https://doi.org/10.58377/slav.2023.5.06

Stylometrická analýza církevněslovanských textů české provenience*

Radek Čech (Brno) – Miroslav Vepřek (Olomouc)

Stylometric Analysis of the Church Slavonic Texts of Czech Origin

The paper presents a pilot study of stylometric analysis of Czech Church Slavonic texts. The aim of the study is to measure similarities / differences among texts based on selected quantitative characteristics. Specifically, the average token length (ATL), moving average type-token ratio (MATTR), and text distances determined by normalized frequencies of the most frequent words (MFW) are applied. For the analysis, we used a corpus of twelve Church Slavonic literary writings attributed (with various probability) to Czech authors in the 10th and 11th centuries. In addition, two more textual sources were added (Codex Suprasliensis and the Life of St. Methodius) to compare the results and get a more complex view of relationships among texts. The results show the plausibility of the application of methods on this specific sample of texts.

Keywords: stylometric analysis, Czech Church Slavonic, token length, lexical diversity, cluster analysis

1. Úvod

Církevněslovanské památky české redakce představují specifický soubor textů, které sice spojuje společný původ v Čechách 10.–11. stol., avšak v mnohých aspektech se navzájem liší, a to na nejrůznějších úrovních (hláskových, ortografických, lexikálních, syntaktických, žánrových atd.). Jednou z možností, jak zkoumat podobnosti / rozdílnosti textů, je aplikace stylometrických metod. Na základě těchto metod lze kvantitativně vyjádřit, do jaké míry jsou si dané texty s ohledem na zvolené metriky podobné, či do jaké míry se liší. Kromě čistě heuristického přístupu, v jehož rámci jsou texty porovnány bez jakýchkoliv teoretických předpokladů a následně jsou zpravidla hledány potenciální interpretace zjištěných podobností a rozdílů, je možné dané metody použít k ověřování domněnek, otázek a hypotéz opírajících se o výsledky předchozích filologických výzkumů. V naší studii přistupujeme k analýze vybraného vzorku církevněslovanských památek oběma způsoby. Na jedné straně očekáváme například to, že by zvolené metody měly odhalit žánrovou blízkost určitých textů. Na straně druhé může heuristický přístup vést k nastolení nových perspektiv či inspirovat k jinému uvažování nad danými texty. V neposlední řadě jsme si vědomi, že konečná podoba analyzovaných textů je výslednicí celé řady faktorů (doba vzniku, vliv originálního textu v případě překladů, autorství, posvátný charakter, vliv redakčních úprav atd.), které na sebe mohou vzájemně působit tak, že výrazně ovlivní či omezí možnosti využití uvedených metod na tento typ textů.

Pro analýzu jsme použili 12 textů s víceméně prokázaným či alespoň zvažovaným archetypem vzniknuvším v přemyslovských Čechách 10.–11. stol. a pro komparaci dva

* Příspěvek byl podpořen v rámci projektu Filozofické fakulty Univerzity Palackého v Olomouci IGA_FF_2022_038 „Bohemistika – od tradice k digitalizaci (od minulosti k současnosti)“.

texty pocházející z jiných redakcí nejstaršího slovanského jazyka. Texty porovnáváme prostřednictvím tří charakteristik: průměrné délky slova (tokenu), klouzavého průměru poměru typů a tokenů v textu a analýzy založené na porovnání normalizovaných četností nejfrekventovanějších slov. Předpokládáme, že na základě těchto stylometrických charakteristik by k sobě měly mít nejblíže texty stejného žánru. Je otázkou, do jaké míry tyto metody zachytí vliv toho, a) zda jde o texty překladové, či původní, b) zda hraje roli doba vzniku (tj. texty ze stejného období by k sobě měly být blíže) a c) společný překladatel (překladatelská škola).

Článek je uspořádán následujícím způsobem. Ve 2. kapitole bude představena metodologie zpracování textů a způsob výpočtu jednotlivých stylometrických indexů. Ve 3. kapitole se budeme věnovat vymezení jazykového materiálu a jeho vlastností, v kapitole 4. prezentujeme výsledky a jejich interpretaci. Studii uzavírá Závěr, v němž nejen sumarizujeme zjištěné poznatky, ale také nastiňujeme možnosti dalšího bádání.

2. Metodologie

Materiál pro stylometrické výzkumy byl získán digitalizací církevněslovanských památek českého původu. Protože se níže charakterizované analýzy zaměřují na zkoumání vlastností textu a nikoli např. hláskoslovných či ortografických specifik, primárním zdrojem digitalizace byly zejména reprezentativní edice csl. památek. Při přepisech[1] nebyly reflektovány diakritické a jiné znaky, které nemají přímý vliv na hláskovou, morfologickou, lexikální a syntaktickou podobu textu, interpunkce a punktace byly sjednoceny jedním znakem. Následně byly pro účely automatizovaného zpracování rozepsány veškeré abreviatury, přičemž doplněné části byly rozvedeny v normalizovaném znění dle úzu *Slovníku jazyka staroslověnského* (*SJS*),[2] nadřádkové litery zařazeny do textu a titly označující číselnou platnost byly ponechány, respektive nahrazeny zástupným symbolem, aby jejich písmenné vyjádření nemohlo být případně nesprávně spojováno se slovními tvary (typicky např. se spojkou – и atd.). Konjektury autorů digitalizovaných edic byly zařazovány do textu. Po důkladném zvážení jsme před samotnými stylometrickými analýzami provedli sjednocení některých variantních grafémů: různé varianty *i* (і, ꙇ, ї, й, и) ve znak jediný (и), varianty *o* (ѡ, ѽ, ѳ, ⊖, ѻ > ѻ), *u* (ꙋ, ꙗ, ѹ > ѹ), *y* (ы, ꙑ > ꙑ) a *ę* (ѧ, ѧ > ѧ). I když se takto stírají případná specifika grafiky konkrétních památek, pro výzkum primárně lexikologický, stylistický, textový, ale v zásadě i gramatický nevedou tyto úpravy k žádnému zkreslení.

Pro vzájemné porovnání textů jsme použili tři charakteristiky: 1) průměrnou délku slova (tokenu), 2) klouzavý průměr poměru typů a tokenů v textu a 3) normalizované četnosti nejfrekventovanějších slov.

Průměrná délka slova (ATL) vyjadřuje hodnotu aritmetického průměru délek všech slov (tj. tokenů) v textu a je počítána v počtu grafémů. ATL může odrážet specifické textové charakteristiky zejména s ohledem na známý vztah mezi délkou slova a jeho frekvencí – čím je slovo frekventovanější, tím je zpravidla kratší. Užití delších slov v textu tudíž

1) Přepisy byly pořízeny spoluatorem této studie Miroslavem Vepřkem (M. V.) a pod jeho vedením též skupinou studentů Filozofické fakulty Univerzity Palackého v Olomouci – Bc. Janou Hauschwitzovou, Bc. Janou Jančíkovou, Bc. Hanou Kačmárovou a Bc. Kateřinou Sommerovou.

2) Jsme si vědomi, že v těchto případech nemusejí být rozepsané tvary v souladu s pravopisným a hláskoslovným charakterem daných rukopisů, většinou opisů v mladších redakcích církevní slovanštiny. Avšak vzhledem ke skutečnosti, že zkracovány jsou frekventované lexémy často se opakující v různých textech a že je v těchto různých textech rozepisujeme podle stejných pravidel, pro automatické analýzy není tento postup nijak zavádějící.

znamená použití té části lexika, která často představuje specifickou tematickou doménu. Užívání delších slov může být také vědomým projevem autora, např. v případně záměrné stylizace či archaizace textu.

Klouzavý průměr poměru typů a tokenů (MATTR) vyjadřuje míru diverzifikovanosti slovníku či tzv. slovní bohatství textu. Čím se jednotlivá slova v textu častěji opakují, tím je hodnota MATTR nižší. Způsob výpočtu MATTR je následující: text je segmentován do překrývajících se bloků (tzv. oken), přičemž je pro každé okno vypočítána hodnota poměru typů a tokenů (TTR). Máme-li např. text o délce 1 000 slov a zvolíme velkost okna 100 slov (tokenů), v prvním kroku vypočítáme hodnotu TTR pro prvních 100 slov. Následně posuneme okno o jedno slovo (token) a vypočítáme TTR pro druhé až sté první slovo textu. Takto postupujeme do té doby, než dosáhneme konce textu. Rovnice pro výpočet MATTR je

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

kde N je počet tokenů v textu, L je velikost okna, V je počet typů v daném okně. Výhodou MATTR je zejména jeho nezávislost na délce textu (více viz Covington – McFall 2010).

Analýza nejfrekventovanějších slov (MFW) je založena na porovnání normalizovaných frekvencí vybraného počtu nejčetnějších slov v textu. Prostřednictvím tohoto porovnání je možné stanovit vzdálenosti mezi texty. V této analýze jsme pro výpočet použili Cosine Delta metodu (Smith – Aldridge 2011) a na základě těchto vzdáleností pak byla provedena hierarchická shluková analýza (srov. Koščová et al. 2017). Cosine Delta metodu jsme mezi různými alternativami (např. Burrow's delta, Eder's Delta, Mannhatan distance) zvolili proto, že při určování blízkosti textů je rozhodující spíše rozdíl ve směru než v délce vektorů (Evert et al. 2017).

Provedli jsme 2 měření, přičemž jsme měnili nastavení počtu analyzovaných slov, abychom ověřili, do jaké míry je metoda citlivá na tento parametr. Konkrétně jsme použili 50 a 200 nejfrekventovanějších slov (dále v textu označujeme jako F50 a F200). Při nastavení F50 hrají rozhodující úlohu pro měření vzdáleností synsémantika, v případě F200 se vedle synsémantik objevuje i vliv nejfrekventovanějších autosémantik. V rámci tohoto typu analýzy je také možné nastavit procentuální počet textů, ve kterých se skupina nejfrekventovanějších slov musí objevit (tzv. culling). V naší analýze pracujeme s hodnotou cullingu odpovídající nule, což znamená, že nijak neomezujeme počet textů, v nichž se daná slova musí objevit. Výpočet vzdáleností a shluková analýza byly provedeny prostřednictvím R balíku Stylo (Eder et al. 2016).

3. Vymezení materiálu zkoumaných textů

Nejrozsáhlejším textem zařazeným do primárních analýz jsou tzv. Besědy sv. Řehoře Velikého na evangelium (dále Bes, nejstarší víceméně kompletní rukopis, tzv. Pogodinský z pol. 13. stol., sestává z 328 listů). Jedná se o překlad z latiny, který byl nejspíše pořízen pro potřeby kulturního centra v Sázavském klášteře v druhé polovině 11. století. Památka představuje základní zdroj pro poznání české redakce církevní slovanštiny, neboť obsahuje řadu bohemismů ve všech jazykových rovinách, a to dokonce i hláskoslovných, přestože se text dochoval v mladších ruskocírkevněslovanských opisech, které byly dle obvyklého úzu upravovány (srov. Vepřek 2022, 33–34). Digitalizovaný text byl připraven na základě

dvoudílné edice Václava Konzala (u druhého dílu ve spolupráci s Františkem Čajkou, viz Konzal 2005; Konzal – Čajka 2006).[3] V materiálu pro stylometrické analýzy ponecháváme text edice i s doplněnými chybějícími pasážemi nejstaršího rukopisu, který je v edici převzat z variantního rukopisu, aby byla zachována pokud možno ucelená podoba textu. Pro některé analýzy byl text Bes rozdělen na dílčí části podle jednotlivých kapitol.[4]

Dále do zkoumaného materiálu zařazujeme csl. překlad apokryfního Nikodémova evangelia (dále Nicod). Rovněž tato literární památka bývá řazena svým původem do prostředí přemyslovských Čech 10.–11. století, i když se objevují odlišné názory kladoucí vznik archetypu na slovanský jih (konkrétně do oblasti chorvatské). Pro účely našeho zkoumání byl pořízen přepis edice Andrého Vaillanta (1968) primárně založené na Novgorodském rukopisu ze 14.–15. stol.

Paraliturgický žánr reprezentují v materiálové bázi tři texty – Modlitba ke sv. Trojici (dále Trin), Modlitba sv. Řehoře (Greg) a Modlitba vyznání hříchů (Conf). U všech těchto tří památek byla prokázána jejich souvislost s prostředím přemyslovských Čech 10.–11. století (Mikulka 2015, nejnověji viz Vepřek 2022, 68–69). Trin je zřejmě výtvorem finální fáze českocsl. písemnictví Sázavského kláštera (Konzal 1991) a obsahuje mj. citace zbývajících dvou modliteb. Digitalizace Trin vychází z přepisu nejstarší doložené kompletní verze dochované v rukopisu z kláštera sv. Kateřiny na Sinaji – sign. SinSlav 13, Greg a Conf uvádíme dle edice M. Vepřka (2013), která je založena na znění z ruskocsl. Jaroslavského sborníku ze 13. stol.

Nejrozsáhlejším souborem textů v analyzovaném korpusu jsou památky hagiografické. Tento žánr představuje „těžiště českocírkevněslovanského písemnictví“ (Mareš 2000, 308), s různou mírou pravděpodobnosti je připsáno českému kulturnímu prostředí osm csl. legend. Šest z nich bylo digitalizováno na základě edic publikovaných Františkem V. Marešem (1979), přehled podáváme v následující tabulce:

Název	Zkratka	Datace a provenience zdrojového rukopisu
Život sv. Václava[5]	Venc	16. stol./Vostokovova red., ruskocsl.
Život sv. Víta	Vit	12. stol./ ruskocsl.
Život sv. Benedikta	Ben	14. stol./srbskocsl.
Život sv. Jiří	Georg	14. stol./srbskocsl.
Umučení sv. Apolináře	Apol	16. stol./ruskocsl.
Život sv. Štěpána, papeže	Steph	15. stol./ruskocsl.

3) Vyjadřujeme srdečné poděkování pracovníkům Oddělení paleoslovenistiky a byzantologie Slovanského ústavu AV ČR, v. v. i., v čele s vedoucím pracoviště dr. Štefanem Pilátem za laskavé poskytnutí elektronického textu edice.

4) Tato segmentace byla provedena v zásadě podle jednotlivých kapitol / homilií, avšak vzhledem k poněkud komplikovanějšímu textovému dochování Bes, kdy jsou např. určité části některých homilií vloženy do pasáží neodpovídajících náležitému pořadí v originálním textu a některé úseky chybějící v primárním rukopisu edice jsou doplněny podle rukopisu variantního, celkový počet jednotlivých segmentů (46) je vyšší než počet homilií (40). Jako součást Bes včleňujeme též těsně připojený text Modlitby za nemocné a část Modlitby za čistotu.

5) Tato legenda, tzv. První stsl. legenda o sv. Václavu, je dochována ve třech verzích (redakcích) – kromě zde zahrnuté Vostokovovy ještě verze z charvátskohlaholských breviářů a tzv. Minejní redakce. Vostokovovu redakci zahrnujeme z toho důvodu, že je dle zjištění V. Konzala (1988, 114–116) nejbližší archetypu z 10. století zejména po stránce textologické a syntaktické.

Nejrozsáhlejším hagiografickým textem je Druhá stsl. legenda o sv. Václavu (tzv. Legenda Nikoľského, dále zkracujeme VencNik). Digitalizovaný text v našem souboru byl pořízen přepisem edice Josefa Vašici (1929) podle ruskocsl. rukopisu Kazaňské duchovní akademie z konce 15. či počátku 16. stol. Legenda o sv. Anastázii (dále Anast) byla digitalizována na základě edice publikované F. Čajkou (2011). Tato edice je založena na ruskocsl. znění z rukopisu Q. I. 320 Ruské národní knihovny, který pochází zřejmě z konce 15. či počátku 16. století.

Mimo rámec památek řazených svým původem do českocsl. písemnictví jsme do analýz zahrnuli také dva další texty, které považujeme za důležité pro komparaci s texty jiných redakcí staroslověnštiny. Jedná se nejprve o Život Metodějův (dále Meth), jehož archetyp je produktem velkomoravské literární školy, neboť vznikl dle většinového přesvědčení velmi krátce po smrti arcibiskupa Metoděje (srov. Vašica 1996, 96). Navíc existuje odůvodněný předpoklad, že tato památka byla známa v Čechách 10.–11. stol., a v neposlední řadě V. Konzal předložil hypotézu o textové a jazykové blízkosti Meth a Venc (Konzal 1988, 119–123). Digitalizovaná verze Meth byla získána přepisem edice nejstaršího rukopisného znění z ruskocsl. Uspenského sborníku z 12. stol. (Kotkov et al. 1971, 188–198).

Konečně jsme do analýz zařadili taktéž digitalizovanou verzi Supraslského kodexu, jenž představuje nejrozsáhlejší rukopis řazený do kánonu staroslověnských památek. Vznikl nejspíše v polovině 11. století v Bulharsku a jako vhodný materiál pro komparaci s českocsl. texty se jeví mimo jiné proto, že jeho obsahem jsou zejména životy svatých a homilie (přesněji se jedná o menologion na měsíc březen). Digitalizovaná verze byla získána úpravou elektronické edice publikované v rámci projektu „Corpus Cyrillo-Methodianum Helsingiense“,[6] úpravy se týkaly transliterace do cyrilice a rozvedení zkratek.

4. Výsledky

Daný vzorek textů vyhodnocujeme dvěma způsoby. V prvním případě pracujeme s Bes jako s uceleným textem, ve druhém jsme jej rozdělili na jednotlivé kapitoly a ty jsme porovnali s ostatními texty. Nejdříve se zaměříme na analýzu prvního vzorku. Z hlediska identifikace specifických žánrových charakteristik se jeví jako nejvhodnější metoda MFW. Na jejím základě (u nastavení F50 i F200, srov. obr. 1 a 2) lze jednoznačné identifikovat shluk textů modliteb (Conf, Greg, Trin). Zajímavým výsledkem je rozdělení hagiografických textů do dvou shluků, což signalizuje vliv dalšího faktoru v rámci tohoto žánru (podrobněji viz níže). Blízkost Supr a Bes může souviset s tím, že oba texty obsahují homilie, u Supr alespoň z větší části. V případě použití metody MATTR (viz tab. 1 a obr. 5) se opět jako specifická vyčleňuje skupina textů modliteb (Conf, Greg, Trin). Všechny tři texty vykazují nejnižší hodnoty MATTR – relativně nízká lexikální diverzita je dána jejich specifickým charakterem, kdy je pro ně typické opakování slov či celých pasáží (např. v Conf opakované vyznání hříchů uvozené slovesným tvarem s předložkou: **сѣгрѣшихъ въ…**). Jak je patrné z obr. 5, mezi ostatními hodnotami MATTR, kromě VencNik a Anast (o této podobnosti pojednáme podrobněji níže), neexistují rozdíly, které by korespondovaly s existencí dvou shluků identifikovaných u MFW. V tomto ohledu tedy hagiografické a ostatní texty nevykazují zásadní rozdíly. Na základě výsledků měření ATL (tab. 1 a obr. 6) není z žánrového hlediska možné identifikovat žádné tendence, a to ani v případě textů modlitebních. Pokud ovšem dáme do souvislosti vztah MATTR a ATL (obr. 7), můžeme sledovat, že texty s vyšší

6) Text je v licenci Creative Commons volně k dispozici (viz Lindstedt 2011).

hodnotou ATL mají tendenci vykazovat vyšší lexikální diverzitu.[7] Z grafu je patrné, že u modlitebních textů se projevují jak nižší hodnoty MATTR, tak ATL. Zda jde o náhodný jev nebo obecnější trend není ale možné z tak malého vzorku odvodit.

U vzorku, v němž jsou Bes rozděleny do jednotlivých kapitol, vytváří soubor textů Bes na základě analýzy MFW samostatný shluk stojící v protikladu k ostatním textům (obr. 3 a 4). Při nastavení F50 se ovšem kapitoly 10 a 46 od ostatních textů Bes vydělují a objevují se v rámci shluku modlitebních textů Trin, Conf a Greg. Specifické postavení kap. 10 a 46 je patrné také u ATL, kdy se jedná o texty s nejmenší průměrnou délkou slova. Pokud dáme do souvislosti vztah ATL a MATTR,[8] na obr. 8 vidíme, že oba texty se nacházejí v největší vzdálenosti od centra shluků ostatních kapitol Bes. K tomuto výsledku se vyjádříme ještě níže.

V rámci všech tří zkoumaných parametrů, tj. ATL, MATTR a MFW, se ukazuje pozoruhodná blízkost textů VencNik a Anast. Na základě MFW se tyto dva texty objevují ve společných shlucích při nastaveních F50 i F200 a bez ohledu na to, zda pracujeme s Bes jako celkem, nebo jeho jednotlivými částmi. U měření ATL v případě zařazení kompletního znění Bes se VencNik a Anast objevují na prvních dvou místech (pro jediné tyto dva texty ATL > 5), u rozděleného textu Bes potom mezi VencNik a Anast vstupuje 11 pasáží Bes. Spojitost VencNik a Anast se jasně projevuje i v případě MATTR, kdy oba texty vykazují jednoznačně nejvyšší hodnotu slovního bohatství ($MATTR_{Anast}$ = 0,86; $MATTR_{VencNik}$ = 0,848).

Domníváme se, že by tato zjištěná blízkost VencNik a Anast mohla podnítit úvahy o vzájemném vztahu obou csl. textů. Připomeňme, že původ překladu VencNik se dle některých názorů klade do období kulturního působení Sázavského kláštera (viz blíže Spurná 2018, 356) a Anast by podle filologických analýz F. Čajky mohla taktéž být spojena s literární činností sázavských mnichů; Čajka dokonce předkládá argumenty v návaznosti na dřívější práci E. Bláhové (1988, 62), že překlad lat. Legendy o sv. Anastázii byl pořízen během vyhnanství slovanských mnichů v uherském Visegrádu mezi léty 1055–1062 (srov. Čajka 2011, 192–193).

Hagiografické texty se jako celek při analýze MFW vydělují do dvou základních skupin: Vit, Meth, Georg a Ben na jedné straně a VencNik, Venc, Anast, Steph a Apol na straně druhé. Tyto dvě skupiny vyplývající ze shlukové analýzy přitom zůstávají beze změny při uplatnění nastavení F50 i F200 a také při zařazení nesegmentovaného i segmentovaného textu Bes. K druhé skupině textů se přiřazuje též Nicod, což je žánrově sice text apokryfního evangelia, ale stylisticky má blízko k legendám, a to mimo jiné i proto, že se jedná především o narativní text.

K dalšímu prověření se nabízí zjištěná spojitost Ben a Georg, které se sdružují do těsného shluku u obou nastavení počtu MFW. V rámci textů zahrnutých do analýz se totiž jedná o jediné dvě literární památky rukopisně doložené v srbské redakci csl. Domníváme se však, že hláskoslovná, pravopisná a snad i morfologická specifika, jež mohou záviset na konkrétním dochování textu, nijak zásadněji neproblematizují výsledky frekvenční analýzy, neboť při bližším prozkoumání dvou set nejfrekventovanějších tvarů nacházíme pouze jednotlivé případy takových forem: např. se jedná o tvar zvratného zájmena v podobě ⷭⷥ × ⷭⷭ (v Ben 14. nejfrekventovanější slovo s frekvencí f = 33, v Georg 3. nejfrekventovanější

7) Kendallův korelační koeficient $\tau = 0{,}57$, p-hodnota = 0,003, jedná se tedy o statisticky významnou (na zvolené hladině významnosti $\alpha = 0{,}05$) středně silnou korelaci.

8) U tohoto vzorku je hodnota Kendallova korelačního koeficientu, $\tau = 0{,}29$, p-hodnota = 0,001. Jedná se sice o slabou korelaci, stále ale statisticky významnou (na zvolené hladině významnosti $\alpha = 0{,}05$).

slovo s frekvencí f = 67), nom. sg. substantiva богъ × богъ (v Ben 97. nejfrekventovanější slovo s frekvencí f = 6; v Georg 18. nejfrekventovanější slovo s frekvencí f = 19) či slovesný tvar (nom. sg. m. ptc. préz. aktiva) глаголѥ × глаголл (v Ben 42. nejfrekventovanější slovo s frekvencí f = 13; v Georg 37. nejfrekventovanější slovo s frekvencí f = 9).

Jak jsme již konstatovali výše, text Bes byl paralelně zařazen do analýz jednak v kompletním znění, jednak rozdělený na dílčí textové úseky, které víceméně kopírují jednotlivé kapitoly / homilie. Primárním účelem sice bylo prověřit, zda výsledky analýz nemohou být ovlivněny faktem, že celý text Bes bezprecedentně přesahuje rozsahem všechny ostatní českocsl. texty (počet tokenů je 93 358, druhý v pořadí text Nicod má 8 651 tokenů), avšak na druhou stranu může rozdělení textu inspirovat k dalšímu zkoumání. Je totiž otázkou, zda je text Bes stylisticky homogenní, nebo mohou eventuální odlišnosti podněcovat např. k úvahám o jediném či naopak různých autorech překladu rozsáhlého díla, přičemž tato otázka, pokud je nám známo, v odborném paleoslovenistickém diskurzu teprve čeká na své podrobnější zhodnocení.

Analýzy, do nichž byl zahrnut segmentovaný text Bes, prokazují přinejmenším plausibilnost zvolených stylometrických metod. Tak např. poslední segment z Bes, který tvoří přídavné texty dvou modliteb – Modlitby za nemocné a část Modlitby za čistotu (blíže viz Čajka 2020), se ve všech sledovaných parametrech (tj. MFW, ATL i MATTR) přibližuje zbývajícím paraliturgickým textům v rámci analyzovaného materiálu, čímž se potvrzuje úspěšnost metod při žánrovém členění textů. Segment č. 10 odlišnou žánrovou charakteristiku od většinového textu Bes nevykazuje, přesná interpretace, proč se řadí k odlišné skupině textů, je tak bez dalšího výzkumu problematická. Snad by zde mohla hrát roli skutečnost, že při poměrně krátkém rozsahu se mohou citlivěji projevit relativně malé změny frekvencí některých slov – např. v Bes10 se vyskytuje několik koordinačních spojení se spojkou *i* (např. се агньць в‹о›жии възємлаи гр҃ѣхъı всего мира иже оубо и смѣрениꙗ своѥго и в‹о›жьствиꙗ х‹рист›ва), což by mohlo konvenovat s některými pasážemi modliteb. Je však nutno poznamenat, že několik dalších segmentů Bes kratšího rozsahu (kolem 500 tokenů) taková specifika neukazuje.

Jak jsme uvedli výše (viz pozn. č. 4), text Bes je ve zdrojovém materiálu pro naše výzkumy částečně složen podle edice z pasáží doplněných z variantního rukopisu (segmenty č. 4, 15, 23, 26 a 45). Zde je nutno konstatovat, že dotyčné segmenty nevykazují žádné bližší vztahy ani v jednom ze tří sledovaných parametrů, což interpretujeme jako další důkaz toho, že výsledky analýz nejsou nijak zřetelně ovlivněny rukopisným dochováním textů.

Za další zajímavé potvrzení smysluplného využití zvolených metod považujeme skutečnost, že analýza MFW přisoudila značnou blízkost (a to při uplatnění F50 i F200) dvěma segmentům Bes, které shodně obsahují kapitolu (homilii) č. 9. Část této homilie se totiž v textu Bes vyskytuje dvakrát, ve druhém případě byla vložena do textu 12. kapitoly a jedná se buď o odlišný překlad téhož latinského originálu, či o značně revidovanou verzi s úpravami tvaroslovnými, lexikálními i syntaktickými.

Nápadné shody ve sledovaných parametrech vykazují také homilie č. 4 a č. 40. Za zmínku zde stojí skutečnost, že obě jsou v edici Bes prezentovány ze dvou rukopisů – základního Pogodinského, kde jsou obě homilie nekompletní, a variantního Uvarovského. Podobně je možno sledovat blízkost dle zvolených stylometrických parametrů u kapitol (homilií) 6. a 18 či 13 a 16. Jsme přesvědčeni, že budoucí analýzy tímto směrem mohou přispět k bližšímu poznání textových specifik Bes.

Dva texty patřící svým původem mimo českou redakci csl. (pomineme-li některé sporné památky jako je např. Georg), tedy Supr a Meth, se vcelku organicky vřazují mezi

ostatní analyzované literární památky a také se spojují do víceméně očekávaných shluků v rámci zkoumání MFW. Za zmínku jistě stojí zjištěná příbuznost Meth a Vit, neboť obě legendy tvoří dvojici při uplatnění F50 i F200, avšak na druhou stranu při analýzách ATL a MATTR nacházíme spíše odlišnosti. Je otázkou, zda výsledky analýzy MFW nemůže ovlivnit skutečnost, že rukopisně obě legendy pocházejí z téhož kodexu (ruskocsl. Uspenského sborníku), avšak výše jsme již několikrát doložili, že rukopisné dochování naše stylometrické analýzy zásadněji nepoznamenává.[9] Připomeňme, že dosavadní filologické výzkumy kladou vznik archetypu Vit spíše do počáteční fáze českocsl. písemnictví (srov. např. Vašica 1949, 162), s čímž by zjištěná blízkost Meth a Vit mohla dobře konvenovat.

V úvodu jsme též nastínili otázku, do jaké míry se do stylometrických analýz může promítat skutečnost, zda se jedná o texty originální či překladové. Z našich výsledků vyplývá, že toto hledisko má spíše nižší vliv, jako podstatnější se jeví charakteristika žánrová. Můžeme to spatřovat při analýze MFW na výsledné blízkosti Bes a Supr (první je překladem z latiny, druhý z řečtiny), dále též u Meth a Vit (první text je originální, druhý překlad z latiny). Ani analýzy ATL a MATTR neukazují na nápadné souvislosti, které by odpovídaly rozdělení zdrojových textů podle jejich eventuálních cizojazyčných předloh. Co se týče Supr, který je obecně považován za stsl. text ovlivněný řečtinou, což se projevuje mimo jiné relativně vysokou frekvencí kompozit (namnoze kalků podle řeckých modelů – viz např. Večerka 2006, 246), poněkud překvapivě nepatří v našem materiálu k textům s nejvyšší hodnotou délky ATL. Kompozita zřejmě nepředstavují natolik frekvenčně dominantní skupinu, která by zásadněji ovlivnila tento parametr, avšak je třeba dodat, že pro bližší zhodnocení by bylo zapotřebí dalšího výzkumu založeného na početně vyšší materiálové bázi.

5. Závěry

Představili jsme pilotní studii, v níž jsme se pokusili prověřit aplikovatelnost zvolených stylometrických metod na specifický materiál csl. textů. V prvé řadě se ukazuje, že při určování příbuznosti textů hraje nejdůležitější roli hledisko žánrové, např. se zřetelně vymezují texty paraliturgické (modlitební), ale v zásadě i další žánry – homilie a legendy. Zároveň se ale domníváme, že zjištěná blízkost textů nevyplývá pouze z obecných charakteristik jednotlivých žánrů, ale může odrážet chronologickou posloupnost vzniku jednotlivých literárních památek. Plausibilitu zvolených stylometrických metod může potvrzovat i silná tendence k vykazování shodných vlastností jednotlivých částí textu Bes, které jsme v části analýz segmentovali na dílčí pasáže.

V úvodu jsme si též položili otázku, do jaké míry může zjišťování příbuznosti textů na základě zvolených stylometrických metod ovlivňovat jejich cizojazyčná předloha. Všechny tři metody poukazují na to, že tato charakteristika není hlavním kritériem pro klasifikaci příbuznosti textů, např. v analýze MFW se shlukují Meth (originální text) a Vit (překlad z latiny) či Bes (překlad z latiny) a Supr (překlad z řečtiny). Analýzy taktéž ukazují, že je možno podpořit některé teze o časové blízkosti vzniku textů (Meth a Vit), případně

9) V této souvislosti jsme provedli důkladné šetření sta nejfrekventovanějších slovních tvarů z obou památek. Můžeme konstatovat, že se mezi nimi nenachází žádná forma, která by byla jakýmkoli způsobem výrazněji ovlivněna potenciálními specifiky konkrétního rukopisného dochování. Nečetné prvky, jež je možno tradičně přičíst charakteristickým rysům východoslovanské redakce csl., jsou přítomny i v řadě dalších památek sledovaného souboru, např. substituce nosovky *ǫ* > *u*/*ʻu* (např. глаголающе), naopak grafické zachování nosového ę (např. сѧ, глаголѧ), z morfologických jevů nestažené tvary adjektiv (např. блаженъіи, свѧтаѥго) apod.

navrhnout nové hypotézy, které by mohly vést ke spojení původu památek do stejného autorského / překladatelského okruhu, jak se ukazuje např. v případě VencNik a Anast, jež vykazují shody ve všech třech sledovaných parametrech.

Výsledky naší pilotní studie jistě podněcují k dalším výzkumům. Zde se nabízí zejména rozšíření materiálové báze o další stsl. a csl. texty z různých redakcí a období nejstaršího slovanského spisovného jazyka. Pro potvrzení zjištěné skutečnosti, že výsledky uvedených stylometrických analýz nejsou zřejmě zásadněji zkresleny rukopisným dochováním textů, se nabízí možnost zapojit do výzkumu znění z variantních rukopisů. V neposlední řadě mohou stylometrické metody přispět k bližšímu výzkumu textu Bes, který představuje zdaleka nejrozsáhlejší literární památku českého původu, a to zaměřit analýzy na jednotlivé kapitoly s cílem vysledovat autorská specifika a kompozici textu. Z hlediska metodologického se jeví vhodná aplikace dalších metod (např. shluková analýza založená na použití dalších vlastností textů, jako jsou délky jednotek, frekvence slovních druhů, gramatických kategorií apod.).

Zkratky citovaných památek:

Anast – Legenda o sv. Anastázii
Apol – Umučení sv. Apolináře
Ben – Život sv. Benedikta
Bes – Besědy sv. Řehoře Velikého na evangelium
Conf – Modlitba vyznání hříchů
Georg – Život sv. Jiří
Greg – Modlitba sv. Řehoře
Meth – Život Metodějův
Nicod – Nikodémovo evangelium
Steph – Život sv. Štěpána, papeže
Supr – Supraslský kodex
Trin – Modlitba ke sv. Trojici
Venc – První staroslověnská legenda o sv. Václavu
VencNik – Druhá staroslověnská legenda o sv. Václavu
Vit – Život sv. Víta

Literatura:

Bláhová 1988: Bláhová, E.: Staroslověnské písemnictví v Čechách v 10. století. In: Bláhová, E. et al. [edd.]: *Sázava. Památník staroslověnské kultury v Čechách*. Praha 1988, 55–69.

Covington – McFall 2010: Covington, M. A. – McFall, J. D.: Cutting the Gordian Knot: The Moving--Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics* 17, 2010, 94–100.

Čajka 2011: Čajka, F.: *Církevněslovanská legenda o svaté Anastázii*. Praha 2011.

Čajka 2020: Čajka, F.: Modlitba za čistotu v rukopisech Josifo-Volokolamského kláštera (GIM, Jeparch 160, 164 a 165). *Bohemica Olomucensia* 12, 2020, 2, 10–32.

Eder et al. 2016: Eder, M. – Rybicky, J. – Kestemont, M.: Stylometry with R: a package for computational text analysis. *The R Journal* 8, 2016, 1, 107–121.

Evert et al. 2017: Evert, S. et al.: Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities* 32 (suppl_2), 2017, ii4-ii16.

Konzal 1988: Konzal, V.: Nejstarší slovanská legenda václavská a její „Sitz im Leben“. *Studia mediaevalia Pragensia* 1, 1988, 113–127.

Konzal 1991: Konzal, V.: Otazníky kolem církevněslovanské Modlitby k sv. Trojici a českých vlivů na literaturu Kyjevské Rusi. *Slavia* 60, 1991, 3, 232–247.

Konzal 2005: Konzal, V.: *Čtyřicet homilií Řehoře Velikého na evangelia v českocírkevněslovanském překladu. Díl první. Homilie I–XXIV*. Praha 2005.

Konzal – Čajka 2006: Konzal, V. – Čajka, F.: *Čtyřicet homilií Řehoře Velikého na evangelia v českocírkevněslovanském překladu. Díl druhý. Homilie XXV–XL*. Praha 2006.

Koščová et al. 2017: Koščová, M. – Čech, R. – Mačutek, J.: Shluková analýza. In: Karlík, P. et al. [edd.]: *Nový encyklopedický slovník češtiny*. Praha 2017 [on-line]. <https://www.czechency.org/> [cit. 15-01-2023].

Kotkov et al. 1971: Котков, С. И. et al. [edd.]: *Успенский сборник XII–XIII вв*. Москва 1971.

Lindstedt 2011: Lindstedt, J.: *Corpus Cyrillo-Methodianum Helsingiense: Corpus of Old Church Slavonic Texts, source [text corpus]*. Kielipankki [on-line] <http://urn.fi/urn:nbn:fi:lb-20140730106> [cit. 28-11-2022].

Mareš 1979: Mareš, F. V.: *An Anthology of Church Slavonic Texts of Western (Czech) Origin.* München 1979.

Mareš 2000: Mareš, F. V.: Církevněslovanské písemnictví v Čechách. In: *Cyrilometodějská tradice a slavistika.* Praha 2000, 256–327.

Mikulka 2015: Mikulka, T.: Ke genezi církevněslovanské Modlitby ke svaté Trojici. *Slavia* 84, 2015, 4, 372–396.

SJS: *Slovník jazyka staroslověnského = Lexicon linguae palaeoslovenicae, I–V*. Kurz, J. – Hauptová, Z. [red.] et al. Praha 1966–2016.

Smith – Aldridge 2011: Smith, P. W. – Aldridge, W.: Improving Authorship Attribution: Optimizing Burrows' Delta Method. *Journal of Quantitative Linguistics* 18, 2011, 63–88.

Spurná 2018: Spurná, K.: Druhá staroslověnská legenda o sv. Václavu ve vztahu k latinské předloze. *Listy filologické* 141, 2018, 355–392.

Vaillant 1968: Vaillant, A.: *L´évangile de Nicodème. Texte slave et latin*. Genève – Paris 1968.

Vašica 1929: Vašica, J.: Druhá staroslověnská legenda o sv. Václavu. In: Vajs, J. [ed.]: *Sborník staroslovanských literárních památek o sv. Václavu a sv. Lidmile*. Praha 1929, 71–135.

Vašica 1949: Vašica, J.: Staroslovanská legenda o sv. Vítu. In: Kurz, J. – Murko, M. – Vašica, J. [edd.]: *Slovanské studie. Sbírka statí, věnovaných prelátu univ. prof. dr. Josefu Vajsovi k uctění jeho životního díla.* Praha 1948, 159–163.

Vašica 1996: Vašica, J.: *Literární památky epochy velkomoravské*. Praha 1996.

Večerka 2006: Večerka, R.: *Staroslověnština v kontextu slovanských jazyků*. Praha – Olomouc 2006.

Vepřek 2013: Vepřek, M.: *Modlitba sv. Řehoře a Modlitba vyznání hříchů v církevněslovanské a latinské tradici*. Olomouc 2013.

Vepřek 2022: Vepřek, M.: *Czech Church Slavonic in the Tenth and Eleventh Centuries.* München 2022.

Masarykova univerzita
Brno

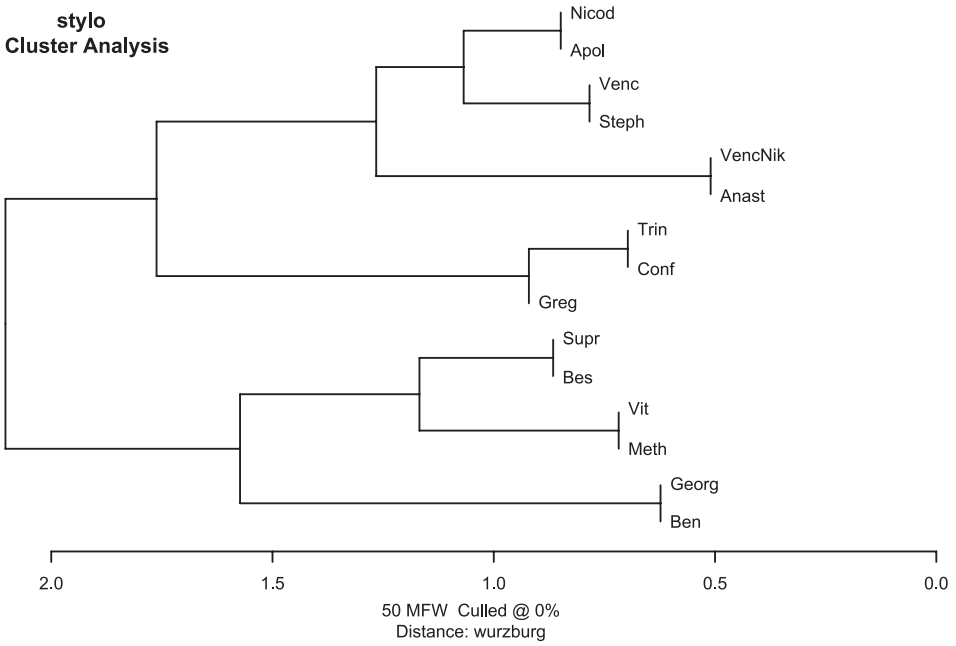
Radek Čech
cechradek@phil.muni.cz
ORCID: 0000-0002-4412-4588

Univerzita Palackého v Olomouci
Olomouc

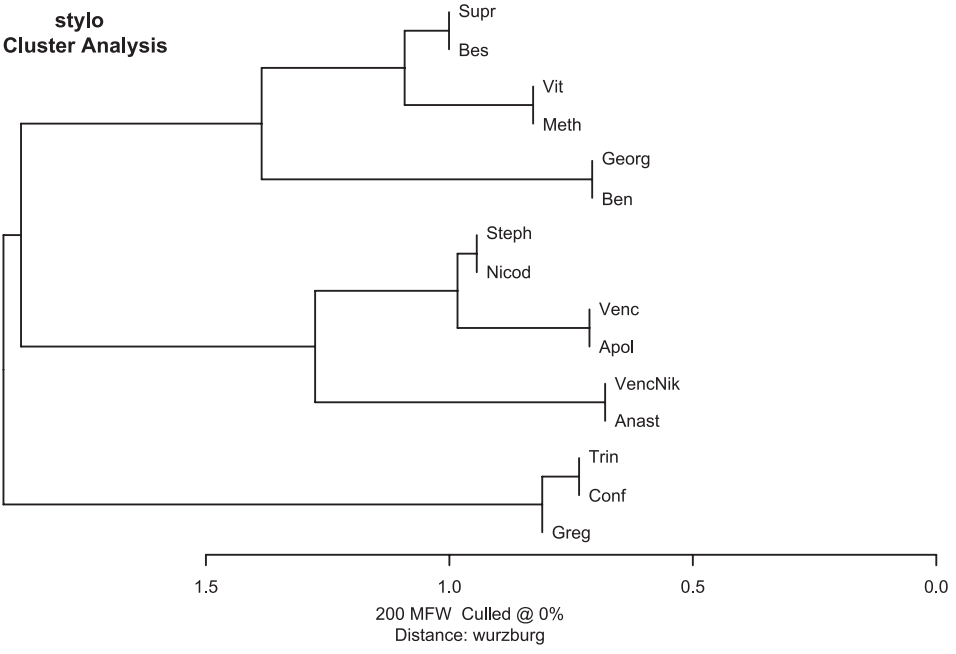
Miroslav Vepřek
miroslav.veprek@upol.cz
ORCID: 0000-0002-2898-5353

text	N	ATL	MATTR	text	N	ATL	MATTR	text	N	ATL	MATTR
Anast	803	5,03	0,848	**Bes_17**	532	5,15	0,776	**Bes_37**	2600	4,97	0,784
Apol	3747	4,77	0,806	**Bes_18**	1369	5,02	0,751	**Bes_38**	2676	5,06	0,794
Ben	3972	4,69	0,793	**Bes_19**	1470	5	0,774	**Bes_39**	5080	5,03	0,769
Bes	93358	4,95	0,778	**Bes_20**	1417	5	0,802	**Bes_40**	2554	5,04	0,79
Bes_01	1554	4,89	0,78	**Bes_21**	1475	5,01	0,789	**Bes_41**	3876	4,89	0,756
Bes_02	1580	4,99	0,778	**Bes_22**	1512	4,89	0,772	**Bes_42**	3018	4,9	0,788
Bes_03	1104	4,83	0,798	**Bes_23**	2338	4,93	0,777	**Bes_43**	4855	4,85	0,77
Bes_04	1524	5,07	0,807	**Bes_24**	3872	5,12	0,777	**Bes_44**	3289	4,88	0,773
Bes_05	799	5,02	0,808	**Bes_25**	4377	5,01	0,792	**Bes_45**	4578	4,91	0,787
Bes_06	1357	5,07	0,789	**Bes_26**	1639	4,81	0,778	**Bes_46**	749	4,7	0,735
Bes_07	479	4,77	0,77	**Bes_27**	2659	4,9	0,777	**Conf**	533	4,69	0,666
Bes_08	189	5,04	0,784	**Bes_28**	609	4,83	0,783	**Georg**	3066	4,67	0,768
Bes_09	1675	5	0,768	**Bes_29**	1866	4,84	0,773	**Greg**	674	4,62	0,75
Bes_10	266	4,7	0,756	**Bes_30**	3154	4,81	0,764	**Meth**	3375	4,98	0,807
Bes_11	250	4,89	0,753	**Bes_31**	2979	4,93	0,763	**Nicod**	8651	4,64	0,776
Bes_12	1684	4,9	0,752	**Bes_32**	2329	4,91	0,767	**Steph**	3852	4,97	0,816
Bes_13	1737	5,16	0,762	**Bes_33**	1022	4,88	0,75	**Supr**	99078	4,71	0,783
Bes_14	1380	5,08	0,802	**Bes_34**	2489	4,94	0,763	**Trin**	3059	4,84	0,708
Bes_15	1700	5,04	0,802	**Bes_35**	3151	4,95	0,756	**Venc**	1380	4,65	0,775
Bes_16	684	4,85	0,728	**Bes_36**	1862	5,07	0,773	**Vencnik**	5973	5,27	0,86
								Vit	2993	4,68	0,775

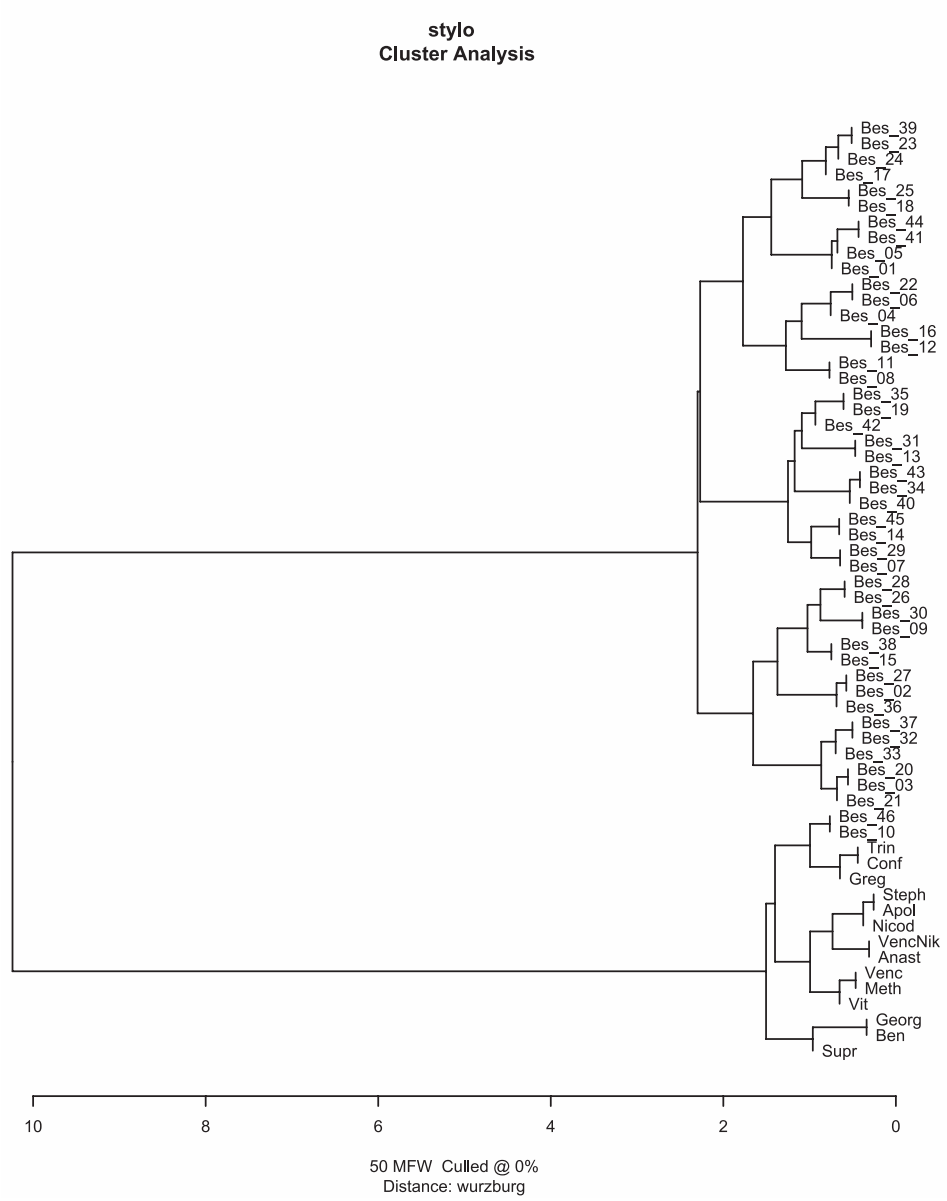
Tab. 1. Délka textu (N), průměrná délka tokenu (ATL) a klouzavý průměr slovního bohatství TTR u zkoumaného vzorku textů.



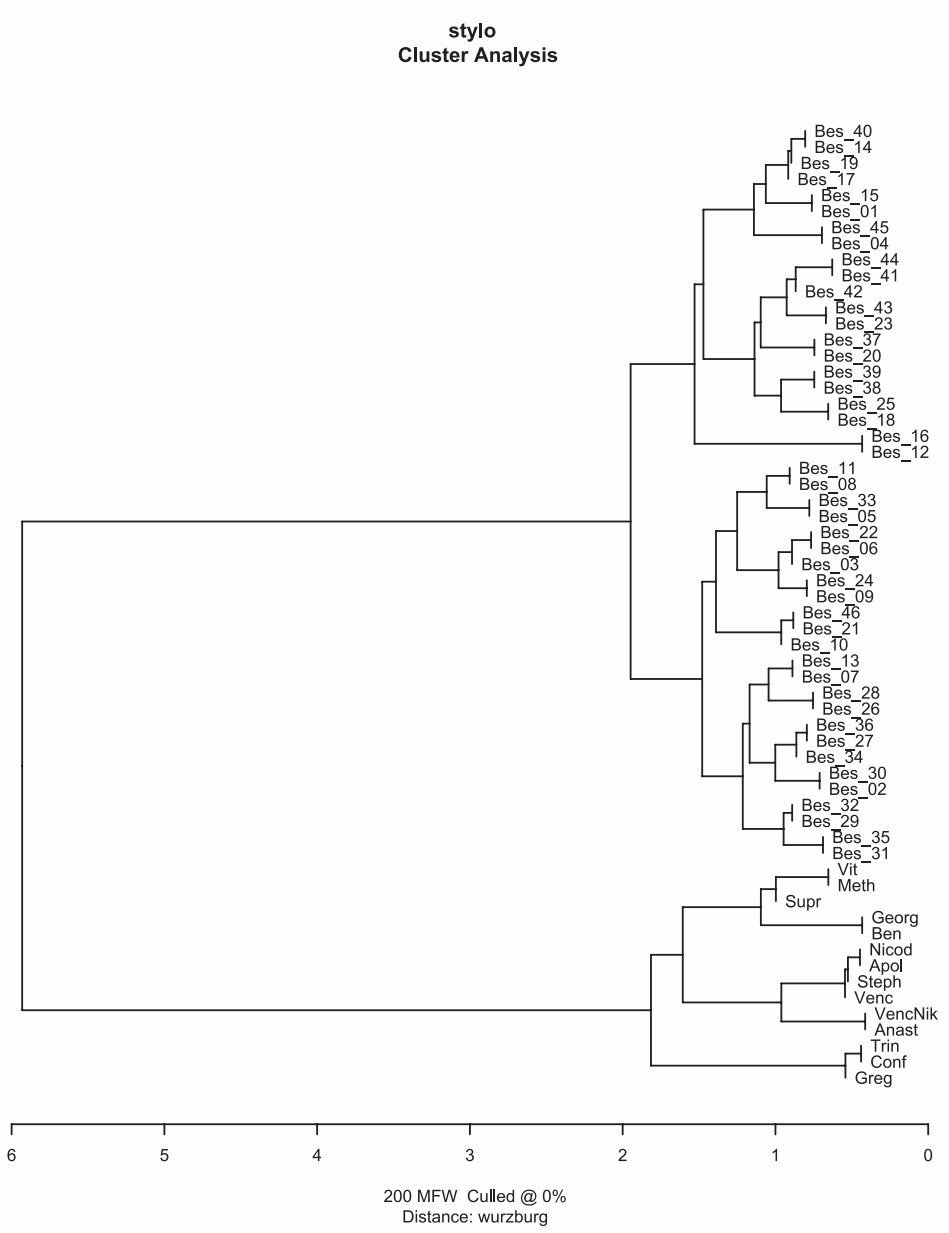
Obr. 1. Shluková analýza založená na srovnání normalizovaných frekvencí 50 nejfrekventovanějších slov.



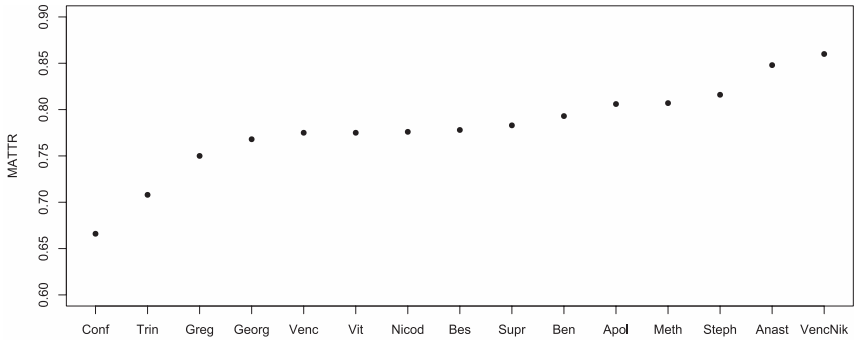
Obr. 2. Shluková analýza založená na srovnání normalizovaných frekvencí 200 nejfrekventovanějších slov.



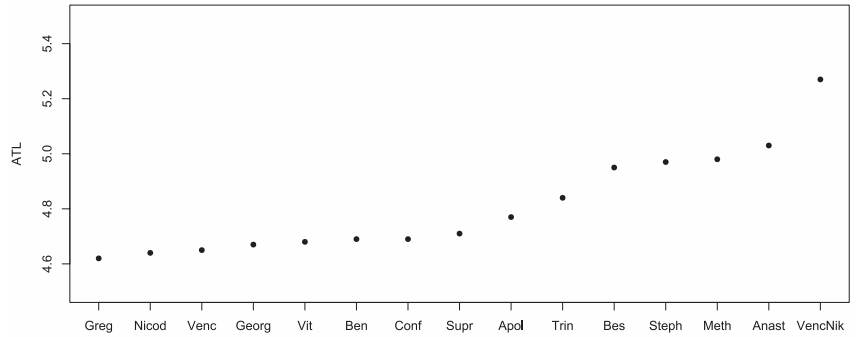
Obr. 3. Shluková analýza založená na srovnání normalizovaných frekvencí 50 nejfrekventovanějších slov.



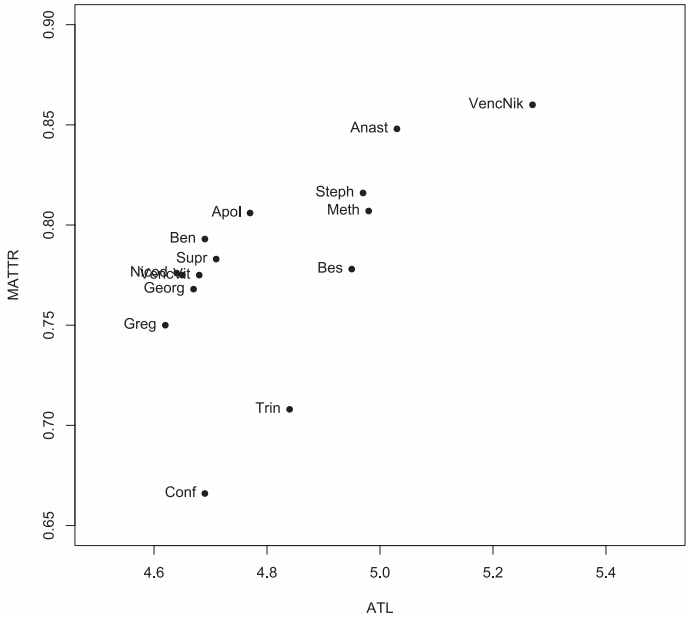
Obr. 4. Shluková analýza založená na srovnání normalizovaných frekvencí 200 nejfrekventovanějších slov.



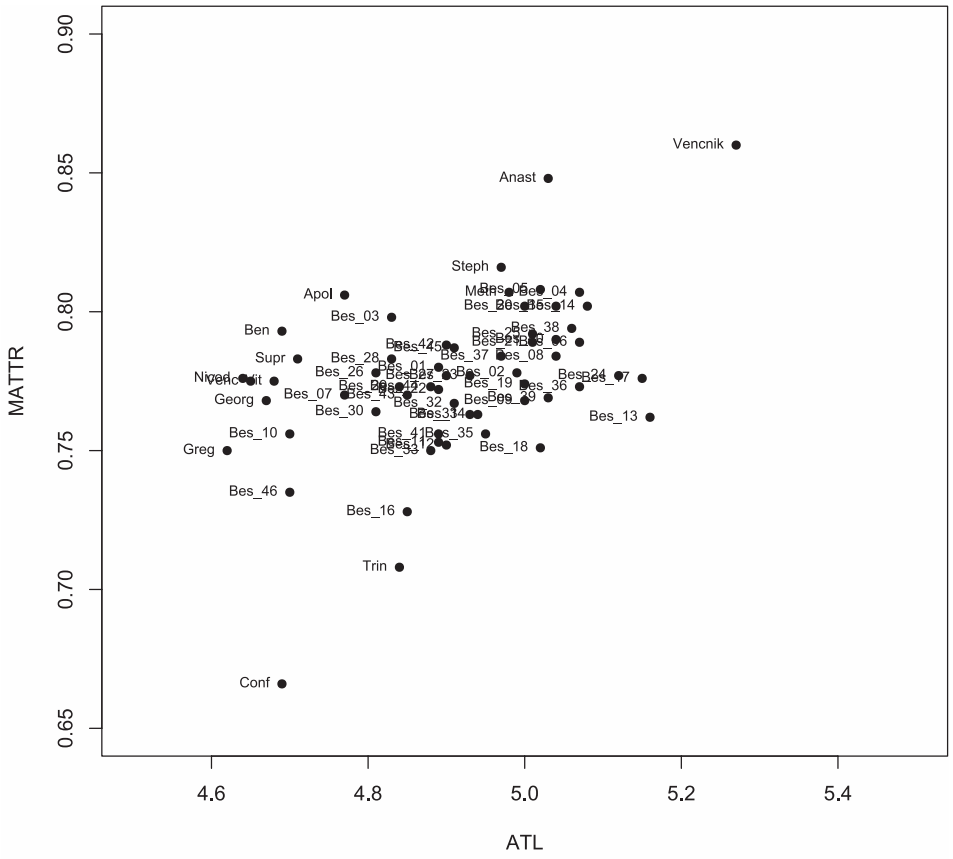
Obr. 5. Hodnoty klouzavého průměru type-token poměru (MATTR).



Obr. 6. Hodnoty průměrné délky tokenu (ATL).



Obr. 7. Vzájemný vztah průměrné délky tokenu (ATL) a klouzavého průměru type-token poměru (MATTR).



Obr. 8. Vzájemný vztah průměrné délky tokenu (ATL) a klouzavého průměru type-token poměru (MATTR).