Filozofická fakulta Univerzity Palackého v Olomouci

Katedra obecné lingvistiky

# Využití sentimentu při určování autorství

*bakalářská diplomová práce*

Autor: Kateřina Chrápková

Vedoucí práce: Mgr. Vladimír Matlach, Ph.D.

**Olomouc**

2023

# PROHLÁŠENÍ

Prohlašuji, že jsem bakalářskou diplomovou práci „Využití sentimentu při určování autorství“ vypracovala samostatně a uvedla jsem veškerou použitou literaturu a veškeré použité zdroje.

V Olomouci dne 28.06.2023

……

Kateřina Chrápková

# PODĚKOVÁNÍ

Ráda bych tímto poděkovala svému vedoucímu práce Mgr. Vladimíru Matlachovi, Ph.D. za jeho odborné vedení a cenné rady. Dále také doc. Mgr. Danu Faltýnkovi, Ph.D. za poskytnutí rad do základů této práce a Bc. Františku Špačkovi za ochotu a poskytnutí nástroje na extrakci superhapaxů a jejich okolí.
V neposlední řadě bych ráda poděkovala svým blízkým, kteří mě podporovali.

# ABSTRAKT

Název práce: Využití sentimentu při určování autorství

Autor práce: Kateřina Chrápková

Vedoucí práce: Mgr. Vladimír Matlach, Ph.D.

Abstrakt: Diplomová práce se zabývá možností využití sentimentové analýzy při určování autorství. V rámci analýzy byly využity hapax legomena – slova v textu s frekvencí rovnající se 1 – a jejich jazykové okolí (dále jen „okolí“), které se může pohybovat od jazykových jednotek na úrovni slov až po úroveň několika vět, dále také nástroj LIWC, který poskytuje data sentimentové analýzy provedené na okolích superhapaxů. Poté byla aplikovaná metoda hlavních komponent, která slouží k získání správného počtu proměnných, a v neposlední řadě nástroj logistické regrese, který byl využit ke stanovení skutečnosti, zdali se dané proměnné dají použít k určení autorství, popřípadě s jakou přesností se dají využít. Celá analýza byla provedena na textech pěti českých autorů, jmenovitě autoři Jaroslav Foglar, Lenka Lanczová, Miloslav Švandrlík, Ludmila Vaňková a Jaroslav Velinský. Pro každého autora bylo dodáno dvacet pět jejich textů, ze kterých byly vyvozovány výsledky pro tyto autory. Tyto výsledky byly následně srovnané s počáteční hypotézou o schopnosti sentimentové analýzy na okolích superhapaxů rozlišit dané autory.

Klíčová slova: určování autorství, hapax legomenon, superhapax, sentimentová analýza, LIWC, logistická regrese, PCA, metoda hlavních komponent

# ABSTRACT

Title: The Use of Sentiment in Determining Authorship

Author: Kateřina Chrápková

Supervisor: Mgr. Vladimír Matlach, Ph.D.

Abstract: The thesis deals with the possibility of using the sentiment analysis as a tool to determine authorship. The analysis uses hapax legomena – words in a text with the frequency equal to 1 – and their surroundings, as well as the LIWC tool, which provides sentiment analysis data performed on the surroundings of superhapaxes. Furthermore, the Principal Component Analysis was employed to obtain the correct number of variables, and finally, the logistic regression tool was used to determine whether the given variables can be used for authorship attribution, or as the case may be with what accuracy they can be used. The whole analysis was conducted on the texts of five Czech authors, namely Jaroslav Foglar, Lenka Lanczová, Miloslav Švandrlík, Ludmila Vaňková, and Jaroslav Velinský. For each author, twenty-five of their texts were provided, from which the results were drawn for the named authors. These results were then compared with the initial hypothesis about the ability of sentiment analysis conducted on the surroundings of superhapaxes to distinguish the authors in question.

Keywords: authorship attribution, hapax legomenon, superhapax, sentiment analysis, LIWC, logistic regression, PCA, Principal Component Analysis

# OBSAH

# ÚVOD

Tématem této práce je analýza možnosti využití sentimentu při určování autorství. Budou zde využity hapax legomena, slova v textu s frekvencí rovnající se 1, a v jejich návaznosti především pak superhapaxy. Na superhapaxech a jejich okolí se bude provádět sentimentová analýza za pomocí nástroje LIWC. Mimo to zde bude zahrnuta i metoda hlavních komponent a logistická regrese, která zde bude hlavním nástrojem pro vyhodnocování analýz hodnot sentimentů v rámci určování autorství.

Podnětem pro toto téma jsou výzkumy prováděné v oblasti určování autorství, hlavně v nedávných letech. Zde se budu věnovat zvážení několika metod, ze kterých budou vybrány ty nejvhodnější pro cíle této práce. Určování autorství obecně spočívá v provádění analýzy schopnosti odlišit určitý počet autorů od sebe navzájem či rozlišit autora, který původcem daného textu být nemůže. K tomuto účelu se využívají rozličné nástroje, u jejichž výsledků se ve většině případů bavíme o tom, jak přesné tyto nástroje jsou při schopnosti odlišení autorů. Pro tuto práci jsou ve výsledku zvoleny hapax legomena, ze kterých jsou odvozené superhapaxy, a jejich okolí. Tato metoda je vybrána z důvodu její aktuality. V nedávné době se hapax legomena využívaly například k odhalování autorů na základě shlukování funkčních slov či ke zkoumání koheze textu. Vzhledem k tomuto vzrůstu využití a k výsledkům daných studií jsem se rozhodla při určování autorství v této práci jako základ využít právě tuto metodu.

Mimo to je součástí tématu také sentimentová analýza a její implikace založené především na studiích Jamese W. Pennebakera, který utvořil samotný nástroj LIWC využívaný během analýz této práce. Sentimentová analýza bere v potaz jednotlivá slova a přidružuje k nim hodnoty vypovídající o jejich sentimentu v několika desítkách různých kategorií. Dále bude v práci také využita metoda hlavních komponent, která se zaměřuje na snížení počtu proměnných na správný počet pro potřebnou analýzu. Nástrojem, ve kterém bude na závěr provedena analýza přesnosti rozlišitelnosti autorů na základě hodnot ze sentimentové analýzy, je logistická regrese. Logistická regrese pracuje s číselnými i binárními proměnnými, což z ní činí při analýze autor A/B v kombinaci s příslušnými vyčíslenými hodnotami sentimentu vhodnou volbu pro postup.

Korpusem této práce budou knihy pěti českých autorů uložené v elektronické podobě v textových souborech – konkrétně se jedná o autory Jaroslav Foglar, Lenka Lanczová, Miloslav Švandrlík, Ludmila Vaňková, Jaroslav Velinský. Korpus se skládá z 25 dostupných děl s délkou alespoň 150 000 znaků od každého z autorů. Jmenný seznam děl autorů je vypsán viz *Příloha č. 1*. Jazykový materiál z děl vyselektovaný později v práci, přesněji superhapaxy

a jejich okolí, bylo nutné dále převést do anglického jazyka z důvodu kompatibility pro zpracování v nástroji LIWC2015.

První část práce se zaměřuje na obecnou teorii, která se týká používaných nástrojů a relevantních výzkumů. Přesněji se pak jedná o témata určování autorství, hapax legomen a na závěr téma samotného sentimentu a sentimentové analýzy. Významnými jmény v této části jsou např. James W. Pennebaker, H. Baayen a Dan Faltýnek. Druhá část pak bude dedikována metodologii a z ní plynoucím výsledkům. Zde se budu věnovat popisu fungování metod, které budou používané v příslušných částech, samotnému konkrétnímu způsobu jejich použití v případě této práce a také výsledkům, které z tohoto použití plynou.

Metodologická část se jako první bude věnovat extrakci superhapaxů a jejich okolí a následně jejich zmíněnému překladu do anglického jazyka. Po tomto kroku bude pro další postup analýzy nutné provést selekci zájmových a relevantních vlastností z nástroje LIWC. Zde bude také prostor na zvážení toho, jaký vliv má na vlastnosti délka textu a jak moc mezi sebou vlastnosti korelují a jsou tak vzájemně redundantní. Krokem finálním bude samotná analýza vybraných vlastností v logistické regresi. V tomto kroku budou zahrnuté dvě fáze a jejich výsledky – fáze analýzy vybraných původních vlastností každé zvlášť a fáze analýzy všech vytvořených komponent metodou hlavních komponent zároveň.

V této práci se nebudu věnovat všem sentimentovým kategoriím nástroje LIWC, kterých je celkově 93, a to z důvodu jejich vysokého počtu a z toho vyplývající časové náročnosti při omezené časové i rozsahové kapacitě této práce. Kategorie, které nebyly využity při této práci, mohou být podnětem pro další výzkum. Navazujícím motivem dalších prací může být také změna rozsahu okolí superhapaxů při jejich extrakci, jelikož je možné tento parametr při dalších analýzách upravit a vyhodnotit tak nové výsledky na základě daného výběru dat.

Cílem této práce je prokázat, zdali je pravdivá hypotéza, která pracuje s premisou, že sentimentová analýza provedená na okolích superhapaxů by mohla dokázat za pomoci metody hlavních komponent a logistické regrese rozlišit autory, popřípadě s jakou přesností je možné tohoto závěru dokázat. Veškeré použité metody a nástroje tak napomáhají tohoto cíle dosáhnout co nejefektivněji a nejkvalitněji.

# TEORETICKÁ ČÁST A CÍL PRÁCE

Hlavním tématem této práce je určování autorství, za jehož účelem zde budou využity hapax legomena, respektive superhapaxy, a sentimentová analýza. Během procesu určování autorství probíhá determinace jednoho či více původců daného textu či textů. Tento koncept se stal populárnějším především z důvodu digitalizace, rozšíření metod strojového učení a rozvoje korpusové lingvistiky. (Mikros, 2009) Obecně lze za tímto účelem využít mnoho metod, přičemž mezi nejčastější patří tzv. stylometrické metody, které pomocí daných vlastností textu měří styl autora a co jej v textu odhaluje či odlišuje od ostatních.

Mezi nejčastěji využívané vlastnosti se dnes zařazují vlastnosti lexika (text jako po sobě jdoucí tokeny), vlastnosti syntaxe (jakým způsobem se skládají věty), strukturní vlastnosti textu (organizace textu), obsahové vlastnosti textu (klíčová slova, specifika obsahu) ad. (Sindhu, a další, 2017) V rámci těchto vlastností jsou poté specifičtější metody, které pomáhají zúžit pole zájmu do menšího okruhu, který je potenciálně schopný odhalit původce zkoumaného textu. Některé specifické metody, které jsou v dnešní době mezi trendy, jsou ku příkladu četnost n-gramů, POS taggování, neuronové sítě, vzdálenosti (Manhattan, Kosinova, Euklidovská, …), Bag of Words (BoW) model (Harris, 1954), Term Frequency – Inverse Document Frequency (tf-idf metoda hodnocení relevance), hierarchické shlukování, způsob použití jazyka, jazyková kompozice apod. (Pennebaker a King, 1999; Luhn, 1957; Jones, 1972; Sindhu, a další, 2017)

Erkan Uyar a další některé ze zmíněných metod využili při své studii určování autorství v roce 2007, kde bylo zjištěno, že největší úspěch mívají kombinace více metod, ne pouze jedné, konkrétně autoři studie uvedli, že z metod, které používali, byla nejvhodnější kombinace vektorové klasifikace, BoW modelu a metody zabývající se funkčními slovy. (Uyar, a další, 2007) O několik let později, v roce 2017, Sindhu a další se také věnovali metodám používaným při určování autorství, přesněji trendům v rámci využívaných metod v dané době. V rámci této studie zjistili, že dobovým trendem bylo použití metody nejběžnějších n-gramů a metody podpůrných vektorů (SVM). (Sindhu, a další, 2017)

Jednou z dalších, moderních metod analýzy textu, které se objevují při zvažování průběhu procesu určování autorství, je analýza zahrnující hapax legomena, tzn. slova s frekvencí v textu právě 1. Tato metoda nabývá v dnešní době navyšující se popularity v této oblasti, jelikož bylo během posledních několika let výzkumníky zjištěno, že dokáže prozradit autora na různých úrovních textu. Již ve své studii z roku 1996 hapax legomena využili H. Baayen s H. van Halterenem a F. Tweedie, avšak do výběru byly zařazeny i kombinace dalších,

odlišných metod. (Baayen, a další, 1996) V této studii primárně porovnávali metody založené na užití slov s metodami založenými na syntaxi. Zjistili tak, že existuje i spolehlivější způsob při určování autorství než pouze užití slov v kontextu – tímto zjištěním byly vzorce syntaxe daného autora, kdy v tomto případě napomáhají k určení autorství funkční slova.

Mimo to byl v této studii také zahrnut pohled na odlišnost textů i na základě žánru, který sám o sobě, jak bylo zjištěno, dokáže autora prozradit. (Baayen, a další, 1996) Důvodem, proč metoda zahrnující funkční slova nebude v této práci využita pro posuzování rozdílů mezi autory, je to, co již zmínili Baayen a další ve své studii – pro takovou analýzu neexistuje dostatek dat, které by měly zaznamenané syntaktické struktury. (Baayen, a další, 1996) V případě této práce se u všech textů jedná o beletrii, ve které nejsou extrémní odchylky především v syntaktické rovině (slaběji i na ostatních jazykových rovinách), jako tomu naopak je například u dvojice lyrická poezie vs. sci-fi román.

Následný bod studie od H. Baayena, H. van Halterena a F. Tweedie se zabýval dalším pohledem na výsledky provedené analýzy a na jejich implikace: „Naše analýzy dále ukázaly, že metody založené na frekvencích velkého počtu typů, ať už typů s vysokou frekvencí, nebo typů s nejnižší frekvencí, jsou podstatně přesnější než metody založené na souhrnné statistice bohatosti slovní zásoby.“ (Baayen, a další, 1996) [vlastní překlad] K tomuto je dále doplněn také fakt, že ač jsou tyto metody spolehlivé, stále mají větší chybovost, než jakou by měla jejich dřívější metoda založená na syntaxi (ke které však z praktických důvodů není lehký přístup). (Baayen, a další, 1996)

Poněkud novější studie byla provedena Danem Faltýnkem a Vladimírem Matlachem v roce 2020. Zde bylo zkoumané především shlukování textů autorů na základě hapax legomenon a následný výzkum prokázal, že tyto metody dokáží určovat autorství s vysokou přesností. (Faltýnek a Matlach, 2020) V této studii byly brány v potaz předpoklady vlastností funkčních slov, která se využívají pro určování autorství již delší dobu. Funkční slova jsou gramaticky podstatná, frekventovaná a přítomna v každém textu, nezávislá na tématu textu a odhalující lingvistické znaky autora, které jím nejsou vědomě zamýšlené. (Binongo, 2003) Bylo však zjištěno, že tyto vlastnosti mají i hapax legomena, čímž se stávají v pořadí dalším vhodným kandidátem pro analýzy za účelem určení autorství. (Baayen, a další, 1996; Faltýnek a Matlach, 2020) Tímto směrem se zmíněná studie také ubírá a testuje tuto metodu na 14 autorech. Výsledky provedené analýzy v této studii prokazují, že použití metody využívající hapax legomena je spolehlivým nástrojem při určování autorství. (Faltýnek a Matlach, 2020)

Studium hapax legomen se tak prokázalo jako efektivní a přinášející hodnotné výsledky. Proto také Dan Faltýnek pokračoval v jejich studiu a v roce 2022 navázal zkoumáním distribuce hapax legomen napříč texty a toho, jestli jsou součástí koheze textu. Autoři studie zjistili, že mimo jiné prostředky textové koheze i hapax legomena na ní mají svůj neodmyslitelný podíl. Podkladem pro tento závěr bylo zjištění, že se hapax legomena vyskytují specificky a částečně i pravidelně napříč celými texty. (Faltýnek a Kučera, 2022) Tato studie tedy naznačuje, že i další metoda zahrnující hapax legomena dokáže být vhodným prostředkem pro analýzu určování autorství, a mimo to zjišťuje, že hapax legomena mají vliv i na samotnou textovou kohezi. Na základě těchto zjištění zde uvádějí pojem „superhapax“, což je slovo v textu, který je rozdělen na určité části, které se v každé takové části vyskytuje maximálně jedenkrát, tedy je pro každou část hapaxem legomenon. (Faltýnek a Kučera, 2022)

Další metoda, která by se dala zařadit mezi populárně využívanou, je metoda sentimentové analýzy textu prováděná za pomocí nástroje zvaného LIWC (Linguistic Inquiry and Word Count). LIWC je aplikace, která se zaměřuje na analýzu emocionálních, kognitivních a strukturálních složek v psaném či z řeči transkribovaném jazyce. Její první verze byla pouze součástí jiné studie, zatímco v dnešní době je to již několikátá verze plnohodnotné samostatné aplikace. (Pennebaker, a další, 2015) Existuje několik verzí (LIWC2001, LIWC2007, LIWC2022), avšak tato práce bude využívat verzi LIWC2015 vzhledem k dostupnosti této verze. LIWC2015 je oproti svým předchozím verzím rozšířen o větší obsah slovníku, který je pro jeho funkčnost klíčový.

LIWC využívá interního výchozího slovníku, ve kterém jsou uloženy tzv. „dictionary words“ (angl. slovníková slova), zatímco slova, která se do programu nahrávají uživatelem jsou nazývána „target words“ (angl. cílová slova). (Pennebaker, a další, 2015) Program pracuje na základě přiřazování cílových slov k těm slovníkovým. Prvně jsou cílová slova načtena do LIWCu, který následně prohledá svá interní slovníková slova, a pokud nalezne shodu, tato slova k sobě přiřadí a „zvýší se příslušná (či příslušné) stupnice slovní kategorie pro toto slovo. Při zpracování cílového textového souboru se rovněž navyšují počty různých prvků strukturní skladby (například počet slov a interpunkce ve větě).“ (Pennebaker, a další, 2015) [vlastní překlad]

Výstupem této aplikace je soubor obsahující 94 kategorií, mezi které patří názvy souborů, *WC* (*Word Count*, angl. počet slov), dále 4 základní souhrnné jazykové proměnné (*Analytic*, tzn. analytičnost, *Clout*, tzn. agresivita či vůdčí schopnosti, *Authentic*, tzn. autentičnost, *Tone*, tzn. emoční tón), 3 obecně popisné kategorie (například *WPS*, tzn. počet slov na větu), 21 standardních jazykových dimenzí (například *ppron*, tzn. procento osobních zájmen v textu),

41 psychologicky založených slovních kategorií (například *affect*, tzn. citově založené procesy), 6 kategorií osobních zájmů (například *work*, tzn. pracovní záležitosti), 5 znaků použití hovorového jazyka (například *filler*, tzn. výplňková slova) a 12 kategorií interpunkce (například *Period*, tzn. počet teček). (Pennebaker, a další, 2015) Všechny tyto kategorie a jejich podkategorie jsou vypsány v tabulce viz Pennebaker, a další, 2015 (článek s názvem *The Development and Psychometric Properties of LIWC2015*).

Slovník LIWC2015 obsahuje cca 6400 slov, které jsou zařazeny do různých kategorií a jejich podkategorií. Jedno slovo může patřit do více kategorií zároveň – například do pozitivních emocí, do pracovních záležitostí, do běžných verb a do analytičnosti. Takové slovo pak zvýší hodnoty své analýzy právě na stupnicích těchto kategorií. Mimo to zde existují i hierarchické struktury – kategorie se svými podkategoriemi, tzn. že například pozitivní emoce budou spadat pod nadřazenou kategorii citově založených procesů. (Pennebaker, a další, 2015) Dále se tento slovník zakládá na kmenech slov, tedy například slovo kalhot*, kde „*“ označuje 0 či více libovolných následujících znaků na dané pozici, bude brát v potaz nejen tvar „kalhoty“, ale i ku příkladu slova „kalhotům“, „kalhotách“ apod., které se se posloupností znaků před hvězdičkou shodují. (Pennebaker, a další, 2015)

Metodu analýzy sentimentu pravidelně využívá například James W. Pennebaker. Prvním takovým příkladem využití sentimentové analýzy je studie od něj a od Marthy E. Francis provedená v roce 1996. V této studii zmíněný pár autorů analyzoval dopady používání různě sentimentově zabarvených slov přímo na zdravotní stav studentů, kteří si psali deníky analyzované touto metodou. Bylo zde zjištěno, že užívání jazyka je přímo spojeno se zdravím studentů, což se projevilo například v počtu návštěv zdravotního centra na univerzitním kampusu v souvislosti s různě sentimentově zabarvenými slovy v denících studentů. (Pennebaker a Francis, 1996)

K tomuto páru autorů se v další, obdobné studii o rok později, v roce 1997, přidala autorka Tracy J. Mayne. Tento výzkum se věnoval traumatickým zážitkům, užívání slov s různým sentimentem a na tom závislým dopadům na zpracovávání traumat. Bylo zde zjištěno, že kognitivní zpracování pomocí využití psaného jazyka, ve kterém se vyskytují právě slova, která implikují vysokou míru kognitivních procesů (slova příčin a vhledu), přímo korelují s lepším fyzickým zdravotním stavem účastníků studie, kteří zažili traumatickou ztrátu svých partnerů. Výsledky tak indikují, že psaný jazyk může být důležitým nástrojem pro zpracování podobných traumat. (Pennebaker, a další, 1997)

Předchozí výzkumy sloužily jako podklady pro další významný výzkum provedený v roce 1999 Jamesem W. Pennebakerem a Laurou A. King, který se týkal užívání jazyka a toho,

jak použití různě sentimentově zabarvených slov ovlivňuje člověka z pohledu psychologie a sociologie. Tato studie po několika zkoumáních nabízí výsledky vypovídající o tom, že užití jazyka dokáže být spolehlivým rozlišujícím prvkem individuality. (Pennebaker a King, 1999) Implikacemi tohoto výzkumu tak autoři přispěli k rozšíření poznatků ohledně schopností sentimentové analýzy a k úpravě a zdokonalení aplikace LIWC, která tuto analýzu umožňuje automatizovat.

Nepochybně zajímavý výzkum podnikl James W. Pennebaker také později s Matthew L. Newmanem, Diane S. Berry a Jane M. Richards, tentokrát autoři přidávají téma, které odhaluje perspektivu zajímající se o detekci lhaní na základě stylu autora. (Pennebaker, a další, 2003) Výsledky této studie se shodují s předpokladem, že lidé, kteří lžou, a lidé, kteří mluví pravdu, jsou kvalitativně odlišitelní. (Undeutsch, 1967, cit. podle Pennebaker, a další, 2003, s. 674) „Naše data naznačují, že lháři mají tendenci vyprávět příběhy, které jsou méně komplexní, méně relevantní pro ně samotné a více charakterizované negativitou.“ (Pennebaker, a další, 2003) [vlastní překlad]

Stejný nástroj, jako v těchto případech využíval James W. Pennebaker a další, jen novější verzi, bude využívat také tato práce, avšak s poněkud odlišným cílem. James W. Pennebaker a další těmito výzkumy prokázali funkčnost a výhodnost použití nástroje sentimentové analýzy a souvislost sentimentu s fyzickým a psychickým stavem člověka. I když se tito lidé nevěnovali přímo určování autorství, dodali důkazy o existenci zmíněné souvislosti, která je předpokladem pro postupy této práce. LIWC má tedy v tomto případě se svou sentimentovou analýzou schopnost napomoci k viditelnosti hodnot popisujících tento vztah.

Vzhledem ke studiím vykazujícím funkčnost metody sentimentové analýzy v nástroji LIWC v rámci souvislosti s fyzickým a psychickým stavem lidského organismu (viz Pennebaker a Francis, 1996; Pennebaker a King, 1999; Pennebaker, a další, 1997) a metody hapax legomen a superhapaxů v rámci určování autorství (Faltýnek a Matlach, 2020; Faltýnek a Kučera, 2022) byly tyto metody pro účely této práce zkombinovány pro zjištění výsledků vypovídajících o vlivu sentimentu při určování autorství. Přesněji metodika zahrnuje věnování se sentimentové analýze na okolích superhapaxů a následné vzájemné porovnání hodnot vlastností sentimentové analýzy, což má za cíl odhalit, zdali jsou jejich rozdíly signifikantní a zdali tak samotný sentiment či jeho část dokáže do určité (a jaké) míry odhalit autora, popřípadě vyloučit autora z výběru možných původců určitého textu.

# METODOLOGIE A VÝSLEDKY

## 1. SELEKCE JAZYKOVÉHO MATERIÁLU

Pro tuto práci byly zvoleny knihy od následujících českých autorů: Jaroslav Foglar (dále už jen Foglar), Lenka Lanczová (dále už jen Lanczová), Miloslav Švandrlík (dále už jen Švandrlík), Ludmila Vaňková, Jaroslav Velinský. Tento výběr je odůvodněn dostatečným množstvím dostupného jazykového materiálu z jejich knih pro požadované množství potřebné k realizaci této práce.

Pro účely této práce bylo potřeba mít ke každému z autorů 25 různých textů. Pro autory, kteří měli více než 25 dostupných děl, jmenovitě to byla Lanczová a Velinský, bylo nutné vyselektovat pouze potřebné množství. Tento postup lze provést například pomocí aplikace QUITA Online. Tento nástroj slouží ke kvantitativní analýze textových dat, především pomocí indexů a Bag of Words. Do nástroje lze zadat hned několik nastavení pro uživatelem požadovanou analýzu, na matematické výpočty používá Python a R. (Katedra obecné lingvistiky, 2016)

Za použití aplikace QUITA Online byl tedy nadále pro každý z textů získán počet tokenů – pomocí tokenizace textů regulérním výrazem „\W+“. Na základě dat získaných tímto postupem byl určen průměr a následně byly rovnoměrně odstraněny texty od tohoto průměru nejvzdálenější a ponechány pouze texty jemu nejbližší. Výsledkem tohoto kroku bylo 25 různých textů v českém jazyce pro každého z daných 5 autorů. Seznam vybraných děl je uveden jako *Příloha č. 1*.

# 2. EXTRAKCE SUPERHAPAXŮ A JEJICH OKOLÍ, PŘEKLAD

Dalším krokem této práce byla extrakce zájmových slov, se kterými se bude dále pracovat. Za tímto účelem byl využit skript, pro jehož následující detailnější popis zde budou vysvětleny určité pojmy, které se budou v popisu vyskytovat:

- pojem „token“ a pojem „typ“ (angl. *type*): filozofie na tyto pojmy nahlíží následovně: „Typy jsou abstraktní formy, jejichž jednotlivé předměty jsou konkrétními tokeny.“ (Baggini a Fosl, 2010) [vlastní překlad] Tuto definici je možné aplikovat i na slova z pohledu lingvistiky, v tom případě typ značí danou jednotku samu o sobě, abstraktně, zatímco token jsou pak jednotlivé konkrétní případy této jednotky realizované v daném textu;
- pojem „Bag of Words“ (též zkráceně BoW): Bag of Words je vektorizace buďto textu, která bude využita právě v této práci, či také jiných objektů (v takovém případě se používá i termín Bag of Features). V případě textové klasifikace BoW představuje číselnou matici, ve které jsou řádky texty, sloupce jsou veškerá ze všech textů známá slova a buňky jsou počty výskytů tohoto slova v daném textu. Nehledí se zde na pořadí slov v původním textu (Qader, a další, 2019);
- pojem „hapax legomenon“: hapax legomena jsou slova v daném textu s frekvencí rovnající se 1;
- pojem „superhapax“: superhapaxem se rozumí slovo v určitém textu, který je rozdělen na stejně velké části. V každé takové části se slovo, které je superhapaxem, vyskytuje maximálně jedenkrát, tedy je pro každou část hapaxem legomenon. (Faltýnek a Kučera, 2022)

V tomto kroku byl za účelem získání zájmových slov využit skript, software Havex, který vytvořil František Špaček (Špaček, 2022). Tento skript je určený k extrahování superhapaxů a jejich okolí. Byly zde zvoleny jednotky několika vět, konkrétně 1 věta před větou se superhapaxem a 2 věty po ní. Tímto způsobem je do analýzy zahrnut kontext, ve kterém se superhapaxy nacházejí, pro větší perspektivu a komplexnost. Zmíněný program funguje v následujících krocích:

1. Ve skriptu se nastaví, na jak velké části (v počtech tokenů) se má text rozdělit a jak velké okolí superhapaxu se má exportovat. Zde byla nastavena velikost části na 6000 slov a pro okolí byla zahrnuta 1 věta před superhapaxovou větou (věta obsahující superhapax) a 2 věty po ní.

2. Do skriptu se načte vstupní soubor a provede se předběžné zpracování (angl. tzv. *preprocessing*), což obnáší čištění přebytečných znaků pro nové řádky, přebytečné tabulátory apod.
3. Text se tokenizuje na slova (interpunkční znaky jsou také považované za tokeny).
4. Tokenizovaný text se převede na „metatokeny“. To znamená, že každý token, který není interpunkčním znakem, bude uložen společně s informací o tokenu jemu následujícím. Takto nebudou interpunkční znaky ovlivňovat další výpočty, zároveň ale bude možné získat informace o pozicích znaků, které ukončují větu.
5. Najde se ideální počet částí, na které se má text rozdělit tak, aby byl počet slov v jedné části co nejblíže zadané velikosti (6000 slov). Postupně se prochází navrhovaná čísla, která indikují, na kolik částí se má text rozdělit. Každým takovým číslem se délka textu vydělí a vznikne tak „mezivýsledek“. Pro něj se vypočítá absolutní hodnota jeho rozdílu. Čím menší rozdíl, tím blíže jsme cílové velikosti části. Toto se opakuje, dokud se nenajde nejlepší dělitel.
6. Text se rozdělí na počet částí o velikosti vypočítané v kroku 5.
7. V každé části se najde pozice prvního znaku ukončujícího větu, jelikož se může stát, že při dělení textu se některé věty rozdělily na dvě části. V tomto kroku se zjistí, které části byly takto rozděleny a kde se nachází jejich konec v následující části.
8. Podle výsledků z kroku 7. se změní velikost části tak, aby každá část byla celou větou se svým odpovídajícím začátkem i koncem.
9. Vytvoří se globální slovník typů z celého textu.
10. Vytvoří se tzv. „Bag of Words“ pro každou část.
11. Odstraní se všechna slova, která mají v kterékoliv z částí výskyt větší než 1, tedy nejsou hapax legomenon pro danou část.
12. Pro každé zbylé slovo se vytvoří suma (součet) výskytů.
13. Vytvoří se seznam superhapaxů seřazený od nejčastějších po nejméně časté.
14. Vstupní text se tokenizuje na věty.
15. Získají se okolí superhapaxových vět v definované velikosti.
    15.1. Pro každý superhapax v seznamu superhapaxů se prochází všechny věty.
    15.2. Pokud věta obsahuje superhapax, uloží se společné s okolím v definované velikosti.

15.3. Až se projdou superhapaxy z jedné frekvenční skupiny (například všechny hapax legomena s výskytem 5), zkontroluje se, jak moc se velikost výstupu liší od 10 % velikosti originálního textu.

15.4. 15.4. Opakuje se od kroku 15.1. a ukládají se okolí do té doby, dokud se velikost výstupu nedostane co nejblíže 10 % velikosti originálního textu. Kvůli konzistenci a objektivitě výsledků se hapax legomena prochází vždy po celých frekvenčních skupinách a kontrola velikosti výstupu proběhne vždy po uložení okolí ke všem superhapaxům z dané frekvenční skupiny.

15.5. Exportuje se výsledek (soubor), který byl svou velikostí nejblíže 10 % velikosti originálního textu.

Pro další zpracování bylo potřeba nově získané soubory (které se vytvořily pro každý ze souborů obsahující díla autorů) zahrnující superhapaxy a jejich okolí přeložit do angličtiny, jelikož aplikace LIWC2015 nepodporuje češtinu. K tomuto zde byl využit program německé společnosti DeepL, která se zabývá umělou inteligencí a pomocí ní vytvořila nástroj strojového překladu založený na zlepšujících se výsledcích matematických a metodologických prostředků neuronových sítí. (DeepL, 2023) Tato práce využívá překladač dostupný na webové stránce https://www.deepl.com/translator. Tento překladač umožňuje přijatelný překlad při vkládání jazykového materiálu i s rozsáhlejšími kontexty, ne pouze při potřebě převodu pouze jednotlivých frází.

Avšak jakožto strojový překlad má i DeepL určitou chybovost. Na vzorku cca 400 slov v souborech každého z autorů jsem manuálně prošla originální i jemu ekvivalentní přeložený soubor a pomocí zaznačování chyb jsem vypočítala procentuální chybovost, která byla poté pro všechny tři autory zprůměrována. Chyby byly selektovány především v lexikální rovině vzhledem k tomu, že ta má na pozdější analýzu největší vliv (oproti například opakovaným předložkám). Tato chybovost vyšla 1,75 % (zaokrouhleně na dvě desetinná místa). V tomto kroku byly také odhaleny další chyby, které však byly způsobeny chybami v původních souborech, nejpravděpodobněji při převodu původních textů do jejich elektronické podoby pomocí nástroje OCR (Optical Character Recognition).

OCR je typ softwaru, který se snaží napodobit jazykové a kognitivní schopnosti lidského oka a mozku, přesněji schopnost rozpoznání písmen a slov. Lidské oko a mozek dokáží rozpoznat různé fonty v různé kvalitě tisku a převést je na slova, která dokáže mozek pochopit. OCR se snaží tyto schopnosti replikovat, i když stále s jistou chybovostí. (Holley, 2009) Právě toto může způsobit chyby podobné těm, které se vyskytují v původních textech pro tuto práci. Na stejném vzorku jsem tedy dále dopočítala také přidanou chybovost v cílových

(anglických) souborech způsobenou právě zmíněnými nedostatky. Když vezmeme v potaz i tyto výsledky, je nutno přičíst procentuální hodnotu 0,19 % (zaokrouhlenou opět na dvě desetinná místa). Celková chybovost překladů zadaných textů je tudíž přibližně 1,94 %.

# 3. LIWC ANALÝZA, ZÁJMOVÉ VLASTNOSTI

Pro analýzu v programu LIWC2015 bylo zapotřebí právě přeložených superhapaxů a jejich okolí. Tato analýza nám poskytne tabulku o 95 sloupcích, které zahrnují různé LIWC vlastnosti textu, a 125 řádcích, které zahrnují názvy využitých děl daných autorů. Získané hodnoty dále využijeme pro selekci menšího počtu vlastností pro analýzu určování autorství. Jako další postup jsem zde zvolila 14 vlastností jakožto zájmových na základě níže popsaných kritérií. Mezi ně byly zařazeny 4 základní, hlavní vlastnosti v programu LIWC2015 – vlastnosti *Analytic*, *Clout*, *Authentic* a *Tone*, zbytek vlastností je doplňkových. Obecně pro všechny vlastnosti platí, že jejich jediný dostupný kvalitativní popis dle toho, jak je uvádí výstup z nástroje LIWC2015, je jejich hierarchie a u některých z nich existuje několik příkladů ilustrativních slov, avšak hlubší kvalitativní porozumění významu jednotlivých vlastností není v manuálu k tomuto nástroji objasněné. (Pennebaker, a další, 2015) Z tohoto důvodu nebylo vhodné mezi zbylými vlastnostmi vybírat dle teoretického popisu.

V návaznosti na předchozí fakta byly ostatní zájmové vlastnosti pro analýzu vybrány dle jejich dostupné hierarchie. Vlastnosti byly selektovány tak, aby byly zahrnuty především ty, které jsou určitým způsobem nadřazené ostatním a částečně tak zahrnují i informace o těch, které jsou jim podřazené. Podřazené vlastnosti jsou s nadřazenými propojené a bylo by nadbytečné používat více detailnějších popisů, pokud dané informace už zahrnuje vlastnost nadřazená. Mimo to byly také vyloučeny ty vlastnosti, které by mohly být zavádějící – například vlastnost *Sixltr*, která uvádí počet slov s šesti a méně písmeny, či vlastnost *Dic*, která uvádí počet slovníkových slov (Pennebaker, a další, 2015). Vzhledem k překladu se může délka slov lišit, jelikož v angličtině například slovo „flowers“ bude mít písmen šest, zatímco český protějšek „květiny“ má písmen sedm, tudíž by do kategorie *Sixltr* na rozdíl od „flowers“ nespadal. Pro slova řadící se do kategorie *Dic* se při převodu do angličtiny může stát, že v českém prostředí by se o slovníkové slovo jednalo, avšak v angličtině již ne, vzhledem ke kontextu a úzům daných jazyků.

Finální vybrané zájmové vlastnosti pro další selekci a analýzu byly: *Analytic*, tzn. analytičnost, *Clout*, tzn agresivita či vůdčí schopnosti, *Authentic*, tzn. autentičnost, *Tone*, tzn. emoční tón, *funct*, tzn. počet funkčních slov, *affect*, tzn. citově založené procesy, *social*, tzn. sociálně založené procesy, *cogproc*, tzn. kognitivně založené procesy, *percept*, tzn. vjemově založené procesy, *bio*, tzn. biologicky založené procesy, *drives*, tzn. hnací síly za určitým cílem či z určitého důvodu, *focuspast*, tzn. orientace na minulost, *focuspresent*, tzn. orientace na přítomnost, *focusfuture*, tzn. orientace na budoucnost. Bližší vysvětlení

manuál nástroje LIWC2015 neposkytuje, tudíž další vztahy mezi jednotlivými vybranými vlastnostmi budou zkoumány dále v této práci kvantitativně právě z důvodu nedostatečného popisu kvalitativního.

Finální tabulka dat s vybranými zájmovými vlastnostmi z tohoto kroku obsahuje 126 řádků, přičemž první je popisný (textový), zatímco ostatní jsou číselné. Tato tabulka má dále 17 sloupců, v nichž jsou již uvedené vlastnosti, dále název díla, zkratka autora („Autor“) a počet slov („WC“) v daném souboru. V prvních dvou sloupcích se nacházejí data popisná, ostatní data ve sloupcích jsou striktně číselná. Alternativní verze této tabulky dále existuje pro účely analýzy v logistické regresi, kde bylo nutné vytvořit tabulky párů autorů, ve které se objevuje kategorická binární proměnná („0“ pro jednoho autora, „1“ pro autora druhého). V této verzi je poté zbytek dat opět číselných.

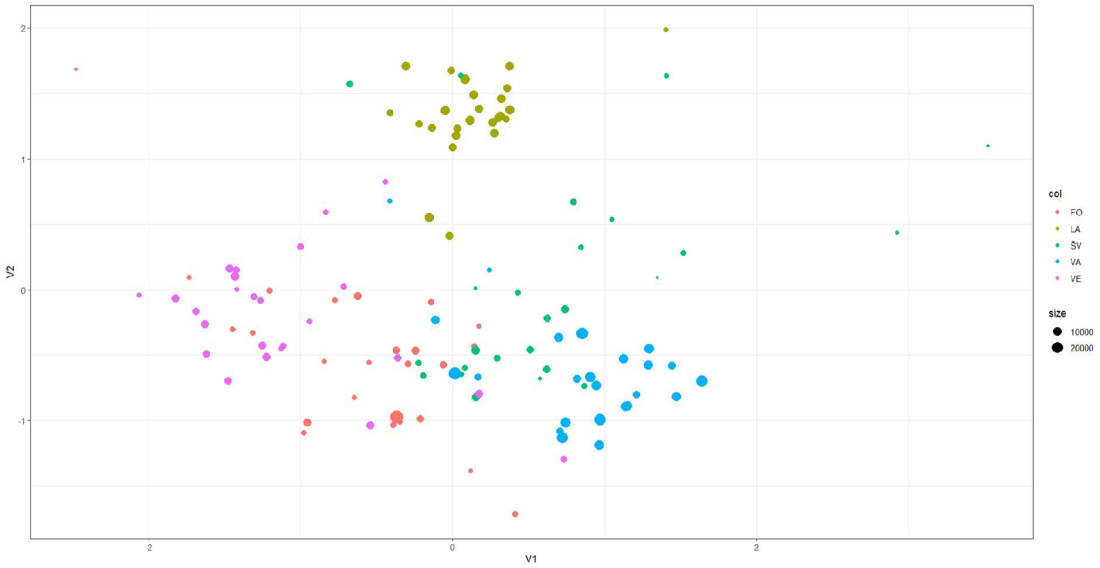
Tato část se dále také věnovala předběžnému zobrazení spojitostí mezi vlastnostmi a jejich rozřazení za pomoci programovacího jazyku R v programu RStudio[1]. Pro chtěné zobrazení byla využita funkce *ggplot()*, což je knihovna pro tvorbu grafů, která uživateli umožňuje upravit vizuální stránku grafu svým potřebám dle několika parametrů a specifik. (Wickham a Grolemund, 2017) Původní data, tj. zmíněná tabulka se všemi texty a všemi vlastnostmi textů kromě *WC* a „Autor“, byla redukována na 2 rozměry pomocí metody UMAP a následně zobrazena do grafu. Redukční technika UMAP funguje na principu konstrukce mnohorozměrné grafické reprezentace a následné optimalizace málorozměrného grafu tak, aby bylo vytvořeno strukturně co nejpodobnější zobrazení. (McInnes, a další, 2018)

Tímto způsobem bylo vyobrazeno rozložení dat v rovině. Jak lze vidět v *Grafu č. 1*, již zde se tvoří určité barevné shluky (clustery). Každá barva představuje jednoho z autorů. Velikost bodů je poté závislá na *WC*, které takto reprezentuje, i když jej metoda nebere v potaz při rozmisťování děl autorů jakožto bodů do grafu, respektive na osy $x$ a $y$. V legendě se vyskytuje „FO“ označující Foglara, „LA“ označující Lanczovou, „ŠV“ označující Švandrlíka, „VA“ označující Vaňkovou a „VE“ označující Velinského. Tyto zkratky a jejich významy platí pro označení autorů při analýzách v celé této práci.

Jak můžeme v *Grafu č. 1* vidět, shluky jednotlivých autorů se částečně překrývají a míchají, což lze přisoudit tomu, že určité své znaky má daný autor společné s autory jinými. I přes tyto překryvy se však v grafu nacházejí jasné, oddělitelné shluky, ať už plně viditelné jako například u Lanczové nahoře (barva khaki), tak i shluky, které se více překrývají a je nutné se na ně o něco více zaměřit, například u Švandrlíka (barva světle zelená), který vstupuje hned do několika

[1] https://posit.co/download/rstudio-desktop/

dalších shluků jiných autorů, avšak při bližším pohledu i on má svůj vlastní rozlišitelný shluk, který se táhne z prostředka dole mírně doleva nahoru. Mimo tyto dva shluky zde ještě vidíme shluk od Foglara (barva lososová), který se nachází dole uprostřed a prostupují jím zástupné prvky i několika dalších autorů, shluk od Vaňkové (barva světle modrá), který se nachází napravo dole a zasahuje až doprostřed, a shluk od Velinského (barva růžová), který se nachází vlevo dole a taktéž mírně zasahuje do středu.




*Graf č. 1: Graf vyexportovaný z RStudia, vytvořený funkcí ggplot() skrze techniku UMAP z matice dat pro vlastnosti textů od autorů, odebrané sloupce „Autor" a „WC"*

Pomocí *Grafu č. 1* tak bylo předběžně zjištěno, že některé z vlastností budou pravděpodobně souviset se schopností rozlišení jednotlivých autorů vzhledem k tomu, že i bez detailnějších kritérií pro určité rozdělení a rozložení hodnot zde nepochybně existuje určitá souvislost, která takto autory spojuje do vlastních skupin. Tyto skupiny jsou v *Grafu č. 1* viditelné pávě jako zmiňované shluky různých barev. Zajímavým zjištěním je také to, že se texty neshlukují podle jejich délky (*WC*), tj. že nejsou vedle sebe pouze body o stejných velikostech, ale míchají se ve vzájemné blízkosti velké i malé body reprezentující dané texty. Toto předběžně naznačuje, že vlastnosti LIWC v okolí superhapaxů nejsou závislé na délce.

# 4. SELEKCE NEJVÍCE RELEVANTNÍCH VLASTNOSTÍ

## 4.1 ANALÝZA VZTAHU DÉLKY TEXTU S OSTATNÍMI VLASTNOSTMI TEXTU

V této části je jakožto první krok nutné eliminovat možnost, že LIWC vlastnosti závisí na délce daných textů. K tomuto budou využity výsledky testu korelací. V případě, že by byla nalezena signifikantní korelace mezi počtem slov a hodnotami některé z daných vlastností, bylo by nutné texty zkrátit na stejnou délku. Pokud by se tato možnost neprokázala, texty mohou být ponechány v jejich původní délce. Kvůli násobnému provádění testů korelace bude použita Bonferroniho korekce.

Tvůrce Bonferroniho korekce byl C. E. Bonferroni. (Bonferroni, 1936) Bonferroniho korekce je jedna z metod, která se aplikuje v případech, kdy A) se na stejných datech provádí několik různých testů (Mundfrom, a další, 2006), či B) kdy se na několika datových setech provádí několik stejných testů. Metoda spočívá v tom, že ve chvíli, kdy jsou prováděny například tři testy, každý z nich bude mít určitou pravděpodobnost vyvolání chyby při testování, tzn. určitý práh signifikantnosti. Ve chvíli, kdy se tyto tři testy provedou zároveň na daných datech, je nutno vzít v potaz vyšší pravděpodobnost vyvolání chyby právě kvůli kombinaci těchto testů. Bonferroniho korekce se snaží této nepřesnosti vyhnout pomocí úpravy prahu signifikantnosti, a to způsobem $\alpha/k$, kdy jako „$\alpha$“ (hodnota alfa) je označen práh signifikantnosti a jako „$k$“ je označen počet testů v případě A), či počet datových setů v případě B). (Mundfrom, a další, 2006)

Pro tuto analýzu byly hodnoty *WC* (angl. počet slov) a hodnoty ostatních vlastností zadány do RStudia a testovány pomocí Spearmanovy korelace za účelem jejich vzájemného porovnání. Výsledkem bylo to, že žádná z vlastností nekoreluje s počtem slov (viz *Tabulka č. 1*). Při tomto výpočtu byla použita právě i Bonferroniho korekce. V tomto případě bylo zvažováno $k$ analyzovaných vlastností, což znamená, že původní hodnota alfa $\alpha_0 = 0{,}05$ byla vydělena právě počtem analyzovaných vlastností $k = 14$ (v tomto čísle není zahrnuta vlastnost *WC*). Z toho vyplývá, že nový práh pro signifikantnost je 0.05/14, tzn. přibližně 0,00357, z čehož vyplývá, že výsledná p-hodnota při této analýze by musela být $\leq 0{,}00357$, aby dosáhla signifikantnosti. Když se s touto premisou podíváme na sloupec „p-hodnota“ v *Tabulce č. 1*, můžeme si všimnout, že žádná z hodnot není $\leq 0{,}00357$, což nám poté ve sloupci „signifikantní ANO/NE“ také značí „NE“ – výsledná korelace není signifikantní.

| jméno vlastnosti | korelace | p-hodnota | signifikantní ANO/NE |
|---|---|---|---|
| Analytic | 0.11 | 0.2 | NE |
| Clout | -0.04 | 0.65 | NE |
| Authentic | 0.02 | 0.81 | NE |
| Tone | 0.12 | 0.17 | NE |
| funct | 0.06 | 0.51 | NE |
| affect | 0.12 | 0.18 | NE |
| social | 0.18 | 0.04 | NE |
| cogproc | -0.08 | 0.37 | NE |
| percept | -0.11 | 0.22 | NE |
| bio | -0.03 | 0.72 | NE |
| drives | 0.18 | 0.05 | NE |
| focuspast | -0.17 | 0.05 | NE |
| focuspresent | 0.07 | 0.47 | NE |
| focusfuture | 0.02 | 0.79 | NE |

*Tabulka č. 1: Test korelace počtu slov s ostatními vlastnostmi (Spearmanova korelace)*

Závěrem tohoto zjištění bylo, že žádná z vlastností LIWC není dle Spearmanova testu korelace signifikantně závislá na délce textu, respektive není dostatečná evidence pro to, že by některá byla. Díky tomuto zjištění je možno odvodit, že není třeba texty zkracovat či jinak brát v potaz hodnoty ostatních vlastností v relevanci k počtu slov, jelikož počet slov neindikuje signifikantní korelaci s žádnou z námi analyzovaných vlastností.

## 4.2 REDUKCE VLASTNOSTÍ PRO PRÁCI S LOGISTICKOU REGRESÍ

Během dalších kroků bude využita logistická regrese pro určování autorství, a to ve dvou případech. V prvním případě (viz kapitola 5.1) bude určování provedeno na základě každé vlastnosti zvlášť (jen vlastnost: *Authentic*, jen vlastnost: *Clout* atd.). V případě druhém (viz kapitola 5.2) bude určování provedeno pomocí všech vlastností zároveň (*Authentic* + *Clout* + ...). Před provedením obou kroků je však nutné se zaměřit na relevanci jednotlivých vlastností, jejich vzájemné odvoditelnosti, redundanci atp. Z toho důvodu bude tento krok dedikován redukci vlastností.

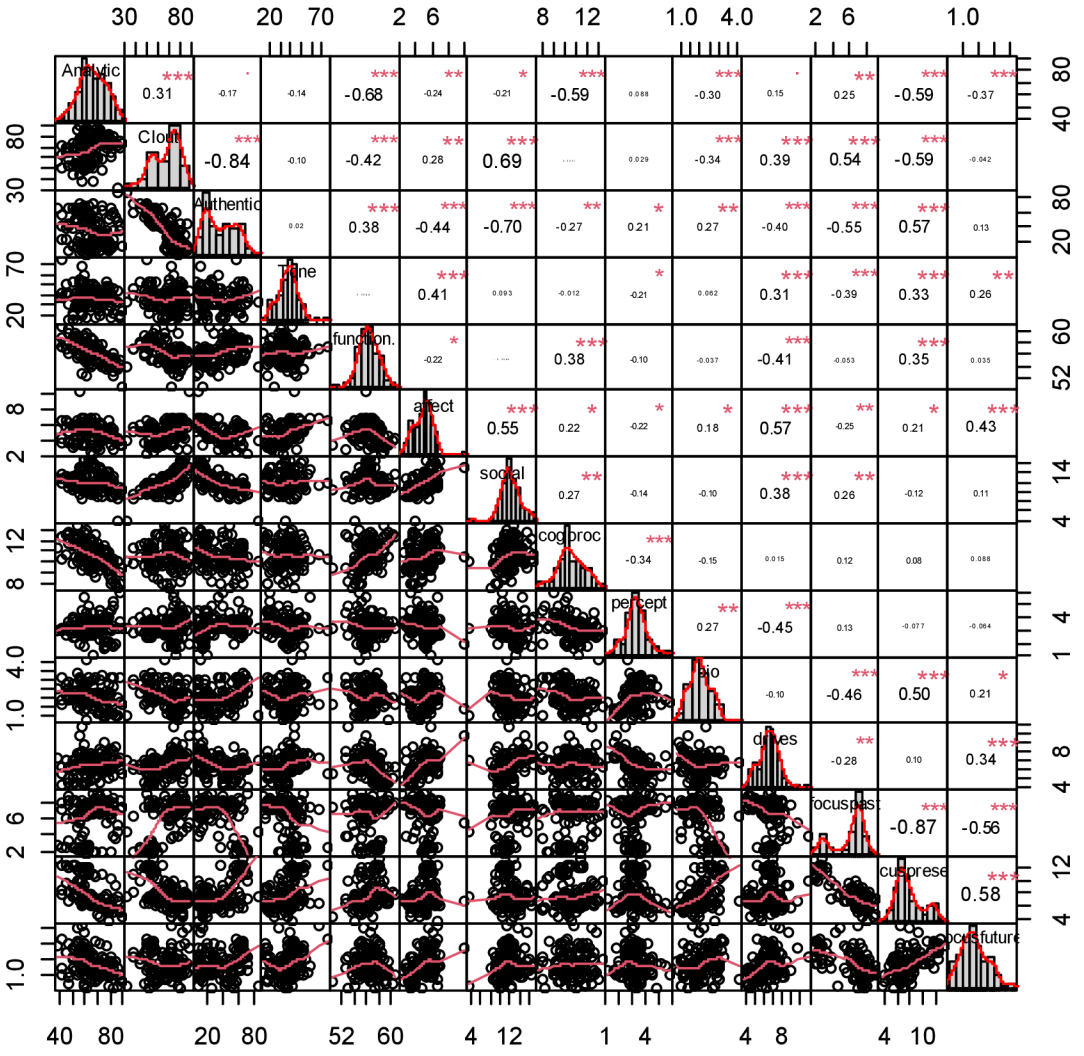
Pro první uvedený případ, tj. určování autorství u každé vlastnosti zvlášť, je nutná redukce, aby nebyly uváděny stejné výpočty pro vlastnosti, které spolu navzájem souvisí, a kde by tudíž výsledky jedné byly odvoditelné od výsledků druhé, popřípadě kde by byly výsledky shodné. Pro případ druhý, tj. určování autorství u všech vlastností zároveň, by došlo implicitně, bez našeho vědomí, k nahnutí a zvýšení důležitosti některé vlastnosti, zároveň zde bude nutné mít pouze 5 vlastností kvůli požadavkům logistické regrese. Oba případy jsou detailněji popsané v následujících kapitolách.

### 4.2.1 REDUKCE VLASTNOSTÍ S VYSOKOU KORELACÍ

Cílem tohoto kroku bylo zjistit, zdali spolu některé z vlastností signifikantně korelují. Pokud by tomu tak bylo, dá se předpokládat, že i výsledky určování autorství spolu signifikantně korelujících vlastností budou přibližně shodné, a tudíž není nutné ponechávat obě vlastnosti, ale postačí pouze jedna z nich. Takto se lze vyvarovat nadbytečné analýze více vlastností, než je ve skutečnosti nutné, čímž předejdeme přítomnosti duplikovaných výsledků, které by mohly buď pozitivně, nebo negativně ovlivnit hodnocení schopností určování autorství. Byla tedy provedena analýza pomocí Spearmanovy korelační matice v RStudiu, jejíž výsledky lze vidět na *Obrázku č. 1*.

V dolní části na *Obrázku č. 1* lze vidět bodové grafy. Tento typ grafu zaznamenává rozptyl dat ve shlucích bodů, kterými je poté proložena křivka pro lepší vyobrazení síly korelace. Toto proložení může směřovat nahoru, což napovídá o tom, že je zde vztah pozitivní korelace, nebo může směřovat dolů, kde tím pádem existuje vztah negativní korelace, či může být křivka přibližně v přímé horizontální poloze, z čehož vyplývá, že zde není vztah korelace, či je zde vztah korelace minimální. Bodový graf s sebou může nést i když ne na první pohled viditelnou informaci o určité souvislosti mezi oběma studovanými proměnnými,

jelikož je schopný ukázat, zdali spolu hodnoty na osách *x* a *y* souvisejí. Jakýkoliv vztah hodnot *x* a *y* je však nutné brát pouze jako korelační, což znamená, že se zde určitá souvislost vyskytuje, ale nemusí být nutně příčinná – nemusí nutně znamenat, že hodnota například z osy *x* předurčuje hodnotu na ose *y*, tzn je její příčinnou. (West a Bergstrom, 2018)




*Obrázek č. 1: Výsledky korelační matice (bodové grafy a hodnoty korelace)*

Při pohledu na *Obrázek č. 1* se v horní části nachází hodnoty samotné korelace a k nim připsané hvězdičky. Číselné hodnoty v mínusu naznačují korelaci negativní, zatímco pozitivní hodnoty indikují korelaci pozitivní. V rámci zaznačených hvězdiček platí, že čím vyšší počet hvězdiček, tím signifikantnější je p-hodnota, přičemž hodnota bez hvězdičky odpovídá p-hodnotě > 0,05, „*“ odpovídá p-hodnotě ≤ 0,05, „**“ odpovídají p-hodnotě ≤ 0,01 a „***“ odpovídají p-hodnotě ≤ 0,001. Když se podíváme na ku příkladu vlastnost *Clout* s vlastností

*Authentic*, v dolní části je vyobrazen bodový graf s proloženou křivkou, která se orientuje prudce směrem dolů, což naznačuje vztah výrazné negativní korelace. V horní části se nám tento předpoklad poté potvrzuje i číselně korelace -0,84 a také za pomoci hvězdiček, které se zde nacházejí tři, tudíž můžeme považovat toto zobrazení v kombinaci s číselným vyjádřením za indikaci signifikantního vztahu korelace.

Naopak například u vlastnosti *social* v kombinaci s vlastností *percept* zde můžeme pozorovat, že v dolní části bodový graf s křivkou proloženou v horizontální linii a v dolní části korelace -0,14 bez přítomnosti hvězdiček společně vykazují to, že tyto dvě vlastnosti spolu předpokládaně nekorelují, nebo korelují naprosto minimálně. Podobně tak lze na *Obrázku č. 1* pozorovat síly korelací jakýchkoliv dvou vlastností, a to jak pomocí bodového grafu, tak i pomocí číselné hodnoty a hvězdiček.

V *Tabulce č. 2* je dále uvedena matice signifikantní a silné korelace dle Evansovy síly korelace. Pro tyto hodnoty jsem zvolila hranici korelace podle Evanse „silnou" (angl. *strong*), což odpovídá hodnotě 0,60 (viz *Obrázek č. 2*) (Evans, 1996, cit. podle Sakamat, a další, 2018). Pokud se hodnota rovnala 0,60 nebo byla vyšší, byla považována právě za silnou. V *Tabulce č. 2* lze vidět binární výsledky 0/1, kdy „0" označuje výsledek, že pro daný případ korelace není takto silná, zatímco „1" označuje výsledek, že korelace je při daném korelačním prahu silná.

| | Analytic | Clout | Authentic | Tone | function | affect | social | cogproc | percept | bio | drives | focuspast | focuspresent | focusfuture |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Analytic | x | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Clout | 0 | x | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Authentic | 0 | 1 | x | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Tone | 0 | 0 | 0 | x | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| funct | 1 | 0 | 0 | 0 | x | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| affect | 0 | 0 | 0 | 0 | 0 | x | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| social | 0 | 1 | 1 | 0 | 0 | 0 | x | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| cogproc | 0 | 0 | 0 | 0 | 0 | 0 | 0 | x | 0 | 0 | 0 | 0 | 0 | 0 |
| percept | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | x | 0 | 0 | 0 | 0 | 0 |
| bio | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | x | 0 | 0 | 0 | 0 |
| drives | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | x | 0 | 0 | 0 |
| focuspast | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | x | 1 | 0 |
| focuspresent | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | x | 0 |
| focusfuture | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | x |

*Tabulka č. 2: Výsledky matice signifikantní a silné korelace dle Evansovy síly korelace (s korelačním prahem 0,60)*

| Correlation Value | Description |
|---|---|
| 0.00-0.19 | Very Weak |
| 0.20-0.39 | Weak |
| 0.40-0.59 | Moderate |
| 0.60-0.79 | Strong |
| 0.80-1.00 | Very Strong |

*Obrázek č. 2: Evansova síla korelace (Evans, 1996, cit. podle Sakamat, a další, 2018)*

V těch případech, kdy výsledek odpovídal „1“, to znamenalo, že daný pár vlastností spolu koreluje natolik silně, že přítomnost obou při dalších analýzách je redundantní, tedy že lze předpokládat, že výsledek jedné z vlastností dokáže potenciálně predikovat výsledek vlastnosti druhé. Z tohoto důvodu poté došlo k redukci právě těch vlastností, které zde dosáhly výsledku „1“, a ponechání pouze vlastností s výsledkem „0“. Tzn. došlo k vyloučení vlastností *Authentic*, *funct*, *social* a *focuspresent* a ponechání vlastností *Analytic*, *Clout*, *Tone*, *affect*, *cogproc*, *percept*, *bio*, *drives*, *focuspast* a *focusfuture*. Díky tomuto kroku tak byl získán zredukovaný počet vlastností, a to přesněji již pouze na množství 10.

### 4.2.2 REDUKCE VLASTNOSTÍ POMOCÍ METODY HLAVNÍCH KOMPONENT

Po analýze vztahu délky textu s ostatními vlastnostmi a zredukování počtu těchto vlastností pomocí korelační matice bylo nutné věnovat se přípravě na další kroky. Dále v této práci bude využita logistická regrese, ve které bude probíhat také analýza několika vlastností zároveň. Logistická regrese má omezení, které spočívá v doporučeném maximu vlastností (proměnných) – 1 proměnná pro logistickou regresi požaduje minimálně 10 pozorování. Při analýze všech vlastností samostatně je toto pravidlo pro daná data bezproblémové, avšak při analýze všech vlastností zároveň je nutno ho brát v potaz. Způsobem vypořádání se s tímto omezením je navazující analýza za účelem snížení počtu vlastností.

K dispozici zde bylo 25 textů pro každého z autorů, tudíž když se bude práce dále (v kapitole 5.2) zabývat porovnáním dvou autorů zároveň, dohromady pro ně bude dostupných 50 textů, tj. pozorování. Z tohoto předpokladu vychází, že pro další zpracování dat je nutná redukce vlastností, tj. proměnných, na počet 5 (50/10 = 5). Vzhledem k tomu, že vlastnosti LIWC2015 nejsou dokonale popsané a stále i z předchozích redukcí jich zbylo příliš mnoho na to, aby je bylo možné takto společně analyzovat skrze logistickou regresi, bylo nutné vytvořit

další filtr, který se bude řídit podle jejich informačního přínosu. K tomuto účelu jsem zvolila metodu PCA (viz dále), která byla aplikována na data opět za pomoci RStudia.

Před aplikací metody PCA bylo nutno do kódu přidat škálování. Každá analýza dat vychází z určitých vztahů mezi danými objekty, které mohou mít různý charakter, různou formu, různé jednotky a různé stupnice, vždy záleží na tom, o jaký typ objektu se jedná. Tyto vztahy mohou „vyjadřovat jejich podobnost či nepodobnost, shodu či neshodu, splynutí určitých vlastností či zaměnitelnost, resp. stupeň podobnosti, shody, zaměnitelnosti atd.“ (Řehák a Loučková, 1983) Běžně má tedy škálování za úkol vytvořit stupnice odvozené od vztahů mezi danými objekty a najít pomocí nich vhodné číselné hodnoty, které tyto vztahy budou reprezentovat. (Řehák a Loučková, 1983) Škálování se provádí právě v těch situacích, kdy by analýza byla zkreslená různými jednotkami daných objektů. Ve chvíli, kdy všechny objekty již mají stejné měřítko vůči sobě navzájem, škálování není nutné.

V případě následně prováděné analýzy však jednotky byly různé a bylo podstatné, aby se tento rozdíl neprojektoval do následného tvoření komponent metody PCA, jelikož by odlišnost jednotek vlastností LIWC mohly zapříčinit nesrovnalosti, až dokonce přímo chyby při dalších krocích analýzy. Škálování bylo provedeno v RStudiu pomocí funkce *scale()*. Poté byla aplikována metoda hlavních komponent (angl. *Principal Component Analysis* neboli PCA), přičemž ta byla provedena se všemi původními analyzovanými vlastnostmi v počtu 14. Metoda hlavních komponent se využívá v případech, kdy při analýze dat existuje velké množství zkoumaných proměnných a je nutné tyto proměnné zjednodušit pro jejich snadnější analýzu. Tato metoda tomu tak napomůže vytvořením umělých proměnných. (Hebák, a další, 2007)

Základy původní metody hlavních komponent uvedl a vysvětlil již K. Pearson v roce 1901. Metoda PCA se v zadané číselné matici (tabulce) pokouší nalézt „skryté (umělé, neměřitelné, latentní) veličiny, označované za hlavní komponenty“ (Hebák, a další, 2007), ty poté vysvětlují variabilitu a závislost daných proměnných. Metoda PCA tak provede analýzu lineárních vztahů původních proměnných pomocí vlastních vektorů, načež vytvoří komponenty, které jsou vzájemně nezávislé lineární kombinace právě oněch původních proměnných.

Petr Hebák a další v knize Vícerozměrné statistické metody (3) uvedli: „…podaří-li se původní proměnné vyjádřit pomocí menšího počtu nových proměnných bez větší ztráty informace, získá se úspornější popis původního systému proměnných, což je užitečné pro interpretaci i pro využití v jiných statistických metodách.“ (Hebák, a další, 2007) Metoda PCA se zaměřuje především na variabilitu původních zkoumaných proměnných a využívá k tomu pozorování jejich rozptylů. Bohužel u této metody nelze plně zaručit

smysluplnost či vysvětlitelnost nových komponent, ale i přesto se metoda již mnohokrát v minulosti prokázala jako velice účinná, pragmatická a spolehlivá. (Hebák, a další, 2007) Běžně metoda hlavních komponent postupuje podle určitých principů, které dodržuje, aby výsledné komponenty fungovaly, jak mají, a byly co nejvíce užitečné. Podobných principů je zpravidla 7.

První je princip, který říká, že nově vytvořené komponenty postupně ztrácejí na důležitosti, tzn. že první komponenta bude mít nejvyšší vypovídající hodnotu vzhledem k datům, zatímco poslední komponenta bude mít vypovídající hodnotu nejnižší. Druhý je princip, který se zakládá na tom, že nově vytvořené komponenty spolu nekorelují, jsou na sobě tudíž nezávislé. Třetím principem aplikovaným při této metodě je princip první komponenty jakožto nejdůležitější – první komponenta by měla vysvětlovat co nejvíce z variability původních dat. Na tento princip navazuje princip čtvrtý, který se řídí tím, že každá další komponenta vysvětluje postupně zbytek variability, což znamená, že poslední komponenta vysvětluje již jen pouhé minimum, které metoda nebyla schopna vysvětlit jiným způsobem. Jako pátý princip, který se aplikuje při použití metody hlavních komponent, je snížení počtu proměnných na adekvátní pro danou analýzu – tzn. hledání správného rozměru dat pro potřebné zkoumání. (Hebák, a další, 2007)

Poslední dva principy ty předchozí doplňují a dokreslují systém tvoření komponent. Šestým principem je tedy tzv. dostatečnost vysvětlení, kdy uživatel této metody by měl mít určitá kritéria pro to, kolik nových komponent by bylo vhodných pro danou analýzu a jakou část rozptylu původních dat by měly cca vysvětlit. Poslední, sedmý princip zdůrazňuje to, že ideálně by měla stačit pouze první komponenta, která by vysvětlovala většinovou variabilitu původních dat, avšak tato situace se naskytne pouze zřídkakdy. Stále to však znamená, že je vhodné mít co nejméně komponent, reálným číslem jsou zde cca tři, přičemž počet nad pět až šest už není příliš efektivní či adekvátní vzhledem k povaze této metody, avšak vždy samozřejmě záleží na okolnostech – na rozsahu původních dat, na počtu původních proměnných, na účelu analýzy apod. (Hebák, a další, 2007)

Metoda hlavních komponent v této části práce tedy tímto způsobem z daných vlastností vytvořila tolik komponent, kolik bylo původně zadaných vlastností, což v tomto případě bylo 14. Vzhledem k tomu, že komponenty metody PCA jsou seřazené od nejvyššího rozptylu (angl. *variance*) po nejmenší, vyextrahovala jsem prvních 5 komponent, tj. 5 synteticky vyrobených vlastností z těch původních tak, aby byly na sobě nezávislé, a k nim následně zjistila, kolik informací celkově zahrnují podle tzv. kumulativní proporce („Cummulative Proportion“ viz *Tabulka č. 3*), která vychází z tzv. proporcí rozptylu („Proportion of Variance“

viz *Tabulka č. 3*), přesněji se postupně sčítá podle toho, kolik komponent z původní analýzy ponecháme.

Zde byl tento počet 5 komponent, tudíž je vhodné se zde zaměřit na výsledek kumulativní proporce pro komponentu PC5, která zde vychází 83,624 % (převedeno na procenta, zaokrouhleno na tři desetinná místa), což tím pádem znamená, že prvních 5 komponent metody PCA je v tomto případě schopno vysvětlit přibližně 83,624 % rozptylu dat. Z toho vyplývá, že ztráta informací je procentuálně přibližně pouhých 16,376 %. Tento výsledek znamená, že těchto 5 komponent dokáže vysvětlit většinu původního rozptylu dat, což je pro tuto analýzu dostačující a vhodné množství. Takto je ztráta informací minimální a je tak možné dále pokračovat v analýze pro určení autorství právě pomocí nově vytvořených komponent (nových vlastností) podle metody PCA.

| | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|
| Standard deviation | 2.0307 | 1.8211 | 1.4695 | 1.15814 | 0.87540 |
| Proportion of Variance | 0.2946 | 0.2369 | 0.1542 | 0.09581 | 0.05474 |
| Cumulative Proportion | 0.2946 | 0.5315 | 0.6857 | 0.78150 | 0.83624 |

*Tabulka č. 3: Výsledky metody hlavních komponent (důležitost prvních 5 komponent)*

Po tomto následovala analýza rotační matice (viz *Tabulka č. 4*). Rotační matice převádí původní hodnoty do zadaného zobrazení, jelikož právě při metodě hlavních komponent se mimo jiné tato metoda může snažit zobrazit data jako shluky bodů v určitém prostoru, kde se vyskytují osy představující proměnné *x* (námi zkoumané hodnoty). V tomto prostoru se pak metoda PCA snaží nalézt takový „úhel pohledu“, kde se ztratí nejméně informací, a ten následně zaznamená a vychází z něj. (Hebák, a další, 2007)

Rotační matice v *Tabulce č. 4* ukazuje ty hodnoty sil vztahů daných komponent metody PCA vůči původním vlastnostem, které jsou buď větší než 0,2, nebo menší než -0,2 (tzn. mimo interval (-0,2; 0,2)), jelikož ty jsou pro další zpracování relevantní. Vztahy, které jsou v daném intervalu, tedy nejsou dostatečně silné, nevypovídají o komponentě tolik, jako to dokáží ty, které mimo interval spadají a jejich síla vztahu je tak výraznější. Nedostatečně silné hodnoty byly proto kvůli irelevantnosti a také pro přehlednost tabulky odstraněny. Tímto způsobem můžeme vidět, které původní vlastnosti jsou nejvíce propojené s danou komponentou a zároveň relevantní – například u PC5 jsou nejrelevantnějšími, či lineárně nejprovázanějšími původními vlastnostmi *Tone*, *drives* a *focusfuture*, zatímco u komponenty PC3 je výrazných vlastností až 5, a to jmenovitě *Analytic*, *funct*, *cogproc*, *drives* a *focuspast*.

| | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|
| Analytic | -0.27 | -0.2 | 0.45 | -0.21 | |
| Clout | -0.42 | | | 0.2 | |
| Authentic | 0.41 | -0.25 | | | |
| Tone | | 0.27 | | -0.21 | -0.83 |
| funct | 0.24 | | -0.51 | | |
| affect | | 0.47 | | | |
| social | -0.23 | 0.35 | | 0.27 | |
| cogproc | | 0.2 | -0.52 | | |
| percept | | -0.24 | | 0.64 | |
| bio | 0.26 | | | 0.5 | |
| drives | | 0.4 | 0.26 | -0.24 | 0.21 |
| focuspast | -0.39 | | -0.27 | | |
| focuspresent | 0.43 | 0.2 | | | |
| focusfuture | 0.21 | 0.32 | | | 0.38 |

*Tabulka č. 4: Rotační matice, relevantní síly vztahů komponent metody PCA a původních vlastností*

Lze tedy vidět, že provázanost komponent s původními vlastnostmi se navzájem nevylučuje, ale ani nepodmiňuje – dvě komponenty mohou mít výrazný vztah se stejnými původními vlastnostmi, ale není to pravidlem, dvě dané komponenty mohou také mít vztahy s naprosto odlišnými původními vlastnostmi. Také počet vztahů, které jsou výrazné, může být, a zpravidla je, pro každou komponentu jiný. V *Tabulce č. 4* můžeme vidět, že komponenta PC1 má těchto významných vztahů 9, zatímco například komponenta PC5 má pouze 4. Je smysluplné předpokládat, že první komponenta, jakožto komponenta s největším informačním přínosem, která vysvětluje největší část dat, bude mít těchto významných vztahů nejvíce, nebo bude mít tyto vztahy nejsilnější.

Zmíněné předpoklady však neznamenají to, že každá další komponenta bude mít významných vztahů méně a méně, postupné snižování informačního přínosu se nemusí promítat rovnoměrně – nemusí být závislé pouze na počtu těchto vztahů, ale také na jejich síle – proto je nutné při zhodnocování výsledných dat pozorovat oba tyto aspekty. Touto metodou tak bylo pro další analýzu získáno 5 komponent s určitým informačním přínosem, které jsou z podstaty metody hlavních komponent na sobě navzájem lineárně nezávislé, a tudíž tvoří vhodný a zajímavý základ pro další kroky této práce.

# 5. ANALÝZA VYBRANÝCH VLASTNOSTÍ POMOCÍ LOGISTICKÉ REGRESE

Pro tento krok bylo zapotřebí vytvořit tabulky párů autorů, u kterých bude v páru vždy jeden z autorů označen jako „0“ a druhý jako „1“, aby byla přítomna binární proměnná, kterou dokáže logistická regrese potřebná pro tuto práci zpracovat. Pro část 5.1 byla taková tabulka utvořena ze zredukovaných 10 vlastností (po provedení analýzy síly korelace dle Evanse v předchozí kapitole 4.2.1), zatímco u kroku 5.2 bylo využito 5 vytvořených PCA komponent z důvodu omezení logistické regrese při analyzování více vlastností zároveň.

Logistická regrese je zobecněný lineární model, který se používá při analýze, která zpracovává dva typy dat – jedním typem jsou data spojitá (zde číselné hodnoty vlastností a komponent) a druhým typem jsou data binární či obecně kategorická (zde informace o autorech textů). Logistická regrese vychází z regrese lineární, avšak lineární regrese vytváří přímku za pomoci funkce $y = ax + b$. (Menard, 2002) Přímka lineární regrese tak zasahuje i pod hodnotu 0 či nad hodnotu 1, což pro binární či kategorické proměnné v podobných situacích jako ta, která bude zkoumána zde, není adekvátní. Logistická regrese namísto přímky využívá křivku vytvořenou logistickou funkcí, která do původní lineární dodá podíl a exponent. Funkce pro logistickou regresi pak vypadá následovně:

$$y = \frac{e^{ax+b}}{1 + e^{ax+b}}$$

(Menard, 2002)

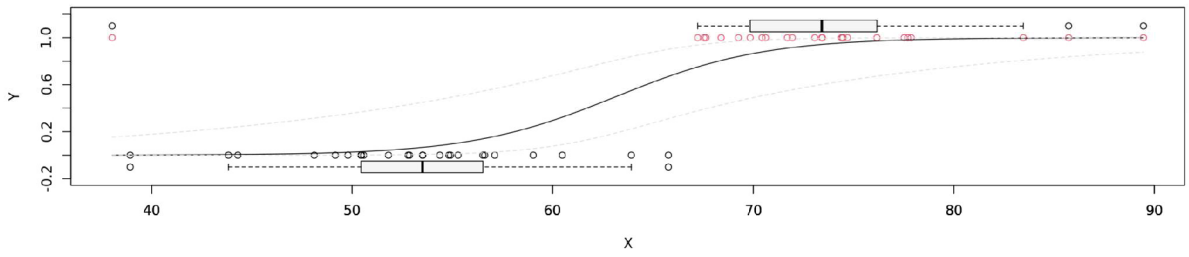
Dodání exponentu do původně lineární funkce zapříčiní „zakulacení“ původní přímky právě do křivky, která analyzovaný vztah vykresluje lépe. Koeficient $a$, neboli koeficient náklonu, v grafickém záznamu ovlivňuje zploštění či rozšíření křivky, zatímco koeficient $b$ ovlivňuje pozici, ve které se v rámci roviny zprava doleva nachází zešikmení křivky. Důležité je poznamenat, že zatímco u lineární regrese se přímka vyskytuje skrze hodnoty i za hodnotami 0 a 1, naopak logistická regrese se jim nekonečně přibližuje jakožto svého minima (0) a svého maxima (1).

Logistická regrese dokáže analyzovat, zdali existuje vztah mezi $x$ a $y$, vztah mezi spojitou a binární proměnnou. Na tento vztah se pak lze dívat jako na pravděpodobnost, s jakou je možné $x$ (nezávislá spojitá proměnná) přiřadit k $y$ (závislá binární proměnná). (Menard, 2002) V případě analýzy v této práci má logistická regrese za cíl ukázat, s jakou pravděpodobností je možné přiřadit hodnotu vlastností k autorovi. Koeficient náklonu nám zde umožňuje sledovat, jak se tato pravděpodobnost (neboli šance) navyšuje či snižuje. (Menard, 2002)

Ve výsledcích aplikací logistické regrese, jako je třeba http://kol-apps.ff.upol.cz/log-reg/, jsou pak tyto výsledky zaznamenané v podobě jako „Navýšení o 1 jednotku v 'x' změní šanci odpovědi '1' X krát“ (Katedra obecné lingvistiky). Toto navýšení je možné uvádět také v procentech s informací o tom, zdali jde o navýšení či snížení této šance.

V grafu logistické regrese, viz *Graf č. 2*, můžeme vidět:

1. samotnou křivku logistické regrese, která je zaznačena tenkou černou čarou;
2. konfidenční pásmo (interval spolehlivosti), který označuje možnost chybovosti křivky, tzn. kde by se teoreticky křivka ještě mohla odchýlit (Hosmer a Lameshow, 2001), v grafu se nachází v okolí křivky, které je označeno světle šedou přerušovanou čarou;
3. body reprezentující každou z hodnot, zde označeny kružnicovými obrysy barev černé pro hodnoty náležící k 0 a červené pro hodnoty náležící k 1;
4. krabicové grafy nad a pod kružnicovými reprezentacemi hodnot, které ukazují distribuci hodnot *x*. Krabicový graf neboli boxplot je typ grafu, který vytvořil John W. Tukey. Tento typ grafu má za účel zobrazit distribuci dat a využívá k tomu své standardní části. První částí je krabice, která se nalézá uprostřed grafu a zaznamenává tzv. mezikvartilové rozpětí (označováno IQR), které obsahuje data od prvního (Q1) až po třetí (Q3) kvartil (okraj krabice označuje hranice pro tyto kvartily). V této krabici se pak uvnitř nachází vertikální úsečka, která představuje medián. Dále jsou v krabicovém grafu tzv. vousy – horizontální úsečky vycházející z krabice, na jejichž konci jsou vertikální úsečky označující minimum (vypočítané jako *Q1 – 1,5 * IQR*) a maximum (vypočítané jako *Q3 + 1,5 * IQR*). Vousy samotné tedy označují data, která se nacházejí mezi minimem či maximem a mezikvartilovým rozpětím. Poslední standardní částí krabicového grafu jsou tzv. outliery – odlehlé hodnoty, které se nacházejí až za minimem či maximem a označují se pomocí kružnic. (Schwabish, 2021)




*Graf č. 2: Ilustrativní graf výsledků logistické regrese*

V *Grafu č. 2* se také dá rozlišit práh (angl. *treshold*) pro třídu „0“ a práh pro třídu „1“. Pokud v logistické regresi spadá hodnota *x* k hodnotě *y*, která je nižší než 0,5, proměnná *x* bude zařazena do třídy autora označeného jako „0“. Pokud je vyšší než 0,5, proměnná *x* bude zařazena do třídy autora označeného jako „1“. Tento práh můžeme takto určit, jelikož jsou obě třídy (oba autoři) zastoupeny stejným pozorováním, v jiném případě by se tento práh musel upravit dle odlišností v pozorováních. Mimo již zmíněné výsledky nám logistická regrese dokáže dát také výsledek toho, v kolika procentech případů byla schopná správně přiřadit „0“ a „1“ k hodnotám *x*, což označuje, s jakou přesností byla schopná rozlišit, které hodnoty náleží ke které z binárních proměnných.

Logistická regrese nám dále také poskytne hodnoty pro chí-kvadrát a p-hodnotu. V tomto nástroji „existuje několik možných způsobů měření rozdílu mezi pozorovanou a očekávanou hodnotou“ (Hosmer a Lameshow, 2001) [vlastní překlad]. Očekávané hodnoty jsou teoretické, zatímco pozorované hodnoty jsou přímo z daných dat. Jedním z těchto způsobů měření rozdílu je v logistické regresi odchylka (angl. *deviance*) a druhým tzv. Pearsonův test dobré shody, a právě z Pearsonova testu dobré shody je vypočítán chí-kvadrát. (Hosmer a Lameshow, 2001; Hebák, a další, 2007) Chí-kvadrát pracuje s nulovou hypotézou (Hosmer a Lameshow, 2001), tedy s hypotézou, že pozorovaný efekt vznikl náhodou (1. lékařská fakulta UK, 2023). Z vypočtené hodnoty chí-kvadrátu následně vychází i výpočet výsledků p-hodnot, které mají za účel se porovnávat s prahem signifikantnosti (hodnota alfa, α) a ukázat tak, zdali daný vztah mezi proměnnými mohl vzniknout náhodou, či zdali je vztah signifikantní (a náhodou tak vzniknout nemohl).

Logistická regrese dokáže také provést výpočty i pro několik nezávislých proměnných (spojité proměnné *x*) zároveň. V takovém případě mluvíme o multinomiální logistické regresi, kde se v modelu počítá s proměnnými $x_1$, $x_2$, $x_3$, … $x_n$, pomocí kterých se poté predikuje odpověď *y*. (Hosmer a Lameshow, 2001) Při práci s logistickou regresí si je však třeba uvědomit, že ani tento nástroj není perfektní a má své limitace, tudíž je vhodné, aby si byl člověk vědom přítomných limitací při zpracování dat tímto způsobem. Proto zde nyní budou představeny některé z problémů vycházejících z daných limitací, které by mohly být relevantní i pro analýzu prováděnou během této práce.

Základ při práci s logistickou regresí stejně jako při práci s jakýmkoliv statistickým nástrojem tvoří nutnost ještě před zahájením analýzy zvážit, které proměnné a hodnoty jsou relevantní a které ne, aby se výsledky poté nejevily zkreslené či nepřesné. (Menard, 2002) Tuto část má na starost sám uživatel, který by měl posoudit, zdali a které jemu dostupné hodnoty by měly být zahrnuty a které naopak nikoliv. Určité zkreslení pak již během samotné

analýzy mohou způsobit tzv. odlehlé hodnoty (outliery), především ve výsledcích p-hodnoty a při determinaci přesnosti přiřazení nezávislých proměnných *x* k závislým proměnným *y*. (Menard, 2002) Tato situace nemusí vždy nutně nastat, ale je vhodné si možnost tohoto zkreslení uvědomovat.

Další problém, který by mohl způsobit zkreslení dat, je tzv. kolinearita či multikolinearita, která může nastat při analýze několika nezávislých proměnných *x* zároveň. Tento fenomén spočívá v tom, že některé z vlastností spolu korelují. Pokud tomu tak je, tak čím vyšší tato korelace je, tím hůře lze z logistické regrese získat unikátní a přesné výsledky pro dané hodnoty. (Menard, 2002) Multikolinearitu lze také popsat jako situaci, ve které se proměnné zobrazené jako datové vektory dají zobrazit v jednorozměrném prostoru, tzn. na stejné přímce, což znamená, že tyto vektory jsou vzájemnou lineární kombinací. (Hebák, a další, 2005) I z tohoto důvodu byl v kroku 4.2.1 proveden test korelace, aby se snížilo riziko toho, že z důvodu korelace vyvstane chyba při tvoření komponent PCA a při následné analýze pomocí logistické regrese.

Mezi další problémy, které mohou nastat, se řadí také situace, ve které se hodnoty pro jednu ze závislých proměnných *y* nemění a jsou nulové, tzn. že každá hodnota pro dané *y* je vždy stejná, a to nulová. Například pokud měříme výsledky v podobě „ano"/„ne" pro různé skupiny lidí, může nastat situace, kdy jedna celá skupina odpoví „ne", čímž bude mít všechny výsledky zaznamenané jako „0", což poté ovlivní analýzy dalších skupin. Takto budou šance ovlivněny buď ve směru nulovosti, nebo ve směru nekonečnosti, což zapříčiní problémy s rozřazením výsledků. (Menard, 2002)

Může také nastat případ tzv. perfektní/kompletní separace (angl. *perfect/complete separation*). Tento stav znamená, že logistická regrese dokáže přiřadit spojitou proměnnou *x* ke kategorické hodnotě *y* s přesností pohybující se blízko 100 %. Pokud je přesnost přímo 100 %, bavíme se o perfektní separaci, při které logistická regrese nedokáže dodat potřebné správné statistické údaje. Pokud je přesnost o něco nižší, odkazujeme se na tzv. kvazi-perfektní separaci. V obou situacích bude mít logistická regrese v důsledku těchto jevů extrémně vysokou chybovost. (Menard, 2002)

## 5.1 ANALÝZA KAŽDÉ VLASTNOSTI ZVLÁŠŤ

Původní vlastnosti zredukované na počet 10 byly po jedné postupně vkládány do logistické regrese společně s informací o tom, zdali daným hodnotám odpovídá autor „0“ či „1“. Vzhledem k množství vlastností bylo nutné opět provést Bonferroniho korekci. V tomto případě zde počet testů určuje počet vlastností vzhledem k tomu, že testy probíhaly na každé zvlášť. Bylo zde tedy prováděno *k* testů, což znamená, že původní hodnota alfa ($\alpha_0$) pro určení práhu signifikantnosti, která je běžně u logistické regrese $\alpha_0 = 0{,}05$, musela být upravena. Bylo nutné ji vydělit právě počtem vlastností, tudíž $k = 10$. Nový práh signifikantnosti (hodnota alfa, $\alpha_1$) byl $\alpha_1 = 0{,}005$, tzn. že pokud bude výsledkem logistické regrese úspěch, tedy to, že p-hodnota pro výsledek chí-kvadrátu je menší než daný práh $\alpha_1 = 0{,}005$, lze říci, že existuje dostatečná evidence o tom, že mezi analyzovanou vlastností a danými autory je vztah, který dokáže odhalit autora.

V *Tabulce č. 5* můžeme vidět úspěšnost určení autora podle každé z vlastností za použití logistické regrese dle nastavení uvedeného v předchozím odstavci, přičemž úspěchy označují zelené „1“ a neúspěchy označují červené „0“. V posledním řádku tabulky je poté zaznamenáno, kolik úspěchů pro dané vlastnosti bylo celkem. Největší úspěšnost lze vidět u vlastností *affect* a *focusfuture*, a to s úspěchem v 8 z 10 případů, tzn. s úspěchem 80 %. Nejmenší úspěšnost má vlastnost *Tone*, a to ve 3 případech z 10, tzn. úspěch 30 %. Tyto výsledky napovídají o tom, že vlastnosti, které by mohly dokázat určit autora v největším procentu případů by byly *affect* a *focusfuture*, naopak nejmenší relevantnost zde má analýza vlastnosti *Tone*.

V posledním sloupci tabulky je zaznamenán počet úspěchů v rámci daných párů, kdy největší úspěšnost má pár Foglar a Lanczová, a to úspěšnost v 8 z 10 případů, tedy 80 %. Naopak nejmenší úspěšnost má pár Švandrlík a Vaňková, a to ve 4 případech z 10, tedy 40 %. Analýza zde tedy naznačuje, že obecně nejvíce rozlišitelní, tedy odlišní, autoři, když se budeme dívat na dané vlastnosti jejich textu, jsou Foglar s Lanczovou, a naopak nejpodobnější si jsou autoři Švandrlík a Vaňková. Další doplnění detailnějšího vhledu do výsledků provedené analýzy za pomoci logistické regrese je možné nalézt v *Tabulce č. 6*, *Tabulce č. 7*, *Tabulce č. 8* a *Tabulce č. 9*. Informace o úspěšnosti analýzy je stále odlišená barevně – zelená označující úspěch a červená označující neúspěch.

| | Analytic | Clout | Tone | affect | cogproc | percept | bio | drives | focuspast | focusfuture | počet úspěchů – páry |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FOG + LAN | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 8 |
| FOG + ŠVA | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 6 |
| FOG + VAŇ | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 5 |
| FOG + VEL | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 6 |
| LAN + ŠVA | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 5 |
| LAN + VAŇ | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 6 |
| LAN + VEL | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 7 |
| ŠVA + VAŇ | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 4 |
| ŠVA + VEL | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 6 |
| VAŇ + VEL | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 7 |
| počet úspěchů – vlastnosti | 7 | 7 | 3 | 8 | 5 | 4 | 5 | 7 | 6 | 8 | |

*Tabulka č. 5: Úspěšnost 0/1 každé vlastnosti zvlášť po analýze logistickou regresí*

Pro to, aby byla analýza vyhodnocena jako úspěšná, bylo nutné, aby výsledná p-hodnota logistické regrese spadala pod práh signifikantnosti. Výsledné p-hodnoty zaokrouhlené na tři desetinná místa můžeme vidět v *Tabulce č. 6* níže. Lze pozorovat, že mnohé z nich, přesněji 47 z 100, se dokonce těsně přibližují k nule (výsledky „~0“). Pro porovnání bylo celkových úspěchů 60 ze 100. Platí zde, že čím nižší je p-hodnota, tím větší je signifikantnost výsledku. Z toho lze odvozovat a porovnávat nejen samotnou úspěšnost mezi jednotlivými vlastnostmi a páry, ale také to, jak moc signifikantní tato úspěšnost je. U vlastnosti *Analytic* můžeme pozorovat, že analýza byla úspěšná u 7 z 10 párů, tedy se dá mluvit o tom, že pozorované hodnoty byly pro oba autory umístěny tak, že jejich následné oddělení nemohlo vzniknout náhodou. U pěti výsledných p-hodnot hodnota odpovídala ~0, tedy v pěti případech se je signifikantnost vysoká a možnost vlivu náhody naprosto minimální.

Pokud se podíváme na vlastnosti, které měly úspěšnost v nejvyšším počtu případů, tak u vlastnosti *affect* pozorujeme, že z celkových 8 úspěchů (z 10 případů) jich 7 odpovídá ~0, a u vlastnosti *focusfuture* z celkových 8 úspěchů (z 10) jich 6 odpovídá ~0. U vlastností *Clout* a *drives*, které mají celkově 7 úspěchů (z 10), tedy jen o jeden méně než vlastnosti s nejvyšším počtem úspěchů, zde můžeme vidět, že u vlastnosti *Clout* všech 7 odpovídá ~0 a u vlastnosti *drives* jich ~0 odpovídá 6. Pouze tyto čtyři vlastnosti mají úspěchů odpovídajících ~0 alespoň 6 či více. Z tohoto se dá usuzovat, že úspěšnost v těchto případech byla vysoká nejen počtem, ale také svou signifikantností. Při pohledu na nejvíce úspěšný pár podle *Tabulky č. 5*, tedy Foglar a Lanczová, z 8 úspěchů je 7 silně signifikantní, tedy odpovídajících hodnotě ~0.

Avšak při pohledu na početně nejméně úspěšnou vlastnost (*Tone*) a nejméně úspěšný pár (Švandrlík a Vaňková) i u nich můžeme vidět, že v těch případech, které byly vyhodnoceny jako úspěchy, byly hodnoty z většiny signifikantní – u *Tone* z 3 úspěchů 2 odpovídaly ~0 a u „ŠVA + VAŇ" ze 4 úspěchů 3 odpovídaly ~0. Lze tedy odhadnout, že pokud již mezi autory lze rozeznat rozdíl, tento rozdíl bude signifikantní.

| | Analytic | Clout | Tone | affect | cogproc | percept | bio | drives | focuspast | focusfuture |
|---|---|---|---|---|---|---|---|---|---|---|
| FOG + LAN | ~0 | ~0 | ~0 | ~0 | 0.07 | 0.002 | ~0 | 0.073 | ~0 | ~0 |
| FOG + ŠVA | ~0 | 0.482 | 0.328 | ~0 | ~0 | ~0 | 0.022 | 0.001 | 0.226 | 0.001 |
| FOG + VAŇ | 0.003 | 0.194 | 0.04 | ~0 | 0.002 | 0.001 | 0.497 | ~0 | 0.65 | 0.609 |
| FOG + VEL | ~0 | ~0 | 0.361 | ~0 | 0.001 | 0.046 | 0.311 | ~0 | 0.029 | 0.001 |
| LAN + ŠVA | 0.262 | ~0 | 0.016 | 0.001 | ~0 | 0.022 | ~0 | 0.014 | ~0 | 0.216 |
| LAN + VAŇ | ~0 | ~0 | 0.022 | 0.085 | 0.078 | 0.134 | ~0 | ~0 | ~0 | ~0 |
| LAN + VEL | 0.288 | ~0 | ~0 | ~0 | 0.038 | 0.085 | ~0 | ~0 | ~0 | ~0 |
| ŠVA + VAŇ | ~0 | 0.013 | 0.507 | 0.134 | ~0 | 0.995 | 0.002 | 0.648 | 0.023 | ~0 |
| ŠVA + VEL | 0.878 | ~0 | 0.088 | ~0 | 0.013 | ~0 | 0.046 | ~0 | ~0 | ~0 |
| VAŇ + VEL | 0.005 | ~0 | 0.003 | ~0 | 0.532 | 0.018 | 0.046 | ~0 | 0.002 | ~0 |

*Tabulka č. 6: P-hodnota každé vlastnosti zvlášť po analýze logistickou regresí*

V *Tabulce č. 7* dále můžeme vidět výsledky chí-kvadrátu pro danou analýzu zaokrouhlené opět na tři desetinná místa. P-hodnoty vycházejí právě z hodnot chí-kvadrátu, přičemž platí to, že čím vyšší je hodnota chí-kvadrátu, tím více se vyvrací pravdivost nulové hypotézy, tedy prokazuje se to, že výsledky dané analýzy nemohly vzniknout pouhou náhodou. Naopak hodnoty chí-kvadrátu přibližující se nule tuto hypotézu pouze potvrzují a ubírají tak na signifikantnosti. (Hebák, a další, 2007)

Další výsledek, který vychází z výpočtů logistické regrese, vypovídá o tom, kolikrát se změní šance odpovědi při navýšení $x$ o 1 jednotku. Tento výsledek je pro ilustraci zaznamenán v *Tabulce č. 8*. Zde je nutné doplnit, že při analýze každé vlastnosti samostatně nebylo provedeno škálování, tzn. s každou z vlastností je zde pracováno v jejích původních jednotkách, které se pro každou vlastnost liší. Z tohoto důvodu lze v tabulce pozorovat, že například hodnoty 1.131 i 13.326 budou přítomny u případů, kdy byl celkovým výsledkem úspěch, ale například hodnota 1.142 bude i u případu, kdy byl celkovým výsledkem neúspěch. *Tabulka č. 8* tak pouze ilustrativně vypovídá o změně šance pro každý z případů, nikoliv o jakémkoliv prahu pro úspěšnost celkové analýzy.

| | Analytic | Clout | Tone | affect | cogproc | percept | bio | drives | focuspast | focusfuture |
|---|---|---|---|---|---|---|---|---|---|---|
| FOG + LAN | 39.037 | 40.312 | 16.952 | 27.872 | 3.271 | 9.409 | 26.959 | 3.203 | 52.484 | 24.263 |
| FOG + ŠVA | 36.039 | 0.494 | 0.956 | 37.846 | 28.106 | 17.773 | 5.226 | 10.556 | 1.469 | 10.222 |
| FOG + VAŇ | 8.834 | 1.686 | 4.227 | 31.029 | 9.801 | 10.599 | 0.461 | 17.777 | 0.205 | 0.262 |
| FOG + VEL | 25.473 | 20.299 | 0.835 | 16.765 | 10.36 | 3.964 | 1.026 | 18.957 | 4.772 | 10.536 |
| LAN + ŠVA | 1.257 | 43.284 | 5.803 | 10.18 | 23.369 | 5.236 | 14.385 | 6.027 | 54.492 | 1.527 |
| LAN + VAŇ | 17.052 | 69.315 | 5.25 | 2.973 | 3.1 | 2.251 | 33.739 | 15.004 | 63.241 | 40.046 |
| LAN + VEL | 1.127 | 13.153 | 24.431 | 61.452 | 4.314 | 2.968 | 37.5 | 39.784 | 69.315 | 50.006 |
| ŠVA + VAŇ | 12.302 | 6.18 | 0.441 | 2.25 | 12.213 | ~0 | 9.936 | 0.208 | 5.191 | 18.078 |
| ŠVA + VEL | 0.023 | 18.734 | 2.902 | 62.483 | 6.206 | 12.505 | 3.984 | 43.972 | 18.664 | 33.763 |
| VAŇ + VEL | 7.94 | 39.379 | 8.918 | 61.137 | 0.391 | 5.582 | 3.983 | 50.364 | 9.755 | 15.974 |

*Tabulka č. 7: Chí-kvadrát každé vlastnosti zvlášť po analýze logistickou regresí*

| | Analytic | Clout | Tone | affect | cogproc | percept | bio | drives | focuspast | focusfuture |
|---|---|---|---|---|---|---|---|---|---|---|
| FOG + LAN | 0.77 | 0.807 | 1.173 | 17.097 | 1.659 | 0.231 | 18.07 | 2.025 | 0.206 | 151.554 |
| FOG + ŠVA | 0.76 | 0.979 | 1.025 | 17.479 | 4.827 | 0.141 | 3.001 | 2.761 | 0.767 | 12.328 |
| FOG + VAŇ | 0.898 | 1.051 | 1.071 | 13.326 | 2.534 | 0.366 | 0.736 | 4.515 | 1.123 | 0.588 |
| FOG + VEL | 0.839 | 0.868 | 0.97 | 0.144 | 2.221 | 0.414 | 1.731 | 0.198 | 1.892 | 0.035 |
| LAN + ŠVA | 1.039 | 1.314 | 0.937 | 5.766 | 5.496 | 0.265 | 0.094 | 2.549 | 6.052 | 0.377 |
| LAN + VAŇ | 1.151 | 2.28E+36 | 0.922 | 2.487 | 1.84 | 0.58 | 0.037 | 6.845 | 59.794 | ~0 |
| LAN + VEL | 1.03 | 1.176 | 0.814 | 0.002 | 1.803 | 3.212 | 0.004 | 0.037 | 1.9E+100 | ~0 |
| ŠVA + VAŇ | 1.131 | 1.128 | 1.017 | 0.565 | 0.318 | 1.002 | 0.181 | 1.142 | 2.138 | 0.01 |
| ŠVA + VEL | 1.005 | 0.857 | 0.956 | 0.006 | 0.506 | 8.479 | 0.252 | 0.038 | 6.322 | 0.002 |
| VAŇ + VEL | 0.922 | 0.742 | 0.898 | 0.003 | 1.189 | 2.367 | 3.388 | 0.03 | 5.199 | 0.002 |

*Tabulka č. 8: Kolikrát se změní šance odpovědi při navýšení x o 1 jednotku*
*při analýze každé vlastnosti zvlášť pomocí logistické regrese*

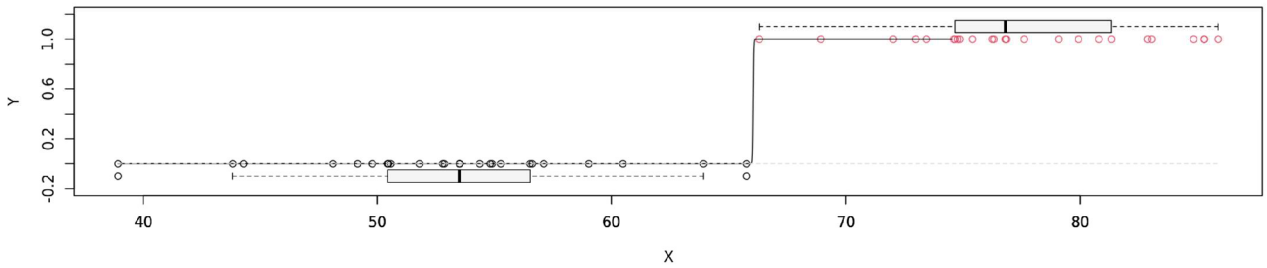
Poslední tabulkou, která dopomáhá vykreslit detailnější pohled na výsledky analýzy každé z vlastností zvlášť, je *Tabulka č. 9*. Zde můžeme vidět, v kolika procentech případů dokázala logistická regrese správně přiřadit „0“ a „1“. Lze zde pozorovat, že některé z výsledků mají nižší procento schopnosti přiřadit „0“ a „1“ než jiné, a i tak spadají v celkové analýze pod úspěch, například u vlastnosti *percept* u páru Foglar + Lanczová (68 %) v porovnání s párem Vaňková + Velinský (70 %). Z tohoto faktu a z výsledků uvedených v předchozích tabulkách lze odvodit, že úspěšnost není založená pouze na linearitě počtu úspěchů/neúspěchů přiřazení „0“ a „1“, ale hraje zde roli i to, jak signifikantně dokáže tyto rozdíly logistická regrese rozeznat.

Můžeme si povšimnout, že v *Tabulce č. 9* jsou dokonce dva výskyty výsledku 100 %, tedy toho, že logistická regrese dokázala odlišit autory „0“ a „1“ se 100% přesností. Tento výsledek je přítomen u páru Lanczová a Vaňková u vlastnosti *Clout* (viz *Graf č. 3*) a u páru Lanczová a Velinský u vlastnosti *focuspast* (viz *Graf č. 4*). Pokud se podíváme i nazpět do *Tabulky č. 6*, v obou případech p-hodnota odpovídá ~0, což znamená, že logistická regrese takto dokáže rozeznat případy jak početně, tak i v rámci signifikantnosti. Z tohoto lze usuzovat, že pro tyto páry jsou dané vlastnosti nejrelevantnější a mají největší výpovědní hodnotu při rozlišení, který autor napsal zkoumaný text.

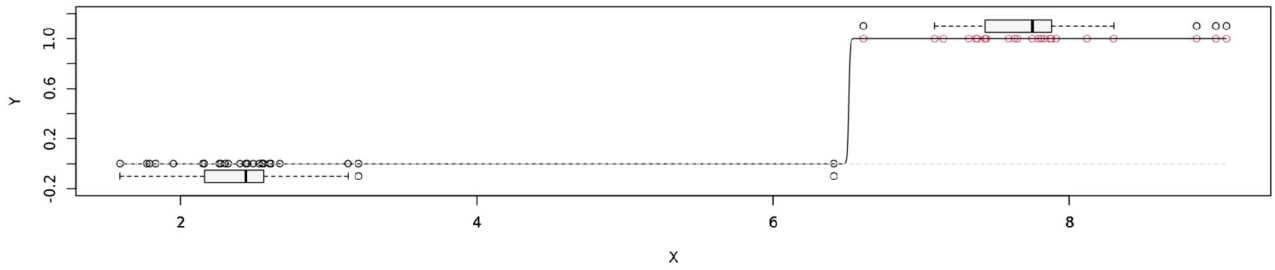
Celkově se v *Tabulce č. 9* nachází 32 hodnot, které přesahují hranici 80 %, tzn. ve 32 případech z celkových 100 dokázala logistická regrese určit autora s úspěšností 80 % či více. Pokud se podíváme pouze na vlastnosti, které v *Tabulce č. 5* dosáhly největšího počtu úspěchů, tedy vlastnosti *affect* a *focusfuture*, v případě *affect* z celkových 8 úspěchů jich 6 přesáhlo hranici 80 % a v případě *focusfuture* z celkových 8 úspěchů tuto hranici přesáhly 4. Pro porovnání s vlastnostmi *Clout* a *drives*, které měly vysoký počet p-hodnot (přesněji 7) odpovídajících ~0, zde úspěšnost přesahující hranici 80 % má u vlastnosti *Clout* 6 případů a u vlastnosti *drives* 3 případy.

| | Analytic | Clout | Tone | affect | cogproc | percept | bio | drives | focuspast | focusfuture |
|---|---|---|---|---|---|---|---|---|---|---|
| FOG + LAN | 90 % | 92 % | 72 % | 88 % | 60 % | 68 % | 84 % | 62 % | 94 % | 84 % |
| FOG + ŠVA | 86 % | 62 % | 58 % | 88 % | 84 % | 74 % | 68 % | 66 % | 62 % | 72 % |
| FOG + VAŇ | 60 % | 50 % | 56 % | 82 % | 70 % | 68 % | 54 % | 76 % | 44 % | 54 % |
| FOG + VEL | 84 % | 84 % | 56 % | 78 % | 66 % | 56 % | 52 % | 78 % | 56 % | 72 % |
| LAN + ŠVA | 52 % | 94 % | 62 % | 66 % | 78 % | 62 % | 80 % | 62 % | 94 % | 64 % |
| LAN + VAŇ | 76 % | 100 % | 66 % | 62 % | 56 % | 60 % | 86 % | 78 % | 98 % | 86 % |
| LAN + VEL | 56 % | 66 % | 86 % | 98 % | 66 % | 54 % | 86 % | 90 % | 100 % | 96 % |
| ŠVA + VAŇ | 68 % | 66 % | 50 % | 54 % | 72 % | 54 % | 74 % | 62 % | 58 % | 76 % |
| ŠVA + VEL | 52 % | 84 % | 64 % | 96 % | 66 % | 76 % | 62 % | 84 % | 76 % | 88 % |
| VAŇ + VEL | 66 % | 86 % | 74 % | 96 % | 56 % | 70 % | 64 % | 90 % | 64 % | 72 % |

*Tabulka č. 9: V kolika % případů dokázala logistická regrese přiřadit „0“ a „1“ při analýze každé vlastnosti zvlášť*




*Graf č. 3: Graf logistické regrese pro pár Lanczová a Vaňková pro vlastnost Clout*




*Graf č. 4: Graf logistické regrese pro pár Lanczová a Velinský pro vlastnost focuspast*

V této části bylo zjištěno, že při zvažování přesnosti určení autorství pomocí jednotlivých vlastností, byly k tomuto účelu se svými výsledky nejrelevantnější následující vlastnosti:

- *affect*, podmíněno výsledky z 10 zkoumaných případů: 8 celkových úspěchů, 7 p-hodnot odpovídajících ~0, 6 případů s úspěšností přiřazení 80 % či více;
- *Clout*, podmíněno výsledky z 10 zkoumaných případů: 7 celkových úspěchů, 7 p-hodnot odpovídajících ~0, 6 případů s úspěšností přiřazení 80 % či více;
- *drives*, podmíněno výsledky z 10 zkoumaných případů: 7 celkových úspěchů, 6 p-hodnot odpovídajících ~0, 3 případy s úspěšností přiřazení 80 % či více;
- *focusfuture*, podmíněno výsledky z 10 zkoumaných případů: 8 celkových úspěchů, 6 p-hodnot odpovídajících ~0, 4 případy s úspěšností přiřazení 80 % či více.

Po zhodnocení předložených výsledků lze usoudit, že nejvhodnějšími samostatnými vlastnostmi, které se dají použít při analýze s cílem určení autorství nejspolehlivěji, jsou vlastnosti *affect* a *Clout*, a to z důvodu vysokého počtu celkových úspěchů, p-hodnot odpovídajících ~0 a případů s úspěšností přiřazení 80 % či více zároveň.

Vzhledem k nedostatečnému popisu samotných vlastností LIWC od jejich původce nelze s jistotou určit, z jakého důvodu jsou úspěšné právě ty vlastnosti, které se takto projevily během analýzy v této práci. Ovšem domnívaným společným znakem by mohla být příznakovost daných prostředků, které tyto vlastnosti zvyšují, či snižují. V jazyce se vyskytuje určitá norma,

od které se na různých rovinách mohou prostředky odklánět a stávat se tak příznakovými. Tyto prostředky mohou pak v textu působit výrazněji. Vzhledem k tomu, že největší úspěšnost zde měly vlastnosti *affect*, tzn. citově založené procesy, *Clout*, tzn. agresivita či vůdčí schopnosti, *drives*, tzn. hnací síly za určitým cílem či z určitého důvodu, a *focusfuture*, tzn. orientace na budoucnost, hlavní zaměření bude na ně. V porovnání je pak vhodné brát do úvahy také vlastnosti, které uspěly nejméně, tedy vlastnosti *Tone*, tzn. emoční tón, a *percept*, tzn. vjemově založené procesy.

Vliv na důvody, proč úspěšné vlastnosti měly signifikantní výsledky, pravděpodobně má specifičnost samotných superhapaxů, a tudíž specifičnost jejich okolí, nikoliv celkového textu. Tato teorie může být podložena tím, že superhapaxy samy o sobě nesou určitou informaci o autorovi a jsou specifické tím, že nejsou v celkovém textu časté. Mohou tak podmiňovat své okolí k tomu, aby mělo určitou formu a používalo určité vyjadřovací prostředky s různým zabarvením, které se v analýze LIWC projeví jakožto hodnoty jednotlivých vlastností. Tyto hodnoty vlastností pak mohou být schopné odhalovat autory, jelikož stojí mimo svou normu, která se vyskytuje v běžných textech.

Některé z úspěšných vlastností by se v textu daly označit za výrazné a částečně zhodnotitelné i samotným čtenářem – například u *affect* se může jednat o slova „šťastný“, „plakala“ apod., což jsou slova, o kterých se dá odhadnout, že v sobě obsahují určitý citově založený proces. O podobný případ by se mohlo jednat u vlastnosti *focusfuture* u slov jako „brzy“, „budu“ apod. U vlastností *Clout* a *drives* již příklady nemusí být tak jednoznačné a viditelné na první pohled. Co však mají tyto vlastnosti společné, je to, že odklon od určité normy může způsobit právě to, že prostředky zesilující či zeslabující hodnoty zmíněných vlastností budou vytvářet výrazný rozdíl.

Ku příkladu se můžeme podívat na vlastnost *focusfuture*. Většinově texty neobsahují velké množství prostředků vyjadřujících budoucnost mimo přímou řeč a přímá řeč v textech nebývá tak častá jako vypravovací styl. Z toho se dá vyvodit, že už malé odklony od běžného používání prostředků v daných částech textu budoucnosti budou výrazné, respektive příznakové. Dále u vlastnosti, jako je *affect*, se může běžně využívat určité množství citově zabarvených prostředků, a pokud se již autor od normy odklání, může se v této oblasti jednat o rozdíly širokého rázu. Tyto rozdíly pak mohou tvořit natolik výrazné odklony od normy, že samy o sobě dokáží odlišit autory od sebe navzájem.

Tuto teorii by mohl podporovat neúspěch vlastnosti *percept* v tom smyslu, že odklon od normy oblasti prostředků vyjadřujících vjemy může být těžší, jelikož tato norma má vzhledem k rozmanitému používání těchto prostředků spíše širší charakter. Naopak však

v opozici k tomuto vysvětlení stojí neúspěch vlastnosti *Tone*, která by měla zaznamenávat emoční tón. V tomto případě by bylo vhodné získat detailnější popis daných vlastností (který však autor programu LIWC v jeho manuálu neposkytuje), jelikož například rozdíl mezi vlastností *affect* a *Tone* by zde velice pomohl kvalitativnímu vysvětlení výsledků vzhledem k tomu, že jejich obecný popis může působit propojeně, avšak výsledky napovídají o tom, že mezi nimi je určitý rozdíl, který je relevantní v rámci určování autorství.

## 5.2 ANALÝZA VŠECH KOMPONENT PCA ZÁROVEŇ

V tomto kroku následovalo vložení všech pěti komponent získaných pomocí PCA do logistické regrese vždy pokaždé pro jeden z párů autorů. Tato prováděná analýza může mít pragmaticky výhodné výsledky díky přidání více proměnných, které mohou zajistit její zpřesnění. Lze zde vycházet z předpokladu, že každá z proměnných má určitý význam pro určení autorství a čím více relevantních proměnných, které nám tyto informace poskytnou, tím spíše bude analýza úspěšná a bude schopná mít vypovídající hodnotu o tom, kterého autora přiřadit ke kterému z textů.

Stejně jako v předchozím kroku byla vytvořena tabulka s informací o autorovi v podobě hodnot „0“ či „1“. I zde bylo žádoucí provést Bonferroniho korekci (viz minulá kapitola 5.1). Nominální práh signifikantnosti byl vzhledem k opětovnému zpracování dat v logistické regresi $\alpha_0 = 0{,}05$. Počet testů se zde odvíjel od počtu párů autorů, tedy zde bylo $k = 10$. Nový práh signifikantnosti byl tedy vypočítán za pomocí $\alpha_1 = \alpha_0/k = 0{,}05/10 = 0{,}005$. V tomto případě tudíž výsledky logistické regrese budou považované za úspěch, pokud p-hodnota bude menší než nový práh signifikantnosti, tedy $\alpha_1 = 0{,}005$, a v situaci, kdy bude překročen tento práh, lze opět odvodit, že zde byla získaná dostatečná evidence o tom, že mezi analyzovanými komponentami a danými autory existuje vztah prozrazující autora.

V *Tabulce č. 10* jsou uvedené hodnoty analýzy kombinace všech komponent za použití logistické regrese, tyto hodnoty označují celkovou úspěšnost zaznamenanou „1“ jakožto úspěch a „0“ jakožto neúspěch, p-hodnotu a úspěšnost přiřazení správných autorů k daným textům. V posledním sloupci poté pro porovnání nalezneme, jaká byla úspěšnost přiřazení autorů k textům samostatné první komponenty PC1. Úspěchů označených jako „1“ je v tabulce 10 z 10 zkoumaných případů, celkově tedy tato analýza byla schopná rozeznat všechny autory od sebe navzájem. Nejen to, ale při pohledu na výsledky p-hodnot v následujícím sloupci, je zřetelné, že tato určení byla vysoce signifikantní, jelikož opět všechny výsledky p-hodnot odpovídají ~0.

V *Tabulce č. 10* je také uvedena zmíněná úspěšnost komponent metody PCA při rozlišování autorů. Tato analýza úspěšnosti byla provedena za pomoci skriptu v RStudiu jak prvně na všech komponentách zároveň, což lze pozorovat ve výsledcích v předposledním sloupci tabulky s názvem „úspěšnost PC1-PC5“, tak poté na samostatné první komponentě PC1, která má své výsledky ve sloupci posledním – s názvem „úspěšnost PC1“. Samotná první komponenta je zde zvolená pro porovnání její úspěšnosti s celkovou úspěšností komponent,

a to z toho důvodu, že ze své podstaty by měla nést největší informační podíl ze všech vytvořených komponent PCA.

V *Tabulce č. 10* tak můžeme ve sloupci „úspěšnost PC1" vidět, že zde úspěšnost správného přiřazení autora vychází průměrně 83,4 %, kdy maximum odpovídá 96 % a minimum 64 %. Můžeme si povšimnou, že počet případů, kdy úspěšnost přiřazení autora odpovídala 80 % či více, bylo 6 z celkových 10. Analýza kombinace již všech prvních pěti vyhodnocených komponent dokáže autora určit s velice vysokou přesností, jak lze vidět opět v *Tabulce č. 10*, tentokrát ve sloupci „úspěšnost PC1-PC5", kde je průměrná procentuální úspěšnost přiřazení autora 95,6 %, přičemž maximum je zde 100% přesnost (která nastala v 6 z 10 případů) a minimum je 84% přesnost, což již naznačuje, že tato metoda je schopná poskytnout spolehlivé a přesné výsledky pro účely určování autorství. Při pohledu na to, kolik případů s úspěšností přiřazení autora odpovídá výsledku 80 % a více, zjistíme, že tomuto kritériu odpovídají všechny výsledky.

Vzhledem k využité metodě hlavních komponent nelze zcela kvalitativně vysvětlit, co analyzované komponenty znamenají samy o sobě, ale lze z provedené analýzy odvodit, že pokud pro daný text jednoho z analyzovaných autorů provedeme analýzu vlastností LIWC a vyextrahujeme vlastnosti *Analytic*, *Clout*, *Authentic*, *Tone*, *funct*, *affect*, *social*, *cogproc*, *percept*, *bio*, *drives*, *focuspast*, *focuspresent*, *focusfuture*, ze kterých poté vytvoříme pomocí metody PCA nové komponenty, již samotné komponenty dají ve většině případů uspokojivé výsledky ohledně přesnosti přiřazení správného autora k danému textu.

Pokud jako kritéria pro spolehlivost použití analýzy pro určování autorství vezmeme stejně jako v předchozí kapitole 5.1 opět vyhodnocené výsledky celkového počtu úspěchů, počtu p-hodnot odpovídajících ~0 a počtu případů s úspěšností přiřazení 80 % či více zároveň, analýza komponent vytvořených metodou hlavních komponent nám poskytne následující výsledky. U analýzy jedné samostatné komponenty – zde komponenta PC1 – se ještě stále objevuje menší úspěšnost přiřazení správných autorů, než by byla žádoucí. Avšak při pohledu na jak počet celkových úspěchů, tak číselného vyjádření p-hodnot a úspěšnost přiřazení správných autorů v případě, kdy byly v logistické regresi analyzované všechny komponenty (PC1-PC5) zároveň, vše naznačuje tomu, že tento výsledek je funkční a vhodný pro odhalení autora textu, či popřípadě jeho vyloučení jako autora určitého textu. Lze tak usuzovat, že samotné komponenty vytvořené metodou PCA ještě budou mít značné nepřesnosti či značnou chybovost při určování autora, ale pokud se použije jejich příslušná kombinace, výsledky se již jeví mnohem spolehlivěji a mající větší přesností.

Příčina takto uspokojivých výsledků může spočívat v tom, že komponenty PCA z celkových původních dat vyextrahují informačně nejdůležitější aspekty, přičemž přitom do analýzy zahrnou jak ty vlastnosti, které samy o sobě mají vysokou přesnost při určování autorství, tak také všechny ostatní zvažované vlastnosti, které tak slouží jako doplňky informační hodnoty. Samozřejmě na výsledky mají také vliv změny v krocích analýzy, které se týkají jejího provedení již nikoliv jen na jednotlivých datových setech zvlášť, ale nyní už také na jejich kombinaci.

| | úspěšnost 0/1 PC1-PC5 | p-hodnota PC1-PC5 | chí-kvadrát PC1-PC5 | úspěšnost PC1-PC5 | úspěšnost PC1 |
|---|---|---|---|---|---|
| FOG + LAN | 1 | ~0 | 69.3146 | 100 % | 96 % |
| FOG + ŠVA | 1 | ~0 | 53.742 | 94 % | 64 % |
| FOG + VAŇ | 1 | ~0 | 59.3936 | 92 % | 68 % |
| FOG + VEL | 1 | ~0 | 34.1762 | 86 % | 82 % |
| LAN + ŠVA | 1 | ~0 | 69.3137 | 100 % | 96 % |
| LAN + VAŇ | 1 | ~0 | 69.3145 | 100 % | 96 % |
| LAN + VEL | 1 | ~0 | 69.3146 | 100 % | 96 % |
| ŠVA + VAŇ | 1 | ~0 | 33.5979 | 84 % | 76 % |
| ŠVA + VEL | 1 | ~0 | 69.3145 | 100 % | 76 % |
| VAŇ + VEL | 1 | ~0 | 69.3146 | 100 % | 84 % |
| počet úspěchů | 10 | | | | |

*Tabulka č. 10: Analýza pomocí logistické regrese pro kombinaci prvních pěti komponent (sloupec 1-3): úspěšnost 0/1 (sloupec 1), p-hodnota (sloupec 2), chí-kvadrát (sloupec 3); analýza provedená v RStudiu (sloupec 4-5): procentuální úspěšnost přiřazení „0“ a „1“ u analýzy všech komponent zároveň (sloupec 4) procentuální úspěšnost přiřazení „0“ a „1“ u analýzy komponenty PC1 (sloupec 5)*

# ZÁVĚR A DISKUZE

Tato práce měla za účel prokázat, zdali by sentimentová analýza provedená na okolích superhapaxů mohla dokázat za pomoci metody hlavních komponent a logistické regrese rozlišit autory, případně jaké přesnosti by mohla dosáhnout. Jako nejrelevantnější kritéria pro určení spolehlivosti rozlišení autorů zde byla zhodnocena kritéria následující:

A. počet celkových úspěchů,
B. množství p-hodnot odpovídajících ~0,
C. počet případů s úspěšností přiřazení autorů 80 % či více zároveň.

První část finální analýzy se zabývala postupem a výsledky, které byly vyhodnoceny z výběru původních vlastností vyextrahovaných z nástroje LIWC. Účelem této části tak bylo zhodnotit schopnost rozlišit autora podle jednotlivých LIWC vlastností. Tato analýza zjistila, že u většiny zkoumaných vlastností nelze říci, že by k tomuto účelu dokázaly sloužit samy osobě jako jednotlivé izolované data sety. Největší úspěch zde vykazovaly vlastnosti *affect* a *Clout*, které vynikaly svými výsledky v rámci právě výše uvedených kritérií spolehlivosti.

Druhá část finální analýzy byla věnována vytváření zcela nových proměnných a jejich následnému zkoumání. Při vytváření nových proměnných byla na původní zájmové vlastnosti vyextrahované z nástroje LIWC aplikovaná metoda hlavních komponent. Primárním zaměřením zde bylo prvních pět komponent, u kterých byla následně provedena analýza na všech z nich zároveň. Výsledky této analýzy ve výše uvedených kritérií spolehlivosti vynikaly svým značným úspěchem. Tento postup se prokázal jakožto velmi přesný a spolehlivý ve schopnosti rozlišení autorů.

Při srovnání analýzy samostatných vlastností s analýzou více vlastností zároveň (za pomocí metody hlavních komponent) z výsledků této práce vyplývá, že druhá varianta je pro uživatele zkoumajícího autorství mnohem vhodnější a spolehlivější. Původní hypotéza odlišitelnosti autorů pro většinu samostatných vlastností nedosáhla dostatečných důkazů pro její potvrzení. V druhém případě, kdy byla analýza provedená na kombinaci vytvořených komponent metodou PCA, se hypotéza odlišitelnosti autorů naproti tomu vzhledem k výsledkům potvrdila. V této práci tedy bylo odhaleno, že varianta společné analýzy několika komponent, které vycházejí z vybraných původních vlastností LIWC, dokáže za pomoci logistické regrese rozlišit autory, a to s vysokou přesností.

Nevyřešenou otázkou výsledků této práce zůstává obtížná kvalitativní vysvětlitelnost komponent vytvořených metodou PCA. Tento problém však vyplývá z všeobecné podstaty dané metody. Metoda hlavních komponent nedokáže vytvořené komponenty zcela smysluplně

vysvětlit nejen v rámci provedené analýzy v této konkrétní práci, ale ani za jakýchkoliv jiných podmínek. Nepřítomnost dostatečného kvalitativního vysvětlení tak předchází možnosti přinést bližší sémantický vhled do výsledků, které jsou nyní přístupné v podobě číselných hodnot.

Poznatky získané v této práci je možné rozvinout v oblastech, kde tato práce rozsáhlejší výzkum již neumožňovala z důvodu jejího rozsahu a časové omezenosti. Podnětem pro další práce tedy zůstává ku příkladu varianta, která by obsahovala větší množství autorů či větší množství textů, v rámci kterých by se podobná analýza prováděla. V této práci bylo zvoleno okolí superhapaxů jakožto 1 věta před větou superhapaxovou a 2 věty po větě superhapaxové, je však také možnost při navazujícím výzkumu toto okolí rozšířit, či omezit a provést stejný postup znovu a výsledky obou analýz navzájem srovnat. Zajímavé výsledky by mohla přinést i analýza samotných superhapaxových vět bez dalšího okolí.

Dalším podnětem je také provedení podobného výzkumu na textech, jejichž originály jsou napsány v jiných jazycích, ideálně v jazyce anglickém nebo v těch jazycích, se kterými dokáže program LIWC2015 implicitně pracovat bez předběžného zpracování. Mohla by také navázat analýza textů v různých jazycích, které nejsou implicitně zahrnuty v LIWC2015 a pro které by následně byly vytvořeny překlady. V dalších částech práce bylo vzhledem k okolnostem zvoleno analyzovat pět komponent vytvořených metodou PCA, avšak při jiném množství původních textů je možné analýzu zúžit na méně, či rozšířit na více komponent. Tímto způsobem by bylo možné dále zjistit, kde nejníže se nachází hranice pro relevantní přesnost rozlišení autorů, nebo také kde se nachází hranice pro rozřazení všech autorů, kdy by všechny výsledky úspěšnosti odpovídaly 100 % (pokud by to vůbec bylo možné).

Tato práce poskytuje vhled do problematiky analýzy týkající se určování autorství za použití superhapaxů, sentimentové analýzy, korelací, metody hlavních komponent a v neposlední řadě také logistické regrese.

# CITACE

**1. lékařská fakulta UK. 2023.** Testování statistických hypotéz. *WikiSkripta.* [Online] 1. lékařská fakulta a Univerzita Karlova, 1. leden 2023. [Citace: 11. duben 2023.] https://www.wikiskripta.eu/w/Testov%C3%A1n%C3%AD_statistick%C3%BDch_hypot%C3%A9z.

**Baayen, H., van Halteren, H. a Tweedie, F. 1996.** Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing.* 1. září 1996, Sv. 11, 3, stránky 121–132.

**Baggini, Julian a Fosl, Peter S. 2010.** *The Pholosopher's Toolkit: A Compendium of Philosophical Concepts and Methods.* 2. Chichester, Anglie: A John Wiley & Sons, Ltd., 2010. 978-1-4051-9018-3.

**Binongo, J. N. G. 2003.** Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance.* 2003, Sv. 16, 2, stránky 9-17.

**Bonferroni, C. E. 1936.** Teoria Statistka delle classi e calcolo della probabilitá. *Pubbl. R 1st. Sup. Sci, Econ. Commer. Fir.* 1936, Sv. 8.

**DeepL. 2023.** Press Release — DeepL Translator adds Korean and Norwegian (bokmål). *DeepL.* [Online] DeepL, 31. leden 2023. [Citace: 6. duben 2023.] https://www.deepl.com/en/press-release.

**Evans, J. D. 1996.** *Straight Forward Statistics For The Behavioral Sciences.* Pacific Grove, CA, USA: Brooks/Cole Publishing, 1996. 9780534231002.

**Faltýnek, Dan a Kučera, Ondřej. 2022.** *Parasyntax jako struktura nízko frekventovaných částí textu Hapax legomenon prostředkem textové koheze.* Olomouc, Česko: Palacký University Olomouc, 2022.

**Faltýnek, Dan a Matlach, Vladimír. 2020.** *Hapax Remains: authorial features of textual cohesion in authorship attribution.* místo neznámé: Preprint, 2020.

**Harris, Zellig S. 1954.** Distributional Structure. *WORD.* 2-3, 1954, Sv. 10, stránky 146-162.

**Hebák, Petr, a další. 2007.** *Vícerozměrné statistické metody (1).* 2. Praha, Česko: Informatorium, 2007. 978-80-7333-056-9.

**Hebák, Petr, a další. 2007.** *Vícerozměrné statistické metody (3).* 2. Praha, Česko: Informatorium, 2007. 978-80-7333-001-9.

**Hebák, Petr, Hustopecký, Jiří a Malá, Iva. 2005.** *Vícerozměrné statistické metody (2).* Praha, Česko: Informatorium, 2005. 80-7333-036-9.

**Holley, Rose. 2009.** How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. *D-Lib Magazine.* 2009, Sv. 15, 3/4.

**Hosmer, David W. a Lameshow, Stanley. 2001.** *Applied Logistic Regression.* 2. místo neznámé: Wiley-Interscience, 2001. 9780471225898.

**Jones, Karen Spärck. 1972.** A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation.* 1972, Sv. 28, 1, stránky 11-21.

**Katedra obecné lingvistiky, UPOL.** Logistická Regrese. [Online] [Citace: 11. duben 2023.] http://kol-apps.ff.upol.cz/log-reg/.

**Katedra obecné lingvistiky, UPOL. 2016.** QUITA Online - Home. *QUITA Online.* [Online] 22. červenec 2016. [Citace: 5. duben 2023.] https://kol.ff.upol.cz/quita/?.

**Luhn, H. P. 1957.** A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development.* 4, 1957, Sv. 1, stránky 309-317.

**Menard, Scott W. 2002.** *Applied logistic regression analysis.* 2. Thousand Oaks, CA, USA: Sage Publications, 2002. 0761922083.

**McInnes a další. 2018.** UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*. 2018, Sv. 3, 29

**Mikros, George K. 2009.** Content words in authorship attribution: An evaluation of stylometric features in a literary corpus. [editor] R. Köhler. *Issues in Quantitative Linguistics.* 2009, Sv. 5, stránky 61–75.

**Mundfrom, Daniel, a další. 2006.** Bonferroni Adjustments in Tests for Regression Coefficients. *Multiple Linear Regression Viewpoints.* 2006, Sv. 32, 1.

**Pennebaker, James W. a Francis, Martha E. 1996.** Cognitive, Emotional, and Language Processes in Disclosure. *Cognition & Emotion.* 1996, Sv. 10, 6, stránky 601-626.

**Pennebaker, James W. a King, Laura A. 1999.** Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology.* 1999, Sv. 77, 6, stránky 1296-1312.

**Pennebaker, James W., a další. 2003.** Lying Words: Predicting Deception from Linguistic Styles. *Personality and Social Psychology Bulletin.* 2003, Sv. 29, 5, stránky 665-675.

**Pennebaker, James W., a další. 2015.** *The development and psychometric properties of LIWC2015.* Austin, TX: University of Texas at Austin, 2015.

**Pennebaker, James W., Francis, Martha E. a Mayne, Tracy J. 1997.** Linguistic Predictors of Adaptive Bereavement. *Journal of Personality and Social Psychology.* 1997, Sv. 72, 4, stránky 863-871.

**Qader, Wisam Abdulazeez , Ameen, Musa M. a Ahmed, Bilal I. . 2019.** *An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges.* Erbil, Irák: IEEE, 2019. 2019 International Engineering Conference (IEC). 978-1-7281-4377-4.

**Řehák, Jan a Loučková, Ivana. 1983.** Klasické mnohorozměrné škálování (Aplikace metody DISTAN). *Sociologický Časopis / Czech Sociological Review.* 1983, Sv. 19, 5, stránky 535-554.

**Sakamat, Norzehan, Khalid, Noor Elaiza Abdul a Azha, Inda Ishadah Nazrul. 2018.** Structuring Elements of Hit or Miss to Identify Pattern of Benchmark Latin Alphabets Strokes. *Indonesian Journal of Electrical Engineering and Computer Science.* 2018, Sv. 12, 1, stránky 356-362.

**Schwabish, Jonathan. 2021.** *Better Data Visualizations: A Guide for Scholars, Researchers, and Wonks.* New York City, NY, USA: Columbia University Press, 2021. 9780231550154.

**Sindhu, Chandra Sekharan, Swain, Siddharth a Gaurav, Mishra. 2017.** *Recent approaches on authorship attribution techniques — An overview.* Kattankulathur, India: Department of Computer Science and Engineering, SRM University, 2017. 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA). stránky 557-566. 978-1-5090-5686-6.

**Špaček, František. 2022.** *Havex.* Software. Katedra obecné lingvistiky, UPOL. Olomouc, Česko.

**Undeutsch, U. 1967.** Forensische psychologie (Forensic psychology). Göttingen, Německo: Verlag für Psychologie, 1967.

**Uyar, Erkan, Bozkurt, Ilker Nadi a Baghoglu, Ozgur . 2007.** *Authorship attribution.* Ankara, Turkey: IEEE, 2007. 2007 22nd international symposium on computer and information sciences. 978-1-4244-1363-8.

**West, Jevin D. a Bergstrom, Carl T. 2018.** *Why scatter plots suggest causality, and what we can do about it.* místo neznámé: arXiv, 2018.

**Wickham, Hadley a Grolemund, Garrett . 2017.** *R for Data Science.* Sebastopol, CA, USA: O'Reilly Media, Inc., 2017. 978-1-491-91039-9.

# SEZNAM ZKRATEK

LIWC … Linguistic Inquiry and Word Count

FOG … Jaroslav Foglar

LAN … Lenka Lanczová

ŠVA … Miloslav Švandrlík

VAŇ … Ludmila Vaňková

VEL … Jaroslav Velinský

cit. … citace

BoW … Bag of Words

tf-idf … Term Frequency – Inverse Document Frequency

# SEZNAM TABULEK, OBRÁZKŮ A GRAFŮ

# PŘÍLOHY

## PŘÍLOHA Č. 1: SEZNAM AUTORŮ A JEJICH DĚL

Jaroslav Foglar – Boj o první místo

Jaroslav Foglar – Devadesátka pokračuje

Jaroslav Foglar – Dobrodružství v temných uličkách

Jaroslav Foglar – Dobrodružství v zemi nikoho

Jaroslav Foglar – Historie svorné sedmy

Jaroslav Foglar – Hoši od Bobří řeky

Jaroslav Foglar – Chata v Jezerní kotlině

Jaroslav Foglar – Jestřábe vypravuj 2

Jaroslav Foglar – Když duben přichází

Jaroslav Foglar – Kronika hochů od Bobří řeky 2

Jaroslav Foglar – Kronika hochů od Bobří řeky

Jaroslav Foglar – Kronika Ztracené stopy

Jaroslav Foglar – Modrá rokle

Jaroslav Foglar – Pod Junáckou vlajkou

Jaroslav Foglar – Poklad černého delfína

Jaroslav Foglar – Přístav volá

Jaroslav Foglar – Soví jeskyně

Jaroslav Foglar – Stezka odvahy

Jaroslav Foglar – Stínadla se bouří

Jaroslav Foglar – Strach nad Bobří řekou

Jaroslav Foglar – Tábor smůly

Jaroslav Foglar – Tajemná Řásnovka

Jaroslav Foglar – Tajemství velkého Vonta

Jaroslav Foglar – Záhada hlavolamu

Jaroslav Foglar – Závod o Modřínový srub

Lenka Lanczová – Čas něhy

Lenka Lanczová – Deváté nebe

Lenka Lanczová – Doteky noci

Lenka Lanczová – Hříšná touha

Lenka Lanczová – Kapky rosy

Lenka Lanczová – Kde končí svět

Lenka Lanczová – Lásky a nálezy

Lenka Lanczová – Létem políbená

Lenka Lanczová – Letní něžnosti

Lenka Lanczová – Manželky, milenky, zoufalky

Lenka Lanczová – Mokrá náruč léta

Lenka Lanczová – Nebe plné hvězd

Lenka Lanczová – Oranžové blues

Lenka Lanczová – Pár minut lásky

Lenka Lanczová – Postel plná růží

Lenka Lanczová – Potížistka

Lenka Lanczová – Půlnoční sny

Lenka Lanczová – Sběratelka polibků

Lenka Lanczová – Souhvězdí labutě

Lenka Lanczová – Třináctá komnata

Lenka Lanczová – Vstupenka do ráje

Lenka Lanczová – Všechno bude fajn

Lenka Lanczová – Zašeptej do vlasů

Lenka Lanczová – Zlodějka snů

Lenka Lanczová – Znamení blížence

Miloslav Švandrlík – Černí baroni

Miloslav Švandrlík – Dejte napít Babinskému

Miloslav Švandrlík – Draculův švagr

Miloslav Švandrlík – Draculův zlověstný doušek

Miloslav Švandrlík – Krvavý Bill a Viola

Miloslav Švandrlík – Nemravná dívka na vdávání

Miloslav Švandrlík – Neuvěřitelné příhody žáků Kopyta a Mňouka

Miloslav Švandrlík – Postrach Posázaví

Miloslav Švandrlík – Rakev do domu

Miloslav Švandrlík – Sexbomba na doplňkovou půjčku

Miloslav Švandrlík – Šance jako hrom

Miloslav Švandrlík – Tlustý muž pod Jižním křížem

Miloslav Švandrlík – Unesli Mňouka, Baskerville!

Miloslav Švandrlík – Úplňky v Kutné Hoře

Miloslav Švandrlík – Válka skřetů

Miloslav Švandrlík – Vražda mlsného humoristy

Miloslav Švandrlík – Žáci Kopyto a Mňouk a akta X

Miloslav Švandrlík – Žáci Kopyto a Mňouk a indiáni

Miloslav Švandrlík – Žáci Kopyto a Mňouk a maharadžova pomsta

Miloslav Švandrlík – Žáci Kopyto a Mňouk a mimozemšťané

Miloslav Švandrlík – Žáci Kopyto a Mňouk a ohnivý myslivec

Miloslav Švandrlík – Žáci Kopyto a Mňouk a ryšavý upír

Miloslav Švandrlík – Žáci Kopyto a Mňouk a únos policajtovy ženy

Miloslav Švandrlík – Žáci Kopyto a Mňouk a vesmírná brána

Miloslav Švandrlík – Žáci Kopyto a Mňouk opět zasahují

Ludmila Vaňková – A byl večer a bylo jitro, den první

Ludmila Vaňková – Byl jednou jeden hrad

Ludmila Vaňková – Cestou krále

Ludmila Vaňková – Cval rytířských koní

Ludmila Vaňková – Černá sága

Ludmila Vaňková – Dítě z Apulie

Ludmila Vaňková – Dotkni se nebe

Ludmila Vaňková – Druhá císařovna

Ludmila Vaňková – Dvojí trůn

Ludmila Vaňková – Hašteření o korunu

Ludmila Vaňková – Jsme jedné krve

Ludmila Vaňková – Kdo na kamenný trůn

Ludmila Vaňková – Mosty přes propast času

Ludmila Vaňková – Nezávislý novinář

Ludmila Vaňková – Od moře k moři

Ludmila Vaňková – Od trůnu dál

Ludmila Vaňková – Orel a had

Ludmila Vaňková – První muž království

Ludmila Vaňková – Příběh mladšího bratra

Ludmila Vaňková – Rab z Rabštejna

Ludmila Vaňková – Roky před úsvitem

Ludmila Vaňková – Svatby jak nemají být

Ludmila Vaňková – Ty jsi dědic svého otce

Ludmila Vaňková – Vězení pro krále

Ludmila Vaňková – Žena pro třetího krále

Jaroslav Velinský – Bestie z Tamberku

Jaroslav Velinský – Bouda

Jaroslav Velinský – Cesta pro vraha

Jaroslav Velinský – Dáma se zeleným slonem

Jaroslav Velinský – Desátý příběh bezděčného detektiva Oty Finka

Jaroslav Velinský – Dívka s copem

Jaroslav Velinský – Divnobraní

Jaroslav Velinský – Fialky pro královnu

Jaroslav Velinský – Hra na smrt

Jaroslav Velinský – Hranice pomsty

Jaroslav Velinský – Krásná vyhlídka

Jaroslav Velinský – Krvavá fontána

Jaroslav Velinský – Mistrova smrt

Jaroslav Velinský – Mosazná postel

Jaroslav Velinský – Našeptavač

Jaroslav Velinský – Poslední tajemství Jana T.

Jaroslav Velinský – Přibližné řešení

Jaroslav Velinský – Případ výjimečné klibny

Jaroslav Velinský – S.I.MAN Dan Young

Jaroslav Velinský – Strašidlo

Jaroslav Velinský – Studená sprcha

Jaroslav Velinský – Velice dlouhé schody

Jaroslav Velinský – Zahrádka smrti

Jaroslav Velinský – Země Thugů

Jaroslav Velinský – Zmizení princezny