde o celkovou distribuci výsledků *PTC*, odpovídá rozložení *TC* a *STC*. Pokud bychom měli určit jednu

z představených metod pro výpočet tematické koncentrace z hlediska diferenciace žánrů či stylů jako nejvhodnější, můžeme použít více hledisek. My budeme vycházet z počtu signifikantních rozdílů v *u*-testu (viz Tab. 18), což koresponduje s koncepcí celé práce, kde mimo jiné porovnáváme různé stylometrické metody z hlediska jejich efektivity diferenciace žánrů.

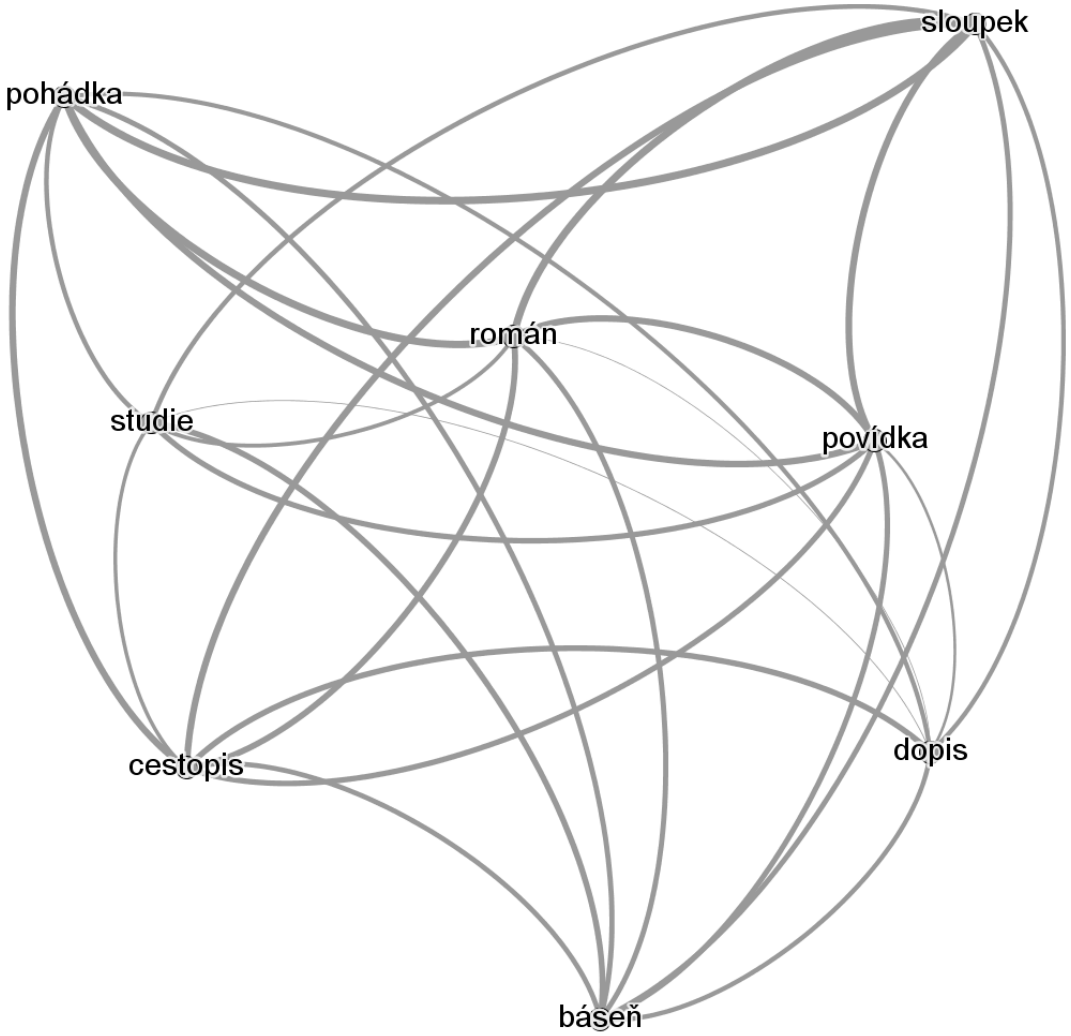


Obr. 38. Porovnání hodnot *TC*, *STC* a *PTC* v různých žánrech

Tab. 18. Počet signifikantních rozdílů dle *u*-testu v jednotlivých metodách výpočtu tematické koncentrace (signifikantní rozdíl $u \geq 1{,}96$, $\alpha = 0{,}05$)

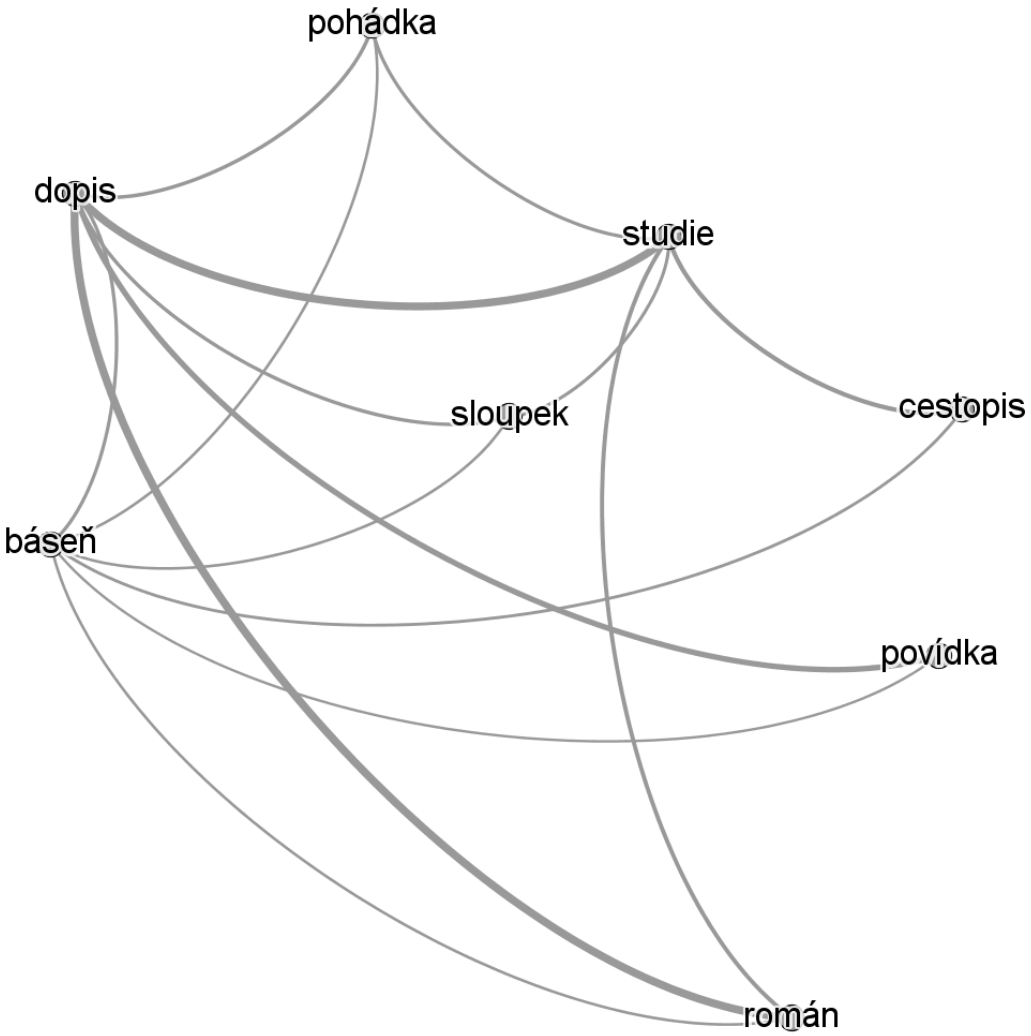
metoda	Počet signifikantních rozdílů
TC	15
STC	14
PTC	14

Na základě zjištěných hodnot nelze s jistotou určit, která ze tří metod je pro naše účely nejvhodnější, neboť se výrazně neliší ani v samotných výsledcích, ani v síle diferenciace žánrů. Vzhledem k tomu, že *TC* je metodou výchozí a dosáhla o jeden signifikantní rozdíl ve statistickém testu více než ostatní, budeme dále pracovat primárně právě s výsledky získanými touto metodou. Abychom lépe vyjádřili vztahy mezi jednotlivými žánry, sestavili jsme síť na Obr. 39.

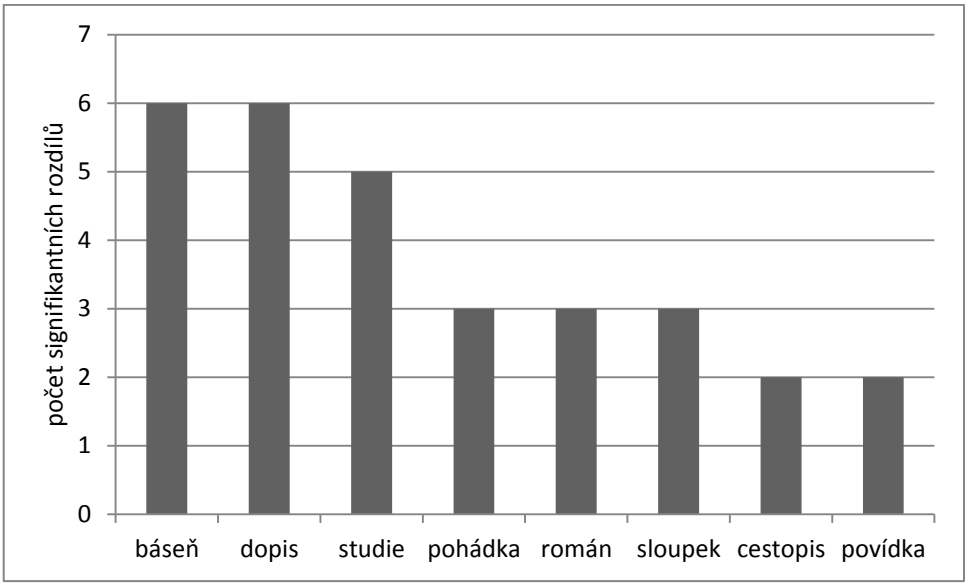


Obr. 39. Síť zobrazující rozdíly mezi žánry v tematické koncentraci (čím širší hrany, tím menší rozdíl)

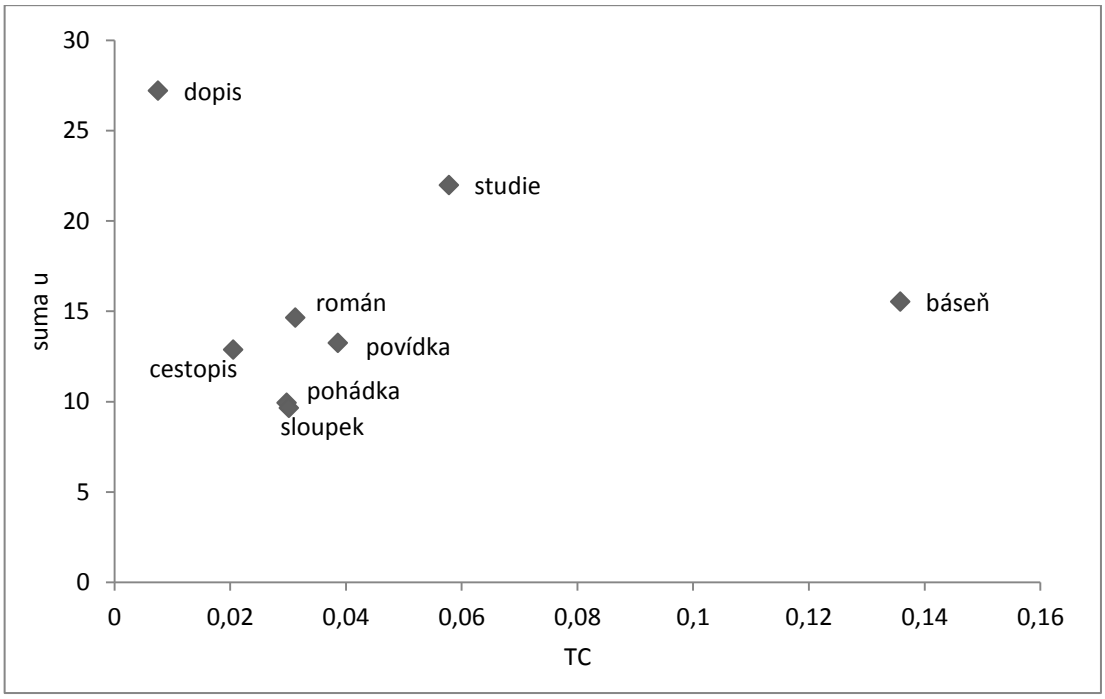
Protože síť na Obr. 39 není příliš přehledná vzhledem k malým rozdílům mezi jednotlivými žánry, vytvořili jsme další síť na Obr. 40, kde jsou spojeny pouze žánry se signifikantními rozdíly, váha vyjadřuje hodnotu výsledku *u*-testu. Pro přehlednost ještě uvádíme graf, který zobrazuje počty signifikantních rozdílů jednotlivých žánrů (viz Obr. 41).



Obr. 40. Síť zobrazující pouze signifikantní rozdíly mezi žánry v tematické koncentraci (čím širší hrany, tím větší rozdíl)



Obr. 41. Žánry seřazené podle počtu signifikantních rozdílů



Obr. 42. Graf zobrazující vzdálenosti jednotlivých žánrů na základě hodnot *u* a *TC*

Z výše uvedených hodnot je patrné, že z hlediska tematické koncentrace textu zaujímá dominantní postavení báseň, která se nejen signifikantně liší od všech ostatních žánrů (kromě studie), ale také dosáhla nejvyšší hodnoty *TC* 0,136. Máme za to, že takový výsledek odpovídá i intuitivnímu předpokladu, neboť právě básně

bývají zpravidla značně tematicky vyhraněné. Jako příklad můžeme uvést Čapkovu báseň *Píseň bez konce* publikovanou v časopisu *Nebojsa*.

"Promiňte, pane, běžím v chvatu,
kdes běží teď o zájem státu.
Až z Vídně jedu, honem, zbystra,
musím to přednést u ministra…"
- Lituji, ale v okamžení
pan ministr tu zrovna není,
pan ministr teď někde řeční,
ti lidé jsou tak za to vděčni!
Schůze je schůze, lid je lid,
pan ministr tam musí být.
"A já zas, pane, mám tu čest
sem důležité zprávy nést.
Řekněte panu ministrovi,
že svědek očitý mu poví…"
- Občane, žel, v tom okamžení
pan ministr tu zrovna není,
pan ministr teď někde řeční,
ti lidé jsou tak za to vděčni!
Schůze je schůze, lid je lid,
pan ministr tam musí být.

Z uvedeného příkladu je zřejmé, že mimo báseň může autor jen stěží dosáhnout tak silné vazby na jediné téma, aniž by tím porušil konvence daného žánru.

Na opačném konci pořadí stojí dopis, který dosáhl taktéž šesti signifikantních rozdílů, hodnota tematické koncentrace (*TC* = 0,008) je za všech žánrů nejnižší. Pokud vyjdeme ze skutečnosti, že šlo o osobní korespondenci, můžeme v těchto poměrně krátkých textech očekávat dominanci fatické funkce. V těchto dopisech tak většinou nejde ani tak o samotný obsah sdělení, jako spíše o udržování kontaktu s adresátem. Proto zde najdeme mnoho zdvořilostních obratů a značnou míru

polytematičnosti (jestli vůbec můžeme mluvit skutečně o nějakém ústředním tématu). Máme tedy za to, že umístění dopisu na posledním místě z hlediska tematické koncentrace odpovídá lingvistickým předpokladům.

Pokud jde o počet signifikantních rozdílů, nelze na Obr. 41 přehlédnout postavení studie, která dosáhla pěti signifikantních rozdílů. Zároveň také vidíme, že v tematické koncentraci zaujímá třetí nejvyšší postavení. Z těchto dat je zřejmé, že studie má skutečně specifické místo mezi analyzovanými žánry. Z lingvistického pohledu ani tyto hodnoty nevybočují z obecných očekávání. Odborné texty ze své podstaty musí být fixovány na jedno či úzký okruh témat, neboť zpravidla zevrubně pojednávají o konkrétní problematice. Je tedy celkem jasné, že autor takových textů se nemůže vyhnout častému opakování stejných slov.

Jestliže se nyní zaměříme na zbývajících pět žánrů (román, povídka, cestopis, sloupek, pohádka), zjistíme, že mezi nimi není téměř žádný rozdíl, a to jak z hlediska samotných hodnot tematické koncentrace, tak z hlediska výsledků statistického testu. Všechny tyto žánry mají k sobě blízko, neboť jde o prozaické umělecké texty, kam můžeme asi bez větších obtíží zařadit i Čapkovy sloupky. Patrně zde nenajdeme žádné závažné lingvistické faktory, podle kterých bychom měli očekávat větší rozdíly mezi těmito texty, alespoň co se týká tematické koncentrace.

Na závěr kapitoly věnované tematické koncentraci můžeme konstatovat, že výsledné hodnoty odpovídají našim předpokladům. Tato skutečnost tak přináší jednak potvrzení našich intuitivních domněnek, ale získaná data také alespoň do určité míry vypovídají o oprávněnosti a užitečnosti celého konceptu měření tematické koncentrace pomocí indexů *TC, STC* a *PTC.*

4.3. Vzdálenosti sloves (*VD*)

Tento indikátor vyjadřuje, kolik tokenů se průměrně nachází mezi dvěma slovesy. Výsledná hodnota *VD* tak určitým způsobem vyjadřuje složitost syntaktické struktury textu. Můžeme předpokládat souvislost *VD* s různými vlastnostmi, například lze očekávat určitý vztah mezi délkou nominální fráze a *VD*, tj. čím delší jsou nominální fráze, tím delší by měly být vzdálenosti sloves. Výsledné hodnoty jsou uvedeny

v grafu na Obr. 43 a Obr. 44, statistické vyhodnocení *u*-testu pak v Tab. 19 a pomocí sítě na Obr. 45. Výpočet *VD* si ukážeme na úryvku z povídky *Historie beze slov* ze sbírky *Boží muka*.

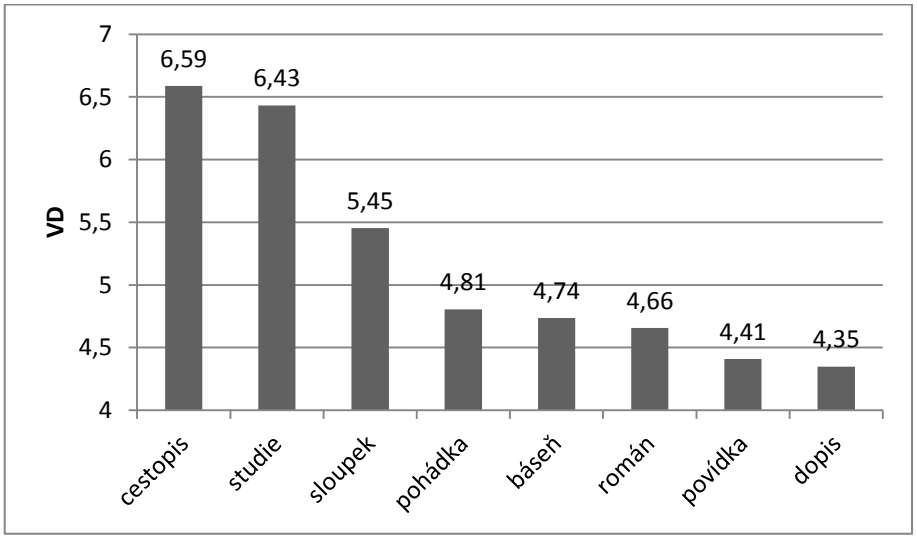
[…]

Jistěže něco **poví**, **myslil** si Ježek; těžko se **hledají** slova pro děj života. Budiž, **počkám**. – Potichu **ulehl** naznak. Slunce ho **udeřilo** v oči a **proniklo** zavřenými víčky; červené a černé kruhy se **roztočily** a palčivě **tančí** před očima.

[…]

V uvedené ukázce najdeme celkem 9 sloves, mezi nimiž se nacházejí sekvence tokenů o následujících délkách: 4, 5, 1, 3, 3, 7, 2. Z těchto hodnot vypočítáme aritmetický průměr, který je konečným výsledkem *VD*:

$$VD = \frac{4 + 5 + 1 + 3 + 3 + 7 + 2}{7} = 3{,}57$$



Obr. 43. Vzdálenosti mezi slovesy v různých žánrech v Čapkových textech

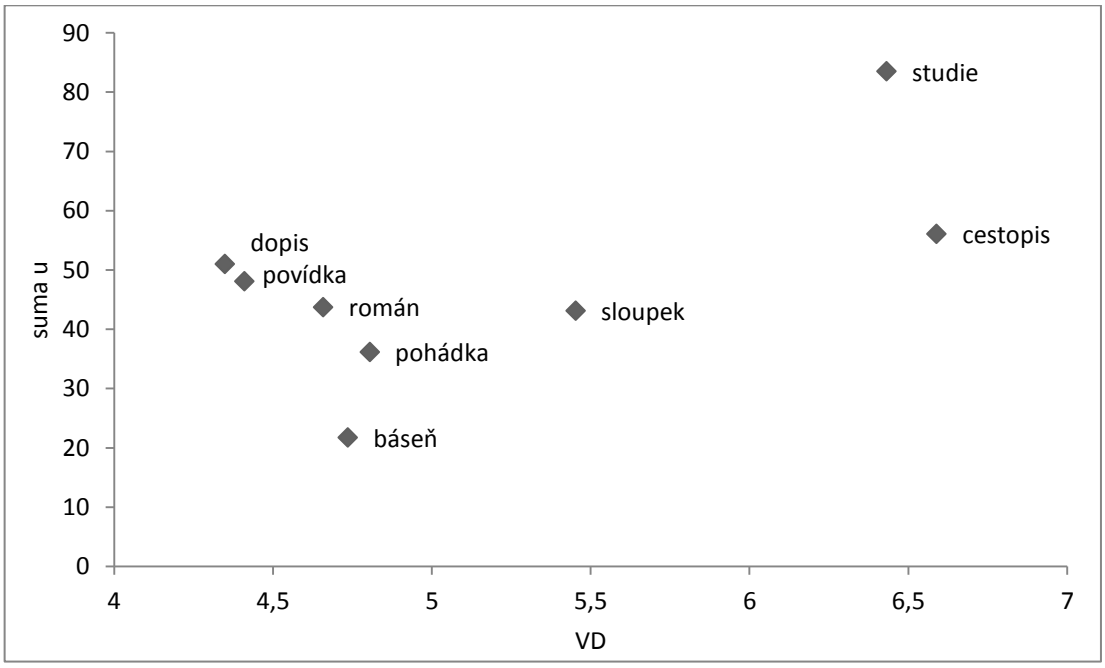
Tab. 19. Hodnoty *u*-testu vzdálenosti sloves *VD* (signifikantní u ≥ 1,96, α = 0,05)

	román	povídka	cestopis	studie	sloupek	pohádka	dopis
román	x						
povídka	**3,36**	x					
cestopis	**10,49**	**11,64**	x				

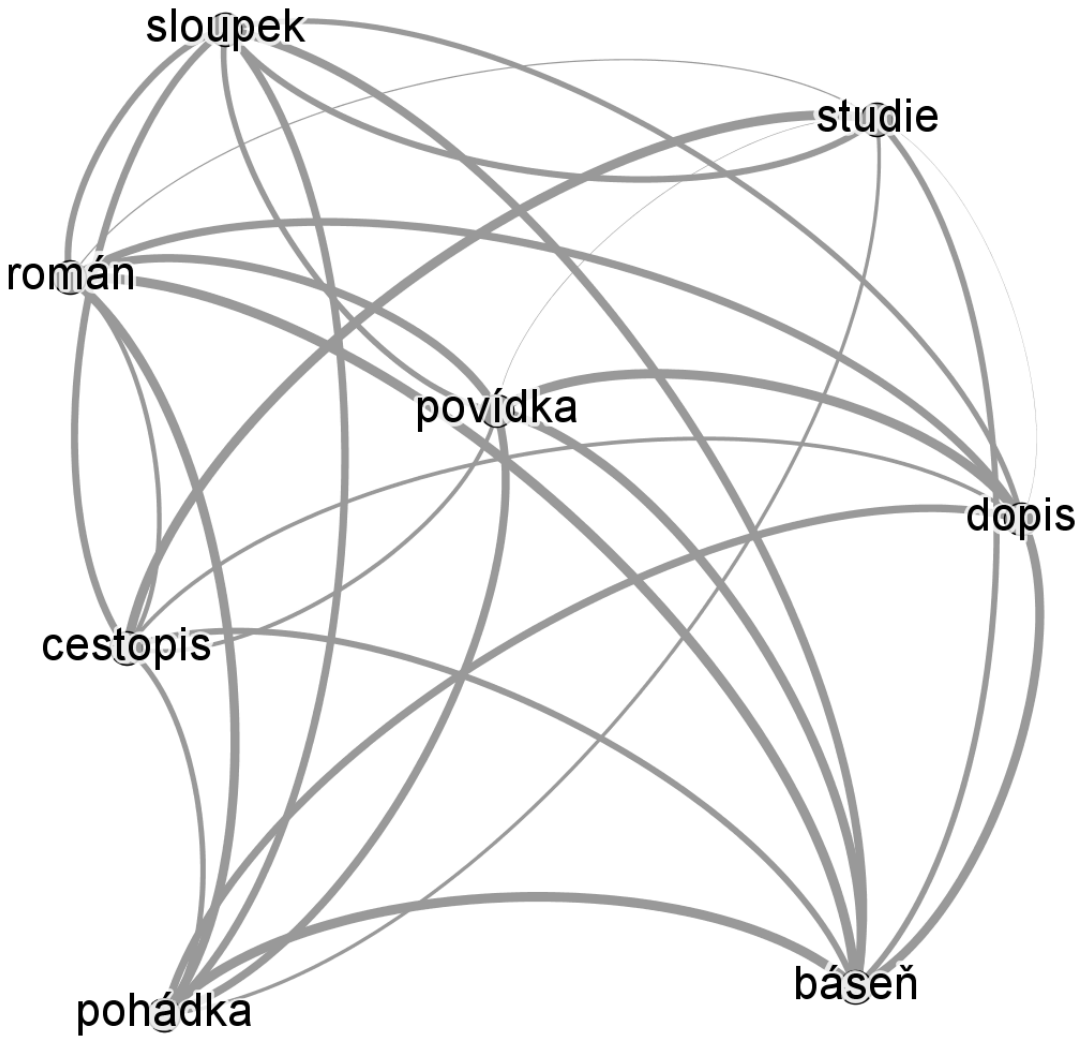
studie	**17,37**	**18,81**	0,79	x			
sloupek	**6,50**	**8,22**	**5,38**	**6,74**	x		
pohádka	**1,47**	**3,75**	**8,96**	**12,81**	**4,51**	x	
dopis	**4,16**	0,75	**11,95**	**19,30**	**8,68**	**4,31**	x
báseň	**0,37**	1,55	**6,88**	**7,66**	**3,10**	0,32	1,84

Tab. 20. Sumy hodnot *u*-testu v jednotlivých žánrech

studie	83,49
cestopis	56,1
dopis	50,1
povídka	48,07
román	43,72
sloupek	43,12
pohádka	36,14
báseň	21,73
celkem	383,37



Obr. 44. Graf zobrazující vzdálenosti jednotlivých žánrů na základě hodnot *u* a *VD*



Obr. 45. Síť zobrazující rozdíly mezi žánry ve *VD* (čím širší hrany, tím menší rozdíl)

Index vzdálenosti mezi slovesy je relativně jednoduchý nástroj, který umožňuje kvantifikovat určitou oblast syntaktické struktury textu. Zatímco většina ostatních indexů použitých v této práci se soustředí spíše na lexikum, *VD* umožňuje alespoň částečně zohlednit syntaktickou stránku. Jak již bylo zmíněno, hlavní předností měření *VD* je jednoduchost, která vyplývá především z toho, že zkoumané texty nemusí být syntakticky označkovány, dokonce nemusíme znát ani hranice vět. Přes jisté zjednodušení se domníváme, že tento index musí korelovat se složitostí syntaktických struktur v 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).

Z Obr. 43 a Tab. 19 je evidentní zvláštní postavení studie a cestopisu, které dosáhly nejvyšších hodnot a také se statisticky nejvíce liší od ostatních. Pokud máme zmínit hlavní důvody takových výsledků, musíme vyjít z charakteru a funkce jednotlivých žánrů. Co se týká studie, je zřejmé, že právě univerzitní odborné texty se vyznačují velmi složitými a dlouhými souvětími. V případě cestopisů hraje nejdůležitější roli popisnost, tedy potřeba co nejpestřeji vylíčit neznámá místa a prostředí. V takových textech nutně klesá frekvence sloves a naopak se zvyšuje frekvence jiných slovních druhů, zejména adjektiv. Tento fakt můžeme snadno ověřit v kap. 4.6 *Distribuce slovních druhů* a v kap. 4.5 *Aktivita a deskriptivita*, protože právě cestopis a studie byly vyhodnoceny jako nejvíce deskriptivní.

Z hlediska klasifikace textů se vzdálenosti sloves jeví jako velmi efektivní nástroj, neboť již z letmého pohledu na Tab. 19 je patrný jednak velký podíl signifikantních rozdílů, jednak také jejich vysoké hodnoty. Konkrétně z celkového počtu 28 rozdílů je 23 statisticky významných, což představuje 82 %. Nelze tedy pochybovat, zvlášť s přihlédnutím ke specifikům našeho korpusu, že tento index má potenciál v automatické klasifikaci textů nebo v určování autorství.

4.4. Průměrná délka tokenu (*ATL*)

Průměrná délka tokenu (average token length) je jednoduchá stylometrická charakteristika, jež vystihuje jeden ze základních prvků obtížnosti textu (readability). *ATL* tak najde uplatnění například ve školství, a to jak při výběru textů pro výuku cizích jazyků, tak i při tvorbě učebnic pro základní školy. Pokud jde o náš výzkum, intuitivně lze předpokládat, že odborné texty by měly mít jistě delší slova než například literatura pro děti. Je totiž zřejmé, že čím je člověk vzdělanější, tím má větší aktivní i pasivní slovní zásobu. Vzhledem k tomu, že nejfrekventovanější slova jsou nejkratší, lze předpokládat u úzce zaměřených textů určených vymezené odborně vzdělané skupině čtenářů vyšší průměrnou délku slova. Výpočet průměrné délky tokenu je velice snadný, stačí znát pouze dvě proměnné (délku jednotlivých

tokenů a jejich počet), z nichž pak vypočítáme aritmetický průměr. Z metodologického hlediska je třeba uvést, že v této analýze pracujeme se slovními tvary, a ne s lemmaty. Toto rozhodnutí pramení zejména z faktu, že distribuce délek jednotlivých tvarů nejsou náhodné,[94] samotnou délku jednotlivých tokenů pak měříme v grafémech.

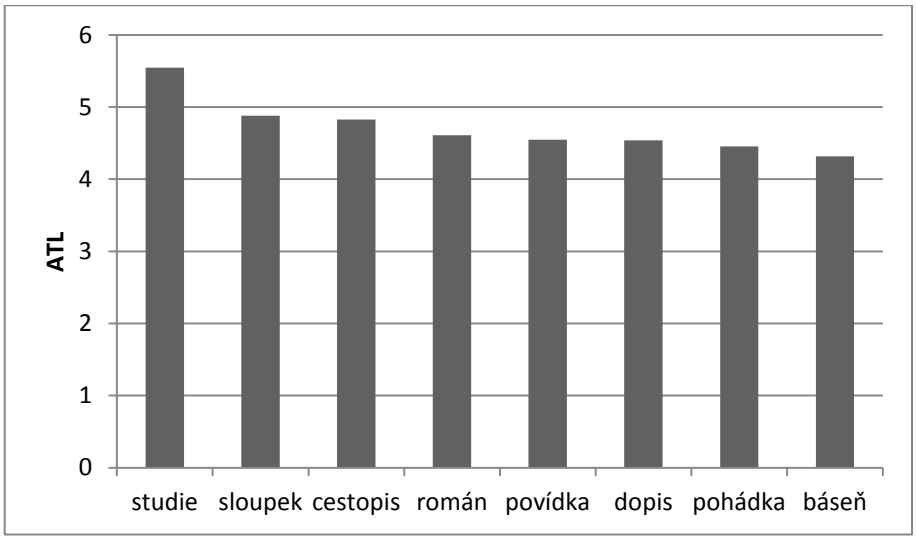
Rovnice 10

$$ATL = \frac{1}{N}\sum_{i=1}^{N} x_i$$

N…počet tokenů

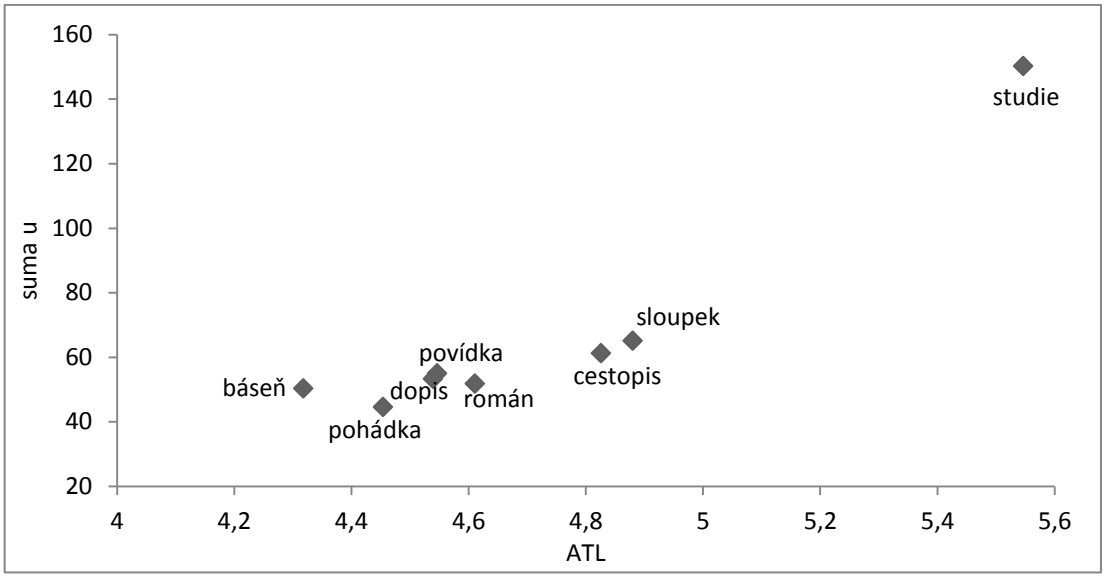
x…délka tokenu

Výsledky průměrné délky tokenu jsou uvedeny na Obr. 46.



Obr. 46. Průměrná délka slova v grafémech v různých žánrech v Čapkových textech

[94] Viz např. Čech, R., Kelih, E., Mačutek, J. (2014).



Obr. 47. Graf zobrazující vzdálenosti jednotlivých žánrů na základě hodnot *u* a *ATL*

Tab. 21. Výsledky *u*-testu v *ATL* (signifikantní u ≥ 1,96, α = 0,05)

	román	povídka	cestopis	studie	sloupek	pohádka	dopis
povídka	**2,23**	x					
cestopis	**6,89**	**9,44**	x				
studie	**24,09**	**26,64**	**18,32**	x			
sloupek	**8,27**	**10,76**	1,63	**16,50**	x		
pohádka	**3,18**	1,91	**7,49**	**19,95**	**8,43**	x	
dopis	**2,32**	0,23	**9,14**	**25,86**	**10,42**	1,73	x
báseň	**4,82**	**3,81**	**8,31**	**18,80**	**9,09**	1,89	**3,64**

Tab. 22. Sumy hodnot *u*-testu v jednotlivých žánrech

studie	150,1673
sloupek	65,11022
cestopis	61,22545
povídka	55,01531
dopis	53,33184
román	51,79354
báseň	50,361
pohádka	44,57307
celkem	531,5777

Průměrná délka slova je velmi jednoduchý nástroj, který signalizuje několik textových charakteristik. Jak již bylo zmíněno výše, jde například o míru obtížnosti

(readability), ale také o určitou řekněme intelektuálnost textu. Získané výsledky takové předpoklady potvrzují, protože právě studie určená úzkému vědeckému okruhu čtenářů zaujímá zcela ojedinělé postavení. U ostatních žánrů již rozdíly nejsou patrné. Jen se snad stručně vyjádříme k textům s nejnižší průměrnou délkou slov, kam patří pohádka a báseň. V případě pohádky je patrné, že autor těchto textů musí přizpůsobit slovní zásobu a tím i délku slov dětem. Pokud jde o básně, tyto texty zaujímají značně specifické místo, neboť délka textu se řídí i jinými zákonitostmi, např. formou verše apod.

4.5. Aktivita a deskriptivita

Aktivita (*Q*) a deskriptivita (*D*) textu jsou poměrně jednoduché stylové charakteristiky. Zatímco deskriptivita je reprezentována poměrným zastoupením adjektiv ke slovesům, aktivita je reprezentována poměrem sloves k adjektivům v textu. Pokud jde o slovesa, zpravidla nejsou do aktivity započítávána stavová slovesa *být, mít, spát* apod. Jak uvádí Čech a kol.[95], je v případě aktivity možné také pracovat s verbálními substantivy (házení, stavění apod.) a v případě deskriptivity lze uvažovat také o zařazení adverbií, kterými se odpovídá na otázku „jak?“. V této práci budeme pracovat pouze s adjektivy a slovesy. Výpočet provedeme pomocí jednoduchého poměru sloves ke slovesům a adjektivům (viz Rovnice 11).

Rovnice 11

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

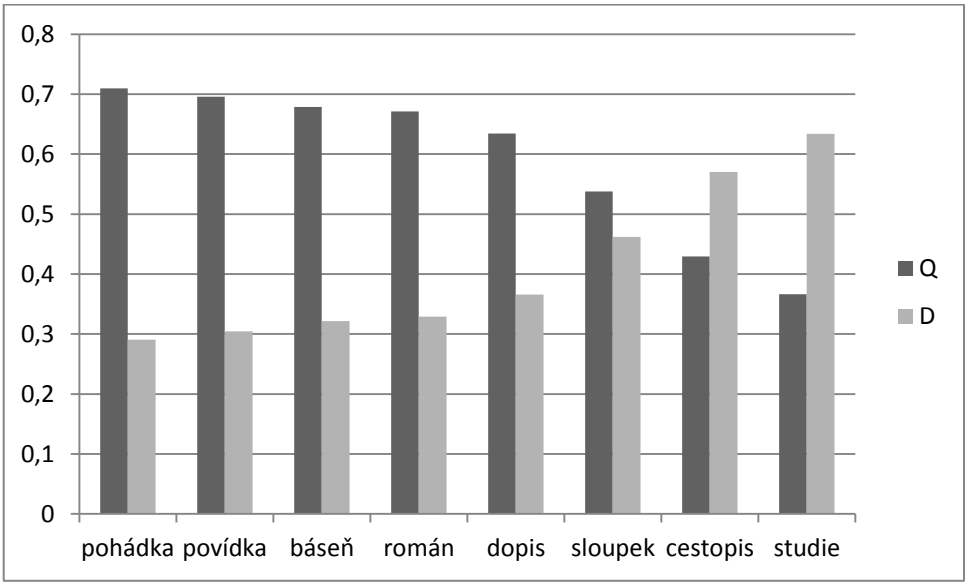
V…počet sloves
A…počet adjektiv

Výpočet ukážeme na povídce *Historie beze* slov ze sbírky *Boží muka*, kde se vyskytuje 198 sloves (*V*) a 89 adjektiv (*A*).

[95] Čech, R., Popescu, I. I., Altmann, G. (2014).

$$Q = \frac{V}{V + A} = \frac{198}{198 + 89} = 0{,}69$$

Výsledky aktivity a deskriptivity jsou zobrazeny na Obr. 48 a Obr. 49, v Tab. 23 pak najdeme hodnoty statistického testu a pro lepší přehlednost také síť na Obr. 50.



Obr. 48. Aktivita a deskriptivita v různých žánrech v Čapkových textech

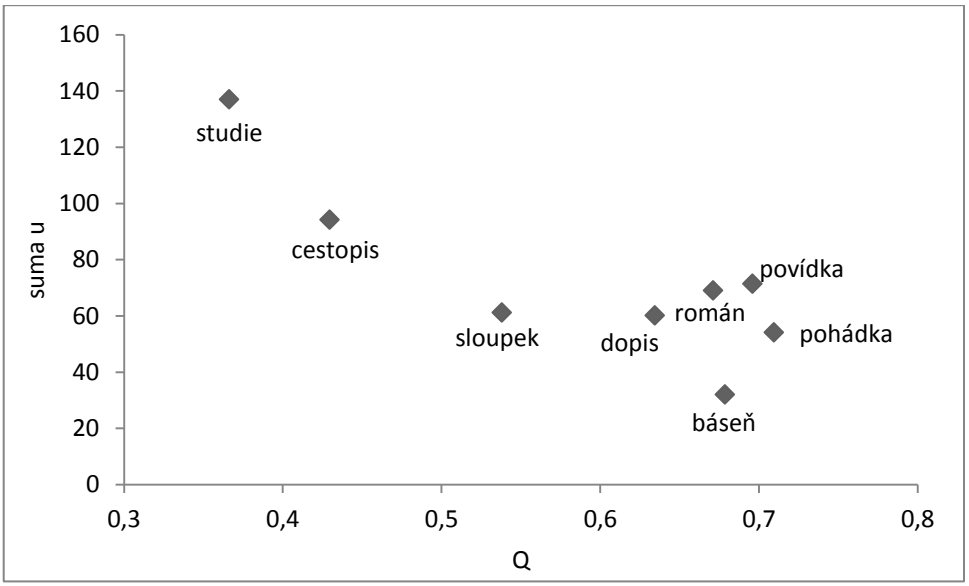
Tab. 23. Výsledky *u*-testu mezi žánry, signifikantní rozdíly ($u \geq 1{,}96$, $\alpha = 0{,}05$) jsou vyznačeny tučně

	román	povídka	cestopis	studie	sloupek	pohádka	dopis
román	x						
povídka	**2,32**	x					
cestopis	**20,06**	**20,83**	x				
studie	**31,03**	**30,73**	**5,25**	x			
sloupek	**9,71**	**10,99**	**7,04**	**12,50**	x		
pohádka	**2,34**	0,80	**15,67**	**20,86**	**9,01**	x	
dopis	**3,25**	**5,09**	**15,46**	**23,77**	**6,53**	**4,32**	x
báseň	0,31	0,70	**9,86**	**12,87**	**5,39**	1,12	1,77

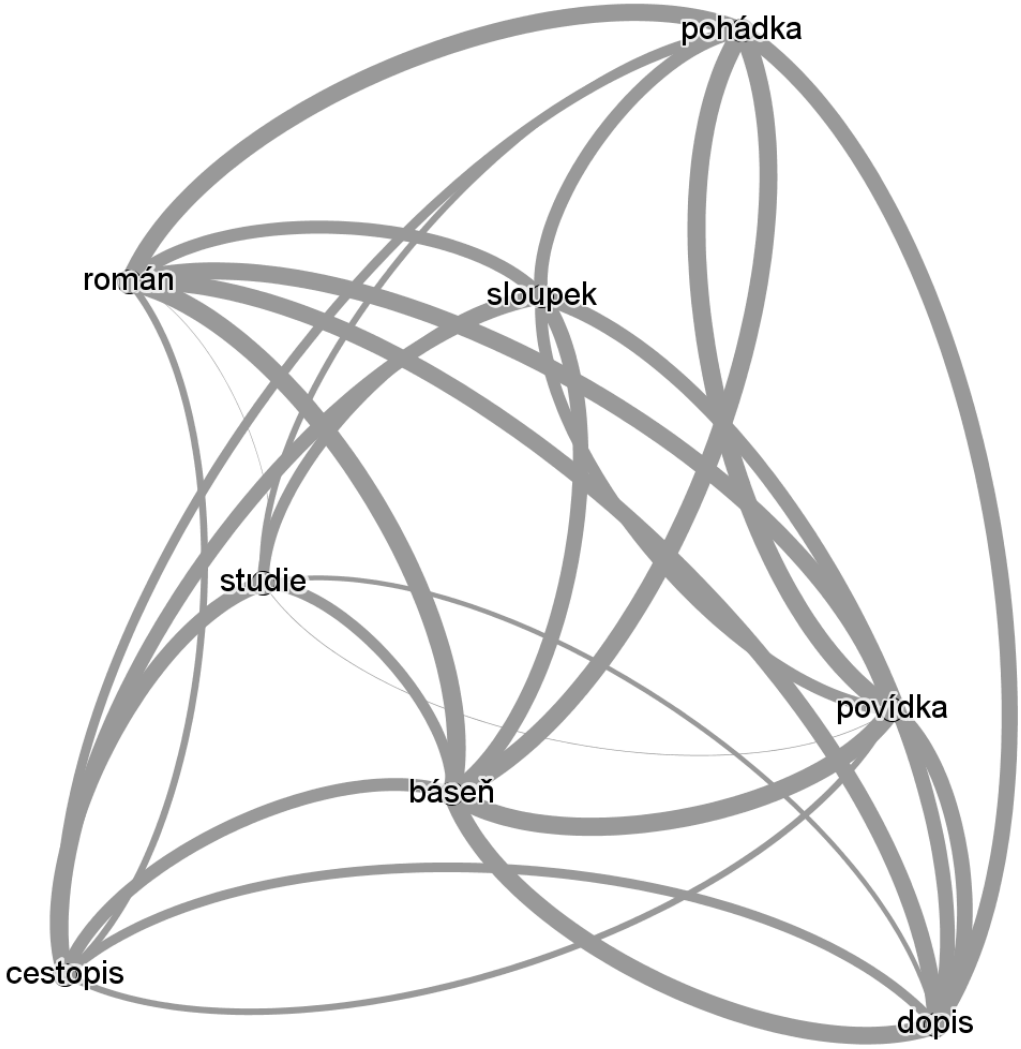
Tab. 24. Sumy hodnot *u*-testu v jednotlivých žánrech

studie	137,01
cestopis	94,17
povídka	71,47
román	69,02

sloupek	61,16
dopis	60,19
pohádka	54,12
báseň	32,02
celkem	579,16



Obr. 49. Graf zobrazující vzdálenosti jednotlivých žánrů na základě hodnot *u* a *Q*



Obr. 50. Síť zobrazující rozdíly mezi žánry v aktivitě (čím širší hrany, tím menší rozdíl)

Hodnoty na Obr. 48 rozdělují žánry na ty, kde je primární příběh, a na ty, které jsou více popisné. Tomuto rozdělení neodpovídá jen dopis, který dle výsledků inklinuje k dějovým žánrům. Zatímco tedy v pohádce, románu či povídce je třeba vyjadřovat děj pomocí sloves, studie či cestopis inklinují k větší četnosti adjektiv pro popis. Báseň pak představuje z tohoto pohledu poměrně specifický útvar, který se bude výrazně lišit u různých autorů. Na pomyslné hranici mezi dějovými a popisnými žánry figuruje sloupek, což je publicistický útvar, který není v tomto rozdělení tak vyhraněný. Z hlediska aktivity se jako nejvíce specifický žánr projevila studie, která má vůbec nejnižší hodnotu a zároveň nejvyšší hodnoty *u*. Tyto výsledky považujeme za očekávatelné, neboť právě u odborných textů je zcela potlačena

dějovost. Pokud jde o cestopis, máme za to, že i v tomto případě výsledky příliš nepřekvapily, protože cestopis na rozdíl od například románu nebo povídky je více zaměřen na popisnost než dějovost, což znamená použití více adjektiv na úkor sloves.

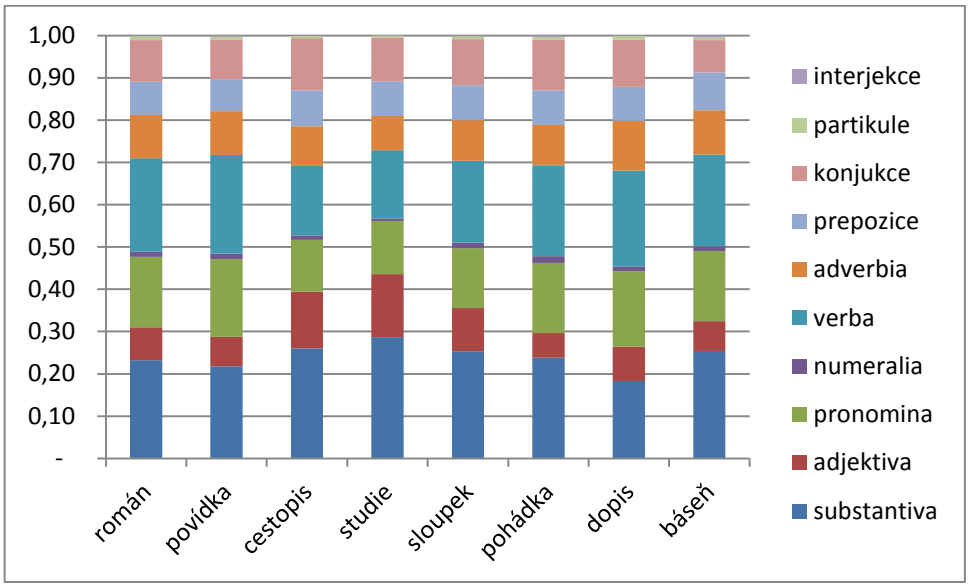
Pokud se podíváme na výsledky z hlediska síly diferenciace žánrů, lze konstatovat, že aktivita a deskriptivita jsou velmi účinné nástroje. Z 28 rozdílů je 23 statisticky významných, což představuje přes 82 %. Je tedy zřejmé, že tato charakteristika je významným indikátorem pro stylometrické bádání.

4.6. Distribuce slovních druhů

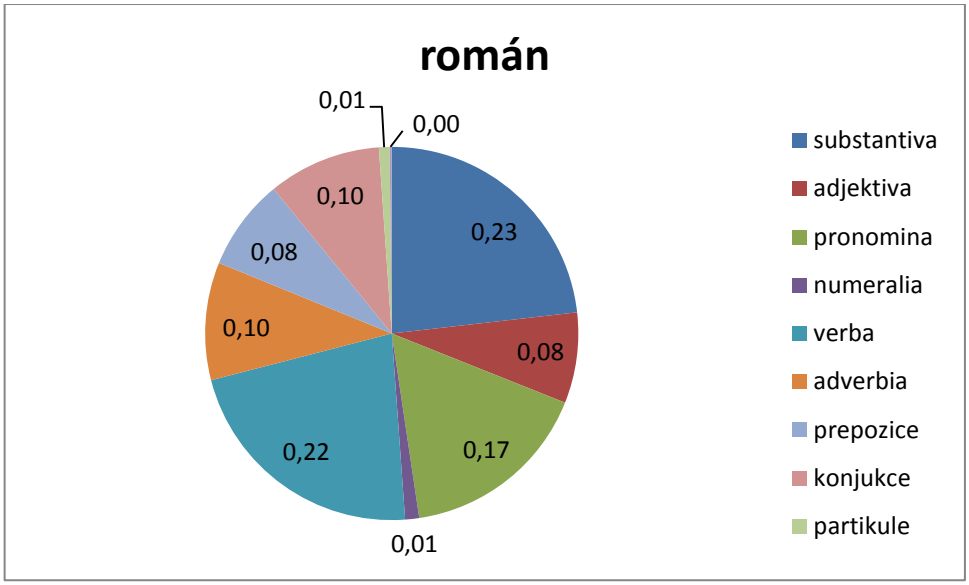
Míra užití jednotlivých slovních druhů je stylometrickou charakteristikou, která nám poměrně jednoduše umožňuje třídit texty. Výhodou těchto měření je úzké sepětí s tradiční lingvistikou a velmi jednoduchá interpretace získaných hodnot. Patrně každý má nějakou představu o rozložení jednotlivých slovních druhů v rámci různých stylů, naším úkolem je ověřit naši intuici na základě empiricky získaných dat. K těmto účelům slouží také různé indexy, kam patří například aktivita a deskriptivita textu, kterým byla věnována předchozí kapitola. Aby byl náš obraz o distribuci slovních druhů v jednotlivých žánrech co nepřesnější, změřili jsme frekvence všech slovních druhů.

Z metodologického hlediska jen poznamenáme, že ačkoliv se v této analýze držíme tradičního rozdělení 10 slovních druhů, jsme si vědomi i možnosti jiného rozdělení. Dále je třeba zmínit, že zpracování dat bylo provedeno softwarem QUITA, jehož POS tagger má úspěšnost přesahující 80% hranici, nicméně určitá chybovost může do jisté míry ovlivnit výsledky. Přesto jsme přesvědčení, že vzhledem k množství dat a totožným systémovým chybám lze získat relevantní výsledky. Pokud jde o určování slovních druhů, je třeba si také uvědomit, že ani ruční zpracování textu, které by bylo z hlediska času nesmírně náročné, by pravděpodobně nepřineslo přesnější výsledky. Dokonce máme za to, že kvůli nesystémovým lidským chybám bychom získali nepříliš vhodná data pro náš účel porovnávání jednotlivých žánrů.

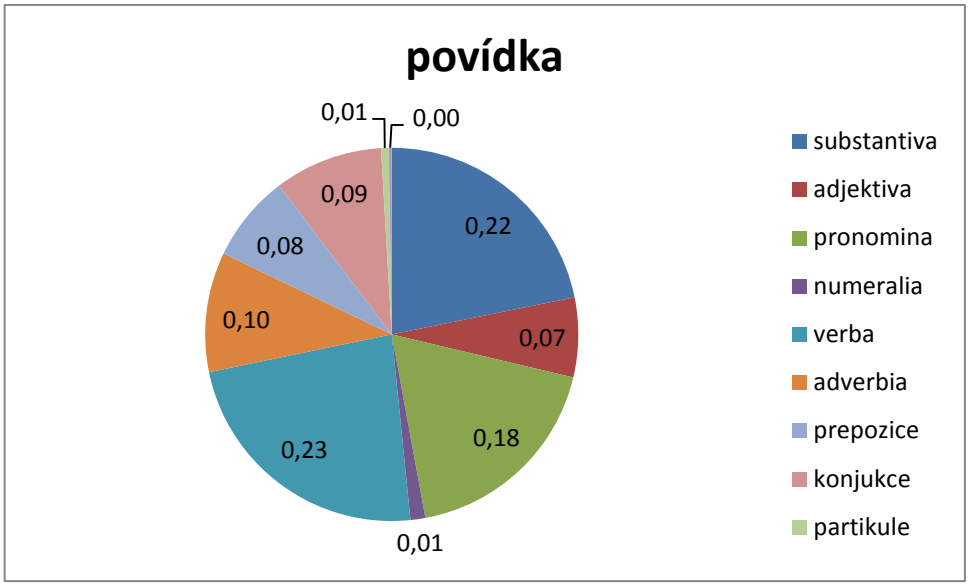
Aby byly získané hodnoty co nejpřehlednější, vytvořili jsme na následujících obrázcích několik grafů, které zachycují jak celkové přehledy, tak detailní výsledky jednotlivých žánrů.



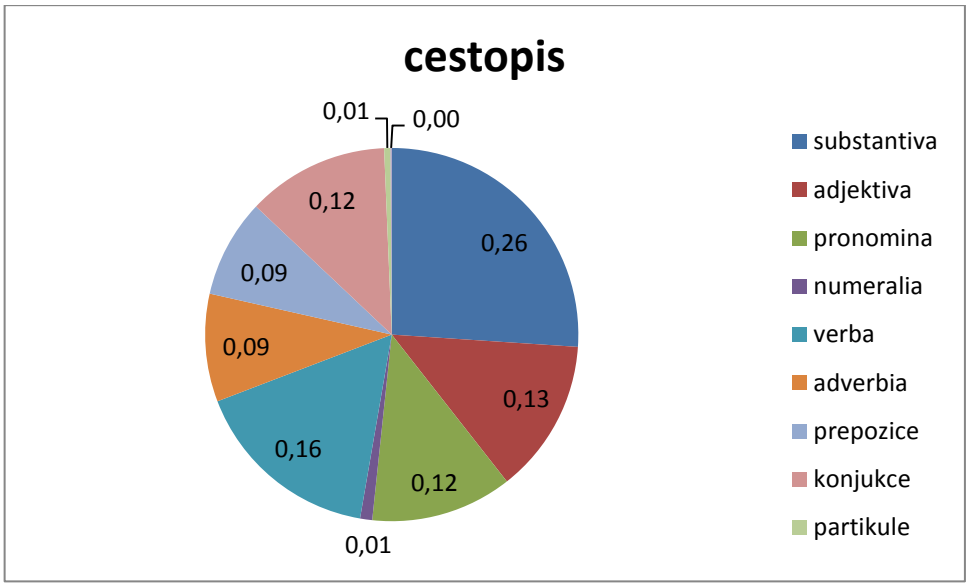
Obr. 51. Přehled distribuce slovních druhů v různých žánrech v Čapkových textech



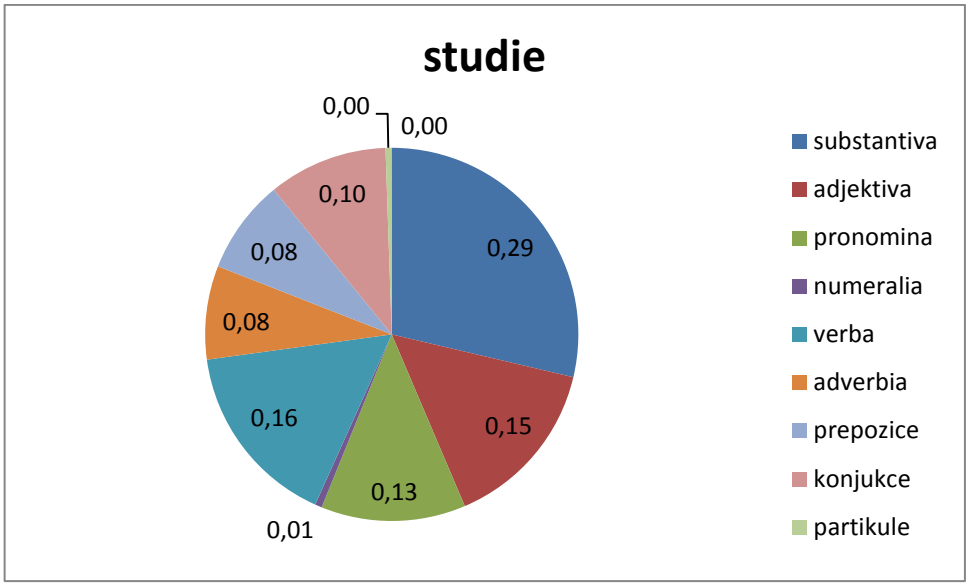
Obr. 52. Proporce slovních druhů v románu



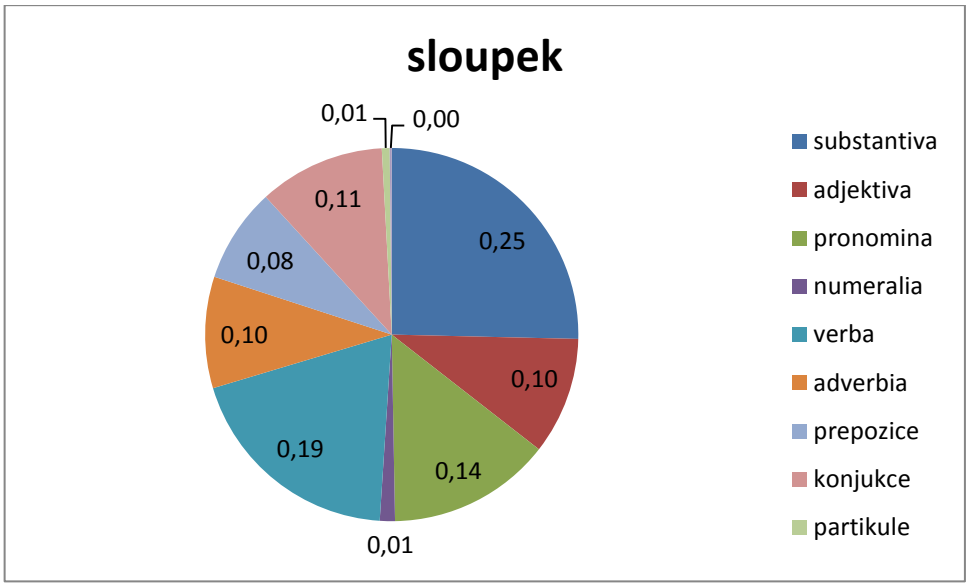
Obr. 53. Proporce slovních druhů v povídce



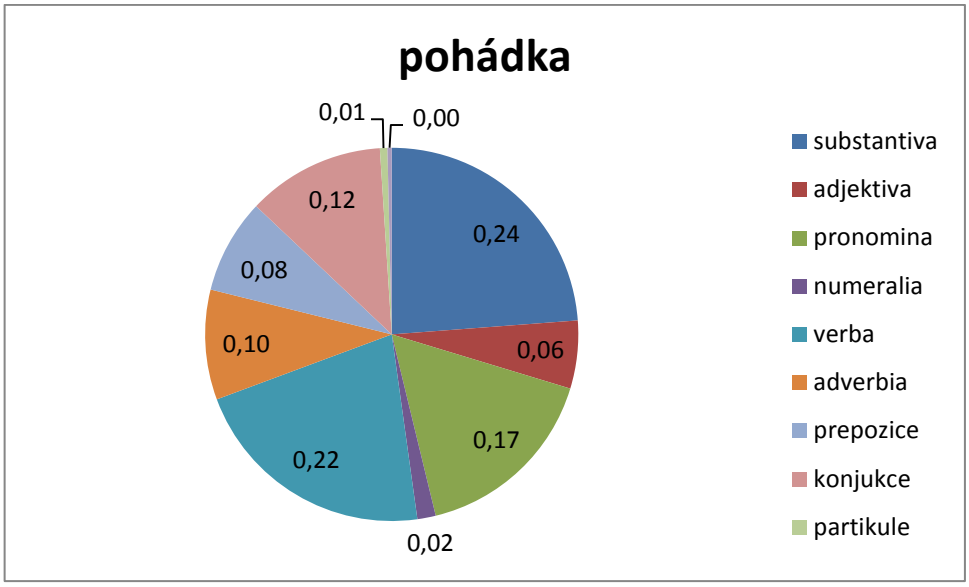
Obr. 54. Proporce slovních druhů v cestopisu



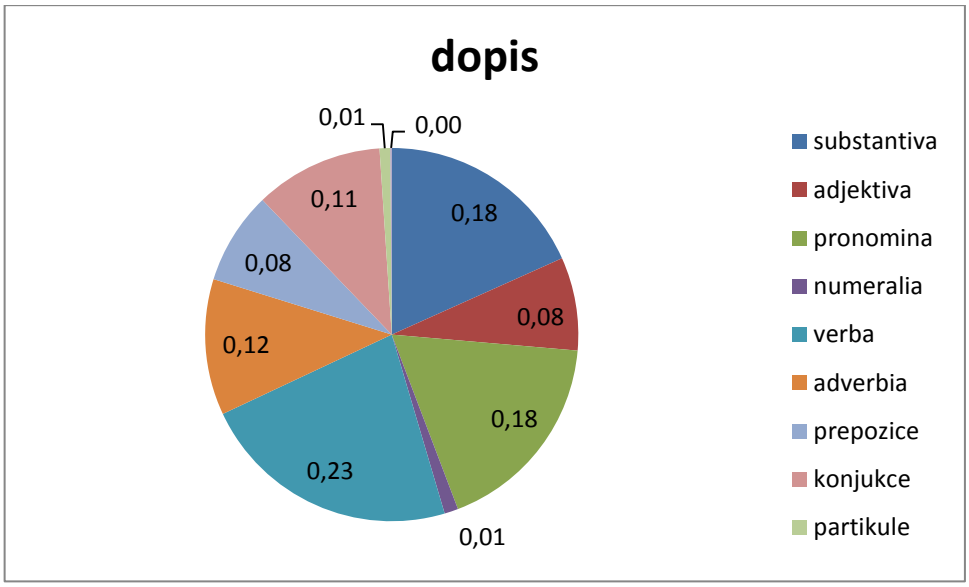
Obr. 55. Proporce slovních druhů ve studii



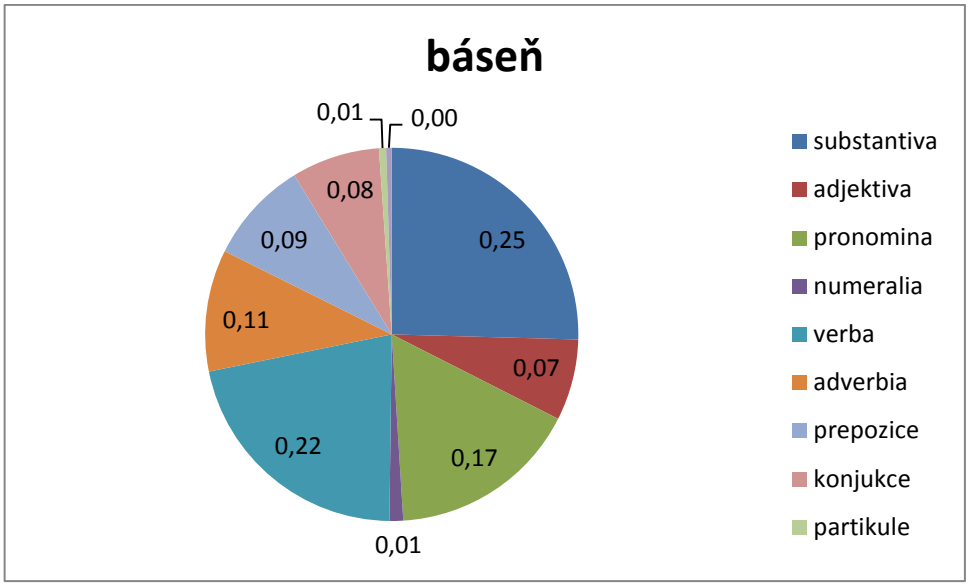
Obr. 56. Proporce slovních druhů ve sloupku



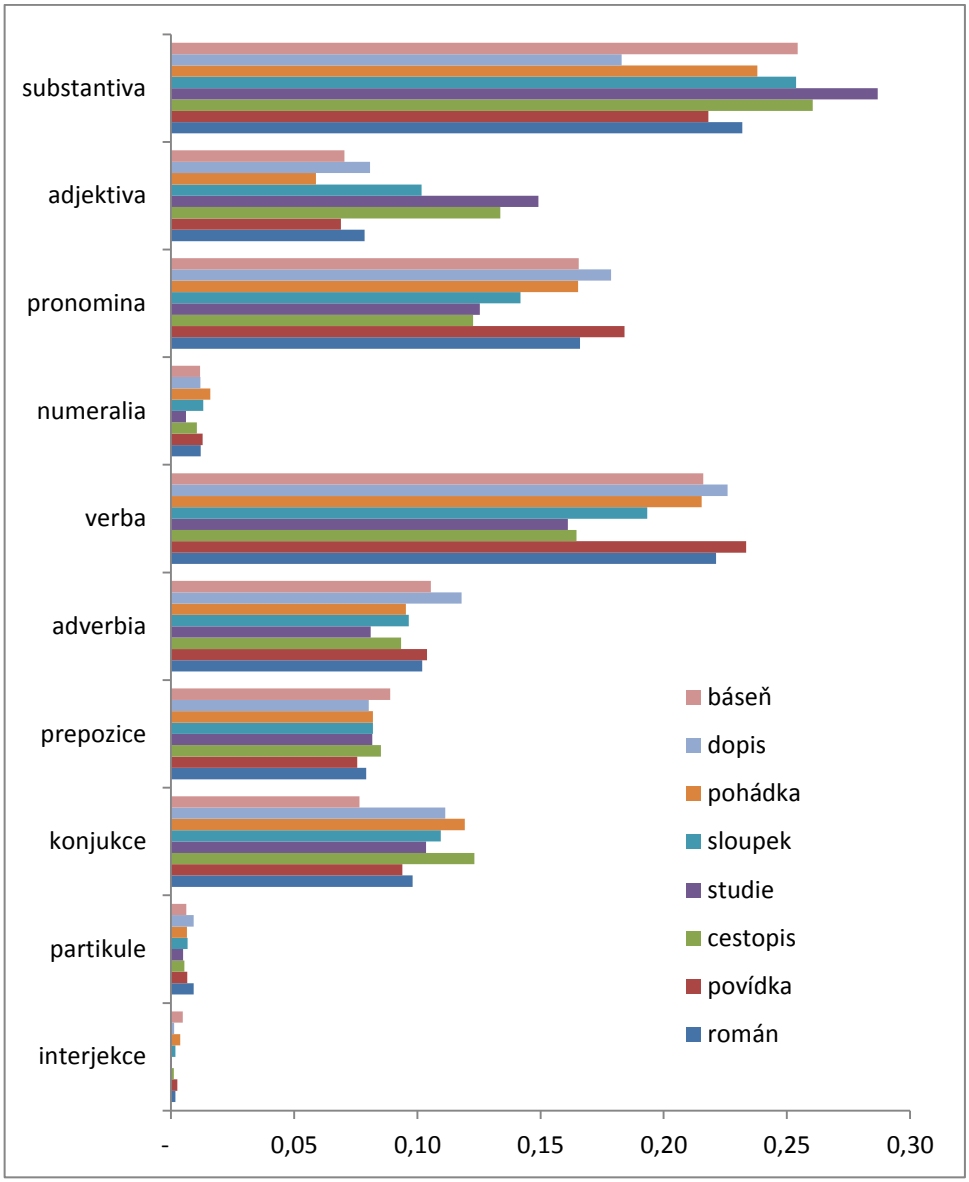
Obr. 57. Proporce slovních druhů v pohádce



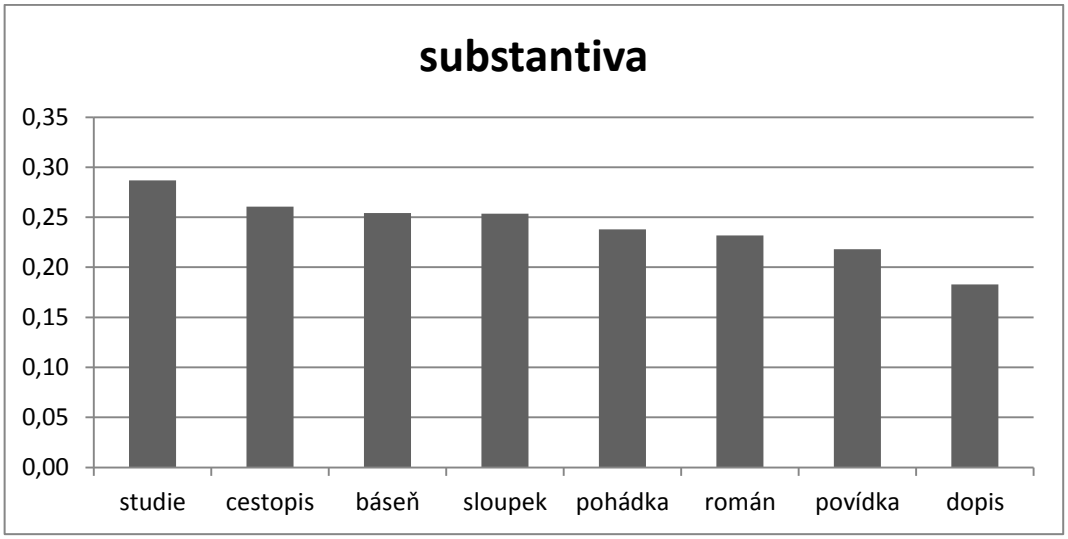
Obr. 58. Proporce slovních druhů v dopisu



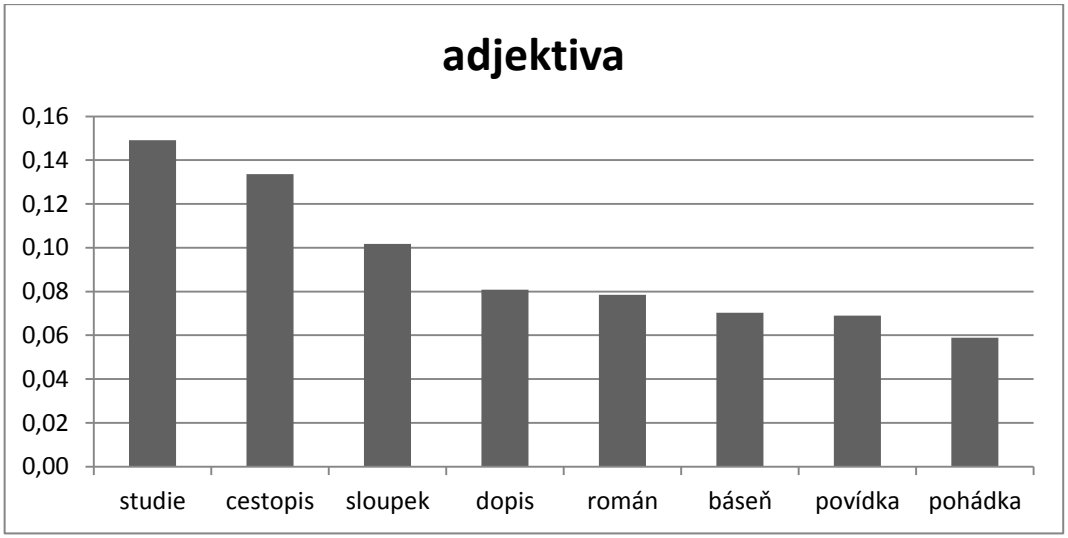
Obr. 59. Proporce slovních druhů v básni



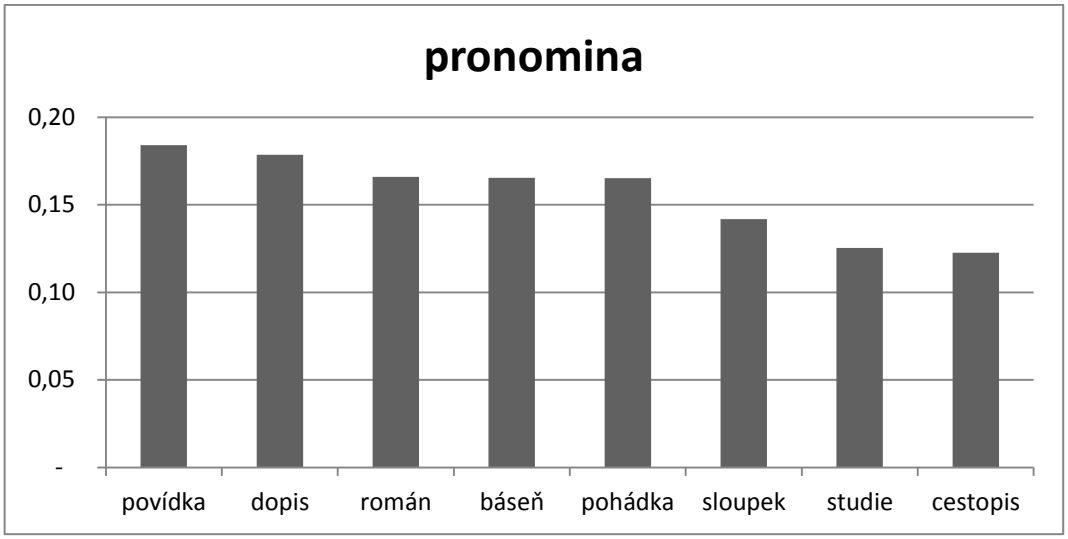
Obr. 60. Proporce slovních druhů v různých žánrech v Čapkových textech



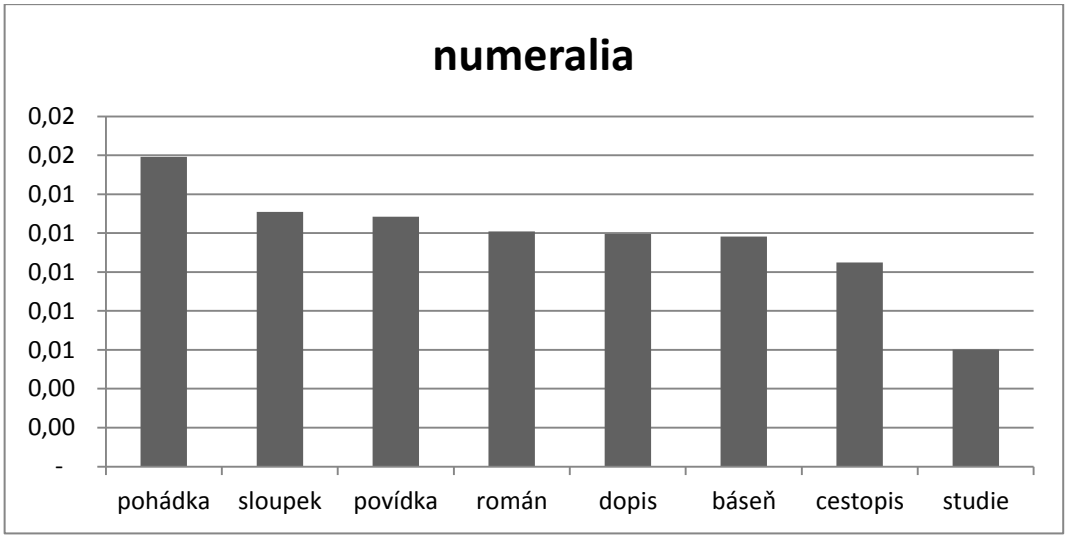
Obr. 61. Proporce substantiv v žánrech



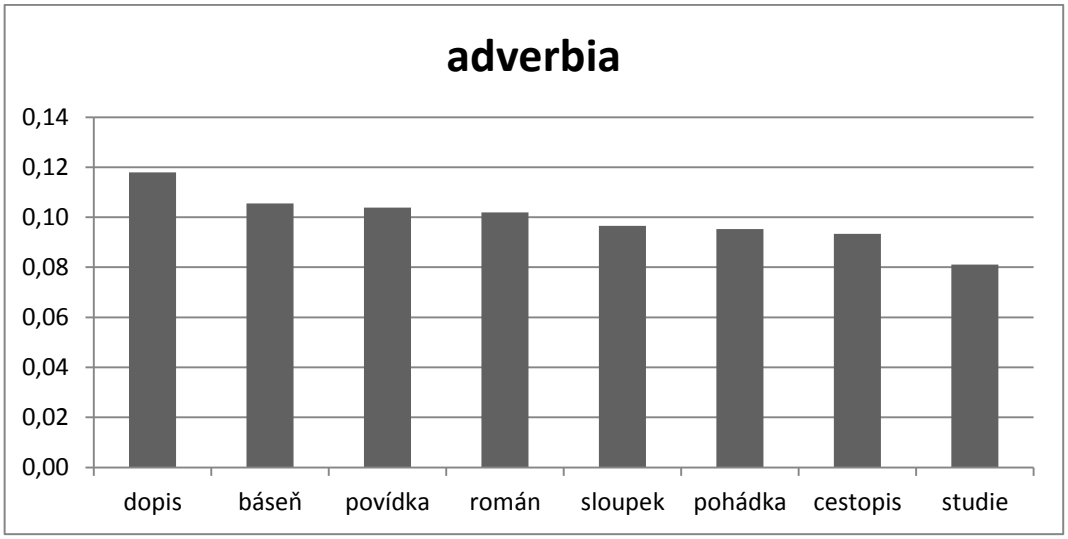
Obr. 62. Proporce adjektiv v žánrech



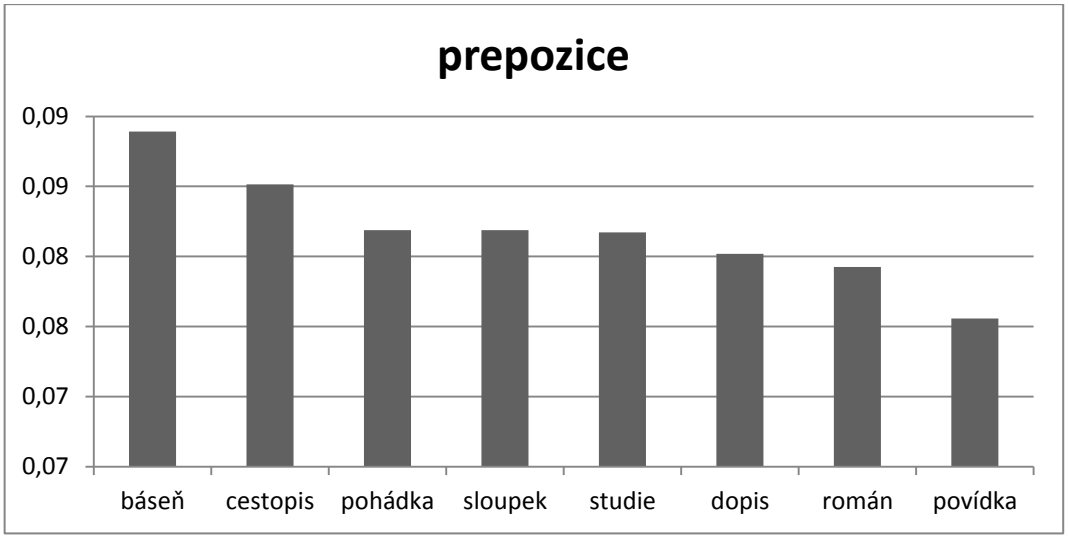
Obr. 63. Proporce zájmen v žánrech



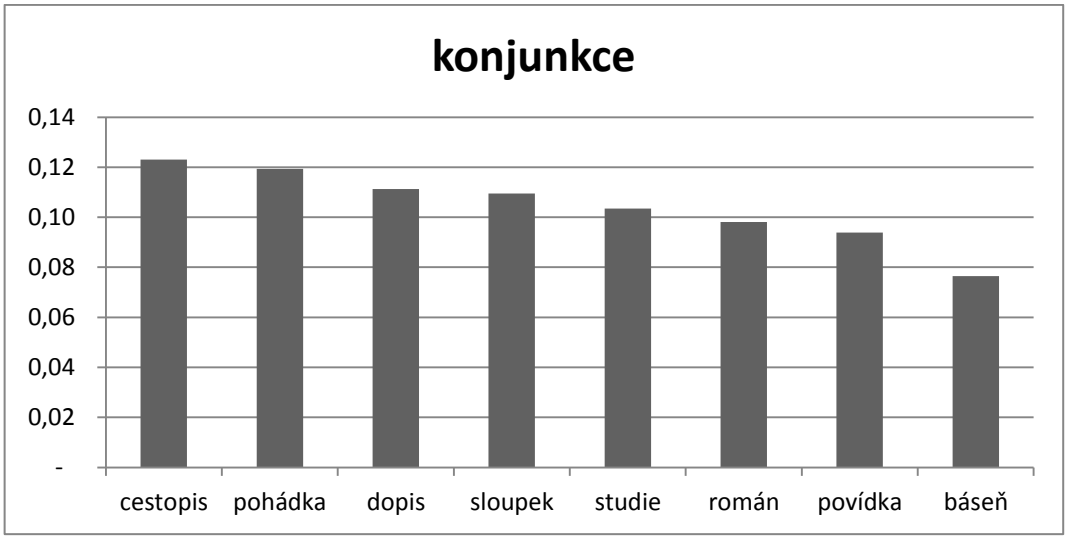
Obr. 64. Proporce číslovek v žánrech



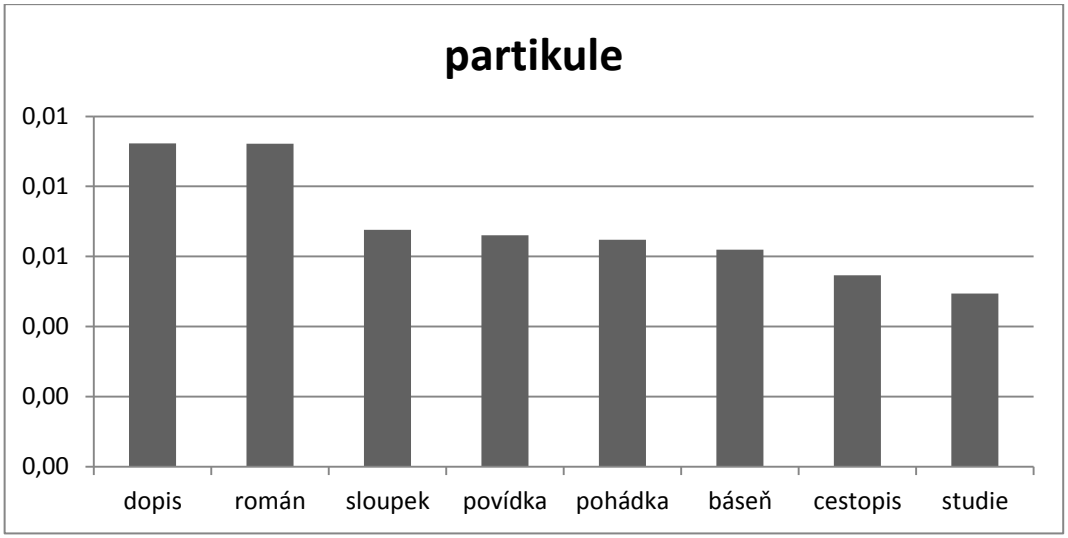
Obr. 65. Proporce příslovcí v žánrech



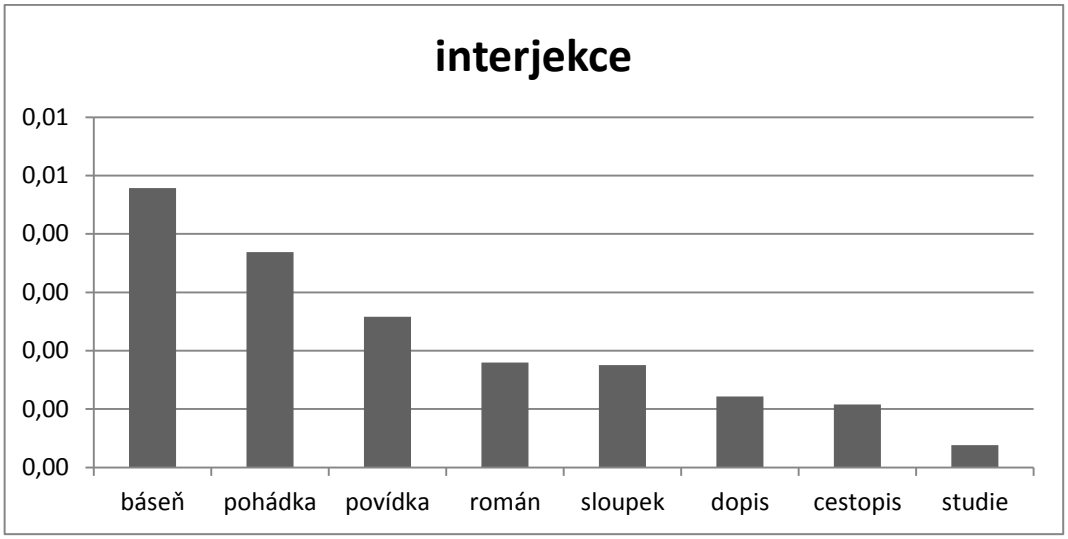
Obr. 66. Proporce předložek v žánrech



Obr. 67. Proporce spojek v žánrech



Obr. 68. Proporce částic v žánrech



Obr. 69. Proporce citoslovcí v žánrech

Ze získaných distribucí slovních druhů jsou patrné jisté diference, nicméně pro porovnání jednotlivých žánrů je třeba použít přesnější metodu. Pro tento účel lze aplikovat x^2 diskrepanční koeficient (C), který se zpravidla užívá pro testování shody měřené distribuce s konkrétním rozdělením.[96] Za hraniční hodnotu C pro stanovení rozdílu jsme určili 0,05. Výsledky jsou uvedeny v Tab. 25, v Tab. 26 pak najdeme sumy hodnot C v jednotlivých žánrech.

Tab. 25. Porovnání distribucí slovních druhů pomocí C ($C \geq 0{,}05$ znamená, že se distribuce liší)

	román	povídka	cestopis	studie	sloupek	pohádka	dopis
román	x						
povídka	0,002516	x					
cestopis	0,023469	**0,052394**	x				
studie	0,027968	**0,063726**	0,006164	x			
sloupek	0,005376	0,017176	0,009322	0,020253	x		
pohádka	0,002862	0,005739	0,039391	**0,062982**	0,015551	x	
dopis	0,003123	0,005503	0,045179	**0,070539**	0,022395	0,0164	x
báseň	0,000293	0,000795	0,006895	0,012718	0,003356	0,002961	0,00736

[96] Srov. Mačutek, J., Wimmer, G. (2013).

Tab. 26. Sumy hodnot *C* v jednotlivých žánrech

studie	0,26435
cestopis	0,182814
dopis	0,1705
povídka	0,147849
pohádka	0,145886
sloupek	0,09343
román	0,065607
báseň	0,034378
celkem	1,104814

Než přejdeme k interpretaci získaných hodnot, jen uvedeme, že problematika distribuce adjektiv a verb byla podrobně popsána v kapitole 4.5 *Aktivita a deskriptivita*, proto se těmto slovním druhům nebudeme již dále věnovat.

Pokud se podíváme na frekvence substantiv, zjistíme, že výrazněji se odlišují zejména studie a dopis. Zatímco studie na prvním místě dosáhla 29% zastoupení podstatných jmen, poslední dopis má 18 %. Tyto výsledky lze vysvětlit tím, že odborné texty jsou popisné a věcné, dějovost je zcela potlačena. Tomu odpovídá i vysoká míra deskriptivity studie (viz předchozí kapitolu). Stěžejní slovní druhy odborné studie jsou tedy substantiva a adjektiva. Dopis je mnohem méně věcný, protože osobní korespondence tíhne zejména k vyjádření emocí, upevnění vztahů apod. Klíčovou roli hraje fatická funkce komunikace, tudíž není třeba tolik používat podstatná jména. O co nižší je četnost substantiv, o to vyšší je v dopisu četnost zájmen, neboť v osobní korespondenci nejsou předmětem zájmu ani tak věci, jako spíše lidé, ke kterým odkazujeme právě osobními zájmeny.

Distribuce zájmen poměrně jasně odpovídá naší intuici. Žánry, kde je pozornost soustředěna na jednotlivé postavy, vykazují vyšší četnost zájmen. Nepřekvapí tak, že povídka, román a dopis se umístili na čelních místech, zatímco studie a cestopis dosáhly nejmenšího zastoupení zájmen.

Adverbia, prepozice, konjunkce, numeralia, partikule a interjekce jsou obecně slovní druhy s poměrně nízkou četností, ne jinak je tomu i v našem korpusu. Vzhledem k malému intervalu, kde se relativní frekvence pohybuje v maximálním rozmezí 0,04, považujeme vyvozování jakýchkoliv závěrů za nepodložené. Grafy na Obr. 64, Obr. 66, Obr. 67 a Obr. 68 tak slouží spíše ilustračně.

Na základě získaných dat tak můžeme pouze konstatovat, že tyto slovní druhy nehrají významnou roli v žánrové klasifikaci, a to jak z důvodu nízkých frekvencí, tak z důvodu malého intervalu.

Pokud bychom měli zhodnotit obecně distribuci slovních druhů v rámci diferenciace žánrů, můžeme na základě získaných výsledků konstatovat, že reálně použitelná jsou pro tento účel pouze substantiva, adjektiva a verba. Ostatní slovní druhy dosahují příliš nízkých relativních frekvencí na to, abychom z nich mohli odvozovat relevantní závěry. Celkově můžeme konstatovat, že měření distribuce slovních druhů se ukazuje z hlediska klasifikace žánrů jako neefektivní metoda.

4.7. N-gramy

N-gramy jsou sekvence obsahující *n* prvků v daném textu, nejčastěji jsou jako jednotky používány grafémy nebo slova, každý si však může zvolit jakékoliv jiné (např. fonémy, slabiky). Přestože lze n-gramy aplikovat na libovolné jazykové jednotky, je třeba poznamenat, že u jednotek, jakými jsou například hreby (hreb, původně označovaný jako agregát, je jazyková jednotka pojmenovaná po jeho autorovi Luďku Hřebíčkovi; hreby jsou jednotky, které odkazují k jedné sémantické entitě)[97] by bylo přinejmenším značně komplikované, nicméně i u těchto jednotek lze teoreticky n-gramy použít. Stejně jako je zcela arbitrární výběr jednotek, tak i délka sekvencí je zcela libovolná (bigramy, trigramy atd.). Například věta „Toto je text." sestává z následujících grafémových bigramů [To], [ot], [to], [o_], [_j], [je], [e_], [_t], [te], [ex], [xt], [t.].

Zásadní nevýhodou n-gramů je skutečnost, že tyto umělé jednotky nemají žádnou oporu v tradičních lingvistických popisech. N-gramy však poskytují oproti tradičním jednotkám také určité výhody, a to především jednoduchou a jednoznačnou segmentaci textu a použitelnost ve většině jazyků. Vůbec nejzásadnější předností n-gramů jsou však výsledky, jakých v současnosti dosahují ve svém výzkumu zaměřeném zejména na určování autorství Mikros a Perifanos.[98] Přesnost přiřazení jednotlivých textů k jejich autorům přesahuje hranici 90 %,[99] což činí z n-gramů pravděpodobně nejspolehlivější nástroj v současné stylometrii. V této kapitole se pokusíme zodpovědět otázku, zda jsou tyto jednotky vhodné také pro diferenciaci žánrů.

Při analýze vyjdeme z výzkumu Mikrose.[100] Konkrétně aplikujeme jeho víceúrovňový autorský profil *AMNP* (Author's Multilevel N-gram Profile), který kombinuje bigramy a trigramy grafémů a slov. Výhodou tohoto modelu je, že zahrnuje více jazykových rovin, konkrétně sémantiku, syntax, morfologii a fonologii (viz Obr. 70). Jakkoliv se *AMNP* může jevit jako příliš uměle vytvořený konstrukt,

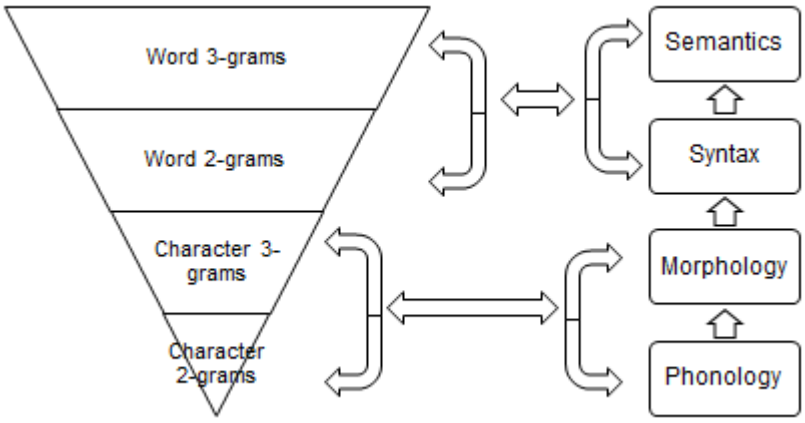
[97] Viz např.:
Hřebíček, L. (1992).
Hřebíček, L. (1995).
Ziegler, A., Altmann, G. (2002).

[98] Viz např. Mikros, G. K., Perifanos, K. (2013).

[99] Viz Mikros, G. K., Perifanos, K. (2013).

[100] Mikrosovi tímto děkujeme za poskytnutí potřebného softwaru a všestrannou pomoc při AMNP analýze.

který je založen na netradičních jednotkách a jejich kombinaci, domníváme se, že jde o významnou charakteristiku textu. Chtěli bychom však uvést, že jsme si zcela vědomi veškerých nedostatků a zjednodušení, které *AMNP* z lingvistického hlediska má. Nicméně pokud projdeme jednotlivé úrovně daného modelu, nelze nevidět jejich souvislosti s jednotlivými jazykovými rovinami. Důležité je také zmínit, že do n-gramů se započítává i interpunkce, což může opět vyvolat mnohé rozpaky. Samozřejmě bude nezbytné aplikovat *AMNP* na texty různých jazyků v dalších výzkumech, ale doposud získané výsledky dokazují, že tento model je s to velice přesně klasifikovat styl textu, a tudíž považujeme za nesprávné odmítat *AMNP* jen kvůli nejrůznějším zdánlivým či skutečným lingvistickým nedostatkům. Zde odkazujeme na kapitolu 3.2 *Jazykové jednotky*, kde se zabýváme problematikou tradičních a nových (umělých) jazykových jednotek.



Obr. 70. Víceúrovňový autorský profil AMNP dle Mikrose[101]

Základními daty pro výpočet jsou relativní frekvence 200 nejčetnějších n-gramů každé kategorie (slovní a grafémové bigramy a trigramy) z celého korpusu, tj. 800 celkem. Tato data jsou použita pro následné vyhodnocení, které lze realizovat dvěma statistickými metodami: Random Forest (*RF*) a Support Vector Machines (*SVM*).[102]

[101] Mikros, G. K.. (2013).
[102] Pro více informací o těchto metodách viz např. Berka, P. (2003).

Výsledky *AMNP* analýzy jsou uvedeny v Tab. 27, Tab. 28, Tab. 29 a Tab. 30, přičemž jsou u každé metody (*RF* a *SVM*) zobrazeny jak absolutní, tak relativní hodnoty. V uvedených tabulkách jsou v každém sloupci zobrazeny počty textů daného žánru. Podle hodnoty v konkrétním řádku poznáme, ke kterému žánru byly jednotlivé texty přiřazeny. Např. v Tab. 27 jsou v prvním sloupci všechny texty románu (celkem 252 textů), při pohledu na jednotlivé řádky zjistíme, že 245 románových textů bylo modelem *AMNP* správně klasifikováno jako román, dále 4 románové texty byly chybně rozpoznány jako cestopis, 1 text jako sloupek a 2 texty jako dopis. V Tab. 28 pak jsou tyto absolutní hodny vyjádřeny procenty. Zjistíme tak, že model *AMNP* např. správně určil 97,22 % románových textů.

Tab. 27. Predikce žánrů v Čapkových textech v absolutních hodnotách dle *RF*

	román	povídka	cestopis	studie	sloupek	pohádka	dopis	báseň
román	**245**	45	16	0	19	20	5	1
povídka	0	**25**	0	0	0	0	0	0
cestopis	4	1	**107**	0	19	2	3	4
studie	0	0	1	**70**	0	0	0	0
sloupek	1	0	3	0	**50**	4	2	4
pohádka	0	0	0	0	0	**0**	0	0
dopis	2	0	3	0	4	0	**83**	7
báseň	0	0	2	0	0	0	0	**8**

Tab. 28. Predikce žánrů v Čapkových textech v procentech dle *RF*

	román	povídka	cestopis	studie	sloupek	pohádka	dopis	báseň
román	**97,22**	63,38	12,12	0	20,65	76,92	5,38	4,17
povídka	0	**35,21**	0	0	0	0	0	0
cestopis	1,59	1,4	**81,06**	0	20,65	7,69	3,23	16,67
studie	0	0	0,76	**100**	0	0	0	0
sloupek	0,4	0	2,27	0	**54,35**	15,38	2,15	16,67
pohádka	0	0	0	0	0	**0**	0	0
dopis	0,79	0	2,27	0	4,35	0	**89,25**	29,17
báseň	0	0	1,52	0	0	0	0	**33,34**

Přesnost klasifikace dané metody můžeme jednoduše vyjádřit poměrem počtu správně klasifikovaných textů ke všem textům.

$$přesnost\ predikce\ (RF) = \frac{počet\ správných\ klasifikací}{počet\ všech\ případů} = \frac{588}{760} = 0{,}77$$

Tab. 29. Predikce žánrů v Čapkových textech v absolutních hodnotách dle *SVM*

	román	povídka	cestopis	studie	sloupek	pohádka	dopis	báseň
Román	**240**	16	6	0	11	5	6	8
Povídka	6	**54**	0	0	0	1	0	1
Cestopis	4	1	**118**	0	14	0	4	6
Studie	0	0	0	**70**	0	0	0	0
Sloupek	1	0	7	0	**64**	8	2	5
Pohádka	0	0	0	0	0	**12**	0	0
Dopis	1	0	1	0	3	0	**81**	3
Báseň	0	0	0	0	0	0	0	**1**

Tab. 30. Predikce žánrů v Čapkových textech v procentech dle *SVM*

	román	povídka	cestopis	studie	sloupek	pohádka	dopis	báseň
Román	**95,24**	22,54	4,55	0	11,96	19,23	6,45	33,33
Povídka	2,38	**76,06**	0	0	0	3,85	0	4,17
Cestopis	1,59	1,41	**89,39**	0	15,22	0	4,3	25
Studie	0	0	0	**100**	0	0	0	0
Sloupek	0,4	0	5,3	0	**69,57**	30,77	2,15	20,83
Pohádka	0	0	0	0	0	**46,15**	0	0
Dopis	0,4	0	0,76	0	3,26	0	**87,1**	12,5
Báseň	0	0	0	0	0	0	0	**4,17**

$$přesnost\ predikce\ (SVM) = \frac{počet\ správných\ klasifikací}{počet\ všech\ případů} = \frac{640}{760} = 0{,}84$$

Přesnost predikce přesahující hranici osmdesáti procent považujeme s přihlédnutím ke všem specifikům našeho korpusu (jeden autor, různé žánry,

segmentace textů apod., viz kapitolu 3.3 *Korpus*) za velmi vysokou. Z hlediska přesnosti predikce můžeme také konstatovat, že se potvrzuje vyšší účinnost statistické metody Support Vector Machines (84 %) oproti Random Forest (77 %).[103]

Pokud se nespokojíme s „pouhou" automatickou klasifikací textů s vysokou přesností a podíváme se detailněji na získané výsledky, můžeme pozorovat poměrně velké rozdíly v úspěšnosti predikce u jednotlivých žánrů. Pro přehlednost uvádíme v Tab. 31 a Tab. 32 žánry seřazené sestupně podle přesnosti predikce.

Tab. 31. Žánry seřazené podle přesnosti predikce (*RF*)

Random Forest	
Žánr	Přesnost predikce v %
studie	100
román	97,22
dopis	89,25
cestopis	81,06
sloupek	54,35
povídka	35,21
báseň	33,34
pohádka	0

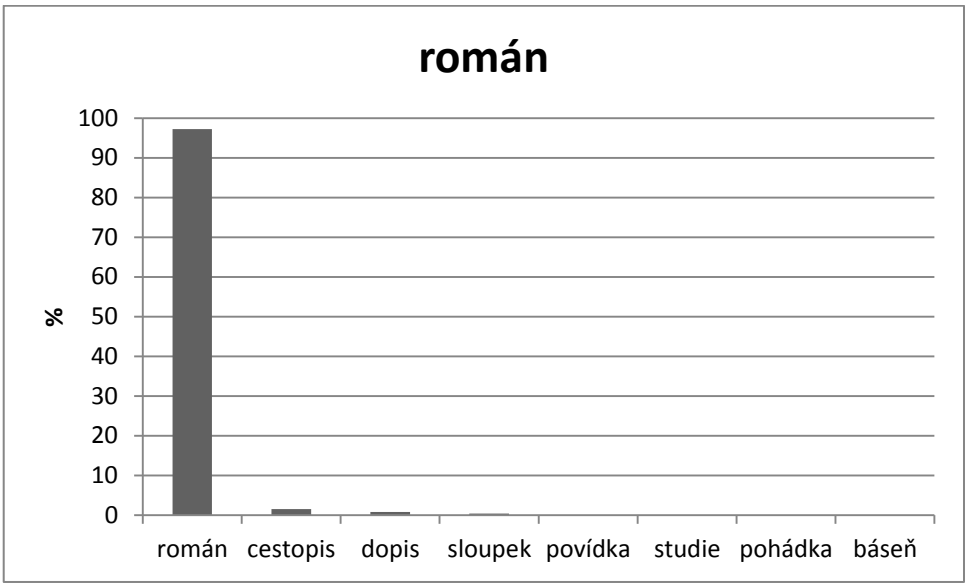
Tab. 32. Žánry seřazené podle přesnosti predikce (*SVM*)

Support Vector Machines	
Žánr	Přesnost predikce v %
studie	100
román	95,24
cestopis	89,39
dopis	87,1
povídka	76,06
sloupek	69,57
pohádka	46,15
báseň	4,17

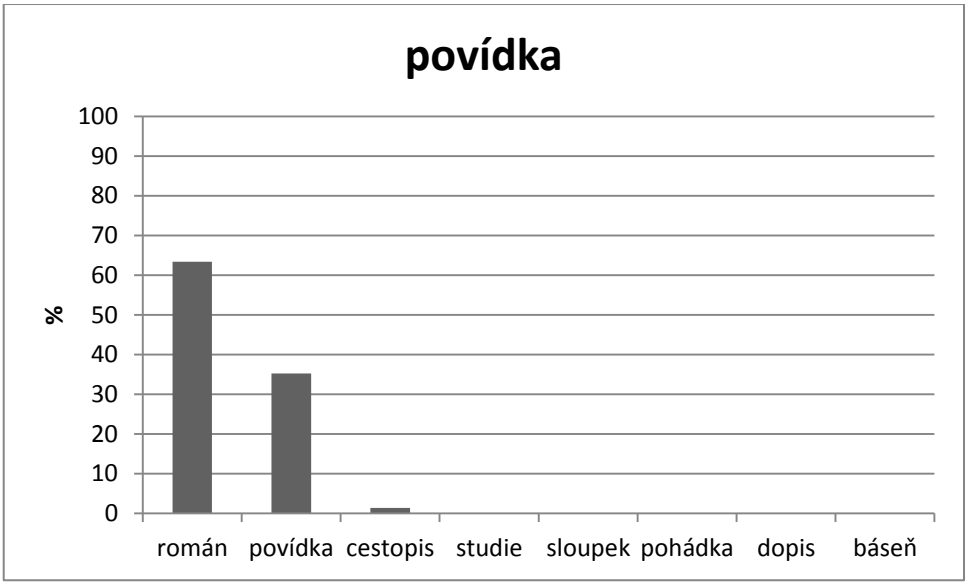
Při pohledu na Tab. 31 a Tab. 32 je patrné specifické postavení studie, která dosáhla u obou metod stoprocentní úspěšnosti klasifikace. Tento žánr se tak jeví z hlediska *AMNP* jako zcela odlišný od ostatních. S jistým odstupem studii následuje

[103] Srov. Mikros, G. K., Perifanos, K. (2013).

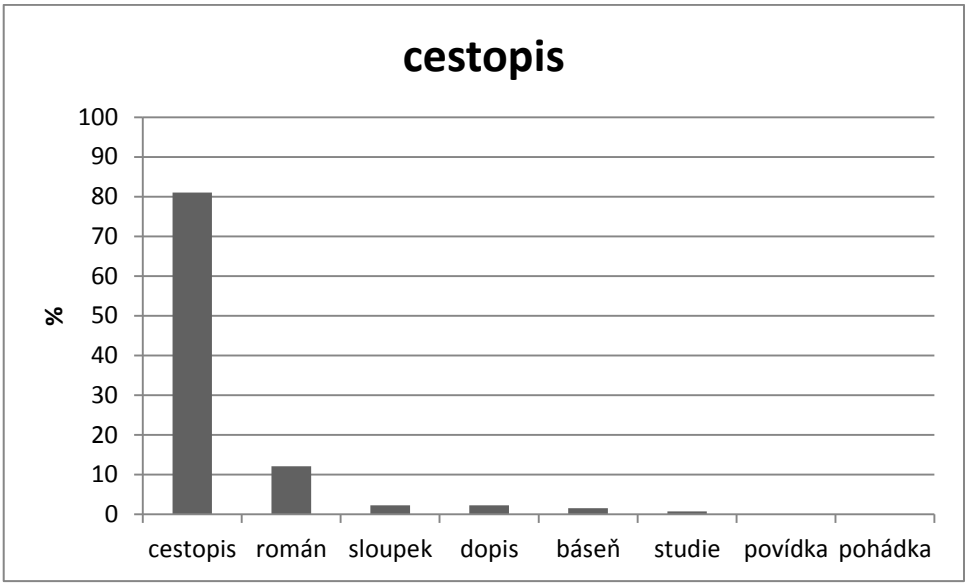
román, cestopis a dopis. Na opačném konci pořadí se umístila pohádka a báseň. Abychom více porozuměli vzájemným vztahům mezi žánry, uvádíme v následujících grafech detailní výsledky každého žánru.



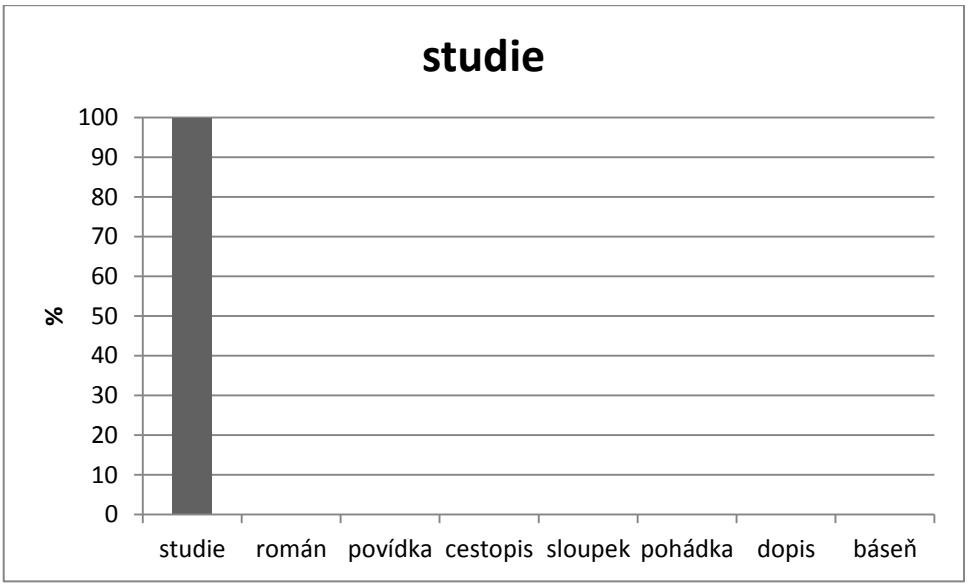
Obr. 71. Automatické přiřazení románu k jednotlivým žánrům dle *RF*



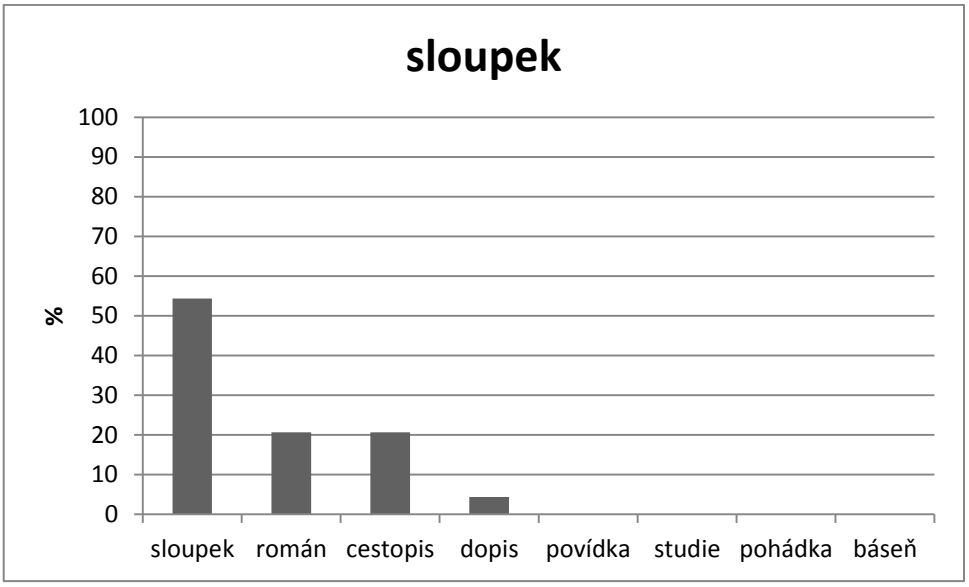
Obr. 72. Automatické přiřazení povídky k jednotlivým žánrům dle *RF*



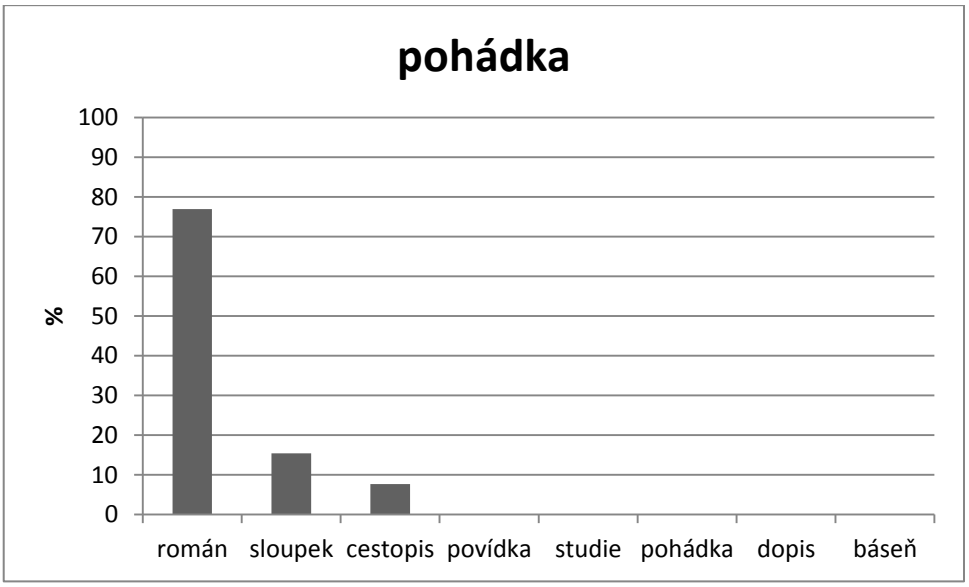
Obr. 73. Automatické přiřazení cestopisu k jednotlivým žánrům dle *RF*



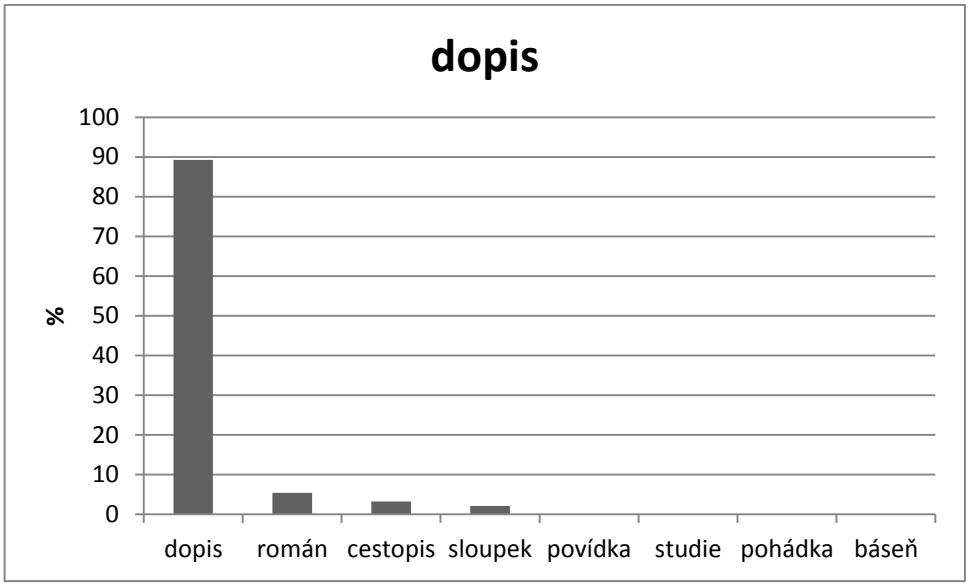
Obr. 74. Automatické přiřazení studie k jednotlivým žánrům dle *RF*



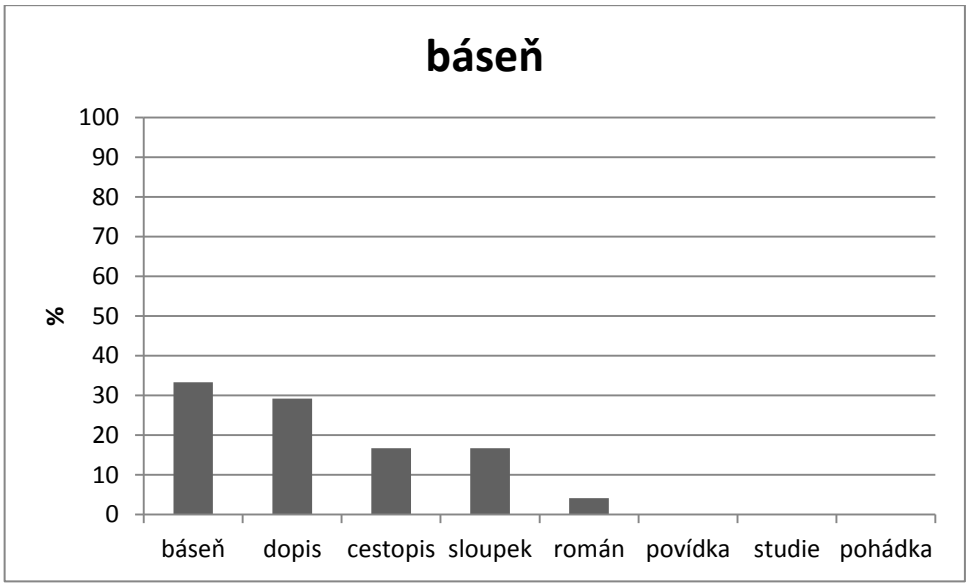
Obr. 75. Automatické přiřazení sloupku k jednotlivým žánrům dle *RF*



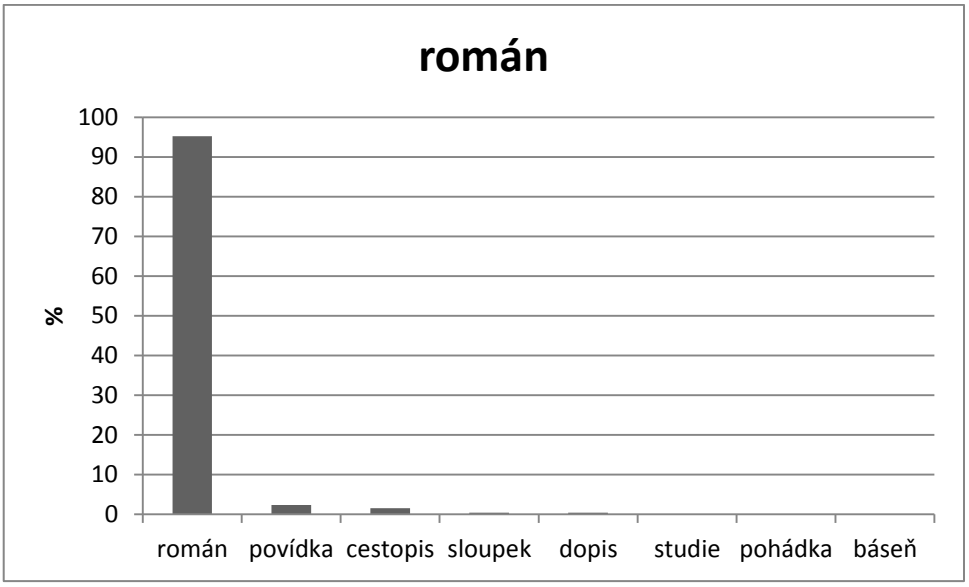
Obr. 76. Automatické přiřazení pohádky k jednotlivým žánrům dle *RF*



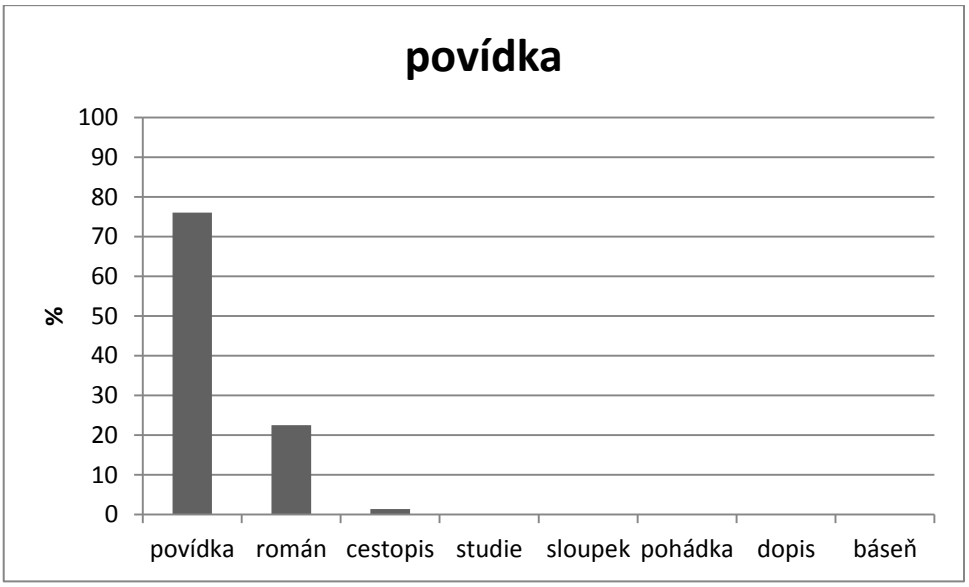
Obr. 77. Automatické přiřazení dopisu k jednotlivým žánrům dle *RF*



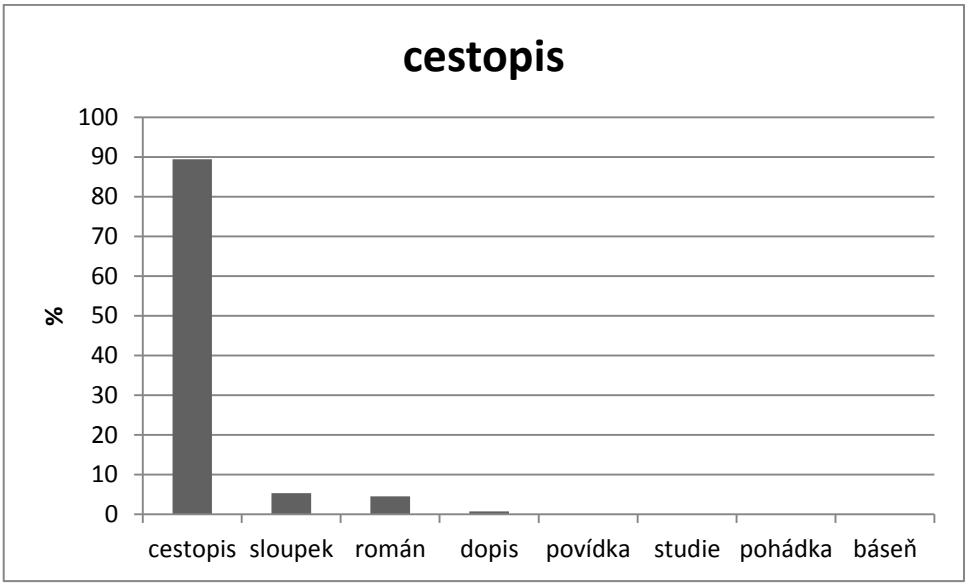
Obr. 78. Automatické přiřazení básně k jednotlivým žánrům dle *RF*



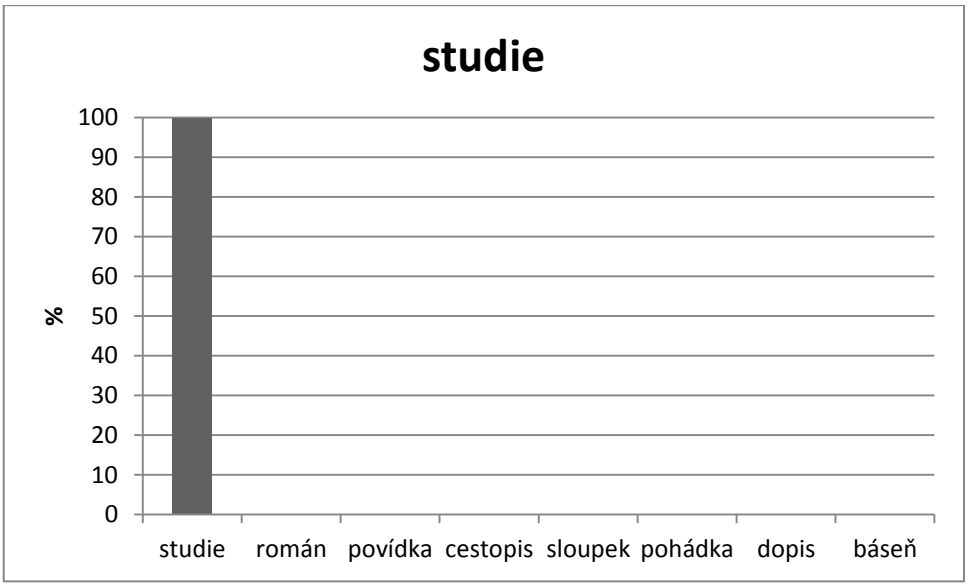
Obr. 79. Automatické přiřazení románu k jednotlivým žánrům dle *SVM*



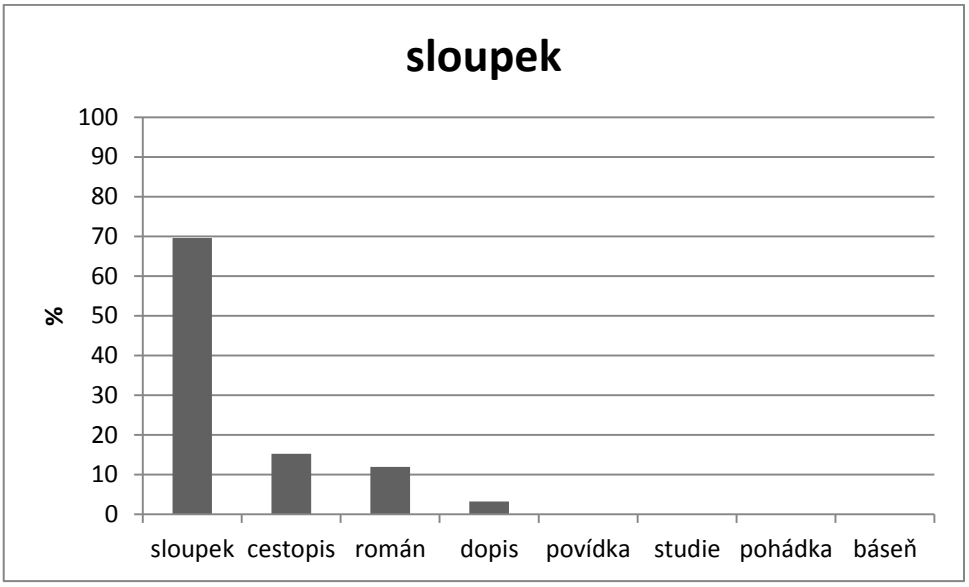
Obr. 80. Automatické přiřazení povídky k jednotlivým žánrům dle *SVM*



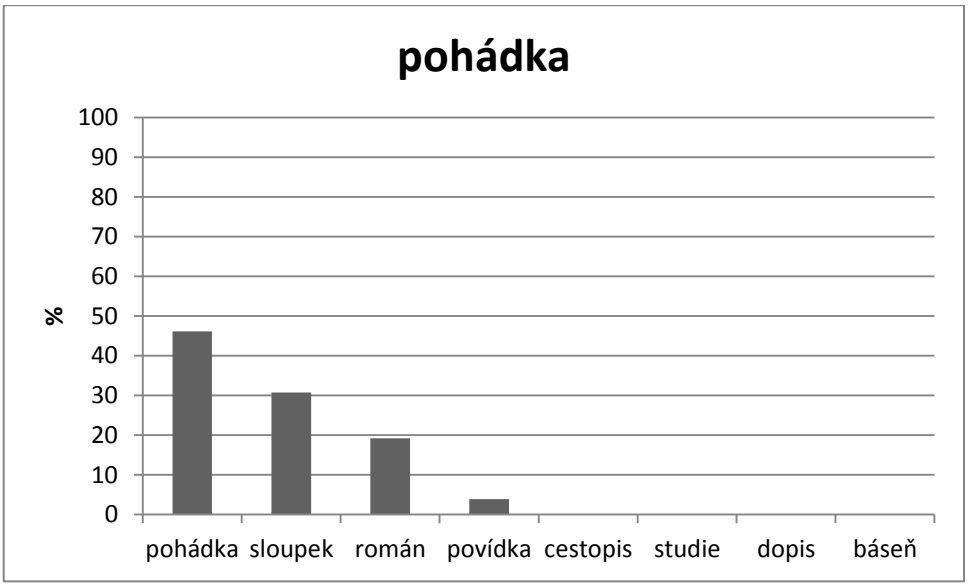
Obr. 81. Automatické přiřazení cestopisu k jednotlivým žánrům dle *SVM*



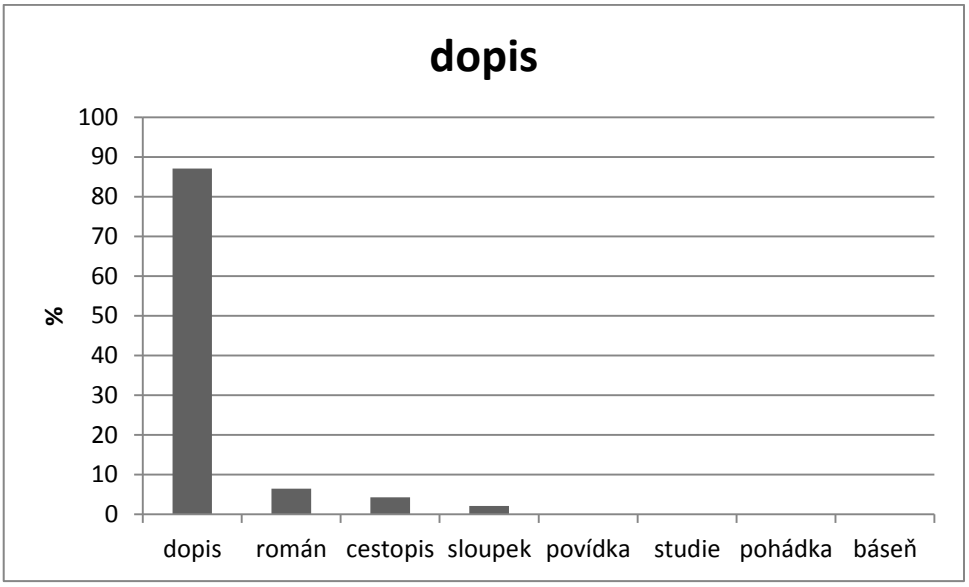
Obr. 82. Automatické přiřazení studie k jednotlivým žánrům dle *SVM*



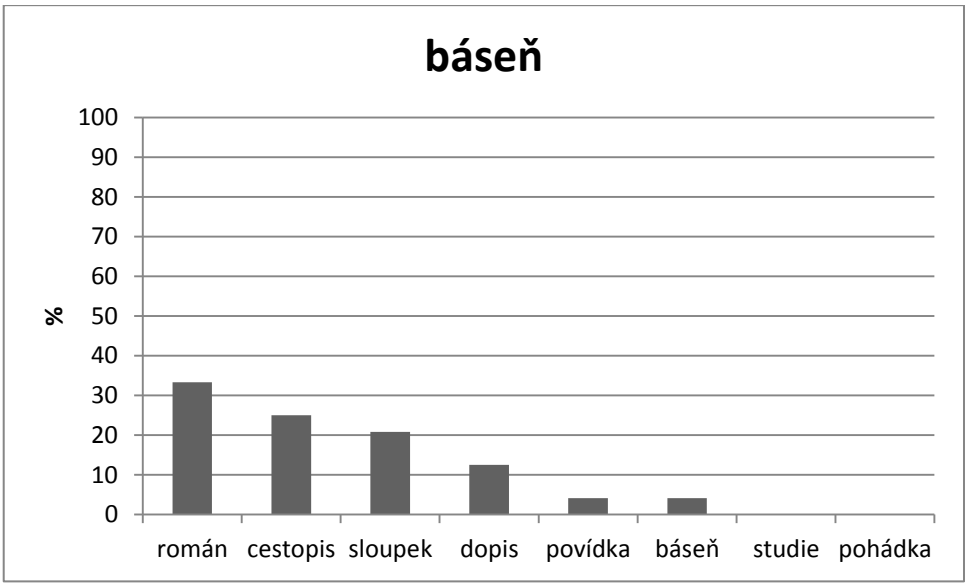
Obr. 83. Automatické přiřazení sloupku k jednotlivým žánrům dle *SVM*



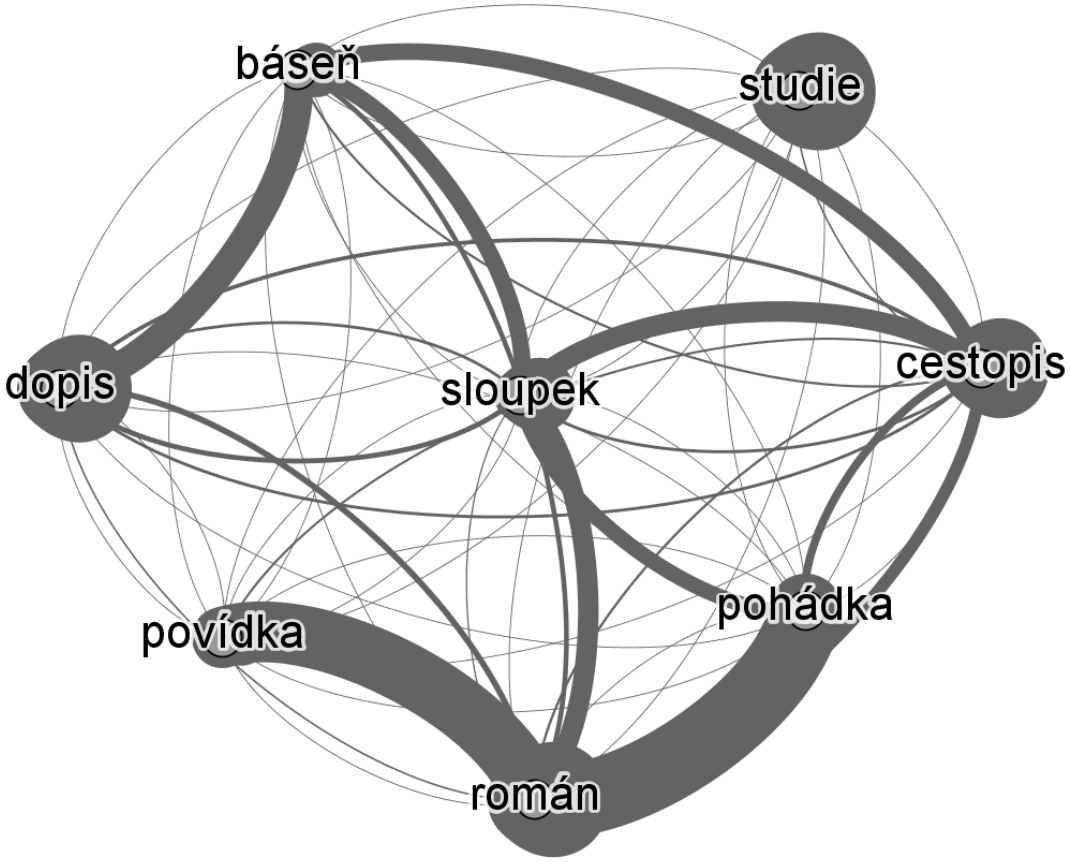
Obr. 84. Automatické přiřazení pohádky k jednotlivým žánrům dle *SVM*



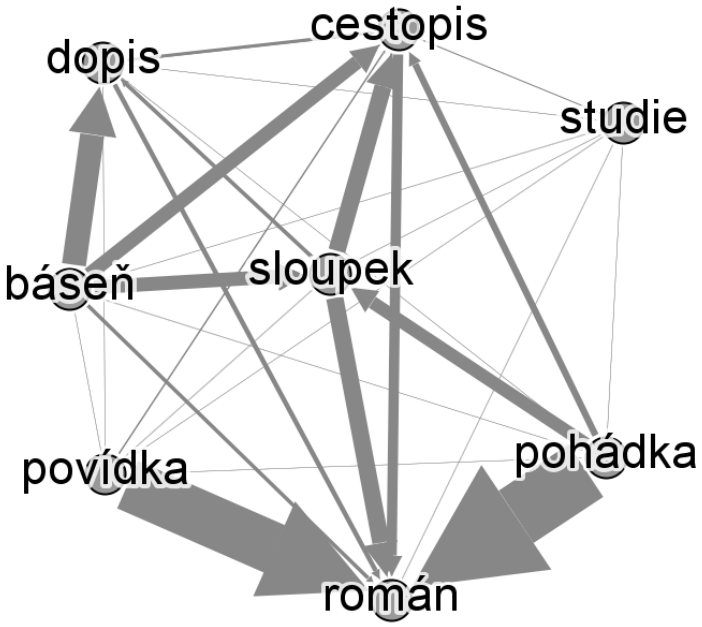
Obr. 85. Automatické přiřazení dopisu k jednotlivým žánrům dle *SVM*



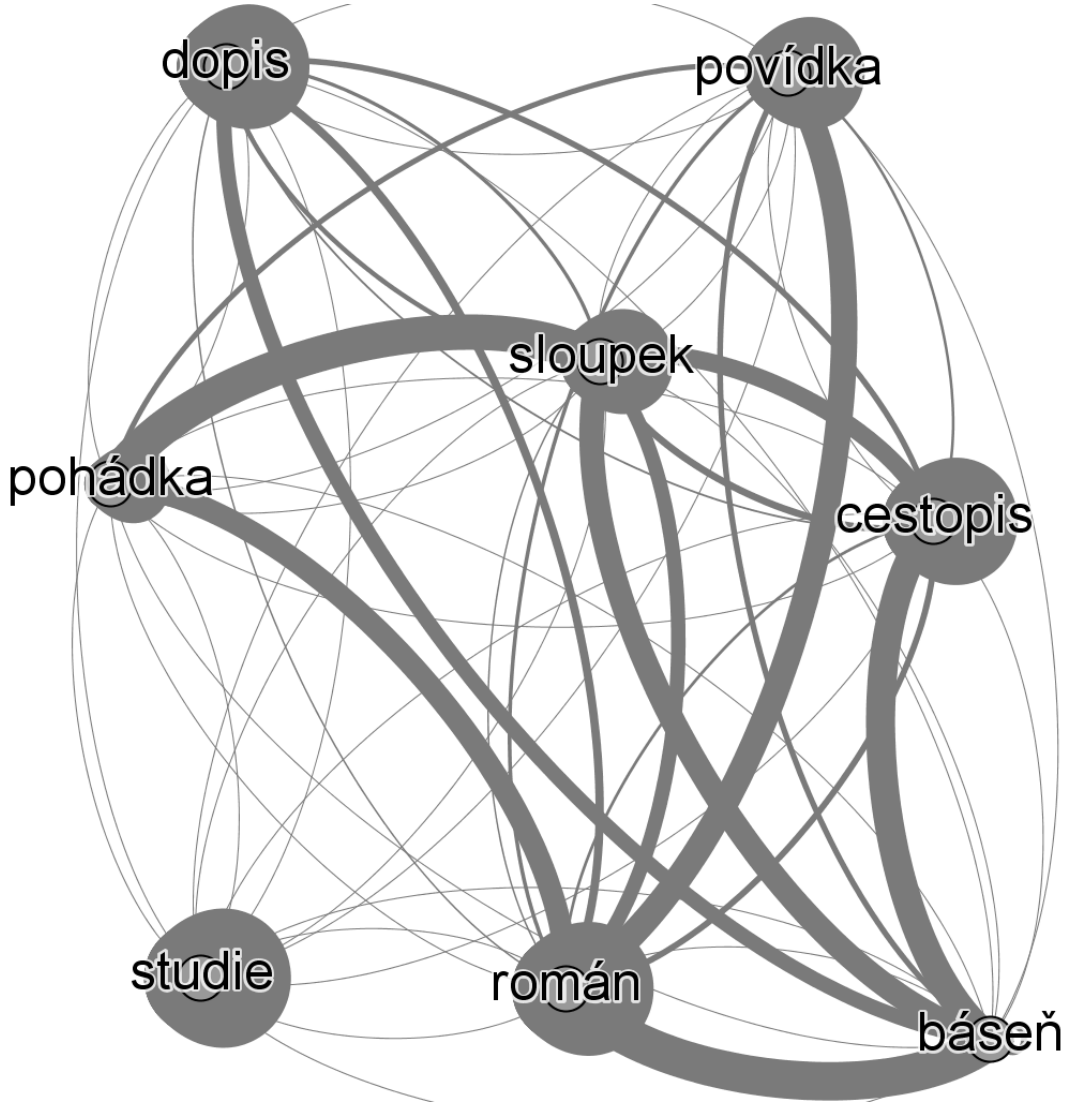
Obr. 86. Automatické přiřazení básně k jednotlivým žánrům dle *SVM*



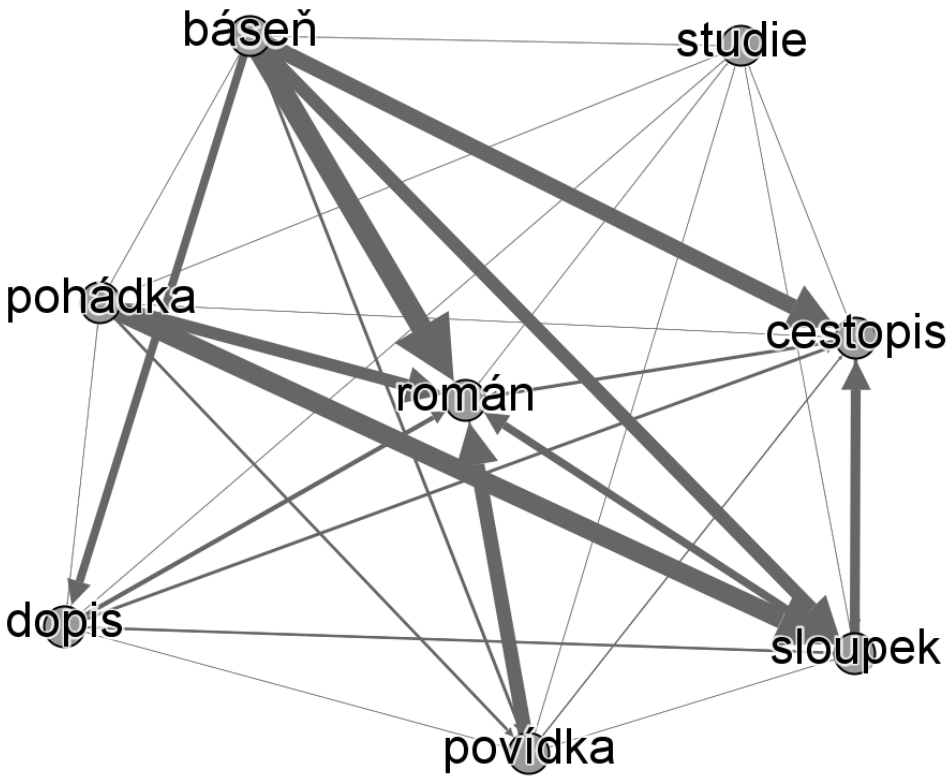
Obr. 87. Automatické přiřazení žánrů dle *RF* (čím širší hrany, tím více přiřazení)



Obr. 88. Automatické přiřazení žánrů dle *RF* s vyznačeným směrem přiřazení (čím širší hrany, tím více přiřazení)



Obr. 89. Automatické přiřazení žánrů dle *SVM* (čím širší hrany, tím více přiřazení)



Obr. 90. Automatické přiřazení žánrů dle *SVM* s vyznačeným směrem přiřazení (čím širší hrany, tím více přiřazení)

Z výše uvedených grafů lze konstatovat několik dílčích závěrů. U obou modelů (*RF* i *SVM*) lze sledovat blízkost povídky a románu, což můžeme poměrně jednoduše interpretovat, neboť oba žánry se vyznačují mnoha společnými prvky, přičemž největší rozdíl je v rozsahu těchto epických textů. Právě délka textu se stává často jediným hlediskem pro rozlišení románu od povídky (popř. od novely). Blízkost těchto žánrů s ne zcela jasnou hranicí komentuje literární vědec Otakar Chaloupka takto: „[…] tentýž text je někdy označován jako povídka a jindy jako novela, či další text někdy jako novela a někdy jako román. Na toto téma byla napsána spousta teoretických knih a studií, jež se někdy shodují, někdy si odporují. V praxi je nejjednodušším (a proto také značně zjednodušujícím) hlediskem, že povídka je krátká, novela delší a román nejdelší.“[104] Nejednotnost literárněvědných badatelů v této problematice ukazuje, že povídka a román mají k sobě velmi blízko a jediným – jakkoli také často nejednoznačným – vodítkem pro rozlišení daných žánrů je jejich

[104] Chaloupka, O. (2007), s. 762.

délka. Tato skutečnost se odráží i v naší analýze, kde i jinak velmi přesná metoda *AMNP* má v mnoha případech problém rozlišit povídku od románu.

Parně nejkomplikovanější je objasnit pozici básně mezi ostatními žánry, neboť právě se zařazením básně měl model *AMNP* největší potíže, a to do takové míry, že metoda *RF* správně rozpoznala báseň jen v 33 % případů, obecně přesnější metoda *SVM* pak dokonce pouze ve 4 %. Takto nízkou úspěšnost přiřazení textů nenajdeme u žádného jiného žánru. Přitom právě básně bychom intuitivně označili jako značně odlišné od ostatních zkoumaných textů. Navíc ze získaných výsledků je zřejmé, že na rozdíl např. od povídky, která byla často chybně přiřazena k jinému žánru (románu), básně nemají blízko k jednomu, ale rovnou k několika různým žánrům. Báseň je tak hojně přiřazována k románu, cestopisu, sloupku či dopisu. Z toho vyplývá, že básně – jakkoliv se mohou jevit jako značně odlišné od ostatních textů – tvoří žánr, který je nejméně specifický a při strojovém zpracování je velmi obtížné jej rozpoznat. Zdá se, že však nejde o anomálii příznačnou pro *AMNP* analýzu, neboť ve většině ostatních analýz v této práci (slovní bohatství, vzdálenosti sloves, průměrná délka tokenu, aktivita, distribuce slovních druhů)[105] se báseň ukazuje jako nejméně specifický žánr. Patrně nejzajímavější na tom je, že zatímco pro člověka by bylo jistě ze všech zkoumaných textů nejsnadnější rozpoznat právě básně, pro uvedené stylometrické kvantitativní nástroje je tomu právě naopak.

Ze získaných výsledků také plyne, že poměrně blízko k sobě mají sloupek a cestopis. Tento fakt pravděpodobně není příliš překvapující, neboť oba žánry bývají řazeny k publicistice, a tudíž lze jistou podobnost očekávat. Posledním žánrem hodným pozornosti z hlediska problematického rozpoznání textů v *AMNP* analýze, je pohádka, která se v některých případech chybně přiřazovala k románu či sloupku. Zatímco v případě blízkosti pohádky a románu budou důvody pravděpodobně obdobné jako u románu a povídky, blízkost pohádky a sloupku může působit poměrně překvapivě.

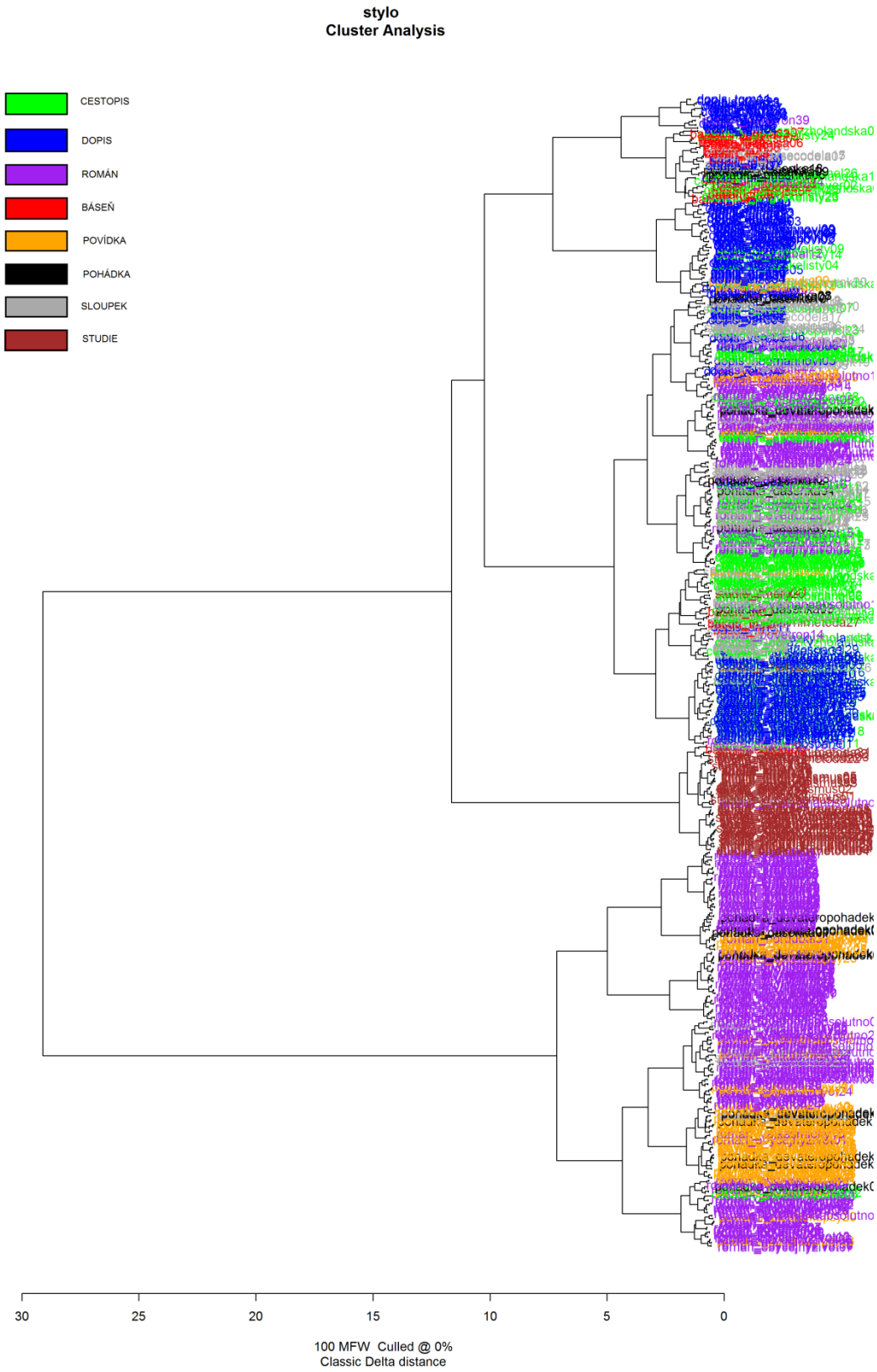
[105] Jedinou výjimkou je tematická koncentrace textu, kde má báseň naopak velmi specifické postavení v rámci zkoumaných žánrů.

4.8. Nejfrekventovanější slova (*MFW*)

Analýza nejfrekventovanějších slov je založena na porovnávání frekvenčních slovníků jednotlivých textů. Prvním krokem *MFW* (most frequent words) analýzy je získání nejčetnějších slov a jejich frekvencí v celém korpusu. Zde je třeba vždy rozhodnout, s kolika slovy se bude pracovat, v našem případě to bude 100 nejčetnějších. Dále je třeba získat frekvenční slovníky jednotlivých textů. Aby byla získaná data porovnatelná, jsou absolutní frekvence přepočítány na relativní. Tato data pak slouží k porovnávání jednotlivých textů. Získané hodnoty se dále použijí pro statistické zpracování. Výsledné hodnoty lze vizualizovat pomocí různých grafů, nejčastěji dendrogramů.

Pokud jde o jednotky, i v této analýze budeme pracovat se slovními tvary, které pokládáme za adekvátní, a to i s přihlédnutím k určitým výhodám práce s lemmaty. Pro více informací k této problematice viz kapitolu 3.2 *Jazykové jednotky*. Výpočet jsme provedli pomocí softwaru *Stylo*[106], který provádí automaticky všechny kroky *MFW* analýzy včetně statistického vyhodnocení a grafické vizualizace. V našem případě jsme použili klastrovou analýzu a výsledky zobrazili pomocí dedrogramu na Obr. 91. Přestože lze získaná data zobrazit také pomocí sítě, museli jsme od takového zobrazení ustoupit kvůli příliš velkému množství textů, výsledná síť by byla totiž naprosto nepřehledná. Kompletní nastavení programu a seznam 100 nefrekventovanějších slov korpusu lze nalézt v příloze této práce.

[106] Eder, M., Kestemont, M., Rybicki, J. (2013).

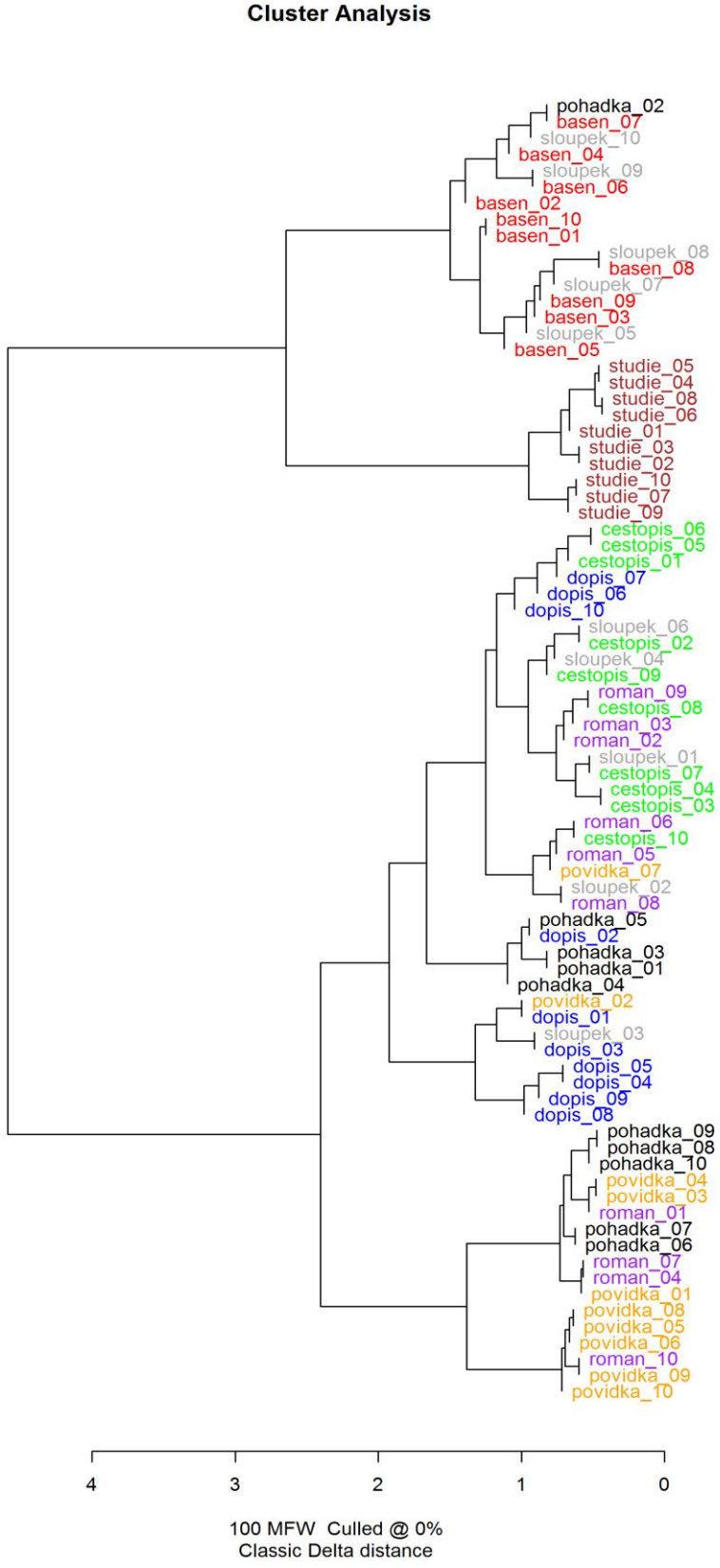


Obr. 91. Výsledky MFW analýzy

Z Obr. 91 je zřejmé, že *MFW* analýza je účinný nástroj pro diferenciaci žánrů. Podobně jako v případě *AMNP* však tato metoda neumožňuje příliš velké možnosti pro lingvistickou interpretaci získaných dat, a to i přes relativní výhodu spočívající v použitých jednotkách (slovní tvary vs. n-gramy). Z hlediska *MFW* však můžeme konstatovat, že romány a povídky mají k sobě nejblíže a tvoří samostatnou větev. Dále lze také sledovat poměrně specifické postavení studie, která se jeví jako žánr značně odlišný od ostatních.

Poměrně zajímavé je, že základní poznatky získané *MFW* analýzou odpovídají výsledkům *AMNP*, kde také román a povídka měli k sobě nejblíže a studie stála zcela mimo ostatní žánry. Zdá se tedy, že vzhledem k různým metodám a dokonce i jednotkám, můžeme tyto výsledky do jisté míry zobecnit. Pokud se na věc podíváme z čistě lingvistického hlediska, můžeme konstatovat, že získaná data odpovídají intuitivním předpokladům, neboť ze zkoumaných žánrů by pravděpodobně každý jazykovědec určil právě román a povídku jako nejbližší žánry. Jistá osamocenost studie pak vzhledem k specifickým rysům odborných textů také odpovídá obecným předpokladům, protože v rámci našeho korpusu jde o jediné (pomineme-li dopis) nebeletristické texty.

Protože Obr. 91 není vzhledem k velkému množství textů (760) příliš přehledný v tom smyslu, že nemůžeme z grafu odečíst konkrétní texty (rozložení žánrů však patrné je), zmenšili jsme náš korpus na 80 textů, tj. 10 textů od každého žánru, abychom viděli detailněji jednotlivé texty, výsledky jsou zobrazeny na Obr. 92. Je však třeba poznamenat, že tato analýza slouží spíše ilustrativně, neboť je zřejmé, že korpus o 80 textech není dostatečně velký na to, abychom z něj mohli vyvozovat nějaké relevantní závěry. Proto také při pohledu na výsledky na Obr. 92 můžeme snad jen konstatovat, že se potvrzuje specifické postavení studie, která má i v takto malém korpusu zcela jedinečnou pozici.

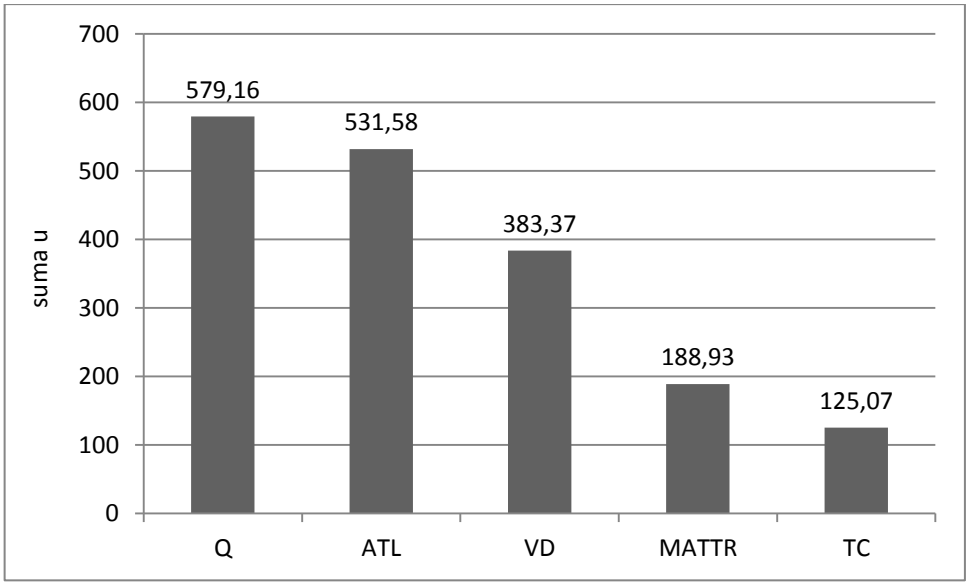


Obr. 92. Výsledky MFW analýzy 80 textů

4.9. Komparace metod

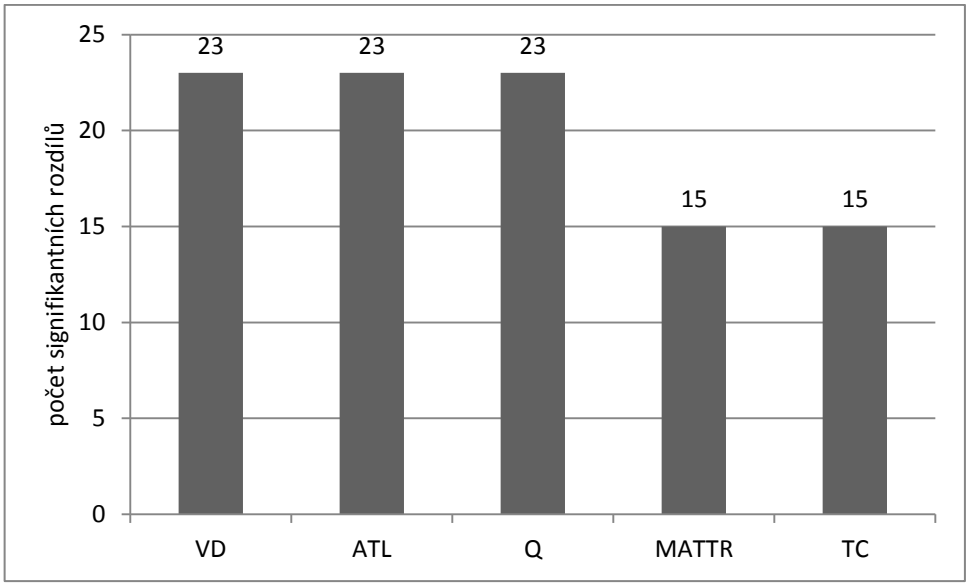
V této kapitole provedeme komparaci metod, které jsme použili v jednotlivých analýzách této práce. Porovnávat je budeme z hlediska jejich efektivnosti při diferenciaci žánrů. Zjistíme tak, které metody jsou vzhledem k žánru inertní a které jsou naopak pro klasifikaci žánrů efektivní. Vzhledem k tomu, že se jednotlivé metody značně principiálně liší, nemůžeme zahrnout do této komparace všechny použité metody, neboť nemáme žádný společný referenční bod, o který bychom se mohli opřít. Vybereme pouze indexy, které jsme v jednotlivých analýzách vyhodnocovali pomocí *u*-testu (*MATTR, TC, VD, ATL, Q*). Právě výsledné hodnoty *u*-testu nám poslouží jako společné hledisko, na jehož základě můžeme provést komparaci.

Nejdříve pro porovnání použijeme sumu hodnot *u*,[107] tedy součet všech hodnot *u* v dané analýze jednoho indexu, výsledky jsou uvedeny na Obr. 93. Na to navazuje graf na Obr. 94, kde jsou indexy seřazeny sestupně podle celkového počtu signifikantních rozdílů v dané analýze.



Obr. 93. Komparace indexů na základě sumy hodnot *u*-testu

[107] V našem případě není třeba sumu nijak normalizovat, vždy pracujeme se stejnými texty.

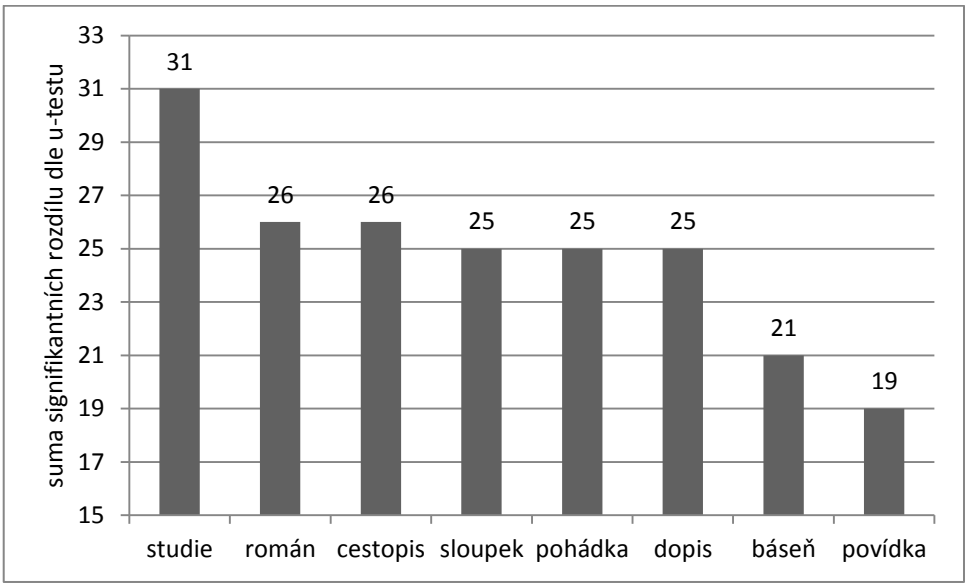


Obr. 94. Komparace indexů na základě počtu signifikantních rozdílů dle *u*-testu

Pří pohledu na Obr. 93 vidíme, že z hlediska klasifikace žánrů se jako nejefektivnější nástroj jeví aktivita následovaná průměrnou délkou tokenu a vzdáleností sloves, jako méně efektivní se ukazuje slovní bohatství a tematická koncentrace textu. Ještě patrnější je rozdělení zkoumaných indexů do dvou skupin na Obr. 94, kde *Q, VD* a *ATL* mají 23 signifikantních rozdílů a *MATTR* společně s *TC* jen 15. Lze však konstatovat, že všechny uvedené indexy nějakým způsobem charakterizují jednotlivé žánry, neboť každý index měl alespoň několik signifikantních rozdílů. Tento fakt hodnotíme jako do jisté míry překvapivý, a to zejména s přihlédnutím charakteru našeho korpusu (jeden autor). Předpokládáme tudíž, že pokud uvedené indexy jsou relativně úspěšně s to klasifikovat jednotlivé žánry i v rámci textů jediného autora, jehož styl navíc můžeme charakterizovat jako poměrně konzistentní, je velmi pravděpodobné, že skupina těchto indexů je důležitým nástrojem pro současné stylistické bádání a že by bylo vhodné budoucí stylistiky obohatit o tyto a podobné indexy jakožto účinné prostředky i pro intersubjektivní klasifikaci jednotlivých stylů a žánrů.

Nyní se detailněji podíváme na jednotlivé žánry, a to z hlediska jejich specifičnosti v rámci našeho korpusu. Jinými slovy: zjistíme, které žánry se nejvíc liší od ostatních a které naopak s ostatními texty splývají. Komparaci provedeme opět na základě výsledků *u*-testu, konkrétně použijeme počet signifikantních rozdílů.

U každého žánru tak při aplikaci *MATTR, TC, Q, VD* a *ATL* rozpoznáme míru specifičnosti jednotlivých skupin textů uvnitř daného korpusu. Výsledky jsou zobrazeny v podobě sloupcového grafu na Obr. 95.



Obr. 95. Komparace žánrů z hlediska počtu signifikantních rozdílů *u*-testu při aplikaci *MATTR, TC, Q, VD* a *ATL*

Na Obr. 95 vidíme sestupně seřazené žánry podle počtu signifikantních rozdílů ve všech sledovaných indexech. Získanou distribuci můžeme rozdělit zhruba do tří skupin. Jedinečné postavení zaujímá studie, následují román, cestopis, sloupek, pohádka a dopis, nejméně signifikantních rozdílů pak mají báseň a povídka. Specifičnost studie v rámci našeho korpusu lze poměrně snadno vysvětlit tím, že jsou to jediné odborné texty mezi uměleckými (s výjimkou dopisu). Jako nejvíce překvapivé hodnotíme postavení básně a povídky. U básně jsme očekávali přesně opačné hodnoty, tedy že se bude výrazně lišit od všech ostatních textů. Tento žánr se nepochybně vyznačuje mnoha specifickými rysy, kterými jej můžeme vyčlenit z množiny ostatních textů, což můžeme demonstrovat i tím, že i při letmém pohledu na jednotlivé texty bychom jistě měli nejmenší práci s vyřazením básní od ostatních textů. Interpretovat postavení povídky, která se zdá být nejméně specifickým žánrem v rámci našeho korpusu, je poměrně složité, neboť na jedné straně to můžeme vysvětlit tím, že povídka je jedním ze základních a nejstarších žánrů vůbec, což Petrů komentuje takto: „[…] najdeme však již ve středověku povídku jako svébytný

literární žánr. […] Z těchto literárních projevů se utvářela povídka v dnešním pojetí jako žánr střední epiky. Její model je poměrně volný a vymezuje se spíše opozicí k modelu románu a novely.“[108] Povídka tedy patří mezi základní epické žánry, což by vysvětlovalo její pozici v naší komparaci. Na druhou stranu to samé by mělo platit i pro román, který se však projevil jako více specifický.

4.10. Poznámka k dramatickým textům

Vzhledem k tomu, že náš korpus je sestaven z díla Karla Čapka, považujeme za vhodné alespoň stručně vysvětlit, z jakých důvodů jsme do výběru nezařadili divadelní hry. Dramata hrají v Čapkově díle nepochybně významnou roli, navíc z hlediska žánrové analýzy by byl výzkum obohacen o další styl. Důvod, proč jsme v této práci nepracovali s dramatickými díly Karla Čapka, tkví v problematickém kvantitativním zpracování těchto textů. V kapitole věnované metodologii jsme uváděli naše pojetí textu a základní východiska zpracování korpusu. Divadelní hry se však několika aspekty výrazně odlišují od ostatních žánrů, a to zejména v:

a) Absenci kapitol. Hry jsou sice rozděleny do několika jednání, scén či obrazů, je však otázka, do jaké míry je tato segmentace kompatibilní s ostatními žánry.
b) Střídání replik.
c) Primární mluvenosti. Drama je na rozdíl od ostatních textů primárně určeno k mluvenému projevu, dokonce někteří lingvisté používají právě divadelní hry jako materiál pro výzkum mluveného jazyka. S tím úzce souvisí další specifikum, a to fakt, že ačkoliv se jednotlivé repliky realizují mezi jednotlivými postavami jako stylizovaný spontánní dialog, tento dialog je předem konstruován jako jeden text pro diváka.

Je tedy zřejmé, že abychom mohli analyzovat dramata a porovnávat je s jinými texty, je zcela nezbytné nejdříve důkladně prozkoumat, jak se tyto texty chovají v nejrůznějších parametrech. Dále je zcela klíčové zjistit, jakým způsobem segmentovat tyto texty, protože není vůbec jasné, který způsob je vhodný v závislosti na komparaci s dalšími styly či žánry. Dokud nebudou vymezena alespoň základní

[108] Petrů, E. (2006), s. 80.

východiska zpracování dramatických textů, považujeme za nevhodné takové texty zařadit do našeho výzkumu.

5. Závěr

Tato práce si kladla za cíl zpracovat problematiku klasifikace žánrů na základě několika experimentálních metod a také ověřit efektivnost a relevantnost daných metod pro klasifikaci textů. Lze konstatovat, že tyto cíle byly splněny, neboť jsme pomocí různých indexů zachytili několik vybraných stylových charakteristik, které jsme kvantifikovali a pomocí statistických testů také interpretovali. Tato práce tak přinesla do českého jazykovědného prostředí nové poznatky, které jsou vždy experimentálně ověřeny. Jednotlivé analýzy nám umožnily nahlédnout blíže do problematiky klasifikace stylů a žánrů z poněkud neobvyklého pohledu kvantitativních měření vlastností textu, což by mohlo napomoci tradiční stylistice a literární vědě. Za největší přínos této práce považujeme to, že jsme umožnili intersubjektivní pohled na zkoumané styly a žánry.

Z hlediska efektivnosti jednotlivých metod při klasifikaci jednotlivých žánrů se jako nejúčinnější metoda ukázal model *AMNP*, který v kombinaci s vyhodnocením *SVM* dosáhl v rámci našeho korpusu přesnosti predikce 84 %, což lze s přihlédnutím k použitým textům (jediný autor) považovat za velmi vysokou hodnotu. Na druhou stranu musíme připustit, že daní za takto vysokou přesnost je velmi problematická jazykovědná interpretace získaných dat, a to hned z několika důvodů (použité jednotky, zkoumání několika jazykových rovin zároveň, poměrně složité statistické vyhodnocení).

Dále můžeme konstatovat, že všechny použité stylometrické metody poměrně účinně rozlišily jednotlivé žánry, a zdají se tak být důležitými nástroji při textových a stylistických výzkumech. Tento fakt považujeme za poměrně překvapivý, neboť korpus omezený na jediného autora nutně minimalizuje veškeré potenciální rozdíly mezi žánry. Přitom Karel Čapek, stejně jako kterýkoliv jiný autor, má svůj individuální styl, který dosahuje jistých modifikací na základě konkrétního žánru jen v omezené míře. V případě Čapka bychom mohli mluvit o publicistickém stylu, jehož nádech lze sledovat ve všech autorových textech. Lze tedy předpokládat, že aplikací použitých nástrojů na texty různých autorů by rozdíly mezi žánry značně vzrostly, a tudíž můžeme konstatovat, že dané metody mohou najít uplatnění jak ve stylistice, tak v literární vědě.

Získaná data ukazují, že čím je použitá metoda složitější, tím zpravidla stoupá přesnost klasifikace textů, ale zároveň klesá možnost lingvistické interpretace. V rámci naší práce je tento fakt nejviditelnější v případě *AMNP*, kde se používá hned čtyř variant n-gramů. Z tohoto důvodu lze tvrdit, že tyto metody jsou vhodnější spíše pro automatické určování autorství (tj. oblast, kde také tyto nástroje vznikly) než pro lingvisticky zaměřenou žánrovou analýzu. Jako příklad může sloužit snad nejznámější stylometrický ukazatel, tj. slovní bohatství, které nám umožňuje detailně interpretovat získaná data, totéž lze uvést i pro vzdálenosti sloves, průměrnou délku tokenu, aktivitu či tematickou koncentraci. Nelze však říct, která metoda je lepší nebo horší obecně, vždy je třeba zvážit její použití na základě konkrétních cílů daného výzkumu.

Pokud jde o jednotlivé žánry, zjistili jsme, že z hlediska zkoumaných stylometrických ukazatelů se jako nejvíce specifická jeví studie, která se ve všech analýzách značně odlišovala od ostatních textů. Tuto pozici studie lze vysvětlit zejména tím, že jde o jediný neumělecký žánr v rámci našeho korpusu (kromě dopisu). Zatímco specifické postavení studie se tedy dalo očekávat, výsledky básně nás naopak překvapily. Ukázalo se totiž, že báseň je nejméně specifický žánr, který daným stylometrickým metodám (výjimku tvoří pouze tematická koncentrace) činí největší potíž při jeho odlišení od ostatních textů. Přitom intuitivně bychom naopak předpokládali, že báseň je značně specifický žánr, který se bude od ostatních výrazně lišit. Ostatní výsledky jednotlivých žánrů nelze zobecnit na všechny použité metody, ale je třeba přihlédnout ke konkrétním dílčím analýzám.

Závěrem je třeba uvést, že získané výsledky a dílčí závěry bude nezbytné podpořit v dalších výzkumech zejména (a) rozšířením analýz o další autory, žánry a jazyky, (b) použitím dalších stylometrických indexů či metod. Pouze skrze dostatečně velké množství relevantních dat můžeme alespoň částečně nahlédnout do mechanismů, jimiž autoři vytvářejí texty jednotlivých stylů či žánrů. Tato práce poskytla v českém kontextu úvodní náhled do problematiky a otevírá prostor dalším analýzám, které nám umožní lépe pochopit principy klasifikace různých stylů.

Kromě nových poznatků přinesl tento text, jako ostatně všechny vědecké práce, několik otázek, jež si zaslouží pozornost dalších výzkumů. Pomineme-li již zmíněné rozšíření o analýzy dalších autorů a metod, nabízí se zde využití získaných výsledků

pro poznání obecných zákonitostí výstavby textu a fungování jazyka obecně. Zejména bude třeba zjistit, zda a do jaké míry jednotlivé indexy vzájemně souvisí. Právě nalezení jazykových zákonů, které nám umožní lépe pochopit a poznat fungování jazyka, je základním cílem kvantitativní lingvistiky a zároveň také tématem našeho dalšího výzkumu. Pokud zůstaneme u popisné stylometrie, velkou výzvu představují dramatické texty, které v sobě skrývají hned několik problematických aspektů znemožňujících jejich analýzu. Zpracování této dodnes neprobádané problematiky si jistě zaslouží samostatný výzkum.

6. Anotace

Název práce: Kvantitativní analýza žánrů

Autor: Miroslav Kubát

Katedra: Katedra obecné lingvistiky Filozofické fakulty Univerzity Palackého v Olomouci

Školitel: Mgr. Radek Čech, PhD.

Počet znaků: 118 534 (včetně mezer a pozn. aparátu)

Počet příloh: 4

Počet titulů použité literatury: 112

Klíčová slova: kvantitativní lingvistika, analýza textu, korpus, Karel Čapek, stylometrie, lexikální statistika.

Abstrakt

Práce zkoumá žánry za použití kvantitativních metod, konkrétně se zaměřuje zejména na indexy frekvenční struktury textu (např. slovní bohatství, distribuce slovních druhů, aktivita textu). Aby bylo zabráněno negativnímu vlivu různých autorských stylů, korpus obsahuje pouze texty jediného autora – Karla Čapka. Výsledky jsou vždy vyhodnoceny pomocí statistických testů a lingvisticky interpretovány. Hlavním cílem práce je experimentálně ověřit obecné stylistické předpoklady.

7. Annotation

Title: Quantitative Analysis of Genres

Author: Miroslav Kubát

Department: Department of General Linguistics,
Faculty of Arts, Palacký University Olomouc

Supervisor: Mgr. Radek Čech, PhD.

Number of pages characters: 118 534

Number of appendices: 4

Number of references: 112

Key words: quantitative linguistics, text analysis, corpus, Karel Čapek, stylometry, lexical statistics.

Abstract

The thesis analyses the genres using the quantitative methods, more specifically it focuses on text frequency structure indexes (e.g. vocabulary richness, POS distribution, activity). In order to avoid the negative influence of different author's styles, the corpus contains only texts written by one author – Karel Čapek. The resulting values are always statistically tested and linguistically interpreted. The main aim of the work is to experimentally verify the general stylistic assumptions.

8. Summary

This work analyses and describes several genres using the experimental methods which are connected to the contemporary quantitative linguistics, especially the stylometry. The research also verifies the effectiveness and the relevance of the used methods for the genre classification. It is important to mention that all results of the analyses in this thesis are always statistically tested and linguistically interpreted. This work brings into the Czech linguistics new findings, which are always supported by results based on the experiment. Individual analyses allowed me to get closer to the issue of the genre classification from a somewhat unusual perspective of the quantitative measurements of the text features. The methodology and the obtained results of this research could help traditional stylistics and literary criticism. This thesis follows up the work of Marie Těšitelová who established the usage of statistical methods in Czech linguistics and brought several studies in this field.

The key part of any quantitative research is a selection of a suitable sample. In this work, the corpus consists of the texts written by one of the most famous Czech writers – Karel Čapek. This corpus offers a wide variety of different genres (novel, short story, fairy tale, travel book, poem, newspaper column, drama, scientific text, and letter). I decided to use a corpus consisting of texts written by only one author to avoid a bias caused by different authors' styles. Each author has a unique writing style which permeates across genres. Thus, a corpus of texts by several authors automatically blocks a relevant evaluation of the results, because it is impossible to discover whether the obtained values reveal the authorship or the genre.

In this work, the following methods are used: moving average type-token ratio (*MATTR*), moving window type-token ratio distribution (*MWTTRD*), thematic concentration (*TC*), secondary thematic concentration (*STC*), proportional thematic concentration (*PTC*), verb distances (*VD*), average token length (*ATL*), activity (*Q*) and descriptivity (*D*), part of speech distribution, author's multilevel n-gram profile (*AMNP*), most frequent words analysis (*MFW*). It is important to mention that the aforementioned methods were chosen because they are not influenced by the text length.

It can be said that all used methods are able to distinguish the genres. This fact is quite surprising because the corpus (limited to a single author) necessarily minimizes any potential differences between genres. Karel Čapek, as well as any other author, has its own individual style which can be changed only to a limited extent. Given that the analysis based on quantitative methods is able to distinguish Čapek's texts among the several genres, it can be assumed that the application of these methods to texts written by various authors has to significantly increase the differences among the genres. To sum up, the quantitative methods may find applications in stylistics and literary criticism.

From the viewpoint of the effectiveness of the individual methods for the classification of various genres, the most efficient method is *AMNP* (author's multilevel n-gram profile). The *AMNP* with the evaluation of *SVM* (support vector machines) achieved within the corpus the highest accuracy of prediction (84 %). On the other hand, I must admit that a tax for such high accuracy is a very problematic linguistic interpretation of the obtained data for several reasons (the used units, examining several language levels at the same time, relatively complicated statistical evaluation).

Finally, it should be noted that the obtained results and the preliminary conclusions have to be supported in further research, especially by (a) extending the analysis of other authors, genres and languages, (b) using other indices or methods. Only through a sufficiently large amount of relevant data it can be at least partially understood the mechanisms by which authors compose texts of individual styles and genres. This work provides the introductory insight into the issue in the Czech linguistics and opens the door for further analyses.

9. Zdroje

- Altmann, G. (1997). The Art of Quantitative Linguistics. Journal of Quantitative Linguistics, 4(1–3), s. 13–22.
- Altmann, G. (2006): Fundaments of quantitative linguistics. In: Genzor, J., Bucková, M. (eds.), Favete linguis. Bratislava: Slovak Academic Press, s. 15–27.
- Altmann, G. (2012). Certain Differences between Qualitative and Quantitative Linguistics. Czech and Slovak Linguistic Review, 2(1), s. 6–15.
- Altmann, G. Altmann-Fitter (software). [Ke staženi na http://www.ram-verlag.biz/altmann-fitter/]
- Altmann, G., Wimmer, G. (1999). Review Article: On Vocabulary Richness. Journal of Quantitative Linguistics, 6(2), s. 1–9.
- Benešová, M. (2011). Kvantitativní analýza textu se zvláštním zřetelem k analýze fraktální. Olomouc (disertační práce). [Dostupné online na http://theses.cz/id/p19fdf/DISERTACNI_PRACE_BENESOVA.pdf].
- Bennett, William Ralph. (1976). Scientific and engineering problem-solving with the computer. Englewood Cliffs, N.J.: Prentice Hall.
- Berka, P. (2003). Dobývání znalostí z databází. Praha: Academia.
- Blecha, I. (2004). Filosofie. Olomouc: Olomouc.
- Caldarelli, G. (2008). Scale-Free Networks: Complex Webs in Nature and Technology. Oxford: OUP.
- Čech, R. (2011). Frequency structure of New Year's presidential speeches in Czech. The authorship analysis. In: Kelih et al. (eds.) Issues in Quantitative Linguistics 2. Lüdenscheid: RAM-Verlag, s. 82–94.
- Čech, R. (2013). Language and ideology: Quantitative thematic analysis of New Year speeches given by Czechoslovak and Czech presidents (1949–2011). Quality & Quantity, 48(2), s. 899–910.
- Čech, R. (2014). Jen popis čísly? Perspektivy korpusové lingvistiky. Naše řeč, 97(4–5), s. 171–184.
- Čech, R. (2015). Text length and the lambda frequency structure of the text. In: Sequences in language and text. (přijato)

- Čech, R., Garabik, R., Altmann, G. (2015). Testing the thematic concentration of text. Journal of Quantitative Linguistics. (přijato).
- Čech, R., Kelih, E., Mačutek, J. (2014). Impact of semantics on case diversification. In: Benešová, M., Kelih, E., Mačutek, J. (eds.) Book of Abstracts QUALICO 2014. Olomouc: Univerzita Palackého, s. 27–28.
- Čech, R., Popescu, I. I., Altmann, G. (2013): Methods of analysis of the thematic concentration of the text. Czech and Slovak Linguistic Review, 3(1), s. 4–21.
- Čech, R., Popescu, I. I., Altmann, G. (2014). Metody kvantitativní analýzy (nejen) básnických textů. Olomouc: Univerzita Palackého v Olomouci.
- Chaloupka, O. (2007). Příruční slovník české literatury. Praha: Kma.
- Chromý, J. (2014). Korpus a reprezentativnost. Naše řeč 97(4–5), s. 185–193.
- Covington, M. A., McFall J. D. (2010). Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR). Journal of Quantitative Linguistics 17(2), s. 94–100.
- David, J., Čech, R., Radková, L., Davidová Glogarová, J., Šústková, H. (2013). Slovo a text v historickém kontextu – perspektivy historickosémantické analýzy jazyka. Brno: Host.
- Davidová Glogarová, J., Čech, R. (2013). Tematická koncentrace textu – některé aspekty autorského stylu Ladislava Jehličky. Naše řeč, 96, s. 234–245.
- Davidová Glogarová, J., David, J., Čech, R. (2013). Analýza tematické koncentrace textu – komparace publicistiky Ladislava Jehličky a Karla Čapka. Slovo a slovesnost, 74, s. 41–54.
- Doležel, L. (1963). Předběžný odhad entropie a redundance psané češtiny. Slovo a slovesnost 24, s. 165–174.
- Dowdy, S., Wearden, S. (1983). Statistics for Research. Wiley: New York.
- Eder, M. (2014). Stylometry, network analysis and Latin literature. In: Digital Humanities 2014: Book of Abstracts, EPFL-UNIL, Lausanne, s. 457–458. http://dharchive.org/paper/DH2014/Poster-324.xml

- Eder, M., Kestemont, M., Rybicki, J. (2013). Stylometry with R: a suite of tools. In: Digital Humanities 2013: Conference Abstracts. University of Nebraska-Lincoln, NE, s. 487–489.
- Eder, M., Rybicki, J. (2009). PCA, Delta, JGAAP and Polish poetry of the 16th and the 17th centuries: who wrote the dirty stuff? Digital Humanities 2009: Conference Abstracts. University of Maryland, College Park (MA), s. 242–244.
- Esteban, M. D., Morales, D. (1995). A summary of entropy statistics. Kybernetica 31(4), s. 337–346.
- Guiraud, P. (1954). Les catactères stitistiques du vocabulaire. Paris: Presses Universitaires de France.
- Guiraud, P. (1959). Problèmes et methods de la statistique linguistique. Dordrecht: Reidel.
- Herdan, G. (1960). Type-Token Mathematics. The Hague: Mouton.
- Heydel, M., Rybicki J. (2012). The stylometry of collaborative translation. Digital Humanities 2012: Conference Abstracts. Hamburg: Hamburg University Press, s. 212–214.
- Hirsch, J. E. (2005). An indicator to quantify an individual's research output. Proceedings of the National Academy of Sciences of the USA 102 (46), s. 16569–16572.
- Hřebíček, L. (1992). Text in communication: supra-sentence structures. Bochum: Brockmeyer.
- Hřebíček, L. (1993). Text as a construct of aggregations. In: Köhler, R., Rieger, B. B. (eds.), Contributions to Quantitative Linguistics, s. 33–39. Dordecht: Kluwer.
- Hřebíček, L. (1995). Text levels. Language construts, constituents and the Menzerath-Altmann law. Trier: WVT.
- Joula, P. (2008). Authorship Attribution. Delft: Now Publishers Inc.
- Juola, P. (2006). Authorship Attribution. In: Foundations and Trends in Information Retrieval, 1(3), s. 233–334.

- Kjell, Bradley, Woods, W. Addison, Frieder, Ophir. (1993). Discrimination of authorship using visualization. Information Processing & Management, 30(1), s. 141–150.
- Kjell, Bradley. (1994). Authorship determination using letter pair frequency features with neural network classifiers. Literary and Linguistic Computing, 9(2), s. 119–124.
- Köhler, R. (1994). Synergetic Linguistics. In: Asher, R. E. (ed.) The Encyclopedia of Language and Linguistics. Oxford, New York, Seoul, Tokyo: Pergamon Press, s. 4454–4455.
- Köhler, R. (2005) Synergetic Linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.) Quantitative Linguistics. An International Handbook. Berlin-New York: Walter de Gruyter, s. 760–775.
- Köhler, R., Altmann, G. (2005). Aims and methods of quantitative linguistics. In: Altmann, G., Levickij, V., Perebyinis, V. (eds.) Problemy kvantitativnoj lingvistiki. Černivci: Ruta, s. 12–41.
- Köhler, R., Altmann, G. (2011). Quantitative linguistics. In: Hogan, P. C. (ed.) The Cambridge Encyclopedia of the Language Sciences. Cambridge: Cambridge University Press, s. 695–697.
- Köhler, R., Galle, M. (1993). Dynamic Aspects of Text Characteristics. In: Hřebíček, L., Altmann, G. (eds.), Quantitative Text Analysis. Trier: WVT, s. 46–53.
- Králík, J. (2013). Srovnání nesrovnatelného. Korpus – gramatika – axologie, 4, s. 48–52.
- Kubát, M. (2013). Kvantitativní analýza žánrů v díle Karla Čapka. In: Lingvistika Praha 2013. [online] Dostupné z WWW: <http://lingvistikapraha.ff.cuni.cz/sbornik>.
- Kubát, M. (2014). Moving window type-token ratio and text length. In: Altamnn, G., Čech, R., Mačutek, J., Uhlířová, L. (eds.), Empirical Approaches to Text and Language Analysis. Lüdenscheid: RAM, s. 105–113.
- Kubát, M., Matlach, V. (2014). QUITA – Quantitative Index Text Analyzer. Poster na konferenci QUALICO 2014.

- Kubát, M., Matlach, V., Čech, R. (2014). Announcement: Quantitative Index Text Analyser (QUITA). Glottometrics, 27, s. 91–92.
- Kubát, M., Matlach, V., Čech, R. (2014). QUITA – Quantitative Index Text Analyzer. Lüdenscheid: RAM.
- Kubát, M., Milička, J. (2013). Vocabulary Richness Measure in Genres. Journal of Quantitative Linguistics 20(4), s. 339–349.
- Laar, M. (1997). Some Quantitative Genre Indices of Academic Writing. Journal of Quantitative Linguistics 4(1–3), 131–134.
- Livio, M. (2002). The Golden Ratio: The Story of Phi, The World's Most Astonishing Number. New York: Broadway Books.
- Lotko, E. (2005). Slovník lingvistických termínů pro filology. Olomouc: Univerzita Palackého.
- Mačutek, J., & Wimmer, G. (2013). Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. Journal of Quantitative Linguistics, 20(3), s. 227–240.
- Markov, Andrey A. (1913). An Example of Statistical Analysis of the Text of “Evgenii Onegin" Illustrating the Linking of Events into a Chain. Bulletin de l, Acadamie Imperiale des Sciences de St. Petersburg, 6(7), s. 153–162.
- Matlach, V. (2014) Kvantitativně lingvistický software. Olomouc (diplomová práce). [Dostupné online na http://theses.cz/id/fz87uj/thesis.pdf]
- Matlach, V., Kubát, M., Čech, R. (2014), QUITA – Quantitative Text Analyzer (software). Olomouc. [Ke stažení na https://code.google.com/p/oltk/]
- McIntosh, R. P. (1967). An indicator of diversity and the relation of certain concepts to diversity. Ecology, 48; s. 392–404.
- Mikros, G. K. (2006). Authorship attribution in Modern Greek newswire corpora. In: O. Uzuner, S. Argamon, J. Karlgren (eds.), Proceedings of the SIGIR 2006 International Workshop on Directions in Computational Analysis of Stylistics in Text Retrieva. Seattle, Washington: ACM, s. 43–47.
- Mikros, G. K. (2007a). Stylometric experiments in Modern Greek: Investigating authorship in homogeneous newswire texts. In: R. Köhler, G. Altmann & P. Grzybek (eds.), Exact methods in the study of language and text. Berlin/New York: Mouton de Gruyter, s. 445–456.

- Mikros, G. K. (2007b). Authorship Attribution Using Discriminant Function Analysis: Exploring Literary Style Variation in Five Modern Greek Novels. Paper presented at the 5th Trier Symposium on Quantitative Linguistics, Trier, Germany.
- Mikros, G. K. (2009). Content words in authorship attribution: An evaluation of stylometric features in a literary corpus. In: Köhler, R. (ed.), Issues in Quantitative Linguistics. Lüdenscheid: RAM, s. 61–75.
- Mikros, G. K. (2013). Systematic stylometric differences in men and women authors: a corpus-based study. In: R. Köhler & G. Altmann (eds.), Issues in Quantitative Linguistics 3. Dedicated to Karl-Heinz Best on the occasion of his 70th birthday. Lüdenscheid: RAM, s. 206–223.
- Mikros, G. K., & Perifanos, Kostas. (2011). Authorship identification in large email collections: Experiments using features that belong to different linguistic levels Proceedings of PAN 2011 Lab, Uncovering Plagiarism, Authorship, and Social Software Misuse held in conjunction with the CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation, 19–22 September 2011, Amsterdam.
- Mikros, G. K., & Perifanos, Kostas. (2013). Authorship attribution in Greek tweets using multilevel author's n-gram profiles. In: E. Hovy, V. Markman, C. H. Martell & D. Uthus (eds.), Papers from the 2013 AAAI Spring Symposium "Analyzing Microtext", 25–27 March 2013, Stanford, California. Palo Alto, California: AAAI Press, s. 17–23.
- Mikros, G. K., Argiri, E. K. (2007). Investigating topic influence in authorship attribution. In: B. Stein, M. Koppel, E. Stamatatos (eds.), Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (Vol. 276). Amsterdam, Netherlands: CEUR, s. 29–35.
- Mikros, George K. (2013). Authorship Attribution and Gender Identification in Greek Blogs. In: I. Obradović, E. Kelih & R. Köhler (eds.), Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO) in Belgrade, Serbia, April 16–19, 2012. Belgrade: Academic Mind.

- Milička, J. (2013). MaWaTaTaRaD. Praha. (Software)
- Mistrík, J. (1985). Frekvencia tvarov a konštrukcií v slovenčině. Bratislava: Veda.
- Mistrík, J. (1989). Štylistika. Bratislava: SPN.
- Mistrík, J. Frekvencia slov v slovenčině. Bratislava: Slovenská akadémia vied, 1969.
- Nagy, G. T. (1998). Journal of Quantitative Linguistics 5(3), s. 232–239.
- Nekula, M. (2002a). Typ(e) a token. In: Karlík, P., Nekula, M., Pleskalová, J. (eds.), Encyklopedický slovník češtiny. Praha: Nakladatelství Lidové noviny, s. 502.
- Nekula, M. (2002b). Text. In: Karlík, P., Nekula, M., Pleskalová, J. (eds.), Encyklopedický slovník češtiny. Praha: Nakladatelství Lidové noviny, s. 489.
- Orwell, G. (2009). 1984. Praha: Český spisovatel.
- Peirce, Ch. S. (1958) Collected Papers of Charles Sanders Peirce. Cambridge: Harvard University Press.
- Petrů, E. (2006). Úvod do studia literární vědy. Olomouc: Rubico.
- Popescu, I. I. (2007) Text ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, K. (eds.) Exact methods in the study of language and text (Quantitative linguistics). Berlin/New York: Mouton se Gruyter, s. 557–567.
- Popescu, I. I., Altmann, G. (2007). Writer´s view of text generation. Glottometrics 15, s. 71–81.
- Popescu, I. I., Altmann, G. (2011). Thematic concentration in texts. In: Kelih, E., Levickij, V., Matskulyak, Y (eds.) Issues in Quantitative Linguistics 2. Lüdenscheid: RAM, 110–116.
- Popescu, I. I., Altmann, G., Grzybek, P., Jayaram, B. D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M. N. (2009). Word frequency studies. Berlin/New York: Mouton de Gruyter.
- Popescu, I. I., Čech, R., Altmann, G. (2011). The lambda-structure of texts. Lüdenscheid: RAM.
- Popescu, I. I., Čech, R., Altmann, G. (2012). Some Geometric Properties of Slovak Poetry. Journal of Quantitative Linguistics 19 (2), s. 121–131.

- Popescu, I. I., Mačutek, J., Altmann, G. (2009). Aspects of word frequencies. Lüdenscheid: RAM.
- Popescu, I. I., Mačutek, J., Kelih, E., Čech, R., Best, K. H., Altmann, G. (2010). Vectors and codes of text. Lüdenscheid: RAM.
- Popescu, I. I., Vidya, M.N., Uhlířová, L., Pustet, R., Mačutek, J., Krupa, V., Köhler, R., Jayaram, B.D., Grzybek, P., Altmann, G. (2009). Word frequency studies. Berlin/New York: Mouton de Gruyter.
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
- Rybicki, J. (2012). The great mystery of the (almost) invisible translator: stylometry in translation. In: Oakley, M. and Ji, M. (eds.), Quantitative Methods in Corpus-Based Translation Studies. Amsterdam: John Benjamins, s. 231–248.
- Rybicki, J., Heydel, M. (2013). The stylistics and stylometry of collaborative translation: Woolf's 'Night and Day' in Polish. Literary and Linguistic Computing, 28(4), s. 708–717.
- Sanada, H. (2013). Thematic concentration in Japanese prose. In: Obradovic, I., Kelih, E., Köhler, R. Methods and Applications of Quantitative Linguistics. Selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO), Belgrade, Serbia, April 26–29, 2012. Belgrade: University of Belgrade, s. 130–140.
- Scott, M. (2013). WordSmith Tools. Liverpool: Lexical Analysis.
- Tešitelová, M. (1972). On the so-called vocabulary richness. Prague Studies in Mathematical Linguistics 3, s. 103–120.
- Těšitelová, M. (1974). Otázky lexikální statistiky. Praha: Academia.
- Těšitelová, M. (1983). Psaná a mluvená odborná čeština z kvantitativního hlediska. Praha: Ústav pro jazyk český ČSAV.
- Těšitelová, M. (1987). Kvantitativní lingvistika. Praha: SPN.
- Těšitelová, M. a kol. (1987). O češtině v číslech. Praha: Academia.

- Tuldava, J. (1977). O kvantitativnych charakteristikach bogatstva leksičeskogo sostava chudožestvennych tekstov. Acta et Commentationes Universitatis Tartuensis 437, s. 159–175.
- Tuldava, J. (1995). On the relation between text length and vocabulary size. In: Tuldava, J. (ed.), Methods in quantitative linguistics. Trier: WVT, s. 131–150.
- Tuzzi, A., Popescu, I. I., Altmann, G. (2010a). The golden section in texts. In: ETC – Empirical Text and Culture Research 4: 30–41.
- Tuzzi, A., Popescu, I. I., Altmann, G. (2010b). Quantitative Analysis of Italian texts. Lüdenscheid: RAM.
- Uhlířová, L. (2005). Quanitative linguistics in the Czech Republic. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), Quantitative Linguistics. An international Handbook. Berlin/New York: de Gruyter, s. 129–135.
- Wetzel, L. (2014) Types and Tokens. In Zalta E. N. (ed.) The Stanford Encyclopedia of Philosophy (Spring 2014 Edition). [dostupné online na: http://plato.stanford.edu/archives/spr2014/entries/types-tokens/]
- Wetzel, L. Type versus Token. (2006). In: Brown, E. K., Asher, R. E., Simpson, J. M. Y. (eds.) Encyclopedia of Language & Linguistics, s. 199–202.
- Wilson, A. (2009). Vocabulary richness and thematic concentration in internet fetish fantasies and literary short stories. Glottotheory 2(2), s. 97–107.
- Wimmer, G. (2005). The type-token relation. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), Quantitative Linguistics. An International Handbook. Berlin/New York: de Gruyter, s. 361–368.
- Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S., Wimmerová, S. (2003). Úvod do analýzy textov. Bratislava: Veda.
- Yule, G.U. (1944). The statistical study of literary vocabulary. Cambridge: Cambridge University Press.
- Ziegler, A., Altmann, G. (2002). Denotative Textanalyse. Wien: Praesens.

10. Příloha

10.1. Nastavení programu Stylo v MWF analýze

corpus.format = "plain"

corpus.lang = "Other"

analyzed.features = "w"

ngram.size = 1

preserve.case = FALSE

encoding = "native.enc"

mfw.min = 100

mfw.max = 100

mfw.incr = 100

start.at = 1

culling.min = 0

culling.max = 0

culling.incr = 20

mfw.list.cutoff = 5000

delete.pronouns = FALSE

use.existing.freq.tables = FALSE

use.existing.wordlist = FALSE

use.custom.list.of.files = FALSE

analysis.type = "CA"

consensus.strength = 0.5

distance.measure = "CD"

sampling = "no.sampling"

sample.size = 10000

length.of.random.sample = 10000

display.on.screen = FALSE

write.pdf.file = FALSE

write.jpg.file = FALSE

write.svg.file = FALSE

write.png.file = TRUE

plot.custom.height = 15

plot.custom.width = 10

plot.font.size = 10

plot.line.thickness = 5

text.id.on.graphs = "both"

colors.on.graphs = "colors"

titles.on.graphs = TRUE

label.offset = 0

add.to.margins = 2

dendrogram.layout.horizontal = TRUE

pca.visual.flavour = "classic"

save.distance.tables = TRUE

save.analyzed.features = TRUE

save.analyzed.freqs = TRUE

dump.samples = FALSE

10.2. Seznam 100 nejčetnějších slov korpusu v MWF analýze

1 a
2 se
3 to
4 je
5 na
6 v
7 že
8 ale
9 jako
10 jsem
11 tak
12 s
13 si
14 co
15 do
16 z

17 já
18 by
19 už
20 k
21 nebo
22 o
23 za
24 i
25 po
26 jen
27 tu
28 jak
29 když
30 ten
31 byl
32 mu
33 ve
34 tam
35 jsou
36 aby
37 není
38 bylo
39 ty
40 ještě
41 řekl
42 pro
43 ani
44 toho
45 něco
46 vám
47 pan
48 nic

49 má
50 ho
51 jeho
52 li
53 u
54 teď
55 od
56 mi
57 pane
58 ne
59 až
60 člověk
61 bych
62 pak
63 než
64 tím
65 tady
66 byla
67 mne
68 své
69 mně
70 ji
71 tom
72 vás
73 být
74 tedy
75 jste
76 jenom
77 snad
78 ta
79 ti
80 ze

81 jsme
82 tomu
83 který
84 kde
85 měl
86 prokop
87 vy
88 které
89 bude
90 nás
91 mezi
92 proto
93 té
94 sám
95 jí
96 ní
97 před
98 bez
99 nám
100 musí

10.3. 200 nejfrekventovanějších slovních tvarů v různých žánrech

#	román	povídka	cestopis	pohádka	studie	sloupek	dopis	báseň
1	a	a	a	a	a	a	a	a
2	se	se	se	se	je	se	se	se
3	to	to	je	to	se	to	to	v
4	na	na	na	na	v	je	že	je
5	je	je	to	je	že	na	je	to
6	v	že	v	v	na	v	jsem	na
7	že	v	jako	tak	to	že	na	že
8	jako	jsem	ale	že	k	ale	v	jen

9	ale	ale	že	ale	nebo	nebo	by	z
10	tak	tak	jsem	si	jsou	z	mi	já
11	co	já	z	do	jako	s	ale	co
12	jsem	si	s	z	o	si	vám	nic
13	s	s	nebo	co	ale	do	tak	my
14	si	co	do	jako	co	tak	o	jsme
15	do	do	jsou	jsem	jen	by	si	nám
16	já	jako	i	když	umění	i	s	nás
17	by	ten	si	ten	není	jako	vás	jsem
18	už	z	tak	řekl	i	co	do	za
19	z	už	co	já	z	o	co	není
20	k	by	už	s	ve	už	už	po
21	za	byl	tu	už	estetické	jsou	k	o
22	mu	řekl	po	za	s	ještě	i	už
23	po	tu	tam	tu	pro	k	z	k
24	o	když	za	pan	tedy	aby	mne	tu
25	jak	za	ve	mu	jeho	za	já	jsou
26	byl	k	k	aby	nýbrž	po	jako	ten
27	jen	po	o	jak	li	ve	za	s
28	tu	pan	by	i	jež	když	bych	i
29	tam	mu	jen	po	by	tu	li	když
30	nebo	jak	ještě	k	estetický	není	jak	do
31	když	pane	ani	jen	estetického	jak	teď	teď
32	i	o	když	ani	tak	ani	jen	puč
33	bylo	jen	není	tam	předmětu	li	snad	pan
34	ty	bylo	tady	nebo	tím	má	pro	ta
35	ten	nebo	jak	vám	než	toho	ani	tam
36	aby	mně	pak	byl	do	jen	ještě	tak
37	ve	ho	aby	pane	předmět	ten	po	ministr
38	prokop	toho	člověk	ty	však	vám	nebo	zas
39	ho	vám	kde	ho	objektivní	musí	než	pro
40	řekl	nic	byl	by	od	něco	jste	té
41	něco	ty	mezi	u	má	režisér	byl	tom
42	pan	tam	toho	toho	jest	pro	ve	bez
43	ještě	ani	až	ještě	za	tam	toho	každý
44	pro	aby	než	má	které	své	bylo	snad
45	nic	něco	ty	ji	tu	který	něco	ale
46	teď	není	mi	ve	může	bude	vy	jak
47	není	i	od	mně	být	u	tu	měli
48	ne	ve	li	jsou	nás	autor	nic	si
49	jeho	člověk	bych	teď	vědomí	já	není	nepsat
50	toho	byla	pro	o	krásy	jsem	aby	lépe
51	má	jste	u	od	vcítění	jsme	mně	sem
52	ani	paní	má	pak	nám	být	když	má

53	u	ta	jeho	bylo	své	které	tom	ji
54	byla	ji	mne	e	vůbec	pak	byste	bitte
55	až	tom	vám	protože	zcela	tím	ten	dnes
56	tady	vás	ten	král	cit	zahradník	mám	až
57	jsou	má	nic	pro	poznání	mu	od	nedá
58	ti	ještě	snad	není	tomu	než	abych	bude
59	bych	u	ze	ti	estetika	od	ne	ach
60	své	jeho	já	nic	něco	nás	které	ti
61	standa	pak	který	povídá	city	ty	být	dělat
62	člověk	pro	země	až	krása	až	jsou	kdo
63	mi	bych	neboť	tedy	atd	jeho	až	by
64	od	mne	bylo	jí	tato	teď	pak	trumfy
65	vám	on	proto	ne	tento	tady	bude	páni
66	jenom	jí	něco	tady	neboť	ze	mnoho	vás
67	li	teď	mají	ta	hodnoty	dále	ty	dušičky
68	ji	snad	lidé	té	subjektivní	tomu	má	mezi
69	jste	od	jenom	kolbaba	bez	tom	tím	však
70	pane	tím	nad	vás	všechny	nám	své	lid
71	mne	ne	moře	bych	při	byl	rád	ze
72	tím	vy	viděl	tomu	proto	nic	protože	li
73	pak	měl	nás	mne	umělecké	člověk	tedy	mně
74	mně	té	které	princezna	díla	ta	také	člověk
75	tom	než	jež	no	toho	bylo	budu	od
76	být	jenom	jsme	byla	lze	mezi	který	spěte
77	měl	tedy	ta	jeho	dle	jenom	tomu	pokoji
78	vy	až	své	jste	ani	také	dnes	věrné
79	ze	proč	bez	měl	po	ji	chtěl	jako
80	jsme	své	být	proto	dílo	nýbrž	měl	ráji
81	kde	jsou	jejich	jenom	jak	my	aspoň	refrén
82	oči	tady	nevím	vy	mezi	nad	ti	stát
83	bude	být	staré	své	estetiky	kde	proto	zkrátka
84	než	víte	tím	kde	soud	kteří	ta	trochu
85	ta	vždyť	mu	tom	býti	pan	pane	schůze
86	vás	ze	asi	zase	zde	lidé	člověk	jezte
87	tomu	protože	také	tím	života	říká	věci	ve
88	který	který	nýbrž	kouzelník	život	tedy	myslím	u
89	sebe	tomu	tedy	něco	aby	ne	přece	ještě
90	jí	před	světa	lotrando	si	mít	u	mne
91	sám	nevím	pod	snad	věci	svou	prosím	byli
92	snad	sám	té	jsme	pak	pane	tam	toho
93	před	mi	ne	člověk	vše	proto	dobře	mrtví
94	mezi	ti	která	jeden	estetická	pod	té	ani
95	nad	li	trochu	před	tom	kdy	jednou	vláda
96	ano	takový	před	tři	nejsou	všechny	příliš	pijte

97	proč	ní	lidí	ní	jenž	jednou	mohl	místo
98	pod	jsme	byla	než	ní	první	nevím	zima
99	bondy	šel	těch	dášeňka	ovšem	obyčejně	prezidente	ba
100	asi	bez	protože	psí	já	neboť	jsme	jí
101	té	všechno	jiné	sidney	sobě	kdyby	bez	píseň
102	které	přece	kolem	takové	proti	která	tě	lidí
103	jsi	prosím	místo	li	podle	snad	která	kam
104	ruce	musí	řekl	mi	tyto	jiné	mít	byl
105	chtěl	viděl	světě	ze	avšak	film	dopis	denně
106	ní	někdo	hory	pod	myšlení	protože	sám	práce
107	svou	bude	vidět	všechno	estetických	vůbec	trochu	adame
108	tedy	mám	tomu	kdy	ze	ní	nám	zrovna
109	hlavou	kdyby	město	víte	sám	víc	byla	jenž
110	všechno	proto	tom	magiá	jej	bych	třeba	práci
111	prý	její	víc	pán	spíše	kdo	ovšem	válce
112	nás	svou	pořád	tě	objektu	jež	život	někde
113	mám	nějaký	skoro	jednou	svět	všech	abyste	také
114	proto	takové	nich	mám	všech	přece	nás	chce
115	kdyby	kde	všude	sám	hodnota	třeba	docela	světa
116	třeba	oči	ulice	musí	vždy	práce	asi	praze
117	kdo	kdo	docela	my	nikoliv	mají	můj	ovšem
118	hordubal	jeden	ji	františek	právě	filmu	kdy	věci
119	taky	chtěl	černé	takový	naše	měl	zase	hraju
120	on	nám	loď	babička	předmětem	někdy	nemohu	všecko
121	nu	sebe	jim	oči	esteticky	jejich	mé	proti
122	nevím	nikdy	ní	řekla	rozumění	před	těch	živly
123	ním	nad	ti	hall	už	byla	tolik	říci
124	člověče	věc	svou	on	hodnot	všechno	ji	ať
125	očima	my	či	který	zvláštní	ti	karel	strany
126	přece	vůbec	mnoho	jůra	každý	dnes	tady	pranic
127	nám	nikdo	abych	zrovna	čistě	těch	víc	eso
128	musí	těch	zase	taky	věcí	jim	svou	nesu
129	víš	které	příliš	ním	u	nich	jsi	kule
130	pepek	věci	vůbec	víš	zároveň	aspoň	vůbec	zde
131	život	komisař	jeden	trochu	skutečnosti	trochu	váš	nyní
132	jednou	mohl	teď	vody	více	bez	psát	jedna
133	prosím	trochu	každý	inu	požívání	podle	kdyby	leč
134	adam	mě	zemi	všichni	vlastní	země	ze	první
135	dál	nás	vše	vždyť	svou	potom	věc	stojí
136	starý	tě	ho	šel	soudy	noviny	buď	přec
137	rukou	pravil	celé	dal	musí	ano	říci	zkušenost
138	bez	dva	vlastně	starý	platnost	ovšem	vaše	tož
139	pořád	abych	PRAVDA	nás	mu	věci	dosud	ty
140	tě	no	člověka	těch	sebe	jeden	psal	stará

141	dobře	případ	sebe	doktor	krásné	jste	čapek	hlavně
142	takové	totiž	stojí	být	čili	nikdo	velmi	musí
143	někdo	chvíli	sám	prý	estetickém	každý	všecko	mu
144	jde	začal	všechno	přece	možno	konečně	krásné	bylo
145	hlavu	třeba	podle	viděl	teprve	tento	tebe	pane
146	dělat	ano	velmi	bez	soudu	nikdy	mezi	jich
147	jim	myslím	někdy	honem	konečně	proč	kterou	letos
148	víc	ke	anglie	vořížek	subjektu	dobře	nyní	létě
149	takový	sem	umění	lidé	filozofie	té	psaní	nich
150	něho	muž	nahoře	ke	světa	vás	pořád	vám
151	mě	doktor	nikdy	jsi	teorie	jaksi	před	velké
152	mohl	tři	všech	hoši	nic	dát	musím	smrti
153	trochu	domů	lodi	cizinec	toto	zase	ní	jenom
154	zase	vše	dole	děti	jich	při	při	jiné
155	dva	asi	hor	hned	něm	kus	kde	nejsou
156	těch	jednou	vypadá	moc	bylo	lidí	skoro	vaše
157	no	rád	krásné	celý	smyslu	může	právě	svůj
158	my	říci	měl	nám	který	mně	píšu	kdyby
159	jo	mejzlík	věci	zvolal	dojmu	ať	sebe	války
160	vždyť	potom	dál	jej	této	takové	totiž	naše
161	carson	den	anglii	pravil	ne	byly	víte	znova
162	skoro	jsi	my	vašek	jiné	nějaký	atd	taková
163	také	aspoň	takové	pes	která	věc	nikdy	všichni
164	věci	celý	přes	soudce	zkušenosti	nemá	lidí	pardon
165	mloci	zase	myslím	jim	estetickou	jiných	jinak	jaro
166	pán	zrovna	nám	proč	zdá	několik	vím	kříž
167	jeden	holub	aspoň	také	pokud	život	radost	krátce
168	kapitán	noci	nejsou	chtěl	sama	muž	dny	být
169	ať	život	chvíli	dělat	např	divadelní	dost	noviny
170	najednou	člověka	vás	nikdo	PRAVDA	musíme	pokud	hráče
171	celý	konečně	kteří	povídám	nich	řekne	práce	prý
172	víte	něho	jednou	přitom	její	například	mu	fuk
173	nikdy	lída	pokud	vzal	ji	tři	všechno	několik
174	tři	tohle	první	zemi	našeho	budou	budete	vídeň
175	sebou	očima	všechny	kočka	hodnocení	pokud	vždyť	udržal
176	aspoň	svého	totiž	ba	požitek	naše	den	atd
177	byly	ono	říci	taková	uměleckého	dělat	slečno	zase
178	lidé	teda	tolik	tuhle	hodnotu	sám	jiné	sám
179	ruku	ach	celý	hlas	totiž	místo	vámi	jednou
180	vůbec	bože	život	které	dále	máme	oddaný	státní
181	něm	ním	jedna	něm	člověka	praví	málo	říká
182	vlastně	hlas	samé	vodník	mají	mi	dr	kde
183	sem	olga	což	třeba	formy	právě	něm	víte
184	potom	hned	jaksi	den	byl	herci	sobě	svět

185	neboť	stalo	takový	země	byla	redakce	zdá	pár
186	nebylo	nebylo	den	sebe	vskutku	jinak	jistě	příteli
187	pravda	povídá	stromy	abyste	zažití	den	musí	nebo
188	ke	najednou	proti	dva	poměr	klára	máte	kra
189	její	jde	mohl	ptal	zkušenost	přitom	kteří	jedno
190	chvíli	jaksi	kdy	sem	jenom	let	kdybych	alles
191	která	dobře	zvláštní	každý	přece	teprve	svůj	sedm
192	docela	lidé	nyní	přišel	nemůže	toto	udělat	mít
193	abych	ruce	oči	bude	jsme	tohle	chci	jedni
194	místo	přišel	nimi	někdy	vkusu	nemůže	což	panu
195	povídá	hlavou	sem	místo	vnitřní	dětí	tož	kříže
196	nebyl	povídám	například	někdo	sice	proti	jaksi	aby
197	stojí	byly	ano	nich	požitku	kterou	anielko	tím
198	sobě	dr	ať	byste	citu	jej	jež	dál
199	nějak	nějaké	nikdo	kdyby	krásu	asi	takové	navrací
200	nyní	dvě	města	jedna	osobní	jde	moje	jaký

10.4. 200 nejfrekventovanějších lemmat v různých žánrech

#	román	povídka	cestopis	pohádka	studie	sloupek	dopis	báseň
1	a	být	a	a	být	a	být	být
2	být	ten	být	ten	a	být	a	ten
3	se	se	se	se	se	se	ten	a
4	ten	a	ten	být	v	ten	se	se
5	on	on	na	on	ten	v	já	já
6	na	já	v	na	estetický	na	ty	v
7	v	na	jako	v	on	že	že	na
8	já	v	ale	já	že	on	by	on
9	že	že	že	tak	na	mít	v	z
10	mít	ale	z	že	jenž	z	na	že
11	by	tak	já	ale	tento	by	mít	mít
12	co	mít	on	mít	k	který	ale	jen
13	jako	pan	člověk	do	nebo	já	on	co
14	ale	ty	s	říci	předmět	ale	který	nic
15	tak	by	mít	z	jako	nebo	tak	pan
16	s	co	nebo	pan	co	s	o	za
17	do	říci	do	ty	já	do	moci	o
18	z	s	by	co	z	tak	co	po
19	ty	z	který	jako	všechno	co	aby	už
20	už	do	co	když	o	i	z	tu
21	k	jako	i	by	ale	jako	s	i

22	svůj	vědět	tak	aby	jeho	svůj	můj	když
23	říci	člověk	svůj	s	umění	o	k	s
24	člověk	už	všechno	už	jen	aby	do	dát
25	vědět	jít	už	za	svůj	všechno	svůj	k
26	pan	k	tu	vědět	i	už	už	svůj
27	za	svůj	po	k	který	k	i	teď
28	aby	tu	tam	tu	můj	člověk	jako	můj
29	po	když	k	všechno	hodnota	muset	psát	ministr
30	moci	za	vidět	takový	mít	moci	chtít	do
31	který	moci	aby	svůj	krása	ještě	tvůj	pán
32	o	aby	za	jak	cit	za	za	puč
33	jak	po	o	i	moci	tento	vědět	člověk
34	všechno	všechno	jen	člověk	umělecký	ty	člověk	stát
35	jít	jak	ještě	po	s	po	li	každý
36	jen	takový	jeho	jít	pro	když	jak	dělat
37	jeho	o	ty	jen	tedy	jiný	teď	všechno
38	tu	který	starý	ani	soud	režisér	muset	jeden
39	tam	jen	jiný	tam	estetika	tu	jen	vědět
40	prokop	jeho	moci	nebo	dílo	jak	snad	tak
41	nebo	nebo	ani	dát	nýbrž	ani	pro	by
42	když	nic	země	moci	by	li	věc	dobře
43	chtít	muset	tento	jeden	život	můj	všechno	tam
44	i	chtít	když	povídat	li	jeho	ani	ty
45	ruka	vidět	jenž	u	objektivní	stát	nebo	jít
46	něco	něco	tady	král	jiný	říkat	ještě	hrát
47	muset	tam	jak	který	každý	dát	po	zas
48	vidět	ani	celý	vidět	tak	zahradní k	nic	muset
49	takový	nějaký	pak	muset	věc	říci	něco	chtít
50	oko	jeden	svět	ještě	svět	jen	než	pro
51	ještě	stát	vědět	chtít	než	něco	říci	[:bitte?]
52	můj	i	říci	hlava	sám	autor	den	kdo
53	nic	dát	jeden	princezn a	do	pan	pan	starý
54	pro	můj	takový	o	od	daleko	tu	nepsat
55	hlava	paní	kde	jeho	člověk	jeden	dát	bez
56	Standa	přijít	mezi	teď	však	chtít	jiný	svět
57	mlok	věc	až	od	objekt	pro	rád	jak
58	dát	kdyby	než	celý	forma	jenž	když	ale
59	stát	oko	moře	doktor	za	takový	takový	sem
60	teď	ještě	od	pán	PRAVDA	tam	tento	dnes
61	ne	u	hora	pak	obecný	nějaký	kdyby	snad
62	tento	říkat	černý	kouzelník	něco	dělat	napsat	ach
63	ani	pak	každý	protože	dojem	každý	od	práce
64	u	pro	li	nic	lze	dobrý	život	trumf

65	dělat	od	pro	e	tu	u	dobrý	spát
66	celý	dělat	u	můj	subjektivní	film	sám	válka
67	až	sám	velký	dělat	krásný	den	krásný	mrtvý
68	tady	snad	stát	pro	příroda	jít	mnoho	až
69	jeden	teď	krásný	oko	aby	věc	hodně	lid
70	sám	ruka	město	voda	vcítění	půda	myslit	přijít
71	od	ne	nic	nějaký	skutečnost	od	dopis	moci
72	nějaký	den	něco	až	vědomí	rok	dobře	tvůj
73	jenom	celý	loď	ruka	zkušenost	pak	také	červený
74	li	tenhle	místo	země	zcela	než	ne	kříž
75	život	než	muset	den	muset	kdyby	pak	jenž
76	starý	doktor	ulice	kočka	bez	velký	až	nést
77	také	myslit	snad	pes	vůbec	až	vidět	velký
78	pán	jenom	chtít	tedy	poznání	vidět	jeden	dušička
79	daleko	dva	dát	ne	vlastní	život	udělat	však
80	kdyby	tedy	neboť	přijít	subjekt	teď	protože	eso
81	jenž	každý	voda	tady	teorie	tady	tedy	li
82	pak	až	nad	starý	možný	ruka	jenž	ráj
83	jiný	někdo	proto	Kolbaba	celý	vědět	prosit	mezi
84	přijít	starý	oko	také	požitek	hodně	dnes	celý
85	říkat	proč	jít	jet	takový	také	nový	věc
86	svět	tady	jenom	Lotrando	představa	celý	přece	jíst
87	kde	jenž	dobrý	babička	jistý	první	celý	jiný
88	věc	bez	hodně	sám	atd	práce	aspoň	střílet
89	než	jiný	les	říkat	neboť	nic	proto	přece
90	dva	povídat	daleko	začít	věda	země	u	říci
91	snad	vždyť	nějaký	no	při	zahrádka	tam	jako
92	dobrý	začít	bílý	každý	obsah	dítě	nějaký	strana
93	druhý	rok	malý	proto	proto	noviny	bez	nebo
94	kdo	protože	můj	vzít	činnost	žádný	láska	pokoj
95	mluvit	udělat	také	kde	filozofie	herec	prezident	věrný
96	udělat	před	věc	něco	dle	starý	jednou	od
97	dívat	muž	bez	psí	ani	svět	velký	vláda
98	před	mluvit	dělat	jenom	poměr	mezi	příliš	vzít
99	nad	tvůj	zelený	tři	po	jenom	práce	pravý
100	Hordubal	najít	ruka	udělat	jak	hlava	jít	zodpovědný
101	mezi	prosit	Anglie	stát	historický	dva	říkat	zima
102	den	hlava	říkat	bílý	mezi	místo	dělat	den
103	ano	život	život	dobrý	fakt	druhý	jeho	jaro
104	každý	li	samý	zase	metoda	nýbrž	daleko	jeho
105	jaký	chvíle	veliký	ocásek	jeden	dobře	trochu	ani
106	pod	přece	hlava	snad	zde	nad	mladý	slovo
107	dobře	pán	anglický	sedět	zvláštní	kde	stát	plenit

108	tvůj	nikdo	cesta	maminka	smysl	muž	třeba	Vídeň
109	proč	případ	skála	dopis	platnost	tedy	čas	hrob
110	bond	rozumět	vypadat	tvůj	jednota	obyčejně	ovšem	refrén
111	hodně	noc	sám	Dášeňka	určitý	ne	pěkný	Adam
112	myslit	kdo	asi	jiný	myšlení	přijít	dovést	schůze
113	Adam	Lída	modrý	hlas	existovat	nový	všecek	zkrátka
114	nový	rád	pěkný	dva	umělec	Klára	svět	pít
115	asi	slyšet	nýbrž	před	pak	jaký	docela	u
116	někdo	nechat	kdyby	Hall	skutečný	sám	tolik	první
117	rok	daleko	národ	dostat	problém	proto	asi	živel
118	místo	dostat	tedy	černý	proti	tenhle	zase	zelený
119	voda	jaký	jaký	než	ovšem	doba	kdy	kam
120	začít	komisař	žádný	druhý	podle	pod	Karel	jarní
121	tedy	žádný	druhý	slečna	psychologický	kdy	možný	místo
122	velký	dobrý	pod	pravit	nikoliv	udělat	tady	daleko
123	prý	hlas	dva	pod	avšak	národ	radost	přítel
124	povídat	okno	jet	Sidney	vnitřní	jeviště	vůbec	vina
125	pepka	druhý	červený	li	dát	jednou	mluvit	vidět
126	slyšet	slovo	ne	svět	libost	neboť	veliký	ba
127	proto	čekat	před	kdy	vysoký	kdo	Čapek	také
128	doktor	proto	trochu	Vašek	lidský	kus	slovo	rok
129	žádný	podívat	protože	kdyby	spíše	veliký	přijít	znát
130	bůh	veliký	podívat	[:magiá?]	souvislost	snad	ruka	trochu
131	bez	kde	kolem	voříšek	nic	redakce	čekat	sloužit
132	sedět	hodně	svatý	nechat	výraz	sto	slečna	píseň
133	veliký	velký	PRAVDA	jednou	otázka	lidský	buď	žádný
134	třeba	pravit	kůň	daleko	vkus	protože	nechat	ještě
135	kapitán	sedět	den	rád	různý	list	každý	tož
136	nechat	vzít	dům	František	hodně	role	dostat	Praga
137	nu	dítě	konec	nikdo	osobní	bez	dosud	muž
138	přece	nad	myslit	rok	vždy	vůbec	málo	říkat
139	rád	dobře	jakýsi	dítě	rozumění	nikdo	přát	bývat
140	práce	ptát	strom	kolega	zdát	filmový	velmi	zde
141	prosit	ulice	dítě	vodník	nutný	několik	holčička	jet
142	tenhle	nikdy	pořád	vrána	právě	hra	věřit	zrovna
143	chvíle	žena	vysoký	dědeček	citový	divadelní	těšit	ať
144	rozumět	tři	jmenovat	umět	požívání	málo	rok	různý
145	vzít	dveře	skoro	noc	konečně	hrát	zdát	jaký
146	země	cítit	okno	místo	individuální	poslední	smět	denně
147	cítit	poslouchat	kráva	hoch	čistě	přece	dva	udělat
148	PRAVDA	vůbec	všude	veliký	esteticky	třeba	jet	pranic
149	jednou	nový	jezero	zvíře	povaha	různý	jaký	sám
150	Štěpán	také	docela	[:jůra?]	daleko	scéna	psaní	vyhrát

151	tři	trochu	lidský	ptát	už	tři	znát	aby
152	tvář	hodina	mluvit	trochu	případ	někdy	mezi	kde
153	Carson	mladý	noc	zvolat	myšlenka	dostat	šťastný	takový
154	noc	spíše	mnoho	cizinec	zároveň	dnes	nyní	srdce
155	pořád	no	přijít	noha	zážitek	pán	pořád	Eva
156	noha	pokoj	býk	jmenovat	u	před	brzy	hlavně
157	dítě	totiž	což	myslit	jediný	tvůj	poslat	pěkný
158	prst	tvář	ostrov	zrovna	líbit	začít	duše	chodit
159	princezna	holub	chvíle	sto	nový	aspoň	odpustit	proti
160	podívat	Olga	nový	drak	daný	námět	při	ovšem
161	nikdo	jet	či	sedm	pohyb	podle	dost	kule
162	dostat	ano	kostel	Ina	velký	trochu	před	ruka
163	malý	třeba	těžký	spát	vztah	pravit	doktor	noviny
164	černý	onen	sever	podívat	kritika	politický	Brno	nějaký
165	Juraj	Mejzlík	kraj	přece	čili	chvíle	tenhle	umět
166	myslet	konečně	zase	pohádka	druhý	oko	právě	leč
167	dlouhý	domů	příliš	vždyť	práce	voda	skoro	všecek
168	zdát	sem	nakreslit	dobře	jev	válka	kde	bratr
169	trochu	hledat	divný	hodně	moment	potom	totiž	psát
170	zase	asi	dívat	jaký	teprve	zkouška	což	zkušenost
171	no	krok	vůbec	prý	princip	slovo	oddaný	nyní
172	jo	jednou	teď	najít	pojem	slyšet	nikdy	prázdniny
173	slovo	místo	sedět	bez	faktor	malý	večer	hráč
174	najít	dívat	bůh	žádný	praktický	stůl	bát	někde
175	vždyť	zůstat	najít	čekat	stát	ovšem	jinak	hlava
176	skoro	dopis	přístav	kdo	dobrý	mladý	atd	dva
177	ať	první	dostat	mladý	část	ano	báseň	několik
178	první	ráno	vlastně	běžet	hodnocení	krásný	noc	veliký
179	najednou	cesta	rok	honem	čistý	pravý	pokud	doba
180	žena	chodit	umění	malý	krásno	psát	cítit	státní
181	ptát	zdát	sto	tisíc	účel	někdo	Praha	karta
182	doba	potom	krása	moc	podmínka	konečně	poslední	prý
183	nikdy	aspoň	první	strážník	mravní	povídat	myslet	jenom
184	kůň	zrovna	tři	letět	úkol	divadlo	starý	hodně
185	dveře	zase	noha	hned	obraz	jaro	přijet	mír
186	aspoň	tento	podle	někdo	ne	jakýsi	žádný	dojít
187	Polana	počkat	někdy	dovést	stávat	těžký	chvíle	stříbrný
188	vůbec	znát	dlouhý	soudce	historie	plný	číslo	ulice
189	vlastně	teda	velmi	proč	stav	proč	vždyť	krátce
190	zůstat	svět	dobře	sednout	onen	nikdy	raději	znova
191	potom	ticho	psát	nebe	životní	kvést	jakýsi	zase
192	sem	brzy	udělat	štěně	oblast	zase	hledět	Udržal
193	neboť	ach	žít	přitom	cíl	rád	číst	řeč
194	ukázat	strašný	dole	rusalka	pokud	konec	list	zůstat

195	rychle	krásný	nahoře	první	proces	při	dávat	pod
196	cesta	Bož	tenhle	velký	například	smět	ostatní	nový
197	mladý	peníz	zvláštní	brzy	zažití	jaksi	dojít	mluvit
198	konec	jaksi	nikdy	ucho	zákon	což	žít	pardon
199	docela	bolest	jméno	tenhle	nějaký	ať	PRAVDA	kůže
200	rameno	najednou	strašný		pouhý	čtenář	jistý	zcela