sciendo
DOI 10.2478/jazcas-2024-0005

OPAKUJI, TEDY JSEM: PARASYNTAKTICKÁ PERSPEKTIVA[1]

MARTINA BENEŠOVÁ – DAN FALTÝNEK – LIBUŠE KORMANÍKOVÁ – ONDŘEJ KUČERA
Filozofická fakulta Univerzity Palackého, Olomouc, Česká republika

BENEŠOVÁ, Martina – FALTÝNEK, Dan – KORMANÍKOVÁ, Libuše – KUČERA, Ondřej: I repeat therefore I am: The parasyntactic perspective. Jazykovedný časopis (Journal of Linguistics), 2023, Vol. 74, No. 2, pp. 477–494.

Abstract: The text presents a theoretical platform and a case study of a new method for authorship attribution based on an author's specific low-frequency lexicon. It will be shown that an author's text is largely context-independent and is constructed by the author's habit based on the regular repetition of certain topics or modes of expression. The author's idiosyncratic way of choosing between synonymous linguistic devices in the text happens at a distance of several word forms or sentence units. This means that texts themselves are constructed using a much wider range of repetitions than expected and that the structure of the text above the level of intersentential linking is determined by a specific group of words (functional but above all content words) obligatorily used by the author in the formulation of the text. The newly introduced method can be used to attribute authorship by relying on the specific linguistic imprint of the author in the text (in this context, we talk about parasyntactic linguistic level). The method is compared with a function-word-based method.

Keywords: low-frequency lexicon, hapax legomenon, text structure, text cohesion, personal content profile, authorship attribution, forensic standard, parasyntax

1. ÚVOD

V tomto článku se zaměřujeme na jazykové povědomí určitého uživatele jazyka. Na způsoby identifikace jednotlivce na základě jeho specifických jazykových návyků a také na povahu strukturace textu, který specifika zacházení jedince s jazykem odráží. Pro tento aspekt vztahu jazyka a jednotlivce formuloval Juraj Dolník koncept tzv. bezznalostní jazykové dispozice. „Racionalita jazyka vo vzťahu ku komunikácii (komunikačná racionalita v tomto kontexte) pôsobí na nositeľa jazyka tak, že ten je účelne vybavený bezznalostnou jazykovou dispozíciou aj jazykovými znalosťami, aby mohol efektívne realizovať svoju komunikačnú intenciu. Efektívnosť tu znamená, že používateľ vynakladá len toľko psychickej energie, koľko je na realizáciu intencie nevyhnutné, pričom energia sa sústreďuje práve na intenčný akt. Zameranosti na intenčné akty je podriadená jednak súčinnosť bezznalostnej jazykovej dispozície s jazykovými znalosťami a jednak súhra neuvedomovaných a viac alebo

[1] Článek vznikl za použití technologie firmy Deepeffects.ai, s.r.o. (https://deepeffects.ai).

menej uvedomovaných jazykových znalostí. Na tomto mieste je primerané v rámci neuvedomovaných jazykových znalostí hovoriť o automatizovaných znalostiach.“ (Dolník 2018, s. 85).

Juraj Dolník k neuvědomovanému použití jazykových prostředků jedincem ještě dodává, že jejich „(...) výrazným príkladom sú automatizované pravopisné znalosti, ale aj štylizačné znalosti, ktoré sa týkajú ustálených formulácií, alebo rutinné jazykové reakcie v schematizovaných komunikáciách.“ (Dolník 2018, s. 85).

Na následujících stranách chceme prozkoumat bezznalostní jazykovou dispozici určitého jedince (k tomu viz dále články Dolník 2019 a Dolník 2021) ve vztahu ke struktuře jím produkovaného textu.

Identifikace mluvčího na základě jazykového ztvárnění textu (oproti písmu, hlasu atd.) je od původu spojována s opakováním určitých prvků. Běžně se setkáváme s formulacemi jako *Karel by řekl...*, *Karel by teď asi poznamenal…* apod. Určité způsoby vyjádření nebo určitá témata si v jazykové intuici evidentně spojujeme s konkrétní osobou – a to na základě opakování. Uvedeme to na několika příkladech:

Románová postava Iana Fleminga James Bond často říká: *Protřepat, nemíchat.* (Fleming 1956, kapitola 9; překlad: *protřepané a nemíchané*, Fleming 2008); své jazykové chování postava navíc reflektuje v odpovědi na stejně znějící otázku (*Shaken or stirred?*): *Vypadám snad, že mi na tom záleží? (Do I look I give a damn?*) (Campbell 2006). Méně známou postavou, která se vyznačuje svým typickým opakováním odpovědi *Ha* na jakýkoliv typ repliky, je Britt-Marie Fredricka Backmana (2016, viz např. s. 61). Když opustíme prostředí světového literárního pole a přesuneme se k domácí produkci, můžeme stejný fenomén pozorovat například v proslulých dramatech Járy Cimrmana. Ve hře *Záskok* se jedná například o slovo *Dals*, které v průběhu děje opakuje postava Vlasta (Smoljak – Svěrák 1994, viz část *Vlasta: vesnický obrázek o jednom dějství*), v případě *Českého nebe* se postava Komenského vyznačuje oslovením *Praotce!* nebo postava Husa slovem *Chlapík!* (Smoljak – Svěrák 2008). Kromě fikčních hrdinů z literárních děl se opakování nevyhýbá ani „skutečným“ lidem. Někteří z nich, kteří pronikli do mediální sféry, nám dali nahlédnout do jejich vlastního užívání jazyka. Vzpomeňme třeba na teleshopping a ikonickou postavu Horsta Fuchse vyznačujícího se větou *A to ještě není vše, přátelé!* (viz reflexe tohoto autorského idiolektismu např. v Janščová 2015).

Výše zmíněné příklady jsme zvyklí spojovat si s konkrétním autorem, tohoto autora určitým způsobem profilují. Slovní tvary v nich obsažené mají zároveň u autora s největší pravděpodobností o něco vyšší frekvenci než v obecném úzu, více se opakují: autor je může pravidelně používat v určitém slovním nebo situačním kontextu; může se jednat o výroky formulující postoje autora; nebo se naopak jedná o textové berličky, jimiž autor vyplňuje text při perspektivizaci výpovědi; mohou to být jevy uvědomované nebo používané nezáměrně. V kontextu opakování určitých

částí textu, na které se jako na autorskou idiosynkrazii zaměřujeme, je podstatné, že se přes pravidelné opakování jedná o jevy řídké v porovnání s ostatním lexikem zastoupeným v textu. V následujícím grafu zobrazujeme opakovaný výrok hlavního hrdiny románu *Kdo chytá v žitě* J. D. Salingera. Výrok *Jestli chcete co vědět* se v románu o délce 68 tisíc slov (v překladu Luby a Rudolfa Pellarových, viz Salinger 2000) opakuje šestadvacetkrát; viz obr. 1.

Obrázek 1: Distribuce slovního spojení *Jestli chcete co vědět* (*If you want to know the truth*) v románu J. D. Salingera *Kdo chytá v žitě* (Salinger 2000; originál: Salinger 1991). V překladu románu o délce 68 110 slov se slovní spojení vyskytuje 26-krát, tzn. že se opakuje v průměru každou 1,5 normostranu (tzn. přibližně každých 250 slov). Graf byl pořízen v programu AntConc (https://www.laurenceanthony.net/software/antconc/). Svislé čáry vyznačují výskyty vybraného slovního spojení v textu, posloupnost slov v textu směřuje v grafu zleva doprava.

O výše vybraných autorských výrocích neuvažujeme jako o jazykových prostředcích, které strukturují text. Chápeme je spíše jako slova výplňková nebo jako autorské licence. Ke strukturaci textu slouží gramatická soustava jazyka, v určování autorství reprezentovaná tzv. funkčními slovy (Binongo 2003), a také slova obsahová, jimiž autor reaguje na aktuální kontext a jimiž vyjadřuje dle své vůle a komunikačního záměru určitý obsah. Výše frekventovaná obsahová slova zároveň vytvářejí izotopickou skladbu textu, tzn. strukturují text na základě obsahových souvislostí (Greimas 1987, s. 66 – 67). Z hlediska analýzy textu jsou ovšem obsahová slova chápána jako prostředek sdružující texty dle tematiky, nikoliv na základě autorských vlastností (Mikros – Argiri 2007). Jestliže proto tato obsahová slova z atribuční procedury vyloučíme a zároveň tak učiníme i s autorskými okazionalizmy, protože jsou sporadické a nedávají nám oporu v určení mnoha textů a jejich částí, musíme se spolehnout na širokou řadu jiných, nepřímých charakteristik autora (funkční slova, n-gramy, kompresní znaky; viz Juola 2006, s. 8; Zhao – Zobel 2005). U těch ale ztrácíme osobní povědomí o jejich přítomnosti v textu a s autorem si je nespojujeme.

2. FUNKČNÍ SLOVA A OBSAHOVÁ SLOVA

Celý autorský text s velkou mírou homogenity (viz v kontrastu k tomu de Roeck a kol. 2004; Sarkar a kol. 2005; Grzybek 2013) prostupuje lexikum s vysokou frekvencí. Jedná se převážně o jazykové prostředky bez věcného obsahu (synsémantika) užívané k formální výstavbě textu a gramatické stavbě věty. Téměř v každé větě některý z nich nacházíme. O těchto prostředcích disciplína určování autorství mluví jako o funkčních slovech (Burrows 2002; 2003; 2007; Binongo 2003). V seznamu českých funkčních slov jsou zastoupeny spojky (*a*, *nebo*, *i*), předložky (*na*, *pod*, *s*/*se*) nebo zájmena (*ona*/*on*/*ono*, *já*, *ty*). Ukazuje se, že určití autoři ve vý-

stavbě textu protežují např. konkrétní pomocná slovesa nebo předložková spojení (Mosteller – Wallace 1964; Baayen a kol. 2002), jiná naopak užívají s menší či minimální frekvencí. To může sloužit k vysoce věrohodné identifikaci autora (Evert a kol. 2017). Pokud pozorujeme takto zřetelný úkaz, že autor se návyky využívat určitá funkční slova s vyšší či nižší frekvencí odlišuje od jiných autorů, a zároveň připustíme, že využívání funkčních slov vědomě nekontroluje, jedná se o ideální stopu zanechanou v textu.

Autorské opakování funkčních slov v určité individuální proporcionalitě je hodnověrně doložený autorský rys textu (viz dále Juola – Baayen 2003). Je ale opakování funkčních slov tím, s čím si spojujeme identitu mluvčích? – pravděpodobně ne. Identitu mluvčího shledáváme spíše v jeho specifických výrazech – výše zmíněných okazionalizmech, které ale nemají povahu text strukturujících prostředků a slouží spíše jako určitý kolorit mluvčího než jako jeho jazykový profil, jak jsme poznamenali. Proč ale autorův profil nepostavit na slovech vyjadřujících určitý obsah? Pomineme-li argument, že tato slova prozrazují tematickou blízkost, nikoliv autorskou identitu (viz citát č. 1 níž), odpověď zní: protože tato složka textu je považována za náhodnou a vzhledem k tomu obsahový profil autora nemůže existovat (viz citát č. 2 níž). Dovolíme si ukázat, že jazykový profil autora existuje, prostupuje text s vysokou homogenitou a s vysokým zastoupením v textu. Co to znamená pro možnosti profilování autora si ale necháme pro pozdější publikace.

Citát č. 1: obsahová slova klastrují texty na základě obsahu, nikoliv autorství, jak píše Patrick Juola (2006, s. 241):
A more serious problem with the (...) approach in particular is that it may permit topic to dominate over authorship. Any two retellings of Little Red Riding Hood will probably incorporate the words "basket" and "wolf"; any two newspaper articles describing the same football game will mention the same teams and star players; indeed, any two reports of football games will probably mention "score" and "winner," and not "basket." Therefore, any measure of vocabulary overlap will find a greater match between documents of similar topic.[2]

Citát č. 2: struktura textu (tzn. přítomnost obsahových slov v textu) je dána kontextem a je ze své podstaty náhodná, pro příklad volíme ukázky z de Beaugrandovy a Dresslerovy publikace (1981):
Linguists of all persuasions seem to agree that a language should be viewed as a system: a set of elements each of which has a function of contributing to the workings of

[2] Závažnějším problémem tohoto přístupu (...) je zejména to, že umožňuje, aby téma převážilo nad autorstvím. Jakékoli dva příběhy o Červené Karkulce budou pravděpodobně obsahovat slova „koš" a „vlk"; jakékoliv dva novinové články popisující stejný fotbalový zápas budou zmiňovat stejné týmy a hvězdné hráče; ve skutečnosti jakékoliv dvě zprávy o fotbalových zápasech budou pravděpodobně zmiňovat „skóre" a „vítěze", a nikoliv „koš". Jakékoli měřítko překrývání slovní zásoby proto najde větší shodu mezi dokumenty s podobným tématem.

the whole. (...) no one would equate this kind of system with the operations of communicating: people are not combining distinctive units in any direct or obvious way. Indeed, empirical tests show that many abstract distinctions are not perceptibly maintained in real speaking and are recoverable only from context (viz kapitola 3).[3]

A young science like linguistics would understandably seek to align itself with older sciences like physics, mathematics, and formal logic. But communication, like any human activity, has its own special physical, mathematical, and logical properties that must not be overlooked. An unduly rigid application of notions from the "exact" sciences could dehumanize the object to the point where the inquiry becomes irrelevant. A formalism is a representation, not an explanation, and a means, not an end. The analysis of formal structures might well fail to uncover the nature and function of an entity in its wider context (viz předmluva).[4]

V následující analýze ukážeme, že kontext je predominován autorským obsahovým profilem a představa o struktuře rozsáhlého textu, jak ji podává teorie textu / pragmatika, není adekvátní (zde zastoupeno citátem č. 2; viz k tomu dále kritéria textovosti – Hirschová 1989). Formální strukturu textu dlouhého rozsahu totiž oproti obecnému nebo zde de Beaugrandem a Dresslerem vyjádřenému názoru dokážeme exaktně vymezit a umíme následně zhodnotit vliv prostředků zapojených do této struktury na okolní text, obecně na text jako takový, včetně vlivu na aktuální kontext. Nedomníváme se, že tímto text jako projev činnosti člověka "dehumanizujeme", jak je uvedeno v citátu č. 2, pouze dokládáme zvykový charakter struktury autorova textu, což naprosto nemusí být překvapující, ačkoliv je to v rozporu s obecnými teoretickými předpoklady i praxí v analýze textu.

3. NÍZKO FREKVENTOVANÉ LEXIKUM

V následujících analýzách proti sobě postavíme přístup určování autorství založený na funkčních slovech a oproti tomu přístup využívající nízko frekventovaných, pravidelně se opakujících slov obsahových. Nebudeme se ale v tomto ohledu zaměřovat na okazionální výrazy typu *protřepat, nemíchat* uvedené v úvodu, pro ana-

[3] Lingvisté všech směrů se zřejmě shodnou na tom, že na jazyk je třeba pohlížet jako na systém: soubor prvků, z nichž každý má funkci přispívat k fungování celku. (...) Nikdo by tento druh systému neztotožňoval s operacemi v komunikaci: lidé nekombinují distinktivní jednotky žádným přímým nebo zřejmým způsobem. Empirické testy skutečně ukazují, že mnohé abstraktní distinkce nejsou v reálném mluvení vnímatelně zachovány a jsou obnovitelné pouze z kontextu.

[4] Mladá věda jako lingvistika se pochopitelně snaží přizpůsobit starším vědám, jako je fyzika, matematika a formální logika. Ale komunikace, stejně jako jakákoli jiná lidská činnost, má své zvláštní fyzikální, matematické a logické vlastnosti, které nelze přehlížet. Nepřiměřeně rigidní aplikace pojmů z „exaktních" věd by mohla odlidštit objekt do té míry, že by se zkoumání stalo irelevantním. Formalismus je reprezentace, nikoliv vysvětlení, a prostředek, nikoliv cíl. Analýza formálních struktur by mohla selhat při odhalování povahy a funkce entity v jejím širším kontextu.

lýzu využijeme široký rozsah plnovýznamových slov tak, abychom prověřili, zda jsou ve svém plošném zastoupení ve slovní zásobě markerem obsahu (tzn. zda klastrují texty na základě v nich obsažených témat, nikoliv autorsky). Obsahová slova by měla být zároveň v textu přítomna vlivem náhodných okolností vyplývajících z kontextu, neměla by proto mít určující autorské vlastnosti, jak naznačují citace na konci předešlého oddílu. Klastrování autorů provedeme tak, že budeme analyzovat lexikum s nejnižším zastoupením v analyzovaných textech. To ovšem neznamená, že tento přístup zcela vyloučí z analýzy slova funkční. V námi analyzovaném vzorku se v případě obsahových slov jedná o rozsáhlou část za klastrování odpovědného lexika. Pro srovnání: v románu *Na cestě* Jacka Kerouaca (1957) mezi za autorské klastrování odpovědnými hapax legomeny funkční slova téměř nenacházíme – velikosti vzorků jsou analogické následující analýze, viz navazující publikace; ve vzorcích klastrujících *Autismus & Chardonnay* (viz oddíl *Analýza*; Selner 2017 a 2019) se mezi vybranými hapaxy naopak vyskytuje 39 % funkčních slov. Naše metoda tedy přítomnost funkčních slov v lexiku determinujícím autora nevylučuje. Takové lexikum je definováno pravidelností svého výskytu a jeho nízkou frekvencí.

Když mluvíme o lexiku s nejnižším zastoupením v textu, myslíme tím tzv. hapax legomena, výrazy vyskytující se v textu pouze jednou. V následující analýze pracujeme s hapax legomeny následujícím způsobem: a) jedná se o slovní formu zastoupenou pouze jednou v určitém autorském textu; b) jedná se o slovní formu zastoupenou pouze jednou ve vzorku autorského textu, když tento text dělíme na dva a více vzorků. K dělení textů na vzorky přistupujeme z následujících důvodů: i) identifikace opakujícího se nízko frekventovaného lexika v autorském textu vyžaduje srovnání nejméně dvou odlišných textů daného autora; pokud u daného autora nemáme dva odlišné texty a určitý text má pro dané účely dostatečnou délku (minimálně 18000 slov před dělením, viz Faltýnek – Benešová – Kučera 2023), dělíme jej na vzorky a v analýze postupujeme totožným způsobem jako při práci s texty odlišitelnými na základě určitých pragmatických kritérií, tzn. přirozené delimitace (trvání konverzace, od první do poslední strany knihy/dokumentu apod.); ii) velké množství autorských textů neposkytuje z hlediska analýzy dostatečné množství textové evidence k prověřování otázek týkajících se autorství, struktury textu apod.; máme na mysli např. izolované emaily, izolované části komunikace na sociálních sítích, blogů atd. Pokud máme pro konkrétního autora k dispozici větší množství takovéhoto textového materiálu, komponujeme jej z toho důvodu do vzorků o podobné délce na základě požadavků na reliabilní výsledky analýzy (Faltýnek – Matlach 2021). V předstihu před oddílem Diskuze zde chceme formulovat, že autorská struktura opakovaného nízko frekventovaného lexika se projevuje bez ohledu na přirozené ohraničení textové produkce autora a její délku (pro validní provedení analýzy je však třeba mít k dispozici výše uvedené kvantum textu). To znamená, že je možné autorské texty sdružovat nebo naopak je dělit pro účely analýz, jak bylo popsáno výše. Předpokládáme, respektive vycházíme z různých ověření tohoto předpokladu

(např. Faltýnek 2020; Faltýnek – Kučera 2022), že pokud máme k dispozici libovolně komponovaný dostatečný rozsah autorova textu, disponujeme zároveň potřebným množstvím jeho preferovaných obsahových slov pro odlišení od jakéhokoliv jiného jedince (v závislosti na čase vzniku a stylu; viz Faltýnek – Kučera 2022). Domníváme se zároveň, že autorský obsahový idiolekt se distribuuje v různých autorových textech homogenně (viz Obrázek 1 a především grafy distribuce hapax legomen v autorském textu v oddíle Analýza).

Pro další výklad definujeme superhapax: Superhapax je taková slovní forma nebo spojení slov vyskytující se pravidelně s nízkou frekvencí v určitém textu nebo vzorcích textu konkrétního autora. Jedná se zároveň o slovní formu identifikující texty nebo vzorky textů s určitým autorem; tzn. že tyto slovní formy jsou odpovědné za klastrování daného autora a nevyskytují se u srovnávaných autorů v podobném zastoupení.

Obsahová a zároveň nízko frekventovaná slova byla již v určování autorství několikrát použita. Slava Katz (1996) detailně prověřuje pravděpodobnost výskytu obsahových slov v textu. R. Harald Baayen a kol. (1996) upozorňují, že hapax legomena a jejich syntaktické vazby mohou sloužit jako signál autorství. Nízko frekventované lexikum zároveň Baayen chápe jako užívané nevědomě (1996, s. 22), podobně jako je tomu u funkčních slov (Binongo 2003). Jsou-li funkční slova z tohoto důvodu ideální stopou autora v textu, jak jsme již jednou vyjádřili, pak i hapax legomena by takové nezáměrně zanechané stopy mohla poskytovat. Obsahová slova užil v určování autorství George K. Mikros (2009, s. 67), mluví o autorsky specifických slovech (Author-Specific Words), která vybírá na základě srovnání autorů v plném rozsahu frekvenčního spektra. Autorsky specifická slova srovnává Mikros s dalšími autorskými charakteristikami jako jsou funkční slova, frekvence délek slov nebo průměrnou délkou věty (Mikros 2009, s. 71 – 72). Ačkoliv se u Mikrose k identifikaci autora ukazují autorská obsahová slova jako výhodnější než zbylé ukazatele, další vývoj v určování autorství tento směr nesleduje (Evert a kol. 2017). G. K. Mikros zároveň pracuje s autorskými specifickými slovy v malém rozsahu textu (50 až 100 slov pro úspěšnou aplikaci, nejvýše pak 500). Malý rozsah vzorku, zahrnutí celého frekvenčního spektra a způsob zpracování (Discriminant function analysis, DFA) způsobují, že neidentifikuje autorský obsahový profil v širším zastoupení autorského idiolektu (graf na s. 70 ukazuje, že při zpracování způsobem DFA dochází ke zhoršení performance nad 500 slovy užitými ke srovnání vzorků; analogickou situaci a její vysvětlení najdete na obrázku č. 6 a v příslušném oddílu ve Faltýnek – Matlach 2021). Jan Rybicki a Maciej Eder (2011) se zaměřují na analýzu autorství s využitím lexika s nižší frekvencí, než je pouze lexikum nejfrekventovanější (právě ve smyslu funkčních slov), v této souvislosti využívají řezy plným frekvenčním spektrem slovní zásoby. Postupují ovšem od nejfrekventovanějších částí k méně frekventovaným v řezech obsahujících 50 až 5000 nejvíce frekventovaných slovních forem (toto

okno postupně posouvají směrem k nižším frekvencím) a analogicky k Mikrosovi tak neanalyzují nízko frekventovanou slovní zásobu samostatně (využívají manhattanskou vzdálenost a z-score; Burrows 2002). Stejně jako u Mikrose tento postup vede ke ztrátě informace o autorovi obsažené samostatně v nejnižším spektru lexika (viz oddíl *Further research* ve Faltýnek – Matlach 2021). Zároveň tito autoři upřednostňují položky s vyšší frekvencí nebo z rozsáhlé části frekvenčního spektra (Rybicki – Eder 2011, s. 5; Mikros 2009, s. 67). K. Amelin a kol. (2018) využívají podobné řezy frekvenčním spektrem analogicky k Rybickému s Ederem (2011) pro trasování rozvoje autorského stylu, jejich volba slovníku ale znovu sleduje vysoké spektrum, i když ve velkém rozsahu zastoupených typů (viz Amelin a kol. 2018, s. 46 a 54; viz také Eder 2015; 2017). J. Savoy (2012) využívá nízko frekventované lexikum k atribuci s výsledky překonávajícími atribuci na základě funkčních slov, zároveň ale pracuje s rozsáhlým spektrem obsahových slov a strukturu nejníže frekventovaného lexika neidentifikuje. Podobným způsobem jako v našem přístupu pracuje s texty a jejich vzorky pro umožnění komparace obsahových slov zastoupených v autorských textech s určitou frekvencí, v našem případě ovšem pouze s hraničně nízkou z hlediska délky srovnávaných textů (Savoy 2012, s. 7). V. Ficcadenti, R. Cerqueti a M. Ausloos (2018, s. 133) využívají hapax legomen jako samostatné charakteristiky autorského idiolektu při porovnání stylu amerických prezidentů, strukturu nízko frekventovaného lexika ale nepozorují. A. Lardilleux a Y. Lepage (2007) upozorňují na informaci obsaženou v hapax legomenech v souvislosti se slovním alignmentem, respektive automatickým překladem (pro vysvětlení v souvislosti s umístěním hapaxů v určitých frázích viz Faltýnek 2020).

4. ANALÝZA

V následující analýze se zaměříme na rozsah opakování superhapaxů v textu. Naším cílem je znovu prověřit hypotézu, že superhapaxy tvoří bez ohledu na téma textu jazykový profil autora (Faltýnek – Matlach 2021). V návaznosti na to se chceme dotázat, jaký rozsah textu je opakovaným obsahovým a funkčním lexikem autora dotčen a v souvislosti s tím jaký vliv má autorské lexikum na reflexi kontextu, tzn. co do našich textů vkládá kontext a co náš zvyk z něj něco zmiňovat. A konečně chceme také v dalších fázích výzkumu zjistit, jak se opakování superhapaxů dotýká výstavby textu, jeho celkové koheze, ve smyslu nejen obsahové (mluvíme o obsahových slovech), ale zároveň formální soudržnosti, včetně vlivu tohoto opakování na použití gramatických prostředků autorem.

4.1 Analyzovaný materiál

K analýze jsme vybrali texty dvou autorů mající blízkou tematiku – reflektují autismus (jelikož je naším cílem ověřovat autorský obsahový profil, chceme srovnávané texty co nejvíce přiblížit obsahově, aby byl výsledek klastrování projevem au-

tora, nikoliv obsahu). V obou případech se jedná o texty vzniklé jako příspěvky na blozích. Pro analýzu jsme kompletní rozsah obou autorských textů rozdělili na vzorky o podobné délce. Autorem prvního souboru vzorků je Martin Selner, texty byly publikovány na blogu dostupném z: http://selner84.blogspot.com/ (navštíveno 15. 12. 2021). Obsah blogu Martina Selnera byl později vydán knižně ve dvou dílech (Selner 2017 a 2019). My jsme pracovali s náhodným pořadím příspěvků na blogu. Druhým autorem je Natálie Ficencová (Zrzavá holka). Soubor jejích textů pochází z blogu dostupného z: https://zrzi.cz/ (navštíveno 15. 12. 2021). Obsah tohoto blogu také vyšel knižně (Ficencová 2021), jako s dostupnými jsme pracovali opět s texty publikovanými na internetu. Ve vzorcích jsou texty řazeny dle pořadí témat rozdělujících příspěvky na blogu (ve směru čtení). Do vzorků jsme nezařazovali příspěvky, které jsou autorčiným překladem jiných textů. Analyzované texty jsme rozdělili na vzorky o velikosti, která odpovídá požadavkům na reliabilní určení autorství na základě nízko frekventovaného lexika (Faltýnek – Matlach 2021). Velikosti vzorků jsou:

Selner 1: 9581 slov
Selner 2: 9552 slov
Selner 3: 9665 slov
Selner 4: 8920 slov
(celkově: 37 718 slov)

Zrzavá holka 1: 9652 slov
Zrzavá holka 2: 9557 slov
Zrzavá holka 3: 9535 slov
Zrzavá holka 4: 13695 slov
(celkově: 42 439 slov)

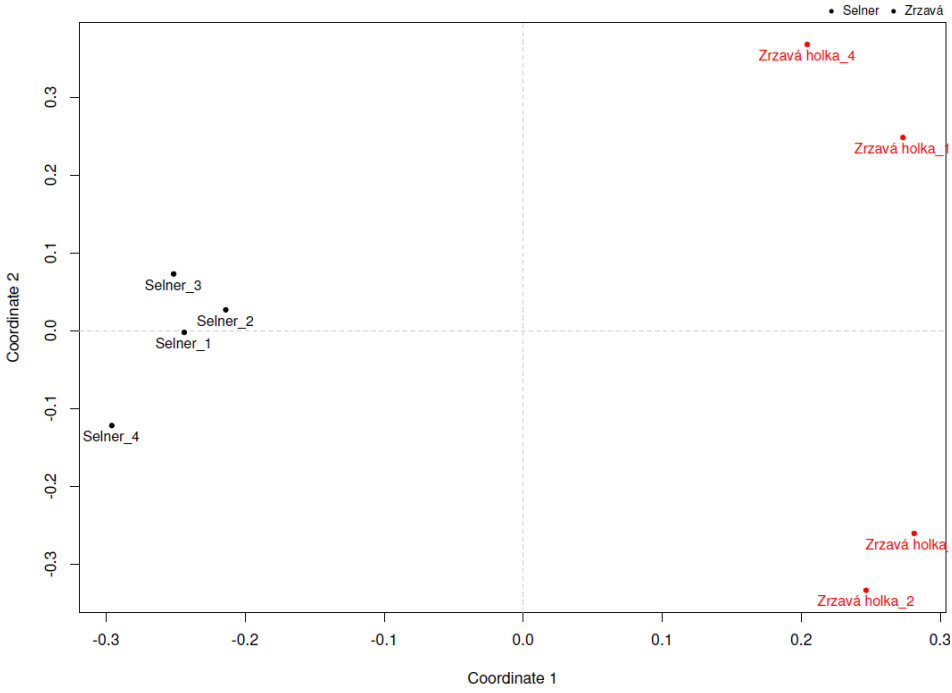
4.2 Klastrování autorů na základě superhapaxů

Primárním cílem tohoto článku je potvrdit hypotézu, že existuje jazykový profil autora. Chceme také s určitostí vědět, pokud by takový profil existoval, jak se jeho jednotlivé součásti zapojují do struktury textu. Předpokládáme totiž zastoupení mnohem širší, než je tomu u okazionálních autorských výrazů uvedených v úvodu (jejich distribuci ilustruje Obrázek č. 1). Pro tyto účely budeme výše popsaný materiál analyzovat na základě nejméně frekventovaných prvků textu, to nám zaručuje, že vybraný soubor slovních forem obsahuje především obsahová slova, a zároveň slova s co nejvyšší mírou nezávislá na obsahu textu.

Jak bylo výše zavedeno, sumáře autorských textů jsou rozděleny na jednotlivé vzorky, ve kterých jsou detekovány hapax legomena. Průnik množin hapax legomen získaných z jednotlivých vzorků definuje množinu superhapaxů daného autora.

Rozsah textu pro úspěšné separování autorských textů do společného klastru v závislosti na délce textu a jazykovém typu je uveden ve Faltýnek – Matlach

(2021; vzhledem k této evaluaci můžeme konstatovat, že délka námi zvolených vzorků bezpečně přesahuje minimální rozsah textu pro úspěšné autorské klastrování, ten vychází z procenta hapax legomen v textu a minimálního množství hapax legomen nutných k úspěšnému odlišení autorských vzorků). Následující graf ukazuje, že nízko frekventované lexikum jednoznačně odlišuje autory vybraných vzorků textu.

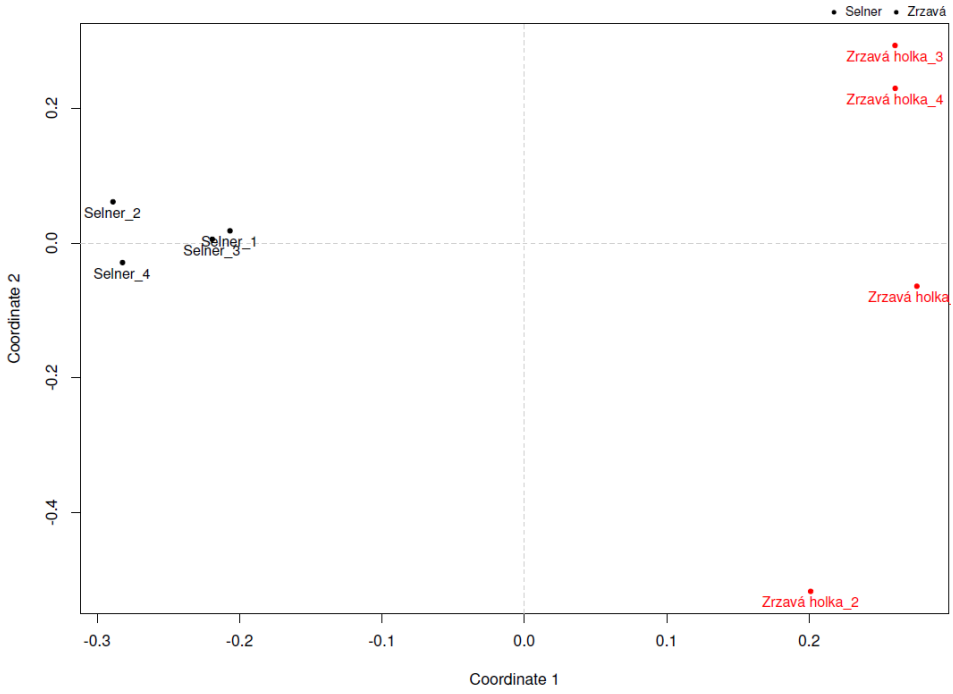


Obrázek 2: Klastrování vzorků blogu *Autismus a Chardonnay* Martina Selnera (dostuné z: http://selner84.blogspot.com/; navštíveno 15. 12. 2021) a blogu *Zrzavá holka* Natálie Ficencové (dostuné z: https://zrzi.cz/2015/12/nebinarita/; navštíveno 15. 12. 2021). Klastrování bylo provedeno na základě hapax legomen přítomných ve vzorcích srovnaných kosinovou nepodobností, zobrazení bylo provedeno formou vícerozměrného škálování (Torgerson 1952).

4.3 Klastrování autorů na základě hapax legomen přítomných v náhodných vzorcích textu

Proto, abychom se ujistili, že autorské klastrování vzorků Martina Selnera a Zrzavé holky z Obrázku č. 2 na základě hapax legomen není dílem náhody, respektive jedinečné konstelace ve vzorcích přítomných, tedy opakujících se hapax legomen, provedli jsme deset klastrování těchto vzorků na základě náhodného výběru 6250 slov z jejich vzorků (původní velikosti viz v oddíle Analyzovaný materiál). Ve všech

deseti případech náhodných výběrů byly vzorky obou autorů jednoznačně separovány v autorských klastrech. Pro ukázku uvádíme výsledek klastrování prvního z náhodných výběrů, modely bag-of-words všech náhodných výběrů jsou dostupné na vyžádání u autorů tohoto článku



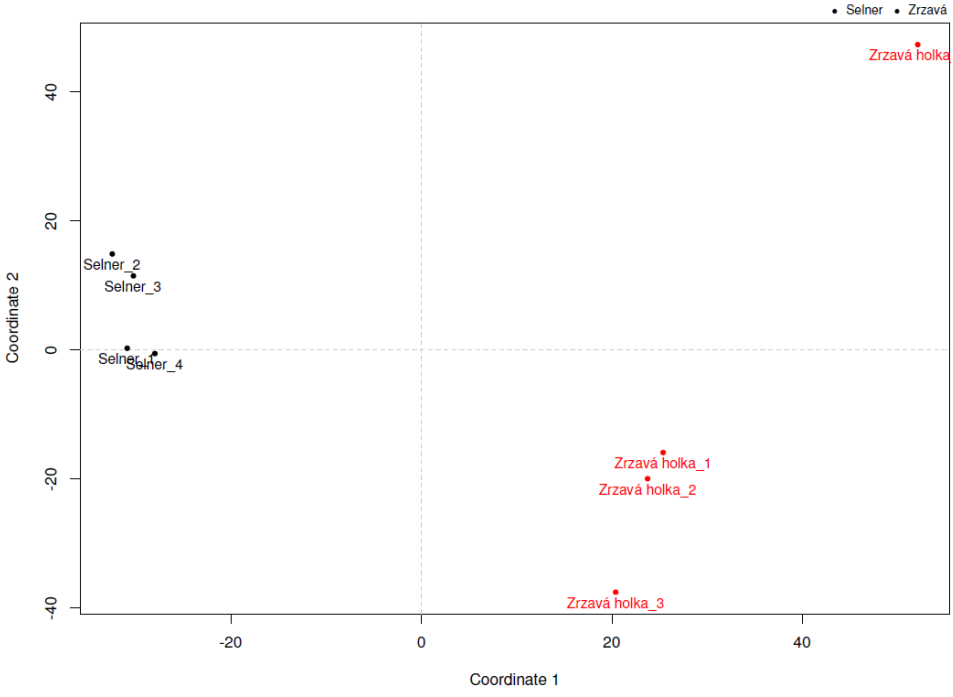
Obrázek 3: Klastrování vzorků blogu *Autismus a Chardonnay* Martina Selnera (dostuné z: http://selner84.blogspot.com/; navštíveno 15. 12. 2021) a blogu *Zrzavá holka* Natálie Ficencové (dostuné z: https://zrzi.cz/2015/12/nebinarita/; navštíveno 15. 12. 2021). Klastrování bylo provedeno na základě náhodného výběru 6250 hapax legomen přítomných ve vzorcích srovnaných kosinovou nepodobností, zobrazení bylo provedeno formou vícerozměrného škálování (Torgerson 1952).

4.4 Klastrování autorů na základě funkčních slov

Pro srovnání s autorskou atribucí na základě hapax legomen jsme provedli autorskou atribuci vzorků Martina Selnera a Natálie Ficencové za užití funkčních slov. Funkční slova jsme vybrali ze 100 nejfrekventovanějších slovních forem v souboru sloučených sumářů textů obou autorů. Z těchto 100 slovních forem jsme odstranili slova obsahová. V následující tabulce uvádíme zbylé slovní formy použité v analýze.

a	aby	ale	ani	asi	by	bych	co
do	ho	i	já	jedno	jen	jestli	jim
jsou	k	které	který	mé	mě	mi	možná
mu	na	nás	ne	nebo	něco	někdo	než
o	od	po	pro	protože	s	se	si
své	tak	taky	takže	ten	tím	to	toho
tom	třeba	tu	ty	u	v	vám	ve
vlastně	všechno	z	za	že			

Tabulka č. 1: Seznam funkčních slov využitých ke srovnání vzorků blogu *Autismus a Chardonnay* a blogu *Zrzavá holka* Natálie Ficencové. Seznam vznikl sloučením všech vzorků do jednoho textu, z jehož frekvenčního seznamu bylo vzato 100 nejfrekventovanějších slovních forem. Z těchto slovních forem byla odstraněna obsahová slova (substantiva, adjektiva, verba a adverbia).



Obrázek 4: Klastrování vzorků blogu *Autismus a Chardonnay* Martina Selnera (dostuné z: http://selner84.blogspot.com/; navštíveno 15. 12. 2021) a blogu *Zrzavá holka* Natálie Ficencové (dostuné z: https://zrzi.cz/2015/12/nebinarita/; navštíveno 15. 12. 2021). Klastrování bylo provedeno na základě srovnání výskytu funkčních slov ve vzorcích, frekvence funkčních slov byla transformována na z-skóre. Srovnání proběhlo formou manhattanské metriky, k vizualizaci bylo využito vícerozměrné škálování (Torgerson 1952). Seznam funkčních slov použitých pro srovnání textů je uveden v Tabulce č. 1.

Analýza obou souborů autorských vzorků (Selner 1 – 4 a Zrzavá holka 1 – 4) za použití funkčních slov prokázala podobné výsledky jako analýza hapax legomen/superhapaxů. Autoři se na základě použití základních textotvorných prostředků odlišili prokazatelně a vytvářejí jednoznačně oddělitelné skupiny.

5. DISTRIBUCE

Na základě předchozích analýz jsme zjistili, že na vybraném jazykovém materiálu se autor jeví jako individualita s ohledem na gramatické či textotvorné prostředky (funkční slova), které je zvyklý v textu užívat, respektive je zvyklý je užívat s určitou proporcionalitou. Zároveň ale zjišťujeme, že autorský profil má autor daný také obsahovými slovy, která v textu používá bez ohledu na kontext, respektive je uvyklý určité obsahy v textu opakovaně zmiňovat, což právě vede k jeho jednoznačné identifikaci. Vzhledem k rozvoji našeho výkladu můžeme nyní formulovat zjištění, že obsahový profil autora nepřipomíná svou distribucí distribuci autorových okazionálních prostředků typu "*protřepat, nemíchat"*, jakými uvozujeme tento článek v prvním oddíle. Autor své preferované obsahy distribuuje frekventovaně a napříč textem, víme to s určitostí právě proto, že se i v rozsahu 6250 slov vždy tematicky identifikuje s jinými svými textovými vzorky (viz aposteriorní metoda ve Faltýnek – Matlach 2021).

Počet zastoupení funkčních slov využitých v analýze výše činí pro vzorky Natálie Ficencové 31,9 procenta (13 529 slov), v případě vzorků Martina Selnera 27,5 procenta (11 666 slov). To znamená že téměř v každém větném celku nacházíme funkční prostředek typický pro autora ve své frekvenční zastoupenosti. V níže uvedených grafech (Obrázek č. 5) uvádíme distribuci šesti nejfrekventovanějších funkčních slov v textech obou autorů (viz Tabulka č. 1).

Obrázek 5: Distribuce šesti nejfrekventovanějších funkčních slov (viz Tabulka 1; jedná se o slova *a*, *aby*, *ale*, *ani*, *asi*, *by*) v textech Martina Selnera (vrchní graf) a Natálie Ficencové (spodní graf). Zleva doprava je vyznačen lineární rozvoj textu, svislé černé čáry označují výskyty vybraných slov. Grafy byly pořízeny v softwaru AntConc (https://www.laurenceanthony.net/software/antconc/). Distribuce slov prokazuje i v dalších nezobrazených částech textu pravidelnou distribuci (k tomu viz de Roeck a kol. 2004 a Sarkar a kol. 2005).

Pro srovnání s distribucí funkčních slov v autorském textu se nyní zaměříme na distribuci autory užívaných obsahových slov prozrazujících jejich totožnost (jsou rozhodující pro výsledky klastrování) a zároveň ukazují specifické preference autorů konkretizovat určitý obsahový aspekt v různých kontextech určených aktuálním aktem komunikace. Vlastnosti těchto slov mohou být nejspíše dále prověřovány z hlediska autorské osobnostní idiosynkracie; výklady k tomu ponecháváme do navazujících publikací.

Obrázek 6 a 7: Distribuce superhapaxů odpovědných za klastrování vzorků Martina Selnera (horní graf) a Natálie Ficencové (spodní graf); soubory superhapaxů jsou pro oba autory odlišné. Pro zobrazení byla vzata hapax legomena vyskytující se ve všech čtyřech vzorcích současně, decisivní roli při autorském klastrování mají ale superhapaxy vyskytující se ve třech, respektive dvou vzorcích textu (těchto superhapaxů je v případě Martina Selnera 4001 a v případě Natálie Ficencové 3817. Popis viz Obrázek č. 5). Grafy byly pořízeny v softwaru AntConc (https://www.laurenceanthony.net/software/antconc/).

Procento zastoupení opakujících se obsahových slov v autorském textu je z hlediska našich vzorků následující: v textu Martina Selnera se hapax legomena určující autorův klastering vyskytují ve 4001 případech (tzn. 9,4 procentech celého textu). To znamená, že se autorův superhapax vyskytuje v cca každém desátém slově v textu. V textu Natálie Ficencové se hapax legomena určující klastering objevují v 3817 výskytech (tzn. v 10,1 procentech celého textu). Klíčový superhapax v jejím textu nacházíme tedy opět v cca každém desátém slově. Pro autora typické obsahové slovo se tak vyskytuje ve vzdálenosti průměrně dvou či tří vět. Zjišťujeme proto, že obsahy, které autor typicky opakuje v textu, ovlivňují výstavbu věty a mezivětné navazování.

6. DISKUZE

Na místě každého desátého slova v textu autor pootočí pomyslným mlýnkem paradigmatu a na dané místo v textu dosadí z mnoha možných funkčně ekvivalentních jazykových prostředků sobě vlastní. Zopakuje se. Vcházíme každou minutou do nových okolností, s novými cíli, ale zapínáme přitom staré registry čárového kódu našich superhapaxů – preferovaných způsobů vyjádření, témat a postojů. Ve struktuře našich textů superhapaxy svou obligatorní přítomností rozvrhují bílá místa, kde se kontext může přihlásit o nová slova. Tvar a barvu mu ale svými superhapaxy nejdřív vyznačíme my.

Jako lidé jsme asi zapamatovatelné individuality. Ale všímáme si – včetně lingvistiky a disciplíny určování autorství – u druhých jen nadpisů a viditelných stop, martini a chardonnay a zrzavé barvy. Nejsme s to dosáhnout povědomí specifické skupiny slov, která činí něčí individualitu takovou, že každý jeden nabývá svou tvář: Spojovat si jednotlivce s tisíci slovními formami, které preferuje, neumíme. Znát někoho naopak asi znamená tušit všechna jeho opakování.

Kdyby byl text strukturován jen naší intencí a neodhadnutelností každého příštího kontextu, zobrazovaly by klastrové analýzy nízkých frekvencí různých autorů pomyslnou termodynamickou smrt. Všechny body by byly ideálně stejně

vzdálené, jen náhodné shody v mikrotématech jednotlivých textů by je nepatrně přibližovaly. Klastrové analýzy nízko frekventovaného lexika ale ukazují nakupené mraky odpovídající jednotlivým autorům. Protože ať už jsme kdekoliv, opakujeme v nízkém lexiku své postoje, témata a způsoby vyjádření. Lingvistika nás v posledních desetiletích naučila dívat se na texty očima velkých korpusů, jazykových dat slitých do hyperreprezentativní množiny tvarů, jejichž frekvence nám dává odpovědi na otázky o povaze jazykových norem, systémů a mluvícího člověka (výhody tohoto přístupu výstižně formuluje Štícha 2001; k následující formulaci viz např. Forster – Chambers 1973). Pravdou ale je, že tato globální reprezentace jazyka nás jako jednotlivce svým stínem nepokrývá. Opakujeme se ve vlastním frekvenčním efektu.

7. DALŠÍ VÝZKUM

V následném výzkumu chceme ukázat, že výskyt superhapaxů, opakujících se nízko frekventovaných autorských slov, je v textu natolik pravidelný a rozmístěný v homogenní disperzi, že nutně ovlivňuje strukturu vět a s určitostí rovinu mezivětného navazování. Mluvíme v této souvislosti o parasyntaxi (viz Faltýnek – Kučera 2022). Chceme později doložit, že koheze textu je určována nikoliv pouze zdola gramatickými vztahy v mezivětném navazování, ale je řízena shora autorským preferovaným lexikem – obligatorně vyskytujícím se v rozsáhlém textu analogicky k obligatorním prostředkům větné skladby. Parasyntax jako jazyková rovina pak představuje specifický výběr v paradigmatu při vytváření syntagmatu jedincem. To zároveň znamená, že stejně jako registrujeme ve větě nutné prostředky její stavby, např. přísudek a jeho valenční doplnění, tak má vyšší textová rovina z hlediska principů svého uspořádání nutně přítomná a zbytek textu určující slova. Tato slova mají určité gramaticko-lexikální vlastnosti, které si vynucují zapojení zbylých slov do textu. Veškeré aspekty s touto vlastností textu vázanou na jednotlivce naprosto vystihuje koncept bezznalostní jazykové dispozice vypracované Jurajem Dolníkem. Vlastnosti textu s tímto spojené pak zahrnují znaky používané v metodách disciplíny určování autorství. Juraj Dolník tyto textové znaky a zároveň disciplínu určování autorství výstižně zařazuje do všeobecné nauky o jazyku a textu. Současně musíme revidovat tradiční kritéria textovosti jakožto určující faktor struktury textu širšího rozsahu, protože jsou částečně nebo ve velkém rozsahu predominovány autorským obsahovým profilem, superhapaxy.

Bibliografie

AMELIN, Konstantin – GRANICHIN, Oleg – KIZHAEVA, Natalia – VOLKOVICH, Zeev (2018): Patterning of writing style evolution by means of dynamic similarity. In: *Pattern Recognition*, roč. 77, s. 45 – 64.

BAAYEN, R. Harald – van HALTEREN, Hans – TWEEDIE, Fiona (1996): Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. In: *Literary and Linguistic Computing*, roč. 11, č. 3, s. 121 – 132.

BAAYEN, R. Harald – van HALTEREN, Hans – NEIJT, Anneke – TWEEDIE, Fiona (2002): An experiment in authorship attribution. In: *Proceedings of JADT, St. Malo*. Université de Rennes, s. 29 – 37.

BINONGO, José Nilo G. (2003): Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. In: *Chance*, roč. 16, č. 2, s. 9 – 17.

BURROWS, John (2002): "Delta": A measure of stylistic difference and a guide to likely authorship. In: *Literary and Linguistic Computing*, roč. 17, č. 3, s. 267 – 287.

BURROWS, John (2003): Questions of Authorship: Attribution and Beyond. In: *Computers and the Humanities*, roč. 37, č. 1, s. 5 – 32.

BURROWS, John (2007): All the Way Through: Testing for Authorship in Different Frequency Strata. In: *Literary and Linguistic Computing*, roč. 22, č. 1, s. 27 – 47.

de BEAUGRANDE, Robert-Alain – DRESSLER, Wolfgang (1981): *Introduction to text linguistics (Vol. 1)*. London: Routledge.

de ROEC, Anne – SARKAR, Avik – GARTHWAITE, Paul (2004): Frequent Term Distribution Measures for Dataset Profiling. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

DOLNÍK, Juraj (2018): Jazykové znalosti a ovládanie jazyka. In: *Jazykovedný časopis*, roč. 69, č. 1, s. 77 – 89. DOI 10.2478/jazcas-2018-0013. Dostupné na: https://www.juls.savba.sk/ediela/jc/

DOLNÍK, Juraj (2019): Stratifikácia používateľov jazyka. In: *Jazykovedný časopis*, roč. 70, č. 3, s. 515 – 528. DOI 10.2478/jazcas-2020-0002. Dostupné na: https://www.juls.savba.sk/ediela/jc/

DOLNÍK, Juraj (2021): Vertikálne a horizontálne jazykové potreby. In: *Jazykovedný časopis*, roč. 72, č. 1, s. 21 – 36. DOI 10.2478/jazcas-2021-0011. Dostupné na: https://www.juls.savba.sk/ediela/jc/

EDER, Maciej (2015): Does size matter? Authorship attribution, small samples, big problem. In: *Digital Scholarship in the Humanities*, roč. 30, č. 2, s. 167 – 182.

EDER, Maciej (2017): Short samples in authorship attribution: A new approach. In: *Digital Humanities 2017: Conference Abstracts*. Montreal: McGill University, s. 221 – 224.

EVERT, Stefan – PROISL, Thomas – JANNIDIS, Fotis – REGER, Isabella – PIELSTRÖM, Steffen – SCHÖCH, Christof – VITT, Thorsten (2017): Understanding and explaining Delta measures for authorship attribution. In: *Digital Scholarship in the Humanities*, roč. 32, č. 2, s. 4 – 16.

FALTÝNEK, Dan (2020): Celkom iste sa príde na to, že niektoré slová sa opakujů doslovne: k Horeckého hypersyntaxi. In: *Jazykovedný časopis*, roč. 71, č 2, s. 185 – 196. DOI 10.2478/jazcas-2020-0021. Dostupné na: https://www.juls.savba.sk/ediela/jc/

FALTÝNEK, Dan – MATLACH, Vladimír (2021): Hapax remains: Regularity of low-frequency words in authorial texts. In: *Digital Scholarship in the Humanities*, roč. 37, č. 3, s. 693 – 715. DOI 10.1093/llc/fqab077. Dostupné na: https://doi.org/10.1093/llc/fqab077

FALTÝNEK, D. – BENEŠOVÁ, M. – KUČERA, O. (2023): How to Use Personal Linguistic Profile to Control People: Deep Sentiment and Content Analysis. In: C. Shei – J. Schnell (eds.): *J. Routledge Handbook of Language and Mind Engineering*. Routledge (v tlači).

FALTÝNEK, Dan – KUČERA, Ondřej (2022): Parasyntax jako struktura nízko frekventovaných částí textu: Hapax legomenon prostředkem textové koheze. In: Z. Popovičová Sedláčková (ed.): *Štýl – komunikácia – kultúra. K stému výročiu narodenia profesora Jozefa Mistríka (1921 – 2000)*. Bratislava: Univerzita Komenského, s. 403 – 416. Dostupné na: https://fphil.uniba.sk/katedry-a-odborne-pracoviska/ksjtk/publikacie/

FICCADENTI, Valerio – CERQUETI, Roy – AUSLOOS, Marcel (2019): A joint text mining-rank size investigation of the rhetoric structures of the US Presidents' speeches. In: *Expert Systems With Applications*, roč. 123, s. 127 – 142.

FORSTER, Kenneth I. – CHAMBERS, Susan M. (1973): Lexical access and naming time. In: *Journal of Verbal Learning & Verbal Behavior*, roč. 12, č. 6, s. 627 – 635.

GREIMAS, Algirdas Julien (1987): *On Meaning: Selected Writings in Semiotic Theory*. Minneapolis: University of Minnesota Press.

GRZYBEK, Peter (2013): Homogeneity and heterogeneity within language(s) and text(s): theory and practice of word length modeling. In: R. Köhler – G. Altmann (eds.): *Issues in Quantitative Linguistics 3*. Lüdenscheid: RAM, s. 66 – 99.

HIRSCHOVÁ, Milada (1989): *Úvod do teorie textu*. Olomouc: Univerzita Palackého.

JUOLA, Patrick (2006): Authorship attribution. In: *Foundations and Trends in Information Retrieval*, roč. 1, č. 3. Dostupné na: https://www.researchgate.net/publication/308073726_Authorship_attribution [datum přístupu 27. 12. 2021].

JUOLA, Patrick – BAAYEN, R. Harald (2003): A controlled-corpus experiment in authorship identification by cross-entropy. In: *Literary and Linguistic Computing*, roč. 20, s. 59 – 67.

KATZ, Slava (1996): Distribution of content words and phrases in text and language modelling. In: *Natural Language Engineering*, roč. 2, č. 1, s. 15 – 59.

LARDILLEUX, Adrien – LEPAGE, Yves (2007): The contribution of the notion of hapax legomena to word alignment. In: *LTC'07*, pp.0. <hal-00252026v1>

MIKROS, George K. (2009). Content words in authorship attribution: an evaluation of stylometric features in a literary corpus. In: R. Köhler (ed): *Studies in Quantitative Linguistics*, roč. 5. Lüdenscheid: RAM, s. 61 – 75.

MIKROS, George K. – ARGIRI, Eleni 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*, roč. 276, s. 29 – 35. Amsterdam, Netherlands: CEUR.

MOSTELLER, Frederick – WALLACE, David L. (1964): *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison Wesley Publishing Company.

RYBICKI, Jan – EDER, Maciej (2011): Deeper Delta across genres and languages: do we really need the most frequent words? In: *Literary and Linguistic Computing*, roč. 26, č. 3, s. 315 – 321.

SARKAR, Avik – GARTHWAITE, Paul – de ROECK, Anne (2005): A Bayesian mixture model for term re-occurrence and burstiness. In: I. Dagan – D. Gildea (eds.): *Proceedings*

of the 9th Conference on Computational Natural Language Learning (CoNLL). Ann Arbor, Michigan: Association for Computational Linguistics, s. 48 – 55.

SAVOY, Jacques (2012): Authorship Attribution Based on Specific Vocabulary. In: *ACM Transactions on Information Systems*, roč. 30, č. 2, článek č. 12, 30 s.

ŠTÍCHA, František (2001): Kritéria gramatičnosti (Korpus jako argument a inspirace). In: *Slovo a slovesnost*, roč. 62, č. 3, s. 161 – 175.

TORGERSON, Warren, S. (1952): Multidimensional scaling: I. Theory and method. In: *Psychometrika*, roč. 17, č. 4, s. 401 – 419.

ZHAO, Ying – ZOBEL, Justin (2005): Effective and scalable authorship attribution using function words. In: G. G. Lee – A. Yamada – H. Meng – S., H. Myaeng (eds.): *Proceedings of 2nd Asian Information Retrieval Symposium*, Berlin, Heidelberg: Springer, s. 174 – 189.

Primární zdroje

BACKMAN, Fredrick (2016): *Tady byla Britt-Marie*. Praha: Host.

CAMPBELL, Martin (2006): *Casino Royale* [film]. Režie Martin Campbell. Velká Británie: Sony Pictures Entertainment.

FICENCOVÁ, Natálie (2021): *Neříkejte autista, to se nehodí*. Praha: CPress. Dostupné z: https://zrzi.cz/.

FLEMING, Ian (1956): *Diamonds are Forever*. London: Pan Books Ltd.

FLEMING, Ian (2008): *Diamanty jsou věčné*. Překlad: Ivan Němeček. Praha: XYZ.

JANŠČOVÁ, Kristína (22. srpna, 2015): *Neváhajte a zavolajte ihneď! Horst Fuchs urobil týchto 9 vecí, aby predal veci v teleshopingu – A to ještě není vše, přátelé!* Získáno 25.12.2021 z: https://fici.sme.sk/c/20057011/nevahajte-a-zavolajte-ihned-horst-fuchs-urobil-tychto-9-veci-aby-predal-veci-v-teleshopingu.html

KEROUAC, Jack (1957): *On the Road*. Chicago: Viking Press.

SALINGER, Jerome David (1991): *The Catcher in the Rye*. New York, NY: Little, Brown and Company.

SALINGER, Jerome David (2000): *Kdo chytá v žitě*. Překlad: Luba a Rudolf Pellarovi. Praha: Volvox Globator.

SELNER, Martin (2019): *Autismus & Chardonnay*. Praha: Paseka. Dostupné z: http://selner84.blogspot.com/.

SMOLJAK, Ladislav – SVĚRÁK, Zdeněk (1994): *Záskok* [divadelní inscenace]. Režie: L. Smoljak, scénář: L. Smoljak – Z. Svěrák. Premiéra 27. 3. 1994, Žižkovské divadlo TGM.

SMOLJAK, Ladislav – SVĚRÁK, Zdeněk (2008): *České nebe aneb Cimrmanův dramatický kšaft* [divadelní inscenace]. Režie: L. Smoljak, scénář: L. Smoljak – Z. Svěrák. Premiéra 28. 10. 2008, Žižkovské divadlo TGM.