Univerzita Palackého v Olomouci
Filozofická fakulta
Katedra obecné lingvistiky

ANALÝZA HLUBOKÉHO A POVRCHOVÉHO SENTIMENTU V AUTORSKÝCH TEXTECH: PŘÍPADOVÉ STUDIE

DEEP SENTIMENT AND SHALLOW SENTIMENT ANALYSIS IN AUTHORIAL TEXTS: CASE STUDIES

Magisterská diplomová práce

Autor: **Bc. Libuše Kormaníková**

Vedoucí práce: **doc. Mgr. Dan Faltýnek, PhD.**

Olomouc

2023

Na tomto místě bych ráda poděkovala vedoucímu mé diplomové práce, doc. Mgr. Danu Faltýnkovi, PhD. za uvedení do této zajímavé problematiky, čas a předané zkušenosti. Poděkování patří i Mgr. Vladimíru Matlachovi, PhD. za statistické konzultace. Dále děkuji Zitce za to, že toho tolik četla, Věrce za její pražský přízvuk a Pepovi, že každé „hele a šlo by“ mění ve skutečnost. Můj velký dík patří rodině, která mě podporovala ve všech mých rozmarech a ostatním přátelům, již mi byli při studiu oporou.

Místopřísežně prohlašuji, že jsem magisterskou diplomovou práci na téma: „Analýza hlubokého a povrchového sentimentu v autorských textech: případové studie“ vypracovala samostatně pod odborným dohledem vedoucího diplomové práce a uvedla jsem všechny použité podklady a literaturu.

V Olomouci dne 11.5.2022 Podpis .....

OBSAH

ÚVOD

'A slow sort of country!' said the Queen. Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!'

- Lewis Carroll, Alice Through the Looking Glass, s. 33

Každý autor na světě má svůj vlastní styl psaní, kterým se odlišuje od druhých a dává nám tím možnost nahlédnout do svého užívání jazyka. Jinými slovy, dává se nám poznat. Lišit se nemusí jen typografií, stavbou vět, ale i různými jazykovými znaky, včetně volby slov. Někteří autoři používají určitá slova nebo slovní spojení častěji než jiní, a i když mohou o svých „oblíbených" slovech vědět a maskovat je či se je snažit změnit, ta, která používají nevědomky, neutají.

Jednou z možných metod určování autorství, je i lexikální analýza, která zkoumá výběr slov a slovních spojení použitých v textu za účelem identifikace autora. Obsahová slova, tedy ta, která nesou význam, se však dají snadno odhalit, nahradit a neříkají nám tolik o autorovi jako spíše o jeho komunikačním záměru a snaze strukturovat text. Mnohem těžší je změnit způsob užívání slov funkčních, což jsou například předložky nebo zájmena. Ty autor již tolik neusměrňuje a poskytuje tím vhled do svého osobního stylu, například do své syntaktické struktury, preference určitých gramatických forem atd. Jinak řečeno, autoři mohou mít soubor svých oblíbených slov, která v textech často používají a která vytváří jejich specifický autorský styl, je však nutné podotknout, že přiřazení autorství pouze na základě volby slov nemusí být vždy spolehlivé, protože při utváření stylu autora hrají roli i další faktory, jako je třeba téma a kontext. Co když jsou ale oblíbená slova zároveň obsahová, a přesto nezávislá na kontextu?

Autorský sentiment, tedy textová analýza zaměřená na identifikaci a kvantifikaci emocionálního tónu v textu, se v poslední době stala důležitým tématem. Jazyk může poskytnout vhled do myšlenek, emocí a chování jedince a odhalit tak jeho citlivá témata, aniž by dotyčný sám chtěl. Sentiment můžeme zkoumat jak na povrchu, přes přímé výroky, tak hluboko v uspořádání textu a stejně dobře k analýze můžeme využít nízko frekventované lexikum oproti vysoko frekventovanému. Porovnávání těchto dvou úrovní nám může přinést poznatky nejen o autorském stylu, ale i o jeho záměrech a dost možná pomůže odhalit nesrovnalosti, které mohou být důsledkem snahy stylizovat text určitým směrem, ač pravda

leží na opačné straně. I když se to nemusí zdát pravděpodobné, každý z nás má slova, která nelžou.

Cílem této práce je podrobněji představit novou metodu, která vznikla na půdě Katedry obecné lingvistiky Univerzity Palackého a která přináší vhled do analýzy autorského sentimentu a poskytuje nám informace o autorovi, jež bychom jinde jen těžko získali. V práci budou mimo jiné také prezentovány experimenty provedené na případových textech, které ověřují účinnost navržené metody a porovnání s některými existujícími metodami.

TEORETICKÁ ČÁST

1 POVRCHOVÝ A HLUBOKÝ SENTIMENT

V této kapitole se budeme věnovat dvěma typům autorského sentimentu. První z nich je dlouho a široce znám i používán v různých analýzách především internetového prostředí, jedná se o sentiment povrchový. Druhý je nazýván hluboký a tento pojem najdeme v několika významech, přičemž jeden význam vzešel z jazykových analýz speciálního typu nízko frekventovaného lexika, které je hlavní náplní této práce. Důležitější pro nás tedy bude právě tento smysl. Musíme ovšem zmínit, že adjektiva „povrchový" a „hluboký" používáme především k odlišení způsobu, jakým byl sentiment textu získán a nejedná se tedy o specifické pojmy snadno nalezitelné ve všech ostatních literaturách, jelikož v nich je tento typ analýzy nazýván pouze jako „sentimentová".

1.1 Povrchový sentiment

Jak bylo zmíněno, tento typ sentimentu (také někdy nazývaný „postojová analýza"; viz Veselovská, 2017) je dobře znám a využívá se především na textech získaných ze sociálních sítí, blogů, recenzí a dalších. Jedná se o proces zpracování přirozeného jazyka, který se používá k určení nálady, postojů nebo emocionálního významu textového obsahu, jelikož způsob, jakým používáme jazyk k vyjádření svých myšlenek a pocitů, ukazuje také naše osobnostní charakteristiky (Godsay, 2015). Jeho nejčastější využití najdeme v oblasti marketingu a podnikání, kde se hodnotí názory zákazníků na daný produkt nebo službu. Sentimentová analýza také umožňuje firmám sledovat, jak se o ní v online prostředí hovoří (Godsay, 2015) a získávat zpětnou vazbu od zákazníků a predikovat trendy (Veselovská, 2017). Kromě politického a podnikatelského využití, se dá se sentimentovou analýzou pracovat i v oblasti sociologie nebo psychologie, například na úrovni analyzování snů (viz studie Nadeau et al., 2006). Dalším užitkem sentimentové analýzy je možnost odhalení autora textu, což může být cenná informace obzvláště v dnešní době fake news (Martins et al., 2021). V současnosti se výzkumy ubírají směrem nejen určit sentiment daného textu, potažmo autora, ale také poznat jeho psycholingvistické chování a vlivem rozvoje internetu a umělé inteligence vznikají nové přístupy, jak této analýzy docílit i na malém rozsahu textového vzorku, jakým jsou například statusy nebo komentáře na sociálních sítích.

1.1.1 Proces analýzy

Pro vytvoření sentimentové analýzy se používají různé techniky, jako je například strojové učení, analýza frekvence slov nebo metody založené na slovnících (lexikální přístup). Tyto metody umožňují identifikovat klíčová slova nebo fráze v textu, které signalizují určitý emocionální význam. Příkladem poslední metody je software LIWC, kterému se více budeme věnovat ve třetí kapitole naší práce. Důležitým faktorem v klasifikačních technikách je určení polarity slova, tj. zda slovo či fráze je pozitivní nebo negativní. Proces analýzy však není zcela jednoduchý a může zahrnovat určité výzvy, jelikož lidé vyjadřují své názory složitým způsobem a často do svého vyjadřování zahrnují rétorické prostředky, jako je sarkasmus, ironie, implikace atd., jež je obtížné analyzovat (Godsay, 2015; Zhou, 2022). Nástroje na provedení sentimentové analýzy zahrnují buď přístup řízeného strojového učení, nebo lexikální metody (Ribeiro et al., 2016).

K sentimentové analýze za pomoci lexikální metody můžeme přistupovat několika různými nástroji. Mezi nejčastější patří lexikony emocí, jako například SentiWordNet, ANEW, WordNet Affect a EmoLex Lexicon. Poslední jmenovaný je nejnovější a oproti jinými verzím spojuje každé slovo s existencí či neexistencí základních emocí podle Plutchikovy teorie, čímž vytváří referenční rámec k analýze vět vypočítávající součet emocí pro každé slovo zvlášť a zároveň zohledňuje fakt, že slova mohou spadat do více než jedné emoce (Martins et al., 2021). SentiWordNet se vyznačuje tím, že spolupracuje s daty v síti WordNet a ten je pak propojen s WordNet Affect, jenž ke slovům připojuje emoční informace (Tabak, Evrin, 2016). ANEW poskytuje soubor normativních emočních hodnocení pro velký počet slov a společně s LIWC a WordNet Affect tvoří jedny z nejvlivnějších lexikonů (Zhang, Provost, 2019). Existuje několik rozdílů mezi těmito přístupy, z nichž jeden se týká kontextu, ve kterém byly vytvořeny. Například LIWC byl původně navržen k analýze sentimentu ve formálně psaných anglických textech, zatímco PANAS-t byl navržen jako psychometrická škála pro webové prostředí (Ribeiro et al., 2016). PANAS-t (Positive and Negative Affect Schedule) bývá často užíván k hodnocení sentimentu v datech ze sociálních médií (především Twitteru) a ukázalo se, že metoda přesně zachycuje pozitivní a negativní nalazení ohledně událostí, o kterých se píše, a lze ji použít pro velké množství dat a dokonce i pro analýzu v reálném čase (Gonçalves et al., 2013) a jedná se o jednu z nejpoužívanějších škál pro hodnocení mikroblogového sentimentu (Ribeiro et al., 2016).

Jak můžeme vidět, nástrojů pro analýzu autorského sentimentu za posledních dvacet let přibylo mnoho a rozvoj se stále nezastavil. V mnoha výzkumech se používají odlišné metody a i přes to autoři dochází k jasným závěrům. Ribeiro a kolegové (2016) ve své studii porovnávají účinnost 24 takových softwarů, jelikož se do té doby tomuto tématu nikdo zcela nevěnoval. Bylo zjištěno, že existující metody se v některých případech liší, tudíž stejný obsah může být interpretován odlišně. Většina nástrojů byla například přesnější ve správném přiřazení pozitivního textu než negativního, což značí tendenci zaujatosti analýzy vůči pozitivitě (Ribeiro et al., 2016). V rámci zjišťování vhodnosti metody pro danou analýzu se hovoří i o tom, zda některé z nich pracují lépe na delších či kratších textech, což je důležité například při práci s daty získanými ze sociálních sítí.

Na základě srovnávací studie bylo zjištěno, že následující metody jsou pro správné určení sentimentu nejspolehlivější: SentiStrength, Sentiment, Semantria, OpinionLexicon, LIWC15, SO-CAL, AFINN, VADER a Umigon; v kontextu sociálních sítí byla nejlepší metodou Umigon, následovaná metodami LIWC15 a VADER (Ribeiro et al., 2016). Tabak a Evrin (2016) v taktéž provedené srovnávací studii (avšak na jiné sadě metod) přišli navíc se zjištěním, že metody, jež přiřazují slova do více kategorií jsou úspěšnější.

Z výzkumu Martinse a kolegů (2021), kteří se zabývali emocionálním profilem autora, vyplynulo, že sentimentová analýza je velmi účinným nástrojem pro rozpoznání toho, komu daný text patří. Se stejným závěrem přišli i Abbasi a kolegové (2022), kteří v analýze zahrnuli i kombinaci algoritmů strojového učení. Kromě identifikace autora nám analýza emocí může přinést i informace v podobě predikce dalšího chování či jiných aspektů, v případě výzkumu Schwartze a kolegů (2016) tomu tak bylo u well-beingu. O prediktivní sílu sentimentové analýzy se zajímal i Onan (2018) a z jeho studie vyplynulo, že propojení softwaru LIWC s „ensemble learning“ zvýší prediktivní výkon na 89,10 %. Predikcí sentimentu sentimentu uživatelů na základě analýzy jejich tweetů se v roce 2022 zabýval také Alsayat a jeho metoda zahrnovala kombinaci deep a ensemble learning pro určování sentimentu tweetů o koronviru, což vedlo ke zlepšení výkonu, jelikož se model „naučil“ rozpoznávat nejen kontext, ale i dosud neznámá slova a hodnotit nové prefixy a sufixy. Svou prací tak volně navázal na dílo Jelodara a kolegů (2020), kteří taktéž v rámci zkoumání sentimentu v tweetech ohledně covidu-19 přistoupili k analýze pomocí hlubokého učení a výpočetních technik. Tato metoda patří mezi nejnovější a s ohledem na současný rozvoj umělé inteligence působí slibně co se budoucnosti nástrojů určených pro sentimentovou analýzu týče. V několika zmíněných výzkumech (viz například Alsayat,

2022; Jelodar et al., 2020) byla využita metoda LSTM, což znamená dlouhá krátkodobá paměť a tento nástroj je využíván i pro výzkum hlubokého sentimentu, jak ho definují někteří autoři.

1.2 Hluboký sentiment

V pojetí ostatních výzkumníků se tento pojem začal objevovat již kolem roku 2013 a jedna ze studií, která se mu v té době věnovala, byla z díla Haeng-Jin Jang a kolegů a zabývala se vytěžováním příčin mezi osobností a postojem pro analýzu inzerátů v sociálních médiích. V jejich pojetí jde o třívrstvý (tedy hloubkový) přístup k analýze postojů zákazníků, kdežto většina soudobých nástrojů poskytovala pouze jedno až dvouvrstvé struktury sentimentu charakterizované polaritou a/nebo kategorizované podle osobního profilu. Mask a Vossen (2012) také ve svém článku hovoří o hlubokém sentimentu, avšak nevysvětlují, jak přesně se liší od do té chvíle proběhlého zkoumání. Hloubková analýza v jejich uchopení je zřejmě metoda propojení vztahů mezi adjektivy, verby a dalšími slovními druhy za účelem získání představy o postoji autora. Většina studií z pozdějších let se zabývá vztahem LSTM a hlubokého sentimentu. Minaee a kolegové (2019) představili model založený na souboru neuronových sítí s LSTM a konvolučních neuronových sítí (CNN), z nichž jedna zachycuje časovou informaci dat a druhá extrahuje lokální strukturu. Tento ensemble model se ukázal jako efektivní pro přesnou analýzu postojů. Se shodným přístupem přišel o chvíli dříve i Huang a kolektiv (2017), podle nichž se tímto způsobem hluboký sentiment nejlépe manifestuje. Taktéž Dong a Melo (2018) jako „deep sentiment analysis“ uvádí analýzu založenou na hloubkových neuronových sítích spíše než by vymezili tento typ sentimentu od jiného. Je však nutné podotknout, že pokud tento pojem překládáme do češtiny, můžeme dostat kromě „analýzy hlubokého sentimentu“ také „hlubokou analýzu sentimentu“, což by se lišilo od našeho pojetí.

V této práci budeme k hlubokému sentimentu přistupovat jinak, než autoři zmíněných studií. Zde tento pojem vlastně zahrnuje vytažení zhruba 10 % textu z celého souboru od jednoho autora. Tato část je vybrána z okolí nízko frekventovaných autosémantik, ke kterým docházíme prostřednictvím analýzy nastíněné ve třetí kapitole. Hluboký sentiment podle našeho názoru vykresluje postoje a psychologický profil jedince lépe než povrchový i vzhledem k tomu, že se jedná o neuvědomované struktury, které ve svém jazyku používáme (Faltýnek et al., 2022). Popisem získání tohoto sentimentu se zabýváme v dalších částech této práce.

2 NÍZKO FREKVENTOVANÉ LEXIKUM

Zde se zaměříme na nízko frekventované lexikum, o kterém byla již několikrát zmínka a vysvětlíme podstatu hapaxů a dalších pojmů, které se k tomuto fenoménu vztahují. Dále představíme nově vytvořený pojem „superhapax“ a způsob, jakým vznikl, a budeme tudíž především vycházet ze studií, které vedly k vytvoření nové metody pro určování nejen autorského sentimentu, ale i autorství samotného a jsou relevantní pro tuto práci.

Pojem nízko frekventované lexikum v sobě zahrnuje ta slova, která se v textu neobjevují tak často jako jiná a přesto mají své specifické vlastnosti, mimo jiné třeba slepovat text, aby držel pohromadě (Faltýnek, Kučera, 2022). Může se jednat o slova typická pro téma, určité slovní druhy, ale také o ojedinělé výskyty lexika, které autor nepoužívá tak často a přesto ho odhalují. Kromě samostatných slov se autor na nízkých vlnách projevuje i v jejich kombinacích – frázích (Faltýnek, 2020). Když hovoříme o nízké frekvenci, máme na mysli nejen slova s jediným výskytem v textu, což jsou hapax legomena, ale také ta, která se v daném korpusu objeví vícekrát, ty se nazývají dis legomena a tris legomena. Dále se již nepočítá, jelikož vícenásobný výskyt není pro etymologii tak relevantní ve smyslu hapaxu (Hladká et al., 2017). Je důležité zmínit, že nízko frekventované lexikum je považováno za neuvědomované, a to buď z pohledu hapaxů (Baayen et al., 1996) nebo funkčních slov (Binongo, 2003). U nízko frekventované lexika nepředpokládáme opakování slovních forem, jež by souvisely s tématem literárního díla nebo jeho stylem, ale usuzujeme na autorskou volbu synonymních výrazů (Faltýnek, 2020). Právě pozice autora je klíčovým prvkem pro analýzu jazykového chování v nízkých frekvencích, jak reflektují Faltýnek a Kučera (2022, oddíl Poznámka na závěr): „*Vrátíme-li do hry autora textu, kterému humanitní disciplíny už před půlstoletím vyhlásily smrt, ačkoliv on je původcem strukturace nízko frekventovaných jevů, čeká na nás zbylých několik desítek procent textu s jeho strukturními vlastnostmi, dosud neznámými.*“

I když je v současné době standardní přistupovat k analýze textu na základě vysoké frekvence (například při identifikaci autora) a zajišťovat tak soulad s jazykovou normativitou, neznamená to, že neexistují přístupy zaměřené na periferní jevy a nízkofrekvenční položky jsou cennou součástí korpusů (Faltýnek et al., 2022). Ukazuje se,

že nízko frekventované lexikum je v textu rovnoměrně rozloženo (Faltýnek, Matlach, 2021) a má ustálenou podobu, která může sloužit jakožto autorský profil (Faltýnek et al., 2022).

2.1 Hapax legomenon

Hapax legomenon je termín, který se používá pro označení slova, které se v daném jazyce, korpusu či textu vyskytuje pouze jednou (Cvrček, 2017). Jedná se tedy o slovo, které je v daném kontextu unikátní a nemá žádné opakování. Často mezi hapaxy můžeme nalézt prosté překlepy, ale ze záměrných užití se zde zároveň nachází mnoho autorských okazionalismů, což jsou slova utvořená vědomě v procesu psaní textu zpravidla určená pro jedno konkrétní použití z důvodu kontextu a ukazují na jazykovou kreativitu (Martincová, 2017). Zároveň mezi nimi v daných textech můžeme nalézt cizí slova, toponyma, vlastní jména, zkratky nebo čísla. Tento jev můžeme pozorovat hlavně na velkých korpusech, ale v běžném životě, kde produkujeme texty v podobě rozhovorů, emailů atd., mají hapaxy spíše povahu obyčejné slovní zásoby (Faltýnek, Kučera, 2022). Korpusová lingvistika tento pojem vtahuje k relativní frekvenci a uvádí termín „semihapax“, což jsou formy vyskytující se v daném korpusu jen několikrát (Hladká et al., 2017).

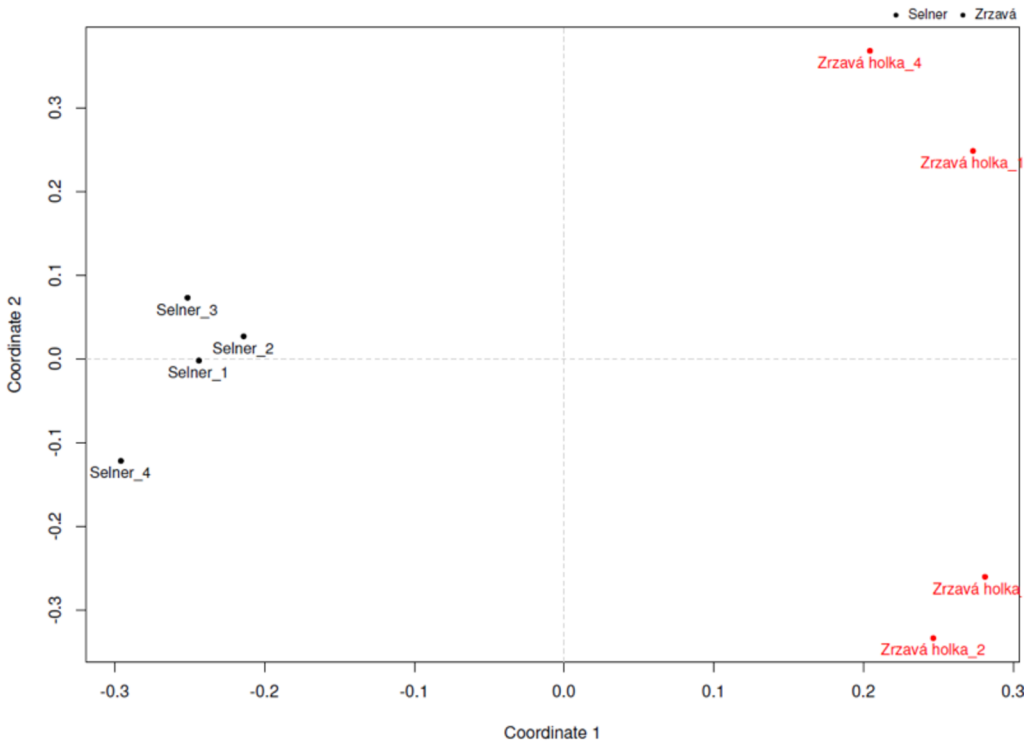
Z hlediska zkoumání jejich užitečnosti v lingvistice proti sobě stojí dva tábory. Na jedné straně slyšíme názory, že kvůli jejich jedinečnému výskytu není možné vytvářet obecné závěry o jejich vlastnostech a využití je tedy sporné (Cvrček, 2017), kdežto oproti tomu stojí výzkumy, které dokazují jejich hodnotu ve forenzní lingvistice v případě určování autorství (viz například Baayen et al., 1996 – který však sílu hapaxu dává do souvislosti s dalšími vlastnostmi daného textu; Faltýnek et al., 2020; Faltýnek, Matlach, 2021). Hlavní příčinou sporů je polemika o obsahových a funkčních slovech, jelikož je bráno za to, že plnovýznamová slova autora spolehlivě nerozeznají, což jak se ukazuje, ale neplatí pro případ hapaxů, jelikož je autor zřejmě opakuje tematicky nezávisle (Faltýnek et al., 2020) jak ukážeme v další části.

2.1.1 Superhapax

V superhapaxech se možná skrývá více, než si myslíme. Jde o slova, kterým se autor prostě nevyhne, jsou svým způsobem jeho oblíbená, byť málo v textu užívaná. Autor je zároveň opakuje pravidelně, nevědomě a v určitých kontextech. Právě z toho důvodu nám

mohou poodkrýt, jaké má autor postoje, komunikační strategie nebo co může prožívat – a proto jsou také důležité, i když jen pro konkrétního autora textu.

Superhapax je slovo nebo slovní spojení vyskytující se vždy jednou v určitých textech nebo vzorcích textu konkrétního autora, ale nikoliv pouze jednou v jeho celém korpusu, jelikož se v textech opakuje s pravidelností. Jelikož se jedná o slovní formu identifikující texty s určitým autorem, nevyskytuje se u srovnávaných autorů v podobném zastoupení (Faltýnek, Kormaníková, v tisku). Co se týče rozložení obsahových a funkčních slov v nízké frekvenci, zatímco většina předchozích přístupů k rozpoznání autora je orientována směrem funkčních slov, ukázalo se, že v nízkých frekvencích s opakováním autor nejen, že volí podobné způsoby výstavby textu, ale s nízkou frekvencí opakovaně pomocí obsahových slov tematizuje určité skutečnosti (Faltýnek, Kučera, 2022). Autorský text je komponován na základě volby mezi synonymními jazykovými prostředky a tak se stává, že se lidé opakují (Faltýnek, 2020). Že superhapaxy mohou velice dobře odlišovat autory textů i případě, že tématika je podobná, jsme ukázali na analýze dvou blogerů – Martina Selnera (Autismus a Chardonnay) a Natálie Ficenové (Zrzavá holka), kteří se ve svých blogových příspěvcích věnovali autismu (Faltýnek, Kormaníková, v tisku). Níže přikládáme graf (Obr. 1) z provedené analýzy superhapaxů. Předpoklad, že hapax legomena dokáží klastrovat autora textu jsme dále ověřili na základě náhodného výběru 6250 slov z jejich vzorků a ve všech případech náhodných výběrů byly vzorky jednoznačně separovány.



Obr. 1: Klastrování vzorků blogu Autismus a Chardonnay Martina Selnera (dostuné z: http://selner84.blogspot.com/; navštíveno 15. 12. 2021) a blogu Zrzavá holka (dostuné z: https://zrzi.cz/2015/12/nebinarita/; navštíveno 15. 12. 2021). Klastrování bylo provedeno na základě hapax legomen přítomných ve vzorcích srovnaných kosinovou nepodobností, zobrazení bylo provedeno formou vícerozměrného škálování (Torgerson, 1952). Převzato z: (Faltýnek, Kormaníková, v tisku).

Z hlediska identifikace autorství se předpokládá, že k jistému odhalení autora práce je zapotřebí 2500 hapax legomen (Faltýnek, Matlach, 2021), a protože podle některých autorů (viz např. Fenxiang, 2010) je jejich obsah v textu kolem 40 %, docházíme k číslu 6250 slov, které značí rozsah celého textu (Faltýnek et al., 2022). Jelikož nárok na superhapax je, že se v textu opakuje alespoň dvakrát, potřebujeme k reliabilní analýze vzorek dvakrát větší. Prozatím není jasně vyhraněno, jak velké, potažmo malé, mohou jednotlivé vzorky víc, ale obecně platí, že delší text přispívá ke stabilnějším jevům.

„Autorské superhapaxy mají velmi proměnlivou podobu v závislosti na délce textu a na počtu srovnávaných textů. Při zvětšování délky analyzovaného textu sice roste počet superhapaxů, některé ze superhapaxů, …, naopak získávají vyšší frekvenci a do autorských profilů rozsáhlejších vzorků nevstupují.“ (Faltýnek, Kučera, 2022, oddíl Diskuze: Rozsah autorsky vázané nízko frekventované slovní zásoby). Nelze však kvůli tomu odmítnout jejich význam, jelikož i tak se jedná o nízko frekventované lexikum, které se jako hapax projeví za určitých podmínek, a každopádně s sebou nese část autorovy osobnosti.

Faltýnek, Benešová a Kučera (2022) ve své studii, která předkládá využití nízko frekventovaného lexika a zejména superhapaxů, předpokládají, že ve srovnání se zbytkem lexika mají superhapaxy specifické postavení, jelikož značí odlišné mentální zpracování částí textu. Zároveň zmiňují, že tyto části textu reprezentují autorovy osobnostní charakteristiky, uchopení okolního světa a prožívání, které do svých děl zahrnuje nevědomě. To znamená, že informace obsažená v superhapaxu není jen ve slově samotném, ale je rozptýlena do okolní výstavby věty a tento efekt je nazýván „Bruntal effect" a pokud se v textu vyskytne místo, kde se superhapaxy nakupí, je vhodné mu při interpretaci věnovat zvýšenou pozornost (Faltýnek et al., 2022).

Nyní se již přesuneme k metodologii práce a případovým studiím, na kterých bude použití sentimentu i nízko frekventovaného lexika více osvětleno.

VÝZKUMNÁ ČÁST

3 PRÁCE S TEXTY

Nyní souhrnně popíšeme, jak přesně jsme postupovali při analýzách získaných textů, jelikož až na pár odlišností (jež jsou zmíněny u konkrétních případů) se jedná o tentýž postup. Také se pokusíme přiblížit programy, které byly využity. V další podkapitole již uvádíme jednotlivé případové studie, na kterých byly provedeny analýzy. Každá z analýz se v něčem odlišuje, a to sice buď v kombinaci metod či v přidání nových, v závislosti na tom, jak přesně jsme zamýšleli na daném vzorku naši metodu otestovat.

3.1 Postup analýzy

Tato metoda je založena na komparaci hapaxů, které se objevují v několika textech od stejného autora. Rozpoznání opakujícího se nízko frekventovaného lexika vyžaduje srovnání nejméně dvou textů jednoho autora (Faltýnek, Kormaníková, v tisku). V první řadě je nutné mít k dispozici texty dostatečné délky a v případě, kdy máme pouze jedno dílo, i to musí být náležité obsáhlé, aby se případně dalo rozdělit do minimálně tří přibližně stejně velkých částí (viz výše v podkapitole Superhapax). K dělení můžeme využít přirozeného konce textu (např. následující kapitola knihy; konec věty), nebo se zaměřit na přesné rozdělení vzorků podle počtu slov. Prozatím nebyl vliv dělení validován (Faltýnek et al., 2022), ale v poslední případové studii se k němu vyjádříme. V některých případech se stane, že nemáme dostatečně dlouhý souvislý text od jedince, což je případ třeba konverzací skrze sociální sítě, a proto v tomto případě nejprve slijeme získaná data do jednoho souboru, který až následně dělíme na vzorky. Samozřejmě v tomto případě delší texty poskytují konzistentnější výsledky, a proto jsme k analýzám v této práci zvolili texty o celkovém rozsahu nejméně 8 000 slov.

Text od jednoho autora tedy nejprve rozdělíme na stejné části. Počet částí se různí, zde pracujeme s rozsahem dělení na 3-10 částí, a to sice z důvodů možných změn, které se mohou potenciálně objevit a samozřejmě i kvůli délce celkového textu. Poté z každé části vyjmeme hapaxy (zde pomocí programu QUITA, a díky jeho možnostem můžeme rychle provést i další analýzy, například zhodnocení frekvencí jednotlivých slov či srovnávat texty různých autorů). Následně přeneseme výsledky do tabulky v Excelu a zjistíme, které hapaxy se opakují ve všech nebo velké většině částí. Tím získáváme superhapaxy. Nyní stanovujeme

cut-off score – tedy hodnotu, jež určuje jak velký rozsah budeme akceptovat, což je v tomto případě počet částí, ve kterých se superhapax objevuje ku celkovému počtu částí. Jelikož je naším zájmem vytáhnout z hloubky slova, která mají nejviditelnější vlastnost se opakovat, zavedli jsme si zde cut-off score 70 %. Pokud tedy například máme v analýze 10 částí textu a superhapax se nám objevuje v 10/10 až 7/10 částech, přijímáme ho pro následnou práci, avšak 6/10 již v tomto případě neakceptujeme.

Následuje opět přímá práce s textem. Získané superhapaxy si strojově v textu vyznačíme a zajímáme se o jejich nejbližší okolí. Může se stát, že v textu námi očekávaný počet daného slova nesedí se skutečností. Tato situace nastává zejména pokud nemáme opakování hapaxu ve všech částech. Jelikož například v částech 1 až 9 se může jednat o hapax a slovo je tam tedy pouze jednou (to jsou takzvané „jedničkové části"), v poslední části, kde slovo hapaxem není („nulová část"), mohou nastat dvě možnosti, a to buď že dané slovo se zde nevyskytuje vůbec, nebo se objevuje dvakrát a vícekrát. V tomto případě hodnotíme, zda není výskyt v těchto „nulových" částech příliš velký, a pokud ano, tento superhapax také z analýzy vyloučíme, jelikož naším cílem je najít slova s nízkou frekvencí, která se však pravidelně opakují po určitém rozsahu. Dalším krokem je zmíněná práce s okolím (tedy odkrývání zmíněného Bruntal efektu).

Při analýze okolí opět ve většině případů používáme strojových metod. Vzhledem k tomu, že v této práci na některých textech testujeme vliv rozsahu okolí na výsledky, bereme okolí v několika podobách. Širší zahrnuje zhruba 39 slov celkem (tedy 20 před superhapaxem, 19 za ním) a jevilo se nám jako nejvhodnější pro následující analýzy, kratší je pak o polovinu menší (19 slov celkem). V jednom případě pracujeme i s dvakrát širším okolím, tedy dohromady 79 slov a k tomuto rozhodnutí jsme došli z hlediska záměru porovnat wordcloudy okolí v několika formách. Samozřejmě pokud se superhapax vyskytuje na začátku nebo konci textu, jeho okolí se liší. V případě rozhovorů zohledňujeme i přerušení otázkou druhé osoby. Tyto rozsahy používáme především pro analýzu v programu LIWC, o kterém se blíže vyjádříme dále, a pro frekvenční analýzu slov. Díky analýze v tomto programu dokážeme získat autorský sentiment, a to na dvou úrovních – pokud použijeme celý rozsah textu, jedná se o „povrchový" sentiment, o němž byla řeč v teoretické části, ovšem když analyzujeme okolí, dostáváme se do hlubokých struktur. V některých případech nás více zajímají data získaná z close-readingu, což je metoda čtení, interpretace a hodnocení témat vyskytujících se kolem superhapaxů a na rozdíl od

předchozího kroku není třeba zde striktně dodržovat rozsah slov, ale spíše se zaměřit na změny, které se tam odehrávají.

V této chvíli se nám již otevírají možnosti, jak dále s analýzou pokračovat. Je možné buď vytvořit wordcloudy, které graficky zvýrazní podstatná slova z okolí, dále se můžeme zabývat frekvenční analýzou přilehlých slov a zjišťovat, zda se vyskytují pouze v nízké frekvenci nebo superhapaxy na jejich počet či rozmístění nemají vliv. Vzhledem k psychologické povaze jazyka se také naskytuje možnost porovnání výsledků s jinými psychometrickými metodami zaměřenými na osobnost. V neposlední řadě se lze zaměřit na ověření pozitivních nebo negativních emocí, které vyvolávají dané superhapaxy jedince.

3.2 LIWC (Linguistic Inquiry and Word Count)

Vzhledem k povaze práce s textem, tedy nutnosti ověřit náboj hlubokého sentimentu, jsme při analýzách (v některých případech) používali také program LIWC, což je nástroj pro analýzu jazyka a textu, který se používá k identifikaci a kategorizaci slov v textu podle jejich vlastností a následně jim přiřazuje psychologický význam. Nyní stručně vysvětlíme jeho vznik a podstatu. Tento program vznikl na základě volání po systému, který by zefektivnil a sjednotil hodnocení textů, jelikož do té doby se zkoumání psychologické stránky textu věnovali pouze proškolení jedinci a i ti se ve svých závěrech lišili, navíc se tento způsob ukázal jako neekonomický. Za vznikem programu v průběhu 90. let stojí americký profesor James Pennebaker, Booth s Marthou Francis a momentálně je nejnovější verze LIWC z roku 2022, avšak analýzy v této práci jsou provedeny na předchozím vydání z roku 2015.

Program byl vyvinut na univerzitě v Austinu a je používán v různých oblastech výzkumu, včetně psychologie, sociologie, lingvistiky a managementu (Nichols, n.d.). Rozpoznává slova a fráze, které spadají do různých kategorií (jichž je přes 80), jako jsou emoce, osobnostní rysy, sociální a kognitivní procesy, autenticita a další, zaměřené i více na lingvistickou podstatu textu, například pomocná slovesa, zájmena a další. Je důležité zmínit, že mnoho kategorií LIWC je uspořádáno v hierarchické struktuře a stejné slovo může být zařazeno do více slovníků (Tausczik, Pennebaker, 2010), například slovo „celebrate“ je jak ve slovníku pozitivních emocí, tak ve slovníku úspěchů (LIWC, n.d.). Velkou výhodou programu je, že také umožňuje analýzu velkého množství textu v krátkém čase a poskytuje informace o skrytých emocionálních a psychologických tendencích v textu. Je možné s jeho pomocí analyzovat nejen dlouhé monografie, ale i příspěvky na sociálních sítích, poezii či

e-maily a další. Slova jsou analyzována pomocí zabudovaných slovníků, což jsou jedny za dvou základních komponent (Tausczik, Pennebaker, 2010). Bohužel ještě není vytvořena verze pro češtinu a tudíž naše texty musíme překládat a až poté analyzovat v softwaru. Dostupné jsou slovníky například pro francouzštinu, italštinu nebo srbštinu (Boyd et al., 2022). V současné verzi z roku 2022 je přes 100 slovníků, které se skládají ze seznamu slov, slovních kmenů, emotikonů a dalších specifických slovních konstrukcí, které byly identifikovány tak, aby odrážely danou kategorii. Například slovník pozitivních emocí obsahuje slova jako „happy“ a „love“, zatímco slovník afiliace obsahuje slova jako „community“ a „together“ (LIWC, n.d.). Počet slov, které do dané kategorie z celkového analyzovaného textu spadnou, se následně přepočítá na procenta. Tento program se běžně využívá ve výzkumech, například ke zjišťování významu v nejrůznějších experimentálních prostředích, včetně určování zaměření pozornosti, emocionality, sociálních vztahů, stylů myšlení a individuálních rozdílů (Tausczik, Pennebaker, 2010). Díky desítkám let empirického výzkumu je možné odhalovat i méně zřejmé vztahy mezi slovy a jejich psychologickými implikacemi. Například lidé, kteří jsou sebevědomější a mají vyšší společenské postavení, mají také tendenci používat slova „you“ v relativně vysoké míře a slova na „I“ v nízké (LIWC, n.d.). Každopádně ani LIWC zatím není neomylný, někdy se může dopustit chyb při identifikaci a započítání jednotlivých slov. Pokud si vezmeme třeba slovo „mad“, které je započítáno ve slovníku hněvu a opravdu může znamenat naštvání, nicméně to nevylučuje možnost, že se objevuje v jiném kontextu „mad about you“, kde je pozitivní, nebo „mad as a hatter“ v případě bláznovství (LIWC, n.d.). Na odstranění tohoto nedostatku se pracuje za využití pravděpodobnostních modelů.

Kategorie LIWC, které v naší práci budeme používat nejvíce jsou: *Analytic, Clout, Authentic, Tone*. Což jsou také první největší ze všech. Dále nás budou zajímat informace obsažené v sixltr (slova delší šesti písmen), posemo (pozitivní emoce), negemo, i, we, you, they. A v neposlední řadě další kategorie, které souvisí s psychologickými procesy, jako třeba **affect** („happy“), **social** („talk“, „they“), **drives** („success“, „danger“, „prize“), **anger**, **cogproc** (kognitivní procesy, např. „think“) (Pennebaker et al., 2015). Kategorie *Analytic* souvisí s tím, jak moc se v textu objevuje formální a logické myšlení v podobě slov jako „justify“ a podobně. *Clout* odráží jazyk leadershipu a souvisí se statusem. *Authentic* zahrnuje upřímnost (např. „honestly“) a *Tone* je zkratka pro emocionální náboj textu a zrcadlí procento negativních či pozitivních emocí (Boyd et al., 2022). Je nutné podotknout, že zde uvádíme příklady z verze LIWC-15, se kterou jsme pracovali, vzhledem k tomu, že u

nejnovější vydání programu má jiné psychometriky a rozdělení některých kategorií. Hlavní změnou v nejnovějším LIWCu je přepracování celkové struktury slovníku rozdělením kategorií na „základní“ a „rozšířené“. Základní slovník zahrnuje většinu již definovaných dimenzí z dřívějších verzí LIWCu, kdežto rozšířený slovník obsahuje značně aktualizované verze kategorií a navíc zavádí řadu nových kategorií i proměnných (Boyd et al., 2022).

Abychom shrnuli využití programu pro naši práci, s jeho pomocí ověřujeme a porovnáváme vlastnosti celého textu versus vytaženého okolí superhapaxů ve zmíněných kategoriích, jelikož okolí superhapaxů dle našeho názoru lépe a zřetelněji odhaluje motivy jedince a způsob, jakým užívá jazyk a zajímá nás jeho psychologický profil. Pro tento program jsme se rozhodli nejen z důvodu dostupnosti, ale také pro jeho širokého využití i v oblasti psychologického profilování autora. Naše rozhodnutí podporují proběhlé výzkumy (viz například Ribeiro et al., 2016; Onan, 2018; Borchers et al., 2021), které tento software zahrnuly do skupiny nejspolehlivějších na základě mnoha charakteristik, a to včetně rizika v podobě malého množství textu.

3.3 Etické hledisko a ochrana soukromí

Z etického hlediska zhodnotíme přístupnost textů a souhlas autorů. Autoři, kteří jsou uvedeni pod pravým jménem, sdíleli své texty popřípadě rozhovory v prostředí internetu, tudíž jsou jejich výtvory volně dostupné každému a je možné na nich provádět analýzu sentimentu bez speciálního souhlasu. Druhou skupinou jsou autoři, kteří poskytli vlastní texty či rozhovory výzkumníkům se souhlasem následujících analýz a jsou tím pádem z důvody ochrany osobních údajů anonymizováni. Z hlediska obsahu zde neuvádíme plná znění textů a na přání některých autorů ukazujeme pouze vybrané pasáže pro ilustraci. Vzhledem k podstatě práce, která se přímo nezabývá určením autorství nýbrž analýzou autorského sentimentu, není třeba uvádět pravá jména a identifikovat osoby.

4 PŘÍPADOVÁ STUDIE I: M. SELNER – AUTISMUS A CHARDONNAY

Autismus a Chardonnay popisuje pracovní zážitky asistenta autistů. Byť je dílo známější jako kniha, původně se jednalo o jednotlivé příběhy, které byly nezávisle na sobě publikovány v blogu. Z těchto textů v článku vycházíme. Příběhy Autismu a Chardonnay (Selner, 2016 - 2019) zachycují náročnost povolání asistenta z pohledu zaměstnance ústavu. Nejčastěji se zde setkáme s komentáři o vztahu autora k dětem a k činnostem běžného dne, nebo taky s líčením zážitků vybočujících ze všednodenní rutiny. Kromě obecných popisů se autor ze svého pohledu vyjadřuje k určitým skutečnostem, které se týkají života dětí s postižením a vlivu jejich specifik na jejich rodinu a okolí. Podrobněji bude obsah rozveden v následující části analýzy, která se soustředí na některé opakované autorovy formulace poodkrývající jeho postoje a prožívání práce s autistickými dětmi. Díky povaze textu jakožto needitovaného a plně autorského výtvoru můžeme ukázat výskyt opakujících se slov s nízkou frekvencí jakožto nevědomých složek textu, z čehož lze usuzovat na obsahové rysy, které jsou danému autorovi vlastní a jsou zároveň nevědomě projevované.

Jak bylo zmíněno, superhapaxy jsou slova, která se v textu autora vyskytují opakovaně, avšak pouze jednou v určitém úseku. Pokud například sebraný text rozřežeme na 10 stejně velkých částí, jsou pro nás podstatná ta slova, která se objevují jen jednou v daných řezech a zároveň se vyskytují ve více než jednom řezu. My jsme pro analýzu stanovili cut-off score 70 %, tzn. že bereme v potaz pouze hapaxy, které se vyskytují nejméně v sedmi z deseti částí Selnerova textu. Jelikož celkový sebraný text byl co do rozsahu dostatečně velký (zhruba 30 000 slov), rozhodli jsme se pro řezy na několik různých částí (v rozsahu od 2 do 10 částí) a ze superhapaxů získaných z těchto řezů jsme poté udělali průnik. Navazujeme tak na zmínku o různosti superhapaxů v případě změny velikosti vzorků (viz Faltýnek, Kučera, 2022), jak jsme uvedli zde na straně 14. Řezy, které se v některých superhapaxech prolínaly, byly řezy na 6, 9 a 10 částí textu. Slova, která byla obsažena alespoň ve dvou z výsledků různých řezů s určenou sedmdesáti procentní hranicí zastoupení, jsme použili pro následující rozbor. Po vybrání superhapaxů, kterých bylo celkem 6, jsme se vrátili v analýze zpět do kompletního autorova textu a zkoumali jsme okolí vybraných slov. Pro porovnání jsme si vyznačili okolí před a za včetně superhapaxu na 39 a 79 slov. Zároveň

jsme se rozhodli vybrané superhapaxy analyzovat ve všech výskytech v textu, tzn. ačkoliv nebyly v určitém vzorku hapaxem, jelikož slova prošla při více než jednom výběru a záleží nám na nízkofrekvenčních slovech obecně. Nejdůležitějším krokem bylo poté vrátit se do textu a z úseků kolem superhapaxů „číst mezi řádky". Pro tyto účely jsme okolí superhapaxů důkladně kategorizovali z hlediska témat, která se zde vyskytují. Chtěli jsme se dotázat na to, zda nízko frekventované, ale pravidelně opakované slovní formy typické pro autora, nemají zároveň své typické okolí a neprozradí nám něco z autorova jazykového podvědomí. Jelikož se jedná o jednu z prvních analýz v rámci této diplomové práce, práce na textu nám poskytla odrazový můstek pro práci na dalších textech a povědomí o tom, jak s okolím a řezáním zacházet.

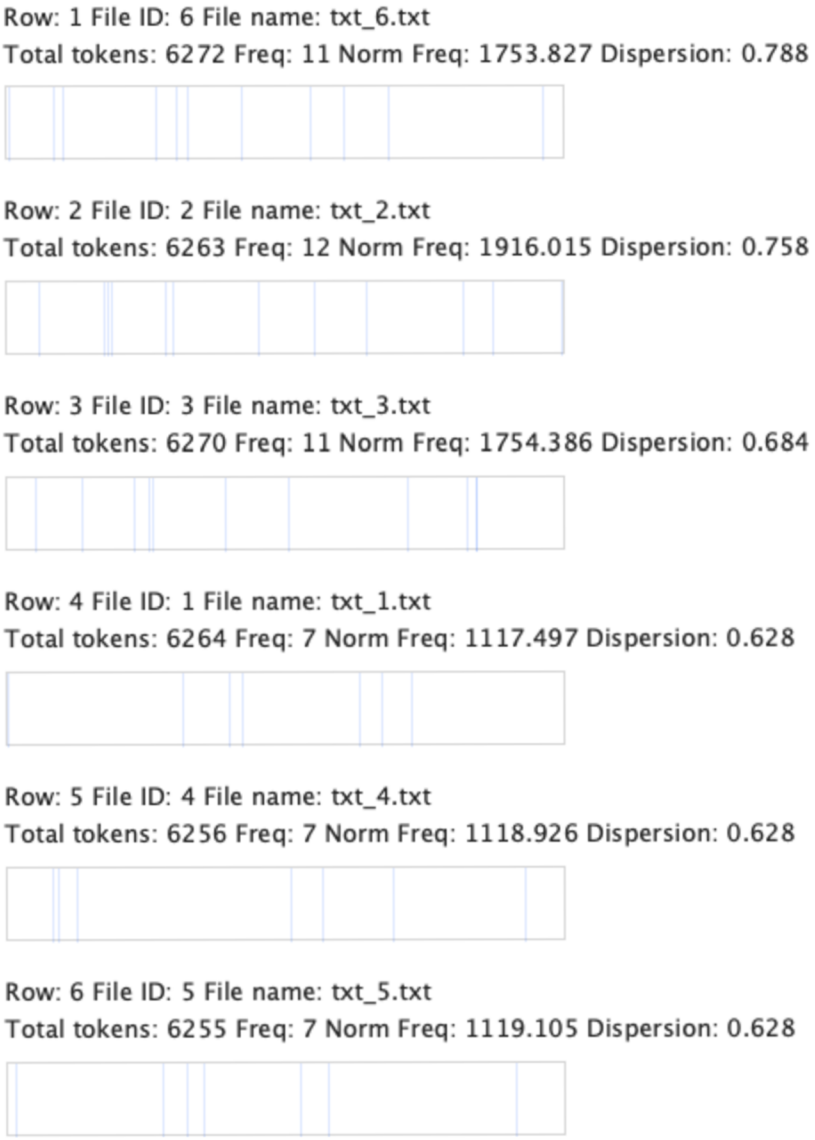
Pro představu uvedeme vysokofrekvenční slova z analýzy celého textu. Uvádíme výběr slov, která se vyskytla na prvních 100 místech ve frekvenční analýze. Na straně obsahových zde byla nejčastější: *děti* (toto slovo bylo dokonce na 10. místě), *dětí, mají, dítě, práce, práci, snažím, autismus*. Oproti tomu funkční, která zabrala přední příčky a byla zastoupena ve větší míře, byla například: *se, na, to, že, už, já, tak, protože, jim, než, něco*.

V nízké frekvenci z okolí 39 slov celkem, uvádíme následující obsahová slova: *děti, jsem, když, tak, někdy, má, kde, hledání, věci*. Z funkčních slov, která se objevila na prvních místech: *se, a, na, to, že, si, s, o, ale*. Pro představu, jak se frekvence změní, zde jsou obsahová slova ze širšího okolí superhapaxů (79 slov): *děti, jsem, jsou, někdy, ještě, stejně, autismus, kde* a funkční: *se, a, na, to, že, si, i ,v*.

Pro srovnání nejprve uvedeme celý rozsah superhapaxů ve zmíněných částech (6, 9 a 10) a poté zvýrazníme stejná slova. Superhapaxy, které jsme získali v řezech na 6 částí (dále R6) jsou: *něčím, patrně, dokážou, zkušenosti, výraz, slovo, dětství, společné, neměli, lavičce, zem, zahradu, mělo, běží, čem, velké, kluk, školy, pojď*. V řezech na 9 částí (R9): *slovo, sice, komu, jenom, jiná, přes, ním, baví, dvě, potkáváme, pohledu, ono, napadlo, neustále, životě, dospělí, zuby, svačinu, cestu, tou, slíbím, hledáme, každého, odchodu, zajímám*. A v 10 částech (R10): *jenom, neustále, dívat, slovo, sice, komu, těžší, nejde, mělo*. Superhapaxy, které se shodovaly ve dvou nebo všech případech a se kterými tudíž pracujeme: ***slovo, sice, komu, jenom, mělo, neustále***.

Na následujícím grafu (obr. 2) můžeme vidět rozvrstvení superhapaxů skrze celý text a na grafu (obr. 3) poté rozložení superhapaxů v jednotlivých vzorcích (zde n=6). Grafy byly pořízeny v programu AntConc.

Obr. 2: Distribuce superhapaxů v Selnerově díle. Modré čáry odpovídají lineárnímu umístění hapax legomen v textu.



Obr. 3: Distribuce superhapaxů ve vzorcích z díla Selnera. Graf zobrazuje umístění hapax legomen, které vzorky klastrují. Modré čáry jednotlivých vzorků odpovídají lineárnímu umístění hapax legomen v textu.

Nyní graficky pomocí wordcloudů přiblížíme dvě různě velká slovní okolí superhapaxů a porovnáme je s celkovým textem.

Obr. 4: Word cloud 100 slov vytvořený z celého rozsahu textu

Obr. 5: Word cloud 100 slov vytvořený z okolí celkem 79 slov kolem superhapaxu

Obr. 6: Word cloud 100 slov vytvořený z okolí celkem 39 slov kolem superhapaxu

Jak si můžeme všimnout, na obrázku 4 převládají především funkční slova, která nám tolik o autorově prožívání neprozradí. Nicméně čím více jdeme do hloubky, tím se nám odkrývá jeho emocionální profil. Porovnejme například slova hledání, radost, domů, vztahu, říct; která najdeme na nízkých vlnách, zatímco z celého textu na nás vyskakují děti a práce. Na následujících řádcích se tento efekt pokusíme ještě více přiblížit tzv. close-readingem, což je propojování informací z okolí superhapaxů – neboli zmiňovaný Bruntal efekt.

Prvním superhapaxem, ke kterému se blíže vyjádříme je „***slovo***“. Tato položka se vyskytla ve všech třech typech řezů zároveň, a z toho důvodu usuzujeme na důležitost témat, která se kolem vyskytují, byť jsou rozmanitá. Počet výskytů tohoto prvku je 7 a ve čtyřech případech se pojí se záporem (*během procházky nepromluví ani slovo*; *Přál možná není to správné slovo*). V jednom případě se okolí tematicky shoduje s okolím jiného superhapaxu (jenom). Jedná se o případ, kdy autor pronese vtip, ale zůstane bez odezvy: *pokusím se vtip, který přijde vtipný jenom mně*; *slovo nemůžu nechci slyšet," pokusím se o vtip. Nikdo se nesměje*. V ostatních částech textu (tedy kolem slov s vysokou frekvencí) se vtip objeví ještě třikrát, ale v žádném takovém případě se už neobjeví zmínka o tom, že se nikdo nesmál.

Dalším příkladem je slovo „***sice***“, které se v textu ze superhapaxů objevuje nejčastěji, a to 13x. Zvláštní vlastností tohoto slova je, že k sobě přitahuje ostatní superhapaxy, o kterých jsme se zmínili výše, ale nezopakovaly se nám ve více analýzách. Například: *my zase u nich. Sice to s sebou nese větší míru zranitelnosti, jinak to ale asi nejde*; *scény. Sice v podprsenkách, ale přeci. Někteří autisté se začnou hned smát. Ze zkušenosti vím*; *Sice nevím, koho někdy něco takového napadlo*; *Sice nedokáže říct, co je za den, ale ví, že dnes jede domů. „Tak pojď se mnou," odpovím mu, protože pro něj jiná varianta*; *domov sice nedokážu nahradit, ale ukážu jim, že existují i jiná místa*.

U „***neustále***“ v blízkosti objevujeme ostražitost: „*neustále sleduje*“ „*neustále rozhlížím*“ „*neustále kontroluje*“ „*dávat neustále pozor*“. Druhou nejčastější spojitostí je úsilí: *Neustále testují hranice / neustále testují vaši lásku / neustále vydávám úsilí.* Jelikož jsme zde již několikrát v teoretické části zmiňovali, že při opakování dochází k nevědomé synonymní volbě, uvádíme zde synonymum „pořád“, které autor využívá v jiných kontextech a objevuje se v textu mnohem více (46 výskytů) a preferuje ho i před „stále“, jež je v textu jen dvakrát. Také kdybychom porovnali volbu ***jenom*** x jen, zjistíme, že „jenom“ má pro autora pravděpodobně citlivější význam, jelikož „jen“, byť by se se superhapaxem zaměnit dalo, a z hlediska jazykové ekonomie by to bylo výhodné, se vyskytuje 143x.

Jedním z nejzajímavějších nositelů Bruntal efektu je „***komu***“. Obvykle se v jeho blízkosti odvolává na blíže nespecifikovanou blízkou osobu ve spojení s mluvením, radostí nebo čekáním. „*nezáleží na tom, kolik slov znáte, ale komu je říkáte*“ „*komu chceme udělat radost, určuje, jakými jsme lidmi*“ „*komu se ještě pořád vyplatí počkat si, až nebudu unavený*“ „*kde s někým, komu jsem se časem naučil věřit*“ „*že to zvládne, že to nemá komu říct*“.

Pokud bychom shrnuli témata, která se objevovala napříč celým okolím, určitě bychom si všimli, že často autor mluví o vztahu a chybění, osamocení, věcech a vysoký počet superhapaxů sousedí s některým z pojmů vyjadřující mluvení. To, že se v okolí tolik jí, není zvláštní úkaz, jelikož je to činnost velice běžná vzhledem k tématu, o kterém píše, nicméně sušenky najdeme procentuálně více v nízké frekvenci, než na povrchu. I když si toho můžeme všimnout v celém textu, z okolí je zřejmá autorova únava z práce, ale zároveň láska k dětem, a tyto dva póly se mezi sebou bijí. Zároveň i když se o asistentovi zmiňuje v textu i mimo okolí, pouze v něm je význam pomoci pro něj: „*že byste sami potřebovali asistenta.*“ „*– tedy najít si svého asistenta nejlépe v podobě servírky*“ / pro srovnání výskyt ve zbytku textu: „*Vyslovit například křestní jména asistentů nedovede*“ „*přezuvky asistenta, se kterým chce jít ven*“.

Pročítání okolí superhapaxů nám tedy přineslo informace, které sice najdeme v celém textu, ale nejsou tak zřetelné a mnohdy zapadnou mezi ostatními tématy, která vykreslují humorné scénky. Na světlo se nám tak z hloubky slov dostává autor.

5 PŘÍPADOVÁ STUDIE II: ROZHOVORY O ZDRAVÍ - 1. ČÁST

V této studii budeme rozebírat text, který pochází z rozhovorů týkajících se zdraví a to jak z pohledu životního příběhu, tak prožívání v době pandemie covid-19. Rozhovory byly provedeny v roce 2022 v Olomouci a participanti budou v rámci zachování anonymity označeni smyšlenými zkratkami. Tato první část se týká zmíněných životních příběhů a do analýzy byl vybrán jako zástupce jeden, nejdelší z nich. Na tomto textu rozebereme opět okolí superhapaxů tak, jak jsme provedli v první studii. Rozdílem od předchozího textu je forma – zatímco Selnerův text byl od počátku skutečně psaný, zde se jedná o přepsaný rozhovor. Analýza nám tedy poskytne vhled do potenciálních odlišností a nástrah mluvy od psaného jazyka. Nejprve stručně představíme práci s textem a poté rozebereme získaná data.

Z textu byly odstraněny otázky a poznámky výzkumníka provádějícího rozhovor a stejně tak označení smíchu či tápání respondenta. Poté byl získaný soubor (celkem 20 264 slov) rozdělen do tří částí po 6746, 6859 a 6651 slovech. Na stejný počet jsme se ho nejprve rozhodli nedělit kvůli zachování konzistence výpovědí oddělených otázkou výzkumníka. Toto rozhodnutí jsme následně ověřili analýzou i stejně rozdělených vzorků (po 6754 slovech) a došli jsme k závěru, že superhapaxy zůstávají naprosto stejné. V případě rozřezání textu na šest částí jsme dostali vzorky o velikostech přibližně 3 377 slov s odchylkou +/- 10 slov. V tomto textu nebudeme dělat průnik superhapaxů mezi řezy, ale zhodnotíme jak se od sebe liší slova a jejich okolí získaná rozdělením na 3 a na 6 částí, jelikož tak můžeme získat náhled na téměř nejnižší frekvenci a tu, která i když je také nízká, se v textu objevuje přeci jen častěji.

Tématem tohoto rozhovoru byl náhled na politické dění v minulém století. Respondent z největší části hovořil o svém životě přibližně od roku 1945 a zmiňoval, jak a v čem ho režim ovlivnil. Mluvil o dopadech na rodinu, svou práci a prožívání. Ve vysoké frekvenci proto mezi prvními 120 najdeme obsahová slova jako: komunista, vedoucí, říkal, lidi, měl, byly, přišel. Ta částečně reflektují téma a kontext textu. Podle frekvenční analýzy provedené na celém textu se mezi „oblíbená" funkční slova řadí: a, to, se, tak, takhle, do, von, já. Frekvenční slova zároveň zaujímala přední příčky frekvence. Analýza nízko frekvenčních slov, kterou jsme provedli pouze na okolí kolem superhapaxů (tzn. zhruba 15 % nebo 6 % textu, vzhledem k počtu řezů 3 nebo 6 – z okolí každého superhapaxu bylo do analýzy vybráno 20 slov před a 19 za ním) poskytla v případě dělení textu na tři části následující obsahová slova: přišel, řek, víte, hlad, vedoucí, říkal, dva a funkční slova: tam, páč, prostě, heleďte, jako. Při dělení na šest částí se objevila „ šel, šest, bylo, měl, říkal" jako obsahová a „ jo, že, já, tam, pro, pak" jako funkční. Pro představu, jak rozhovor vypadal, přidáváme úryvek.

„*To sem Vám chtěl říct, že to vodpovídalo politický situaci, my sme tam dostávali noviny Práce, Rudý právo, zadarmo. V novinách prošla zpráva, že Adenauer je pozvanej do Moskvy. To byl rozhodující bod v kriminále, páč i na tom Jáchymovsku ty bachaři nebyli tak blbí, páč viděli, kolik tam maj retribučáků, Němců, esesáků a normálních vojáků, takže zvláště ti esesáci budou pečlivě každej krok komentovat a vyprávět u výsleších až budou doma, že voni pudou domů a taky šli. Čili, to zacházení tam úplně změklo, ty byli posraný, ti bachaři, jo. To vedení se na tom Nikolaji se uplně změnilo, tam přišli nějakej Slánský, předtím se menoval, takovej nějakej raubíř a tendleten uplně řval na šachtě, že nás vobíraj vo prachy, že by měli dávat víc, že vod nás chtěj hodně práce a vod těch civilů ne, páč tam fáral- i civilové, jo. Uplně votočený a pochopitelně votočený i tady. Než sem tam přišel do to , tak tam bylo pár mladejch kluků ze Slovenska, vod tý Bílý legie, jo. To byli ty katolíci, nějaký ty maďarský, pravověrný, tak tam popichovali. Voni je prostě přeložili jinam, toho dozorce, kterej prostě s se nima kočkoval, toho osvěťáka, toho uplně zrušili, taky odešel jinam*

Na Jáchymově ste měl nárok na návštěvu jednou za ¾ roku. Návštěvu na ¾ roku. Tady kdybych chtěl, tak každej tejden. Ale ta ženská, do Opavy se dostat. A domů. Dopis taky. Ale co každej den psát? Kdykoliv se mě to napadlo, dopis,

Těsně, to je- skutečně kopíruje tu dobu, páč tady ty komunisti si uvědomovali, že se s ním musej nějak vyrovnat"

Superhapaxy získané ze řezů na tři části

Z programu QUITA jsme získali hapaxy obsažené ve vzorcích. Ukázalo se, že celkem 26 hapaxů se opakuje ve všech třech částech a tedy jsme je vyjmuli pro následnou analýzu. V tomto případě nebylo potřeba ověřovat skutečný výskyt v textu, jelikož jsme brali pouze 3 ze 3 a tudíž tomu odpovídal i skutečný výskyt těchto slov v celém rozsahu. Vzhledem k tomu, že některá slova byla přepsána hovorově, kontrolovali jsme i výskyt pravopisných variant (například vodpoledne – odpoledne -> zde nebyla nalezena opačná varianta; špatnej – špatný -> zde špatný bylo také užito jako nespisovný tvar „je to špatný“). Slovo pokuď se v textu vyskytovalo i ve variantně pokud, a protože nebylo možné provést kontrolu rozdílu ve výslovnosti, z analýzy jsme ho vyřadili. Také jsme procházeli pravopis shody podmětu s přísudkem (u slova znali) a neukázalo se, že by byla nutnost vyřadit některé z těchto slov. X20 označuje číslo, které participant užíval. Při projití přepisu jsme však narazili i na přepis této číslovky slovy a z hodnocení sentimentu okolí jsme ji vyřadili. Počet hapaxů se tak snížil na 24.

Superhapaxy: *naši, dopadlo, sranda, špatnej, občan, skoro, vodpoledne, znali, X20, radši, neviděl, někam, spolupráci, zavolal, podepsat, kterou, výsledky, pokuď, krásnej, kluk, zajímal, seděj, moskvy, kluků, abyste, opavě*

Můžeme si povšimnout, že slova, která autor rozhovoru používá na své nízké frekvenci nejsou v převaze funkční, ale obsahová a liší od těch, která strukturují text na vysokých vlnách. Nyní rozebereme blíže sentiment, který se vyskytuje v okolí některých superhapaxů

Sentiment okolí superhapaxů

Naši: okolí tohoto superhapaxu se vztahovalo ve všech třech případech k českému národu. Také se v blízkosti slova vyskytují názvy ostatních zemí. V prvním případě se jedná o „Sovětští přátelé“, v druhém je to „česká rafinérie…japonská rafinérie“ a v posledním se před slovem vyskytuje „Vatikán“ a za ním „Slováci a Poláci“

Sranda: i když se na první pohled zdá, že by sentiment tohoto slova měl být pozitivní a vtipný, autor ve všech třech případech zmiňuje nějaký průšvih. V prvním výskytu se například jedná o útěk z vojny „*strašná **sranda**, veselí, že se mu podařilo, byli sme veselí, když se mu podařilo uletět dostali sme pak nějakou jako nakládačku prostě na vojně*.“

Radši: dalším superhapaxem je radši. Nejzajímavější na jeho okolí je kolokace se slovesy, která se týkají pohybu. Jedná se o slova chodil, šel a vystoupím. Sentiment je zde negativní, jak si můžeme všimnout z okolních vět: „*Von říká, než abych měl žalovat, tak* ***radši*** *vystoupím z KSČ*“; „*…měli malé hospodářství, ale otec to nějak prošustroval,* ***radši*** *chodil do hospody, za války zemřel a teďko bylo pohraničí…*“; „*V Praze. Až sem přišel do Hu-, jo, takhle se třásli ti kádrováci rudej, tak sem* ***radši*** *šel než dostanou infarkt*“.

Krásnej: na tomto superhapaxu je nejzajímavější nespisovná varianta, jelikož se sice vztahuje k mužskému rodu, ale nekončí na „ý“. Jediný případ v textu, kde se nevyskytuje nespisovný výraz, ale je dodrženo pravidlo, je výraz „krásný rukopis“, který se nachází zcela mimo okolí. Zde je nejbližší okolí následující: „*v tom lágru byl* ***krásnej*** *bazén“; „špičkovej časopis, zahraniční politika,* ***krásnej*** *článek“; „patrový postele a byl tam jeden* ***krásnej****, hezkej kluk, nejlíp voblečenej*“

V posledním případě si všimněme, že za SH krásnej se vyskytuje i SH kluk. Kumulace superhapaxů se v textu objevila několikrát, další výskyt byl ve spojení „špatnej občan“. Pokud předběhneme, tento fenomén se objevuje i pokud zkombinujeme SH získané ze 3 i 6 částí (viz „dobře dopadlo“ „neviděl…svatej…při“). Tento spoluvýskyt superhapaxů, především v analýze tří řezů, se v textu často objevuje v rozmezí jednoho řádku. Z okolí některých kumulací si lze povšimnout rozhořčení, pomoci a převratu událostí. Především se dá ale vyčíst pohrdání nastaveným režimem. „*A vod tý doby sem byl* ***špatnej občan****, byl sem reakcionář a to, se moje myšlení uplně totálně vobrátilo*“ ; „*Tak za ním šel už takhle z dálky, mu ruku a takhle mu poplácal po rameni, že to* ***dobře dopadlo****, jo. Víte, proč to prdlo to Rusko?*“ ; „*a když jelo auto s chlebem, z Čech, tak ho* ***neviděl*** *a to u nich byl* ***svatej****, no. Takže takhle to začalo a* ***při*** *něj, když něco potřeboval pomoct, tak vždycky*“.

Kluk: v návaznosti na předchozí odstavec rozebereme okolí tohoto SH. Z analýzy nízké frekvence vyplynulo zajímavé zjištění, a to sice autorův postoj k mužům. Jak je vidět, 2 z jeho vlastních slov jsou ***kluk*** a ***kluků***, a v obou případech je v okolí zmínka o mládí nebo mít někoho na starost. V porovnání s označením žen se také dá říct, že je sentiment výrazně pozitivnější.

Superhapaxy získané ze řezů na šest částí

Z programu QUITA jsme v případě šesti vzorků textu vytěžili 5 hapaxů, které se opakují v šesti nebo pěti částech a vzhledem k tomu, že skutečná frekvence těchto slov se blížila počtu řezů (tzn. byla 5, 6 nebo 7), nemuseli jsme žádný superhapax vyřadit, a to ani z důvodu možné záměny za spisovný tvar.

Superhapaxy: *dobře, tejdnů, při, udělali, svatej*

Tejdnů: V tomto případě je v okolí ve všech případech časový odhad. V šesti z celkových sedmi výskytů se před SH objevuje slovo „šest" a v sedmém je to „pět". Z okolí se dá usuzovat, že tento superhapax je spojen se zdravotním stavem autora. Často kolem hovoří o nemoci, práci a penězích.

Svatej: V okolí se nachází zmínka o letectví, potažmo létání (celkem 3x) a střílení (2x). V nadpolovičním výskytu je i „Německo". Je patrné, že tento superhapax má pro autora význam hlavně, když vzpomíná na vojenskou službu.

Udělali: Dvakrát se za slovem objevuje kolokát „tanec" a odkazuje na řešení potíže. V jednom případě se v okolí objevila slova, která kdyby byla užita v jiném pádu či rodu, tak se také stanou superhapaxy („*stála, z komínu se přestalo kouřit. A to byl jedinej dodavatel benzínu, jinej nebyl pro armádu. Čili v tu ránu* ***udělali*** *patent, že tadleta fabrika je* ***svatá****. No čili já si z nich dělal* ***srandu****, já to věděl, já toho*").

Při: Jediné funkční slovo, které se v této analýze objevuje. Třikrát je spojeno s výslechem, a ve všech případech s autoritou, která s autorem nezacházela dobře, většinou se jednalo o souvislost s válkou (Německem) nebo vojenskou službou. V tomto případě také nacházíme kumulaci více SH: „*auto s chlebem, z Čech, tak ho* ***neviděl*** *a to u nich byl* ***svatej****, no. Takže takhle to začalo a* ***při*** *něj, když něco potřeboval pomoct, tak vždycky (jméno), todleto a todleto, takhle, vona byla na tom, na tý svobodárně*".

Dobře: Tento SH je jediný, který se v celém textu objevuje přesně šestkrát. Dvakrát se vztahuje ke zdravotnímu stavu autora. Dvakrát jde o dobře umět (německy) či znát a jednou je tohoto slova užito v kontextu, že autor „udělal dobře". Poslední výskyt je pak spoluvýskyt superhapaxu „dopadlo" a je zde naznačeno naštvání.

SOUHRN

Pokud porovnáme témata, která nám vzešla ze dvou analýz stejného textu, je patrné, že v případě, kdy jsme text rozřezali na šest částí (dále R6) a z nich vytáhli superhapaxy, jsou témata v okolí konzistentnější a lépe vykreslují citlivá místa. Nedá se říct, že analýza tří úseků (dále R3) by nepřinesla žádná zjištění, nicméně reflektuje větší procento textu, a možná z toho důvodu se okolí různí. Nejčastěji zmiňovaným tématem v R6 byla vojenská služba. Jednalo se nejen o službu státu po válce, ale autor zmínil i situaci při ní. Dalším velmi častým motivem bylo zdraví a nemoc jedince. V okolí R3 se často objevilo téma kluků a vojenství, ale zajímavějším byla volba oněch slov. Například „radši" se pojilo vždy se slovesem pohybu, „Moskvy" bylo vždy spojeno s „do" a jménem, nikoliv například „z". Stejně tak nespisovná výslovnost „krásnej", jelikož v případě, že autor řekl „krásný", nejednalo se o SH ani blízké okolí jiných vlastních slov. Také u SH „naši" si všímáme, že se jedná pouze o vztah k národu. Posledním příkladem je „dopadlo", které nikdy neznačí skutečnost, že něco padalo, ale přeneseně konec události.

V obou případech jsme kvůli opakování témat v okolí navíc zkoumali i frekvenci slov. Mezi slova, která se objevují pouze (nebo téměř výhradně) kolem superhapaxů, patří například: pivo, hlad, bratr, strýc, věznice. Na tématu rodinných příslušníků je zvláštní i to, že slovo „táta" (byť se z vyznění celkového rozhovoru jeví jako jedno z důležitých témat) se v nízké frekvenci vůbec nevyskytuje, kdežto „otec" je téměř v polovině případů užit v hlubokém sentimentu. Pivo si autor zásadně dává na svých nízkých vlnách a stejně tak hlad nenajdeme v textu na jiných místech. Co se týče odvozeného „hladovej", i zde je zajímavé poukázat, že v doslovném významu se vyskytuje v okolí SH, ale metaforický význam najdeme na hladině. Také můžeme vidět hru se synonymy. V textu najdeme například slova věznice, lágr, kriminál, ale jen věznici uvidíme blízko superhapaxů. Jak jsme již nastínili v případě SH „tejdnů", také slovo šest má pouze své nízké zastoupení a právě kolem citlivých slov. Posledním takovým případem je „lítat" a všechny jeho tvary. Je patrné, že lítání pro autora znamenalo důležitý úkol a na nízké frekvenci o něm často mluví, nicméně toto vyjádření se neobjevuje na povrchu, i když by se to dalo podle tématu očekávat.

Provedená analýza nám tedy poskytla náhled na neuvědomované jazykové chování autora, které je specifické pro jeho důležitá témata a podala více podstatných informací, než které by poskytla vysokofrekvenční analýza.

6 PŘÍPADOVÁ STUDIE III: ROZHOVORY O ZDRAVÍ - 2. ČÁST

V této studii, která tematicky i formou navazuje na předchozí, se nebudeme zaměřovat na okolí a podrobný popis superhapaxů, ale ukážeme, jak hluboký sentiment může promítnout osobnost jedince. Do této studie bylo zařazeno celkem pět rozhovorů od tří žen a dvou mužů, kteří hodnotili svůj život od okamžiku, kdy se dozvěděli o vypuknutí pandemie covid-19 a rozsahově se jednalo o délku mezi 8 a 14 tisíci slovy.

Zde jsme opět provedli analýzu superhapaxů jako v předchozích částech s tím rozdílem, že nyní jsme již nezkoušeli dělat víceřezová porovnání, nýbrž jsme se soustředili jen na rozdělení textu do tří částí a podstatnější pro nás bylo hodnocení okolí. Do analýzy bylo vzato pět rozhovorů a počet SH byl po odstranění potenciálních chyb 12, 22, 12, 12, 6. Kolem těchto superhapaxů jsme vysekali dva typy okolí. První zahrnovalo poměrně úzkou výseč textu a jednalo se o 10 slov před a 9 za. Druhé bylo stejného rozsahu, jaký jsme použili předchozí studii a jednalo se tedy o 20 před a 19 za. Tato okolí byla následně přeložena do angličtiny a vložena do programu LIWC pro zhodnocení sentimentu. Vzhledem k příliš nízkému počtu slov v případě velmi úzkého okolí, jsme do následující korelační studie zahrnuli pouze porovnání celého textu a okolí „20-19“. Rozdílnost velikosti jsme původně udělali kvůli porovnání sentimentu na nízkém a velmi nízkém rozsahu textu. Participantům byl také administrován inventář NEO FFI a jeho výsledky jsme tak porovnávali se sentimentovou analýzou získanou prostřednictvím LIWCu. Níže popíšeme podstatu inventáře a jak předchozí studie pracovaly s potenciální korelací s výsledky jazykové analýzy v LIWCu.

6.1 Inventář NEO FFI

Pětifaktorový model osobnosti představuje v současnosti jedno z nejuznávanějších rysových pojetí osobnosti. I když světlo světa spatřily jeho revidované verze, stále je tato validní a užívanou metodou. Největší předností tohoto modelu, je především jeho kódování v přirozeném jazyce. Kromě oficiálního názvu NEO FFI je tento inventář známý také pod pojmy Big Five nebo OCEAN, což je zkratka počátečních písmen pěti dimenzí. Pětifaktorová struktura osobnosti byla formulována na základě lexikální analýzy adjektiv, která se jevila jako nejvhodnější pro popis vlastností osobnosti jedince, ve spojení s faktorovou analýzou díky níž bylo odhaleno pět hlavních kategorií (Hřebíčková, Urbánek, 2001), nicméně pozdější výzkumy prokázaly, že i ty se dají dále rozdělovat do menších celků (Hřebíčková, 2004).

Pro českou adaptaci této metody se používá pět faktorů pojmenovaných podle Costy a McCrae: neuroticismus, extraverze, otevřenost vůči zkušenosti, přívětivost a svědomitost (Hřebíčková, Urbánek, 2001). Inventář NEO-FFI se skládá celkem z šedesáti položek, které jsou dále rozřazené po dvanácti do zmíněných pěti dimenzí. Respondent u každé položky poté vybírá, do jaké míry ho předložené výpovědi vystihují na pětistupňové Likertově škále (0 – vůbec nevystihuje, 1 – spíše nevystihuje, 2 – neutrální, 3 – spíše vystihuje, 4 – úplně vystihuje). Přibližně polovina položek je skórovaná v obráceném bodovém pořadí (inverzně), což slouží k ověření konzistence odpovědí a potenciálnímu podchycení lhaní. Celkový čas vyplňování je kolem 15 minut a normy jsou dostupné pro muže i ženy ve věku 15-75 let. Inventář je hojně využíván pro zkoumání korelací s ostatními fenomény jak v klinické psychologii, tak například v poradenství a někdy i při volbě povolání nebo v psychologii práce a organizace (Hřebíčková, Urbánek, 2001). V následující tabulce (Tab. č.1) popisujeme jednotlivé dimenze a uvádíme k nim příkladové položky, přičemž uvádíme i Cronbachovo alfa pro koeficienty reliability (tzn. míra spolehlivosti testu) podle příručky Hřebíčkové a Urbánka (2001).

Tab. 1: Škály a některé vybrané položky NEO FFI

Škála	Popis škály a položky	α
Neuroticismus	Tato škála zjišťuje individuální rozdíly emoční stability (lability). Jedinci, kteří mají skóry v této dimenzi vysoké, vykazují vyšší nervozitu, omezenou schopnost zvládat stresové situace a úzkostnost. Naopak stabilní jedinci jsou klidní, bezstarostní a sebejistí. - *Málokdy pocítím strach nebo úzkost. (inverzní)* - *Často mě rozčílí, jak se mnou lidé jednají.* - *Obvykle si nedělám starosti. (inverzní)*	0,81
Extraverze	Extraverti jsou společenští, aktivní, hovorní a optimističtí. Introverze v tomto případě není pravý opak, ale spíše nepřítomnost extraverze. Introverti jsou zdrženliví, nezávislí a vyrovnaní. Škála souvisí i s interpersonálním kontaktem. - *Mám rád/a kolem sebe mnoho lidí.* - *Raději bývám sám/sama než ve společnosti. (inverzní)* - *Často žiji v rychlém tempu.*	0,81
Otevřenost	Otevřenost vůči zkušenosti souvisí se zvídavostí, vnímavostí, touhou vyhledávat nové zážitky, experimentováním, bohatou fantazií a nekonvenčním přemýšlením. Škála mimo jiné také souvisí s estetickým cítěním a intelektuálností. - *Toužím po poznání a vědomostech.* - *Nerad/a ztrácím čas denním sněním. (inverzní)* - *Neměním vyzkoušené způsoby, jak něčeho dosáhnout. (inverzní)*	0,67
Přívětivost	Jedinci s vysokým skórem jsou altruističtí, spolupracující, neradi soupeří, vyjadřují pochopení a jsou nekonfliktní. Nicméně nedávají najevo negativní emoce a nebojují za své zájmy. Osoby s nízkou přívětivostí jsou egocentričtí a soutěživí. Tato škála úzce souvisí se sociální desirabilitou. - *Často se dostanu do sporu se svou rodinou nebo spolupracovníky. (inverzní)* - *Pokud někoho nemám rád/a, dám to dotyčnému najevo. (inverzní)* - *Raději bych s ostatními spolupracoval/ než soupeřil/a.*	0,72
Svědomitost	Svědomitost reflektuje vztah k pracovním výkonům, respektive sebekontrolu v oblasti organizace a aktivního plánování. Vysoký skór značí cílevědomost, spolehlivost, pevnou vůli, vytrvalost, ale i pedantnost a workoholismus. Nízký skór značí nestálost a nedbalost. - *Své věci udržuji v pořádku a čistotě.* - *Promarním mnoho času, než se pustím do práce. (inverzní)* - *Usiluji o dokonalost ve všem, co dělám.*	0,80

6.2 Výsledky analýzy NEO FFI a LIWC

V této podkapitole již popisujeme výsledky korelační analýzy dvou použitých metod. Následující tabulka (č. 2) ukazuje percentily participantů v jednotlivých škálách NEO FFI (použity byly normy pro muže a ženy zvlášť) a tabulka č. 3 skóry v kategoriích LIWC. Jelikož kategorií v LIWC je přes 80, uvádíme níže pouze některé, které zároveň vykázaly korelaci v kladném či záporném směru. V této tabulce u každého participanta ukazujeme výsledky získané z analýzy celého textu (označení „C“) a zároveň i z okolí superhapaxů (označení „O“), pro lepší představu o rozdílnosti.

Tab. 2: Výsledky participantů v inventáři Big Five

NEO_FFI M / F	skóre M1	skóre M2	skóre F3	skóre F4	skóre F5
Neuroticismus	89	92	16	29	95
Extraverze	16	46	35	65	52
Otevřenost	99	58	3	99	3
Přívětivost	81	9	51	73	4
Svědomitost	94	26	83	29	58

Tab. 3: Ukázka výsledků participantů v kategoriích LIWC

	M1C	M1O	M2C	M2O	F3C	F3O	F4C	F4O	F5C	F5O
Analytic	17,9	18,4	12,0	8,5	10,9	10,5	14,7	14,6	8,3	8,6
Clout	33,0	31,6	49,2	35,1	64,9	74,7	30,8	32,4	41,1	66,5
Authentic	69,2	54,7	69,2	75,3	46,0	35,0	76,7	81,3	77,9	71,0
Tone	42,7	43,3	28,2	40,7	59,1	62,2	79,3	82,4	36,0	30,0
WPS	26,2	28,4	112,6	52,2	23,6	23,0	18,4	14,9	84,1	99,8
Sixltr	14,7	14,9	12,2	12,4	12,8	12,7	13,0	12,8	11,1	8,8

Pro analýzu dat jsme využili programu R Studio a použili statistický test Spearmanův korelační koeficient, jelikož nelze předpokládat normální rozdělení dat a počet pozorování je velmi nízký (n=5). Z výsledků testové statistiky vyplynulo, že pozitivní či negativní korelaci o hodnotě alespoň .75 s některou škálou NEO FFI vykázalo 45 kategorií analýzy LIWC v případě korelací na datech z **celého** textu; a 52 kategorií v datech získaných z **okolí** superhapaxů. Pro přehled níže uvádíme tabulky korelací. Tabulka 4 zobrazuje hodnoty získané z porovnávání LIWC provedeného na celém rozsahu textu s NEO-FFI a tabulka 5 pak vykresluje korelace získané z porovnávání výsledků LIWC analýzy provedené na okolí superhapaxů opět s NEO-FFI.

Celkový text:

V případě ***neuroticismu*** byla zjištěna pozitivní korelace s kategoriemi: WPS (word per sentence), ppron (personal pronouns), netspeak. Negativní korelace se ukázala pokud šlo o: Tone, affect (affective processes), posemo (positive emotion), percept (perceptual processes).

Extraverze nejlépe pozitivně korelovala s: health, ingest (ingestion) a negativně s: quant (quantifiers), female (female references), hear, body, drives, work

V dimenzi ***otevřenosti*** jsme shledali nejvíce korelací. Pozitivní: Analytic (analytical thinking), sixltr (> 6 letters), ipron (impersonal pronoun), prep (preposition), adj (common adjectives), compare (comparisons), interrog (interrogatives), cogproc (cognitive processes), insight, cause (causation), differ (differentiation) a space. Negativní: Clout, shehe (3rd person singular), adverb (common adverbs), conj (conjunctions), verb (common verbs), number, social (social processes), male (male references), bio (biological processes), achieve, focus future, relativ (relativity), motion, time, home a money.

Přívětivost pozitivně i negativně koreluje s kategoriemi, které již byly objasněny u otevřenosti

Poslední dimenze, ***svědomitost*** – vykázala jen pozitivní korelace, opět názvy již byly upřesněny. Dokonce v hodnotě 1 se shodovala s kategorií affiliation.

Kategorie	Neuroticismus	Extraverze	Otevřenost	Přívětivost	Svědomitost
Analytic	0	0	0,949	0,9	0
Clout	0	0	-0,791	0	0
Tone	-0,8	0	0	0	0
WPS	0,8	0	0	0	0
Sixltr	0	0	0,791	1	0
ppron	0,8	0	0	0	0
shehe	0	0	-0,791	0	0
ipron	0	0	0,791	0	0
prep	0	0	0,791	0	0
adverb	0	0	-0,949	0	0
conj	0	0	-0,791	-0,9	0
verb	0	0	-0,791	0	0
adj	0	0	0,791	1	0
compare	0	0	0,791	0	0
interrog	0	0	0,949	0,8	0
number	0	0	-0,791	0	0
quant	0	-0,9	0	0	0
affect	-0,9	0	0	0	0
posemo	-0,9	0	0	0	0
social	0	0	-0,949	0	0
female	0	-0,8	0	0	0,8
male	0	0	-0,791	0	0
cogproc	0	0	0,949	0	0
insight	0	0	0,791	0	0
cause	0	0	0,791	0,9	0
differ	0	0	0,791	0	0
percept	-0,9	0	0	0	0
hear	0	-0,9	0	0	0
bio	0	0	-0,791	0	0
body	0	-0,9	0	0	0
health	0	0,8	0	0	0
ingest	0	0,872	0	0	0
drives	0	-0,8	0	0	0,8
affiliation	0	0	0	0	1
achieve	0	0	-0,791	0	0
focusfuture	0	0	-0,791	-1	0
relativ	0	0	-0,949	-0,8	0
motion	0	0	-0,949	0	0
space	0	0	0,949	0,9	0
time	0	0	-0,949	-0,8	0
work	0	-0,8	0	0	0,8
home	0	0	-0,949	-0,8	0
money	0	0	-0,791	0	0
netspeak	0,8	0	0	0	0
nonflu	0	0	-0,791	0	0

Tab. 4: Spearmanova korelace (treshold .75) výsledků v jednotlivých kategoriích testů NEO-FFI a LIWC – celkový rozsah textu

Okolí superhapaxů:

V případě ***neuroticismu*** byla zjištěna pozitivní korelace s kategoriemi: WPS (word per sentence), family, space. Negativní korelace se ukázala pokud šlo o: Tone, auxverbs (auxiliary verbs), affect (affective processes), posemo (positive emotion), percept (perceptual processes), feel, achieve.

Extraverze nejlépe pozitivně korelovala s: Authentic, Dic (dictionary words), ppron, relig (religion) a negativně s: negate (negations), negemo (negative emotions), discrep (discrepancy), hear, bio, health, risk.

V dimenzi ***otevřenosti*** jsme opět shledali nejvíce korelací. Pozitivní: sixltr, i (1st person singular), ipron, cogproc (cognitive processes), insight, power. Negativní: Clout, you (2nd person singular), shehe, article, conj, social, male, relativ (relativity), motion, time a money.

Přívětivost pozitivně korelovala s: Analytic, sixltr (úplná shoda), we (1st person plural), drives a work. A negativně se pak projevila ve spojitosti s: conj, male, relativ, motion a time.

Poslední dimenze, ***svědomitost*** měla nejméně korelací ze všech pěti dimenzí. Z výsledků můžeme vidět, že v pozitivním směru se jednalo o anx (anxiety), hear a affiliation. V opačném směru šlo o korelaci s Authentic a netspeak.

Plné názvy kategorií byly převzaty z psychometrického manuálu pro LIWC 15 (Pennebaker et al., 2015; dostupné z: https://www.liwc.app/static/documents/LIWC2015), kde lze dohledat pro lepší představu i slova, která do jednotlivých kategorií (slovníků) spadají.

Můžeme si všimnout, že v některých případech se sice výsledky obou korelačních analýz překrývají, avšak v případě analýzy LIWC nízkého lexika se přidávají rozmanitější kategorie, často spojené s psycholingvistickými vlastnostmi.

Nejvíce korelací se potvrdilo ve vztahu s dimenzí *otevřenost vůči zkušenostem*. Významnou míru vztahu této dimenze a kategorií LIWC vykazuje následujících 17 veličin v analýze s nízkým lexikem a až 29. Zároveň stejně v obou případech na tom byla *svědomitost*, která vykázala nejméně korelací s LIWC kategoriemi.

Kategorie	Neuroticismus	Extraverze	Otevřenost	Přívětivost	Svědomitost
Analytic	0	0	0	0,9	0
Clout	0	0	-0,949	0	0
Authentic	0	0,8	0	0	-0,8
Tone	-0,9	0	0	0	0
WPS	0,9	0	0	0	0
Sixltr	0	0	0,791	1	0
Dic	0	0,9	0	0	0
ppron	0	0,8	0	0	0
i	0	0	0,949	0	0
we	0	0	0	0,9	0
you	0	0	-0,791	0	0
shehe	0	0	-0,791	0	0
ipron	0	0	0,791	0	0
article	0	0	-0,791	0	0
auxverb	-0,8	0	0	0	0
conj	0	0	-0,949	-0,8	0
negate	0	-0,975	0	0	0
affect	-0,9	0	0	0	0
posemo	-0,9	0	0	0	0
negemo	0	-0,8	0	0	0
anx	0	0	0	0	0,975
social	0	0	-0,791	0	0
family	0,8	0	0	0	0
male	0	0	-0,949	-0,8	0
cogproc	0	0	0,791	0	0
insight	0	0	0,791	0	0
discrep	0	-0,9	0	0	0
percept	-0,9	0	0	0	0
hear	0	-0,8	0	0	0,8
feel	-0,8	0	0	0	0
bio	0	-0,9	0	0	0
health	0	-0,9	0	0	0
drives	0	0	0	0,9	0
affiliation	0	0	0	0	0,9
achieve	-0,9	0	0	0	0
power	0	0	0,791	0	0
risk	0	-0,791	0	0	0
relativ	0	0	-0,949	-0,9	0
motion	0	0	-0,949	-0,9	0
space	0,821	0	0	0	0
time	0	0	-0,949	-0,9	0
work	0	0	0	0,9	0
money	0	0	-0,791	0	0
relig	0	0,975	0	0	0
netspeak	0	0	0	0	-0,821

Tab. 5: Spearmanova korelace (treshold .75) výsledků v jednotlivých kategoriích testů NEO-FFI a LIWC – okolí superhapaxů

6.3 Dílčí diskuze výsledků

Některé z korelací, které naše studie odhalila, byly zjištěny i v předcházejících výzkumech, jiné jim odporovaly či nebyly prokázány. Nutno podotknout, že porovnávané studie nejsou z českého prostředí a je možné, že zde hraje roli i jazyk, jelikož texty použité v této byly přeloženy do angličtiny. Také ostatní výzkumy používaly ke korelaci analýzu celého textu, kdežto my se především soustředíme na okolí superhapaxů (k jehož výsledkům vztahujeme tuto diskuzi), což může mít vliv na míru korelace i objevení nových spojitostí a nebereme to jako limit ale výhodu. O souvislost osobnostních charakteristik s výsledky analýzy LIWC se do dnešní doby zajímalo již mnoho výzkumníků (mezi prvními samotní tvůrci LIWC, viz Tausczik, Pennebaker, 2010). Spojitost chování a mluvy, jejíž výzkum jde stále kupředu, rozebírají také Boyd a Schwartz (2021) nebo Bettis (2021). Myšlenka korelace mezi sebe-posuzujícími osobnostními charakteristikami (jaké měří například NEO-FFI) a volbou slov, je velice podobná pojetí souvislosti hlasu nebo výrazu obličeje s chováním (Koutsoumpis et al., 2022). Studie Koutsoumpis a kolegů (2022) uvádí metaanalýzu současného poznání v této oblasti a autoři dochází k několika důležitým bodům: Sebehodnocení osobnostních rysů sice významně koreluje s jazykovými kategoriemi, ale velikost účinku je relativně malá; hodnocení osobnostních charakteristik od druhých se vztahuje k jazykovému chování lépe; objevují se moderátoři, kteří vztah kategorií ovlivňují (blíže specifikujeme dále v diskuzi).

Pokud porovnáme naše výsledky v oblasti extraverze s těmi, k nimž došli Chen a kolektiv (2020), v podstatě docházíme k podobnému závěru, a to sice, že ač se tak v některých jiných studiích uvádí, míra extraverze nekoreluje dostatečně silně s pozitivními emocemi nebo sociálními slovy. Jak jsem zmínili, nejvíce korelací se nám objevilo u dimezne Otevřenost vůči zkušenostem, což jde proti zjištění Tackman et al. (2020), kde nejvíce korelátů obsahovala extraverze. V tomto výzkumu se také objevila negativní korelace extraverze s Analytic, což se v našem výzkumu neprokázalo. Baek a Ihm (2021) zjistili korelaci mezi extraverzí (stejně tak Qiu et al., 2012), avšak v našem případě se vysoká korelace mezi dlouhými slovy potvrdila v dimenzích otevřenost a přívětivost, přičemž v dimenzi otevřenosti je v souladu se zjištěním Mehl a kolegů (2006). V oblasti neuroticismu docházíme ke stejnému závěru jako Biel et al. (2013) a shledáváme negativní korelaci vysokého neuroticismu s pozitivními emocemi. Používání zájmena „já“ mělo ve studii Qiu et al., 2012 korelaci s neuroticismem a používání slov souvisejících s prací souviselo se

svědomitostí, nicméně ani jeden vztah u nás na hladině vyšší než .75 nebyl prokázán. Důvodem neshod může být také způsob, jakým byl zkoumán jazyk, jelikož v našem případě se jednalo o rozhovory o zdraví, mezi kategoriemi health a extraverze byla (při analýze okolí superhapaxů) potvrzena silná negativní korelace, kterou žádný z užitých zdrojů nevykazuje. Zajímavé však je, že při analýze celého textu byla korelace pozitivní, což značí o jiném jazykovém užívání. Kategorie money vykazují v této studii s osobnostní dimenzí otevřenost vůči zkušenostem negativní korelaci, oproti výzkumu (Schwartz et al., 2013), který vykazuje v případě významnou korelaci v opačném směru. Stejně vysoce relevantní je prokázaná negativní korelace kategorie social, která vykazuje vysokou shodu podpořenou dvěma srovnávacímu výstupy (Schwartz et al., 2013).

Na heterogenitu ve velikosti efektu mají dle metaanalýzy vliv následující moderátoři: velikost vzorku, pohlaví, věk, délka textu, rok vydání, jazykový režim, synchronnost, verze LIWC, složení vzorku, jazyková formálnost) (Koutsoumpis et al., 2022). Celkově tato zjištění ukazují, že - alespoň do jisté míry - osobnostní rysy lze skutečně měřit z textu a kategorie LIWC pomáhají vysvětlit, jak osobnostní rysy souvisejí s mluveným a psaným textem. Vzhledem k tomu, že v našem souboru je malý počet pozorování a tudíž není možné pracovat s p-hodnotou, neuváděli jsme žádné hypotézy a náš výzkum byl v tomto případě deskriptivní. Zároveň je otázkou, zda všechny texty byly dostatečně dlouhé pro výběr nízko frekventovaného lexika. Nicméně je patrné, že okolí superhapaxů koreluje s testem osobnosti více, než celkový rozsah souboru. Dalším limitem je využití neparametrického testu, který se vztahuje k malému N, avšak shoda výsledků s některými jinými studiemi značí (například Tackman et al., 2020 – také využili Spearmanův korelační koeficient, nicméně na větším vzorku), že při výzkumu na větším vzorku by se mohlo jednat o slibnou metodu. Máme nyní na mysli zjišťování osobnostních charakteristik pomocí jazykové analýzy z okolí citlivých autorských slov.

Na závěr bychom chtěli dodat, že zjištěné výsledky mají dozajista svůj význam obzvláště kvůli tomu, že srovnávají s osobnostními charakteristikami okolí nízko frekventovaného lexika, jež dle našeho přesvědčení odráží osobnost robustněji než rozsáhlé texty, které jsou i z časového hlediska náročné, avšak je potřeba výzkum zopakovat na větším množství participantů, ověřit hypotézy, které vznikly a výsledky validovat.

7 PŘÍPADOVÁ STUDIE IV: E. HOLMES – SOUDNÍ PROCES

Následující studie se zaměřuje na analýzu textu, které pochází ze soudního líčení s E. Holmes, které probíhalo v polovině minulého roku. Jelikož se jedná o opět jiný druh textu, který je specifický v délce úseků, jelikož se jedná mnohdy o velmi krátké úseky, avšak poskytuje výhodu jazyka, jelikož není třeba ho z důvodu analýzy v LIWC překládat. V této studii se krátce zaměříme na popis některých superhapaxů a graficky znázorníme, jak vypadá hluboký sentiment celého textu oproti vysekaným nízkým frekvencím.

Celkový soubor po odstranění otázek a výpovědí ostatních zúčastněných tvořilo 67 407 slov, což je velmi solidní základ. Při analýze superhapaxů jsme text rozdělili do sedmi vzorků po zhruba stejném rozsahu. Získané hapaxy, které se opakovaly alespoň v pěti částech byly: *ready, single, mostly, comment, primarily, hired, show, invest, ask, taken, opposed, since*. Při zpětné projití textu jsme pro close reading vyřadili *show*, jelikož se v celém souboru vyskytlo více než 20x a tedy nesplňovalo podmínky ve stejné míře jako ostatní slova, avšak toto slovo bylo ponecháno pro analýzu sentimentu, jelikož se jedná stále o nízko frekventované lexikum. Může to být zapříčiněno i různými způsoby jeho užití (např. *to show* vs. *show up*). Jak jsme zmínili, jelikož se jednalo o odpovědi na otázky soudu, nebyly jednotlivé výpovědi tolik rozsáhlé a vzhledem k tomu, že jsme chtěli ošetřit soudržnost okolí, nezařazovali jsme přesahy přes odpověď. Celková velikost okolí superhapaxů činila 5424 slov. Na obrázku (obr. 7) níže vidíme rozložení superhapaxů skrze celý text pořízené opět v programu AntConc.

Obr. 7: Rozprostření superhapaxů v celém textovém souboru

V tabulce (tab 6.) dále uvádíme pro jednotlivé superhapaxy také výsledky LIWC analýzy sentimentu. Tentokrát pouze pro čtyři nejhlavnější kategorie, na které se budeme soustředit i v grafech dále.

Filename	Analytic	Clout	Authentic	Tone
Ask.txt	10,34	66,17	41,42	75,05
Comment.txt	30,66	30,18	81,13	58,50
Hired.txt	40,61	85,10	33,82	78,68
Invest.txt	45,31	81,22	18,00	50,59
Mostly.txt	62,53	50,00	65,95	72,80
Opposed.txt	61,94	72,93	28,07	44,41
Primarily.txt	38,77	80,62	22,88	72,05
Ready.txt	37,01	70,13	59,40	92,79
Since.txt	37,10	27,20	85,99	31,21
Single.txt	58,13	69,73	49,28	47,47
Taken.txt	34,14	68,29	22,71	55,71

Tab. 6: Hlavní LIWC kategorie pro jednotlivé superhapaxy

Můžeme si všimnout, že největší emoční tón je zaznamenán pro okolí „*ready*“. Uvedeme pro lepší představu níže několik příkladů z tohoto okolí.

„ *I believe that Walgreens understood the lab wasnt* ***ready*** *to go live yet because we hadnt gone live, and they were pushing us really hard to go live as soon as possible*.“

„*We were investing a lot in R D and operations and hiring people and wanted to get* ***ready*** *for these rollouts and launches. And I needed to make sure that we werent going to have to either change our operations or that we were going to run out of cash*.“

Nejmenší náboj je pak patrný u slova „*since*“.

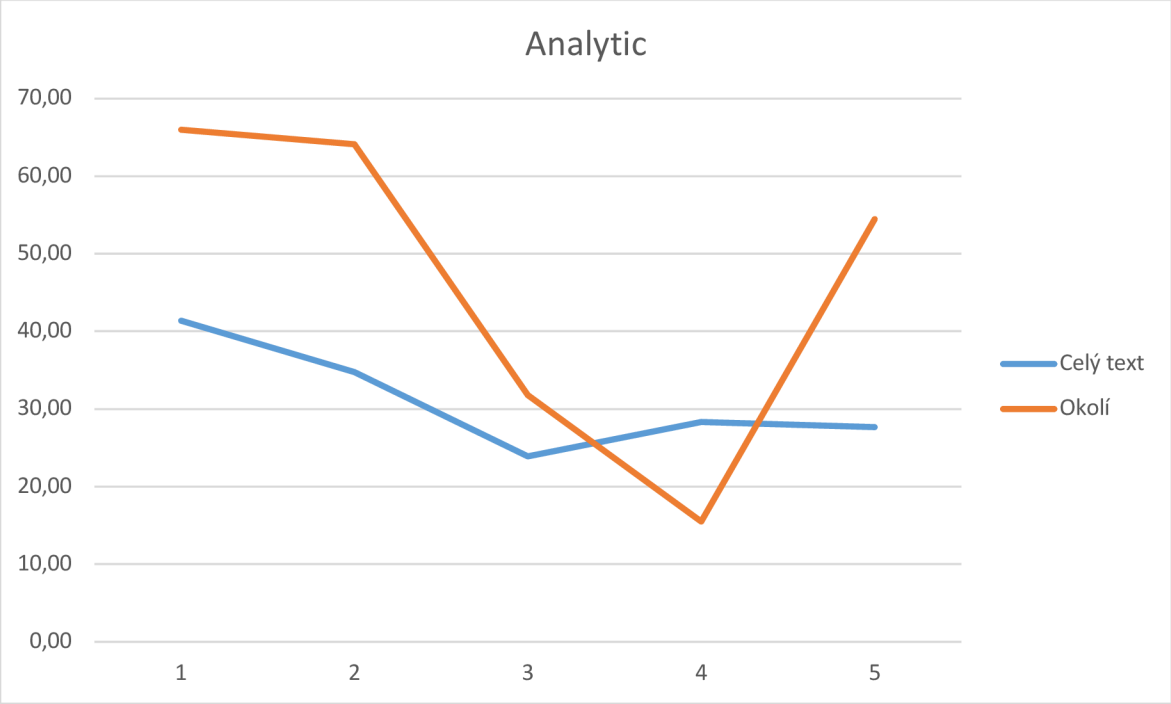
„*Im assuming that* ***since*** *she says the same assumptions, this is based on something else. I don ,t know what it was*.“

„*But we havent had any substantive discussions* ***since*** *he left left the company*.“

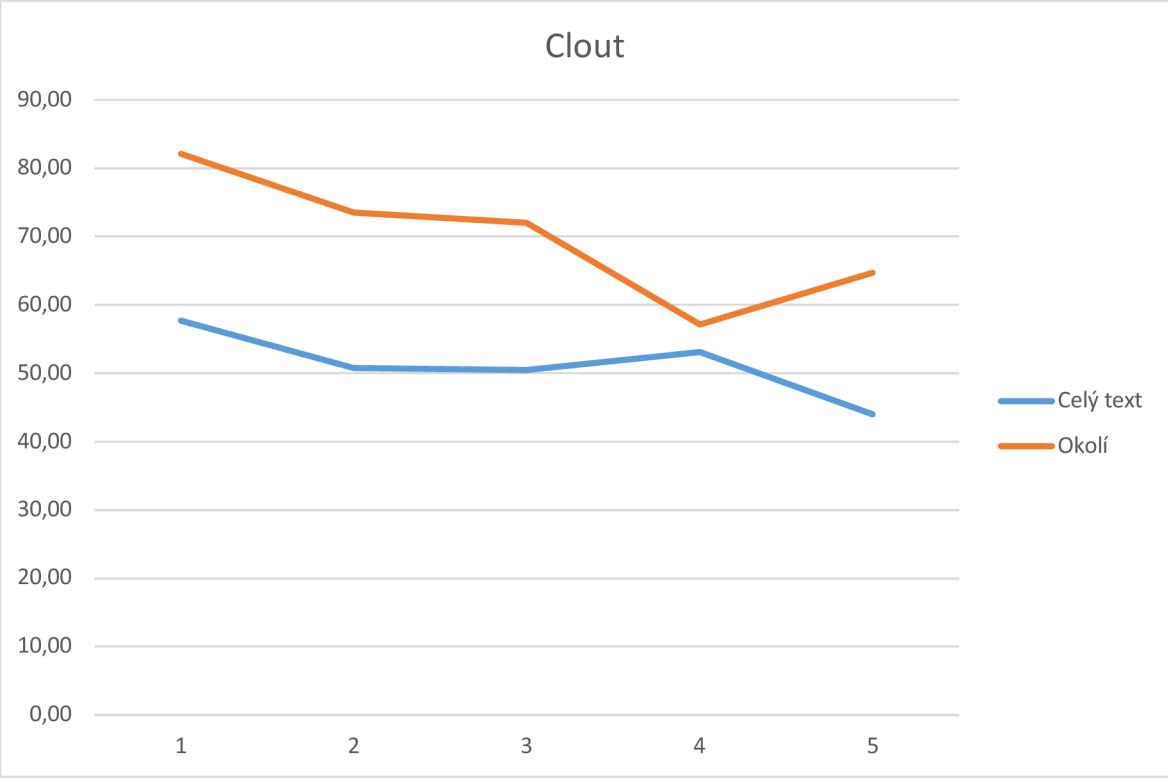
Provedena byla také frekvenční analýza okolí stejně jako v druhé případové studii v této práci a zjistili jsme, že v porovnání se zbytkem textu kolem svých citlivých slov hovoří například o „Sunny“, což byl spoluobviněný, následně „run“ „running“ „sample(s)“ „decison“ a „january“.

Nyní již představíme grafy pro jednotlivé kategorie LIWC: Analytic, Clout, Tone a Authentic, a porovnáme, jak se liší jejich průběh v textu pokud zahrneme celý rozsah vůči tomu, když zahrneme pouze vysekané okolí.

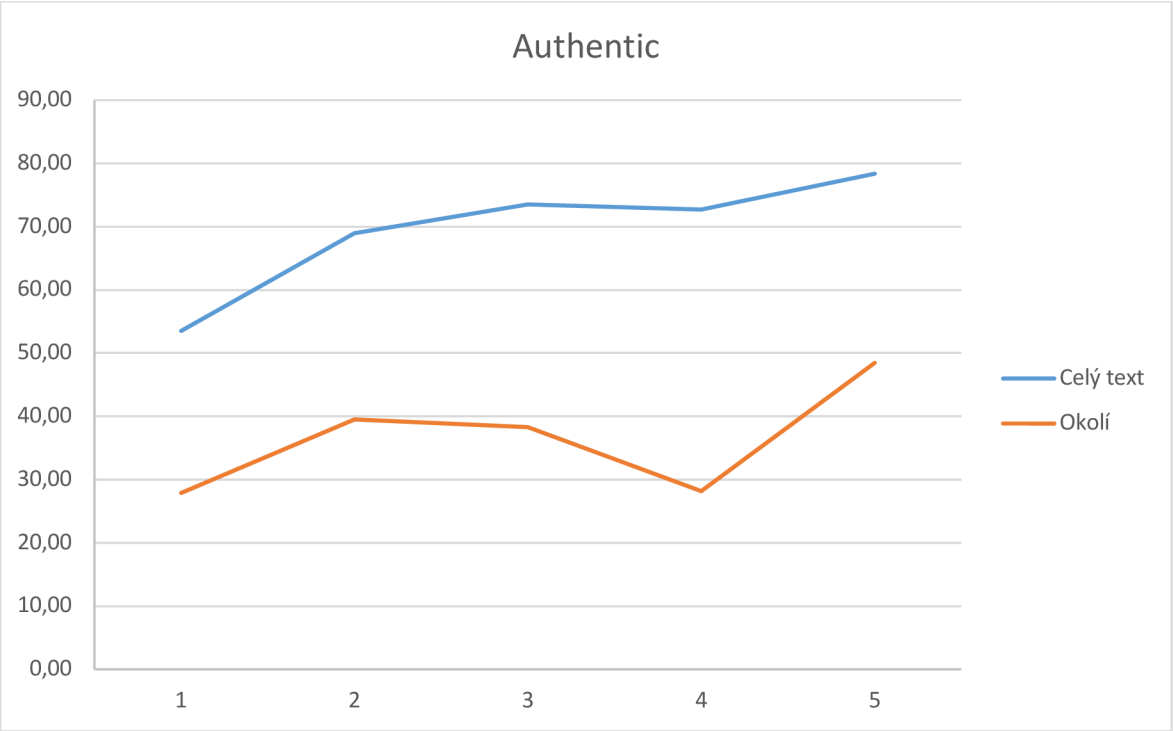
V grafech vidíme rozložení textu tak jak jde popořadě za sebou a to i v případě okolí. Text i okolí byly rozděleny do pěti úseků, v celém rozsahu se jednalo o úseky v délce 13500 slov a v případě okolí šlo o 1000 slov na úsek.



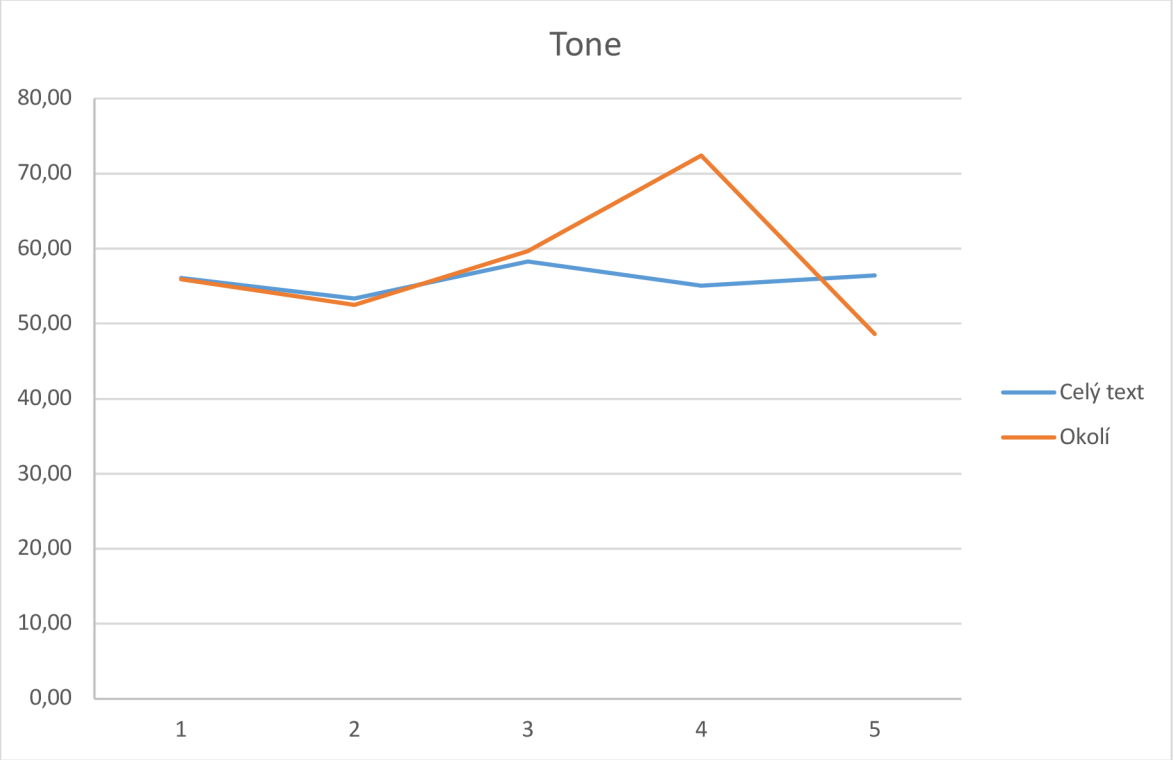
Graf 1: Sentiment kategorie Analytic v celém textu v porovnání s okolím



Graf 2: Sentiment kategorie Clout v celém textu v porovnání s okolím



Graf 3: Sentiment kategorie Authentic v celém textu v porovnání s okolím



Graf 4: Sentiment kategorie Tone v celém textu v porovnání s okolím

Z grafů je patrné, že křivky okolí kopírují sentiment celého textu, ale jsou citlivější a tudíž se pohybují v extrémnějších hodnotách. Zajímavé jsou výsledky v kategorii Tone (graf 4), jelikož jak je patrné, sentiment si zde téměř zcela odpovídá až na jednu část, která je z analýzy okolí prokazatelně emočně nabitá. Také si lze povšimnout, že byť se celý text projevuje vysokou autenticitou na povrchu, v nízkých frekvencích tomu zřejmě tak není. Poslední dvě kategorie – Analytic a Clout – také vykazují vyšší hodnoty v případě okolí, než je tomu u celkového textu.

Podíváme-li se blíže na změny, které se v křivkách odehrávají napříč všemi kategoriemi, můžeme si povšimnout, že ve čtvrtém úseku dochází k extrémním změnám. Zatímco Analytic, Clout a Authentic vykazují značný propad, stoupá hodnota Tone, což značí zvýšenou emocionalitu. Také křivky Authentic a Clout se od druhého úseku chovají podobně co se výkyvů týče. U Analytic je patrné, že nejvyšší hodnoty vykazuje na začátku textu a poté klesá. Zaměříme-li se blíže na křivku Authentic, můžeme si všimnout, že oproti ostatním případům je extrémně nízká. Vzhledem k faktu, že tato kategorie měří upřímnost a jedná se o data získaná ze soudního líčení, může nám to o vyslýchaném prozradit své. I to může souviset s tím, jak jsme jako lidé zvyklý se jazykově chovat a hlavně stylizovat. Hlídat si mluvu tak, aby šla podle našich plánů, je ze začátku zdá se lehké, nicméně čím víc a čím delší dobu člověk hovoří, tím méně si hlídá jazyk.

8 PŘÍPADOVÁ STUDIE V: „DANA A DITA“

V této poslední studii srovnáme dvě autorky a zaměřovat se budeme na ověření sentimentu získaných superhapaxů, respektive na to, zda pro nás některá „naše slova“ nastavují text do pozitivního nebo negativního emočního pólu.

Vzhledem k faktu, že texty vznikaly v rozdílných časových obdobích a jsou dokonce i jiné povahy (kombinace přepsaného rozhovoru a původně psaného textu), zajímá nás, zda aktuální kontext ovlivňuje autorský styl (jak píší Orgoňová, Bohunická, 2018 in Faltýnek, Kučera, 2022), nebo má nízko frekventované lexikum stabilní povahu.

Korpus Dany zahrnuje dva typy textových souborů. Jedním z nich jsou seminární práce, eseje, protokoly a další materiály, které byly psány primárně pro školu, avšak neobsahují přímé ani nepřímé citace, vyjma práce s učebnicí. Doba vzniku se datuje od roku 2021 po současnost. Druhou část tvoří přepis rozhovorů z osobní komunikace a volný proud myšlenek. Celkový soubor obsahuje 23 424 slov a rozhodli jsme se pro jeho rozdělení do 3 částí. Nepřistupovali jsme k dělení tematicky, nýbrž dle počtu slov, avšak mluvené přepisy byly před analýzou seřazeny za sebe. Získali jsme 12 superhapaxů, u kterých bylo poté opět vytaženo okolí a které byly tentokráte dále přeloženy do angličtiny kvůli práci s LIWCem.

V případě Dity se jedná o soudržný korpus, který zahrnuje jen úvahy a motivační dopisy, jejichž časový rozptyl je od roku 2019 do dneška. Do této analýzy jsme zahrnuli 15768 slov a text byl též rozdělen do 3 vzorků. Zde se objevilo 24 superhapaxů, s nimiž jsme pracovali stejným způsobem jak bylo popsáno výše.

Pro přehled přidáváme tabulky se superhapaxy obou autorek:

Superhapaxy - Dana				
výsledek	rok	dní	mohly	kterým
něm	obor	malý	pět	tímto
profese	orgány			

Tab. 7: Superhapaxy získané ze tří vzorků autorky - Dana

Superhapaxy - Dita					
milovat	později	slov	každého	nemůže	lidé
podle	prvním	mohly	hned	několik	otázku
vedení	bývá	věděla	rádi	poznat	všech
nejlepších	malý	přátelé	bezpochyby	neví	sdělit
tzn	seminární				

Tab. 8: Superhapaxy získané ze tří vzorků autorky – Dita

Z tabulek si můžeme všimnout, že se obě autorky shodují v SH „mohly“. Nicméně sentiment kolem tohoto slova je odlišný, jak ukážeme níže. Dále z analýzy vyřazujeme slova, která se objevila pravděpodobně vlivem tématu prací. U Dany jsou to: orgány, u Dity: tzn, seminárních.

V této případové studii se však nezaměřujeme na ruční procházení okolí a souvisejících témat, jde nám především o ověření sentimentu, se kterým daná slova autorky používají. Proto byla provedena analýza hlubokého sentimentu v programu LIWC. Uvedeme nyní postupně u obou autorek hodnoty, které jsme z programu získali a následně se přesuneme k metodologii.

DITA

Filename	Analytic	Clout	Authentic	Tone	i	negate	affect	posemo	negemo	anx	anger	sad	affiliat
každého	17,90	47,12	7,20	25,77	1,44	5,76	3,60	1,44	1,44	0,72	0,00	0,00	2,16
vedení	54,88	90,44	63,54	39,52	3,08	1,54	3,85	2,31	1,54	1,54	0,00	0,00	5,38
otázku	51,03	50,00	95,02	25,77	5,88	2,94	4,41	2,21	2,21	1,47	0,00	0,00	2,94
rádi	23,57	87,95	11,50	98,20	4,14	2,07	8,97	6,90	1,38	0,00	0,00	0,00	4,83
poznat	82,16	85,86	48,70	99,00	3,36	0,00	9,40	8,05	1,34	0,00	0,00	0,00	6,71
hned	65,77	53,05	67,14	97,81	6,87	0,76	8,40	6,87	1,53	1,53	0,00	0,00	4,58
několik	91,85	61,18	73,80	38,37	5,67	0,71	4,26	2,13	1,42	1,42	0,00	0,00	5,67
mohly	72,24	47,01	18,73	94,38	4,48	3,73	8,96	6,72	2,24	0,00	0,75	0,00	4,48
podle	83,11	77,44	64,86	51,40	0,68	1,37	4,11	2,74	1,37	0,00	0,00	0,68	4,11
nemůže	23,37	25,38	42,51	25,77	2,65	5,30	3,97	1,99	1,99	0,66	1,32	0,00	1,32
věděla	14,91	80,77	54,89	38,67	7,97	3,62	6,52	3,62	2,90	0,00	1,45	0,00	5,07
lidé	21,43	75,91	22,19	14,89	0,00	3,91	3,91	1,56	2,34	0,00	0,00	0,00	0,78
bezpochyby	52,71	64,46	3,84	15,37	1,48	2,22	3,70	0,74	1,48	0,74	0,00	0,00	0,74
všech	29,30	70,70	91,54	76,14	6,12	2,72	6,80	4,76	2,04	0,68	0,00	0,00	6,12
nejlepších	82,31	64,16	50,10	78,80	1,45	2,90	4,35	3,62	0,72	0,00	0,00	0,72	3,62
sdělit	54,47	84,33	44,41	98,03	0,00	1,55	6,20	5,43	0,00	0,00	0,00	0,00	7,75
malý	59,57	70,36	32,77	39,41	2,29	3,82	5,34	3,05	2,29	0,00	0,76	0,00	6,11
bývá	28,25	40,96	95,67	15,10	6,11	3,05	0,76	0,00	0,76	0,00	0,76	0,00	4,58
později	60,46	79,60	61,54	51,59	2,76	2,07	4,14	2,76	1,38	0,69	0,00	0,00	2,76
neví	38,60	59,26	74,76	3,45	0,78	3,12	2,34	0,00	2,34	0,78	0,00	0,78	2,34
prvním	58,86	12,22	93,59	85,60	8,22	2,74	5,48	4,11	0,68	0,68	0,00	0,00	0,00
milovat	34,38	75,79	24,10	77,29	9,09	1,40	5,59	4,20	1,40	0,00	0,70	0,70	7,69
slov	12,13	58,84	15,56	97,47	4,48	5,97	6,72	5,97	0,75	0,75	0,00	0,00	3,73
přátelé	27,59	74,00	62,17	25,77	6,43	2,86	2,86	1,43	1,43	0,00	0,71	0,71	5,00

Tab. 9: Vybrané LIWC kategorie pro zjištění hlubokého sentimentu

V tabulce 9 bychom chtěli upozornit na několik superhapaxů, které se významně projevují v určitých kategoriích. První je slovo „poznat“, kde pozorujeme největší hodnoty v kategorii Tone, affect a posemo, což svědčí o velmi pozitivním náboji okolí tohoto superhapaxu. Vysoký Tone mají ještě „rádi“ a „sdělit“. Nicméně „sdělit“ neskóruje na prvních úrovních ani v posemo ani v affect. Zdá se tedy, že poznat by mělo mít nejpozitivnější okolí, avšak když se podíváme zpět do textu, zjistíme, že celkové vyznění hraje spíš ve prospěch „sdělit“, alespoň co se emočního náboje týče. „*Určitě není lehké volit správná slova, umět* ***sdělit*** *své myšlenky a názory citlivě a adekvátně k dané situaci a kontextu*“ x „*Myslíme si, že po takto vedeném výcviku by se skupina měla možnost zase více* ***poznat****, stmelit a nadále by mohla dobře fungovat*“. x „*Se budu na zádech škrbat asi v triku, ne? Můžou být* ***rádi****, že si občas vezmu aspoň spodní díl. Jejda… teď mi došlo, že tohle možná nebudeš číst jenom Ty, co?*“

Naopak do nejvíce negativního pólu spadla slova „neví“, „lidé“ a „věděla“. A ač se zde jeví jako bez přítomnosti emocí, zaujalo nás slovo „bývá“. Porovnáme-li okolí těchto čtyř příkladů, dojdeme k závěru, že „bývá“ rozhodně bez náboje není.

„*že to byla vždycky procházka růžovým sadem. Ale tak už to v životě* ***bývá****, nežijeme v utopii a i na to je potřeba se připravit. Když si vzpomeneme, kolikrát jsme na všechno nadávali*,“

„*Když se koukneš zpátky do tvý životní historie, nepřijde ti, že je všechno až děsivě propojený? Někde ještě člověk* ***neví****, co to má být a proč se to stalo, někde je to ale to, co bylo předtím nepochopitelný*,“

„*Byla taková naše adoptivní mamka, vždycky za nás bojovala, i když* ***věděla****, že jsme si za pár problémů mohli sami. Nikdy nás nezklamala a snažila se být pro nás tou nejlepší třídní, což rozhodně byla*.“

„*Je až fascinující, jak člověk, pokud má kuráž nebo pokud je mu tak trochu jedno, co si o něm* ***lidé*** *myslí, protože už je zkrátka nikdy neuvidí, dokáže obalamutit kdekoho. Stačí jen vypadat přesvědčivě, tvářit se seriózně a působit*“

Ze zmíněných příkladů jsme vybrali k otestování primingu (= fenomén, při kterém vystavení určitému podnětu ovlivňuje reakci člověka na následující podnět; zde vystavení superhapaxu by znamenalo ovlivnění hodnocení sentimentu dané věty) tato slova: sdělit; bývá; poznat a přidali mohly, jelikož jak uvedeme dále, se jedná o superhapax společný pro obě autorky, leč v jiném kontextu.

DANA

Filename	Analytic	Clout	Authentic	Tone	i	negate	affect	posemo	negemo	anx	anger	sad	affiliat.
dní	93,26	78,45	46,48	70,24	1,57	0,00	3,94	3,15	0,79	0,00	0,00	0,00	3,15
kterým	61,36	90,70	4,84	15,44	0,00	1,47	3,68	1,47	2,21	0,74	1,47	0,00	8,09
malý	65,33	39,41	31,58	16,21	3,36	2,68	11,41	5,37	6,04	0,67	3,36	0,00	2,68
mohly	63,99	80,77	16,76	93,61	2,90	1,45	5,80	5,07	0,72	0,00	0,00	0,00	7,25
něm	61,44	78,21	3,03	83,48	0,65	0,00	6,49	4,55	1,30	0,00	0,65	0,65	1,95
obor	70,34	66,48	37,87	93,02	2,13	2,84	4,26	4,26	0,00	0,00	0,00	0,00	2,13
orgány	96,32	55,48	1,33	25,77	0,00	0,00	0,69	0,00	0,00	0,00	0,00	0,00	0,69
pět	78,15	50,00	63,19	93,41	3,60	0,72	7,19	5,76	1,44	0,72	0,00	0,00	2,16
profese	96,46	45,62	63,54	84,10	1,10	2,20	5,49	4,40	1,10	0,00	0,00	1,10	2,20
rok	80,14	32,48	78,62	7,96	4,55	3,03	3,03	0,76	2,27	1,52	0,00	0,76	1,52
tímto	60,32	79,59	4,05	54,07	3,01	1,50	3,01	2,26	0,75	0,00	0,00	0,75	1,50
výsledek	80,01	33,62	68,81	25,77	4,23	3,52	4,23	2,11	2,11	1,41	0,70	0,00	2,11

Tab. 10: Vybrané LIWC kategorie pro zjištění hlubokého sentimentu

U tohoto výběru jsme kromě hodnot v LIWC (tab. 10) pozorně sledovali i užití v textu, jelikož se jednalo o spojení mluveného projevu a seminárních prací. Zde jsme na straně vysokého Tone, affect a posemo identifikovali slova: „pět“, „malý“, „mohly“. Nicméně „malý“ se ukázalo být také nejvíce negativním okolím. Mezi další slova s vyšší negativitou patřily „rok“ a „kterým“. Poslední zmíněné vykazovalo též nejvyšší Clout.

„*No. Ale. No a mně asi tady došly otázky, ale jdu se podívat, kolik to má minut 36 ale. prvních* ***pět*** *se nenahrávalo ale tak to nevadí, takže 30 minut co bych ještě tak mohla říct teďka se jdu učit, protože.*“

„*prosinci na tu na ten open class a ty vado já jsem přišla, jak kdyby přesně kdybych tančila třeba jenom* ***rok****, víš, jakože jsem si úplně říkala, ty vole všichni říkají to, co oni si nezapomíná. No, tak já*“

„*to je jedno kde a našli jsme šlapku a byli jsme přesvědčený, že to je šlapka z toho kola, na* ***kterým*** *on jezdí, že nás sledoval. Tam mu ta šlapka upadla a my jsme tu šlapku vzali a dali jsme*“

Do primingové úlohy emočního hodnocení vět jsme se rozhodli vybrat „kterým“ a „rok“, jelikož „pět“ se nám podle okolí nejevilo jako vhodný příklad, nicméně jsme ho použili pro jiný typ úlohy, který popíšeme v odstavci metodologie.

Zmínili jsme, že mezi superhapaxy se objevila i slova, která se u autorek shodovala. Níže ukazujeme, jak se styl jejich použití u obou liší.

DANA:

*„tebe ošklivý a milujou tě nejvíc pod sluncem. I kdyby si udělala, nevím co a kočky jsou prostě takový sviňáci **malý**, agresivní nebo nejsou agresivní, ale takový vychcaný. No takže peníze nebo láska, peníze a láska? Ne nevím, vím ale.“*

*„pocit, že nás chce unýst, takže jsme pak chodili po Žižkově a hledali jsme stopy, který by nás k němu **mohly** jako vést a. Vím, že jsme si založili deník, kam jsme teď dávali ty důkazy a jednou jsme šli na Žižkově prostě.“*

DITA:

*„‚by se jako muselo stát, aby ho kopla do zadku?‘ Automaticky a bez přemýšlení jsem bleskurychle udala odpověď: ‚Řeknu ti **malý** tajemství, chlapče. Víš, jestli brutálně zamilovaná ženská něco není schopná překousnout, tak je to nevěra, ne bankovní loupež.‘“*

*„z mateřský člověk nevyžije, viď. A hlavně jsem zjistila, že nevyžiju, respektive nepřežiju bez akademický půdy a bez toho, abychom spolu **mohly** nadávat na všechny úkoly, práce, akademiky, senát, dokumenty…“*

8.1 Metodologie

K ověření toho, zda autorská slova skutečně mají emoční význam pro jejich majitele jsme použili vytvořený dotazník na platformě GoogleForms. Dotazník přidáváme pro úplnost do přílohy č.3. Sestaven byl z pěti otázek, které obsahovaly cílová slova a distraktory. První a třetí otázka se zaměřovala na pozitivní sentiment, druhá a čtvrtá na negativní. Poslední otázka obsahovala prostý výběr slov, která participanta zaujmou. Následovala „poznámka“, která byla určena pro identifikaci cílových autorek.

První otázka obsahovala slovo „sdělit“ a byla mířená na Ditu. Druhá otázka byla taktéž mířená na stejnou autorku a jednalo se o slovo „bývá“. Třetí otázka byla zacílená na Danu, avšak citlivé slovo bylo „mohly“, což je i jeden ze superhapaxů Dity. Ve čtvrté otázce bylo slovo „kterým“ a tato položka byla směřována na Danu. V poslední otázce jsme mezi výběrem slov zahrnuli superhapaxy obou autorek a distraktory: pět (Dana), jenom (X), mohly (Dita, Dana), lavička (X), poznat (Dita), vědět (X), pouze (X).

Dotazník byl rozeslán nejprve mezi kontrolní skupinu, kterou tvořilo 13 participantů v odpovídajícím věku a vzdělání autorek, abychom ověřili, zda některé vytvořené nejsou obecně vnímané jako emocionálně nabité, ale potenciální výsledek nebyl ovlivněn

heterogenitou skupiny. Poté byl dotazník předán autorkám, které do poznámky uvedly své jméno.

Očekávali jsme, že autorka Dita bude skórovat do jejích vybraných slov ve větší míře, než autorka Dana, z důvodu původu výchozího textu, z nějž byly superhapaxy získány. Dále se předpokládalo, že ve třetí otázce autorky vyberou shodnou odpověď z důvodu stejného superhapaxu. V poslední otázce jsme očekávali, že autorky vyberou právě dvě svá slova.

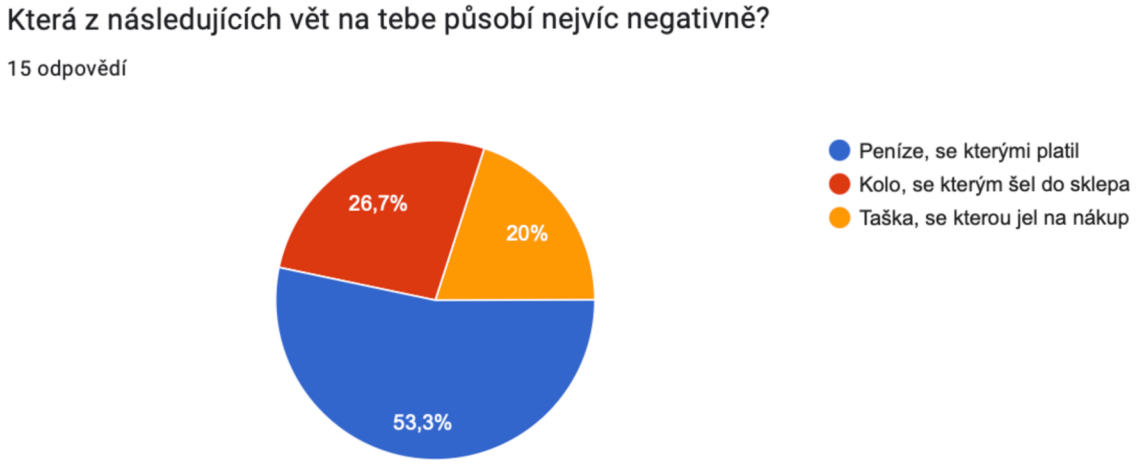
8.2 Výsledky

Níže přikládáme grafy rozložení odpovědí na jednotlivé otázky pro celý soubor. Autorkám se jednotlivě budeme věnovat posléze.

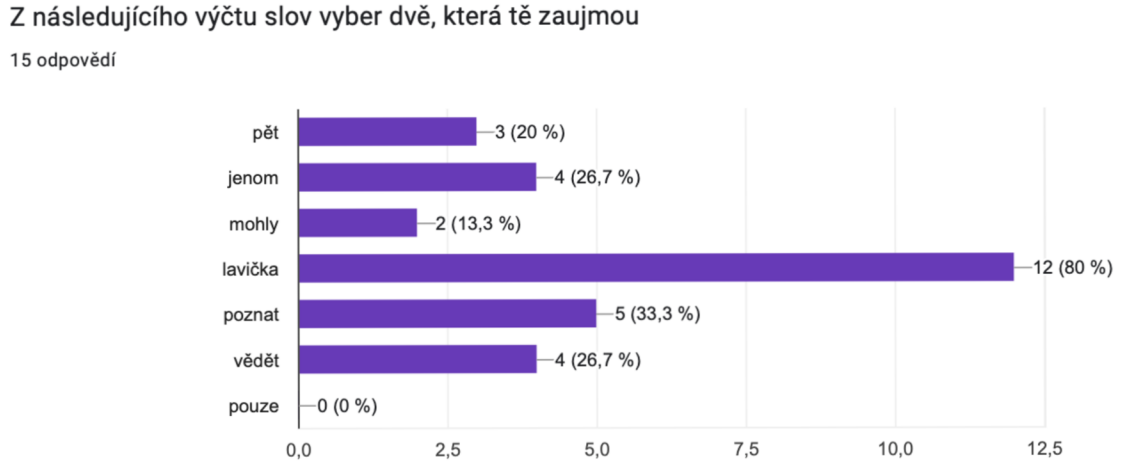
Graf 5: Rozložení odpovědí všech participantů na první otázku

Graf 6: Rozložení odpovědí všech participantů na druhou otázku

Graf 7: Rozložení odpovědí všech participantů na třetí otázku



Graf 8: Rozložení odpovědí všech participantů na čtvrtou otázku



Graf 9: Rozložení odpovědí všech participantů na pátou otázku

Z grafu 1 můžeme vidět, že rozložení odpovědí je rozprostřené mezi všechny položky, nicméně převládá „Chtěla bych ti něco říct“. Rozhodli jsme se položku neupravovat, jelikož se nejednalo o převládající výběr cílového slova a v případě, že by autorka i přes převahu této položky vybrala tu svou, jednalo by se o silnější ukazatel. Tímto způsobem jsme postupovali i u ostatních otázek. Ve třetí otázce (graf 7) si lze všimnout, že sice převládá výběr cílové položky, avšak podobné množství respondentů zvolilo jinou variantu. U poslední otázky (graf 9) jednoznačně převládá slovo „lavička“, proč tomu tak může být zhodnotíme v dílčí diskuzi. Nyní okomentujeme výběry jednotlivých autorek.

DITA:

1. Chtěla bych ti něco **sdělit**
2. A jak už to **bývá**,....
3. **Mohly** jsme si vybrat
4. Peníze, se kterými platil
5. Lavička, **poznat**

Z odpovědí vyplývá, že v prvních dvou otázkách, které byly na autorku přímo cíleny, vybrala v obou případech položku se svým superhapaxem. Taktéž ve třetí otázce zvolila své slovo, byť nebylo přímo určeno pro ni. Čtvrtá otázka žádnou položku s jejím slovem neobsahovala a v poslední otázce zvolila jeden svůj superhapax „poznat“.

DANA:

1. Chtěla bych ti něco povědět
2. A jak už to bývá,....
3. **Mohly** jsme si vybrat
4. Kolo, se **kterým** šel do sklepa
5. **Pět**, lavička

První dvě otázky nezahrnovaly autorčiny superhapaxy. Ve třetí a čtvrté otázce vidíme, že se ve výběru v obou případech objevila autorčina nízko frekventovaná slova. V páté otázce vybrala jeden svůj superhapax – „pět“.

8.3 Dílčí diskuze výsledků

V obou případech vyšlo najevo, že autorky preferují položky se svými superhapaxy, ačkoliv v případě Dany jsme tento jev původně kvůli nekonzistenci textů nepředpokládali.

V poslední položce kromě jednoho svého superhapaxu bylo zvoleno slovo „lavička“, které bylo preferováno i u ostatních participantů a nezáměrně se stalo atraktorem. Může to být způsobeno přílišnou konkrétností slova a také jiným slovním druhem, než jsou ostatní položky. Podstatné zjištění však je, že ze všech ostatních distraktorů autorky zvolily své slovo bez ohledu na otázku „zda je emocionálně nabito“.

Ta důležité považujeme zmínit, že posléze proběhla s autorkami diskuze nad výběrem odpovědí a Dita svou volbu komentovala následovně „*Já nevím, jako nepřišlo mi to nejpozitivnější, ..., ale je to něco, co bych řekla já.*“. Dana na výběr lavičky odpověděla „*tak tam si hned představíš něco, vzpomínku třeba.*“ O výběru slova „pět“ uvedla, že netuší, prostě se jí líbilo. Toto zjištění dozajista může pomoct v následujících výzkumech tohoto tématu.

Vzhledem k tomu, že ani jeden z textů nevznikl v krátkém časovém úseku, ale v průběhu několika let, můžeme usuzovat, že některá citlivá slova pro nás mohou být v čase stabilní. Do dalšího výzkumu navrhujeme zjistit minimální rozsah textu, který se dá k tomuto účelu použít, nebo také rozdíly mezi mluveným a psaným textem a určitě výzkum provést na větším vzorku účastníků. Dále doporučujeme znovu otestovat, zda je stabilita v čase platná, a pokud ano, kam až může sahat a jak se liší výsledky současných superhapaxů s těmi minulými. Tyto výsledky slouží spíše jako odrazový můstek a deskripce jevu, než hodnoty, které lze brát za obecnou pravdu. Každopádně tato studie poodkryla další tajemství superhapaxů z hlediska autorského hlubokého sentimentu.

9 DISKUZE

Vzhledem k tomu, že problematika superhapaxů není do detailu prozkoumaná a jedná se o nově vzniklou metodu, rozhodli jsme se každou analýzu pojmout lehce odlišně. U Martina Selnera jsme zkusili prozkoumat pouze okolí slov, která se objevila ve více než jedné analýze vzorků. Netvrdíme, že pokud by se analýza ubírala jiným směrem a například hodnotila pouze superhapaxy vzešlé z řezů na šest částí, výsledky by neukázaly jiné nebo další skutečnosti. Nedomníváme se však, že by přístup průniku superhapaxů měl být špatným, ba naopak nám může poskytnout vhled do nízko frekventovaného lexika z jiného úhlu pohledu, jelikož opakování slov na nízkých frekvencích poskytuje podstatné informace samo o sobě. I když je text převzatý z některých povídek sdílených na blogu (viz Selner, 2016 - 2019), a tudíž do něj nebylo zasaženo editorskou rukou, má určité odchylky od přepsaného mluveného slova. Je možné, že autor přidával příspěvky v jiném pořadí, než ve kterém je psal, mohla se vyskytnout i prodleva v dopisování určitých odstavců a podobně. Nicméně specifické užívání jazyka by nemělo záviset na časové prodlevě, a pokud jedinec má slova, která používá v určitém tématu nevědomě, i jejich zastoupení v textu nebude známkou stylizace. Zároveň nepředpokládáme, že by seřazení textu mělo na výsledek značný vliv, jelikož se jedná o nízko frekventované lexikum a tudíž jeho výskyt napříč textem nebude vysoký. Jak píší Faltýnek a Kučera (2022), pokud se rozhodneme analyzovat vzorky o různých velikostech, může se stát, že některé z nich ztratíme, nicméně jak se ukázalo i u Selnera, superhapaxy se mají tendenci kupit u sebe nebo u důležitých témat a proto jiné řazení textu sice přináší nová poznání, nicméně neztrácíme kvůli němu podstatné informace. V teoretické části jsme zmiňovali také poměr obsahových slov vůči funkčním. I když je bráno za fakt, že podle funkčních spolehlivě poznáme autora, ve výběru superhapaxů vidíme, že na nízkých frekvencích autoři nevědomě užívají autosémantika. Pokud bychom mohli spojit výsledky naší analýzy se skutečností, z příspěvků autora víme, že za nějaký čas po dopsání odešel ze své původní práce kvůli syndromu vyhoření, což odpovídá tématům a emočnímu náboji vět, které se vyskytovaly v okolí superhapaxů. Autor často hovořil o odchodu z práce, únavě, chybění a hledání. Nacházíme tedy soulad v nízko frekvenčním vyznění textu s predikcí budoucího chování, nicméně je potřeba zjištění brát s rezervou, jelikož ani ve vysoké frekvenci se těmto tématům nevyhýbá, byť jsou řidší a vyznění je pozitivní.

V druhé případové studii jsme se soustředili na zcela jiný typ textu, a to sice transkript mluveného slova. Přináší nám to nestylizované užití jazyka jedincem, což s sebou přináší hodnotu autentičnosti, nicméně také nástrahy v analýzách. Přepisy rozhovorů mohou reflektovat užívání jazyka v řeči, jako například v případě „ste“ „votevřel“ nebo „sem (jsem)“, popřípadě mohou být přepsány i nespisovné výrazy jako „mladej“, což může pozměnit jazykovou analýzu, pokud přepis není sjednocen a někde dochází k přepisu nespisovné formy do spisovné nebo není dodržena soudržnost přepisu (ne)spisovnosti v celém textu. Nicméně i pokud se v jazyce autora objeví spisovný výraz, zatímco do té chvíle v rozhovoru používal jeho hovorovou variantu, může to nést význam. Stejně tak u slov jako „ste“ atd. nepředpokládáme, že by měla být nízko frekventovaná a tudíž by neměla do analýzy superhapaxů zasáhnout. Pro další analýzy rozhovorů tedy zůstává otázkou, zda text přepsat do spisovné formy, nebo zahrnout i fonetické změny mluvy. S tím se pojí i otázka I vs Y na konci sloves v minulém čase. V případě, že se využívá automatický přepis, může dojít ke ztrátě rozlišení měkkosti. Také ve slovech „ačkoli“ x „ačkoliv“ a jim podobným se může stát, že přepis není jednotný. Zde je pak nutné posoudit, pokud se taková slova objeví jako superhapaxy, jaký je výskyt jejich protějšků a zda je možná záměna. V našem případě jsme tento jev neshledávali jako častý ve většině takových situací, pokud jsme usoudili, že se nemusí jednat o nízko frekventované lexikum, bylo dané slovo vyřazeno a proces okomentován. Nicméně v této případové studii jsme se soustředili na porovnání superhapaxů, které vznikly ze dvou typů analýz, a to sice z rozdělení autorského textu na tři nebo na šest vzorků. V prvním případě nám vyplynulo, že v okolí se vyskytují zajímavá témata, nicméně oproti rozdělení textu na šest částí nejsou tolik konzistentní. Byla zjištěna i kumulace některých superhapaxů a kolokací jednotlivých slov. Při použití frekvenční analýzy jsme zjistili, že některá slova se vyskytují pouze v určitém blízkém okolí superhapaxu a zřejmě tedy superhapaxy mají jisté charakteristiky, které odráží naše jazykové chování a tím pádem by mohly souviset s naší osobností (viz např. Bettis, 2021; Boyd, Schwartz, 2021). Ukázalo se, že superhapaxy přitahují jiné jazykové chování (viz například zmíněné létání nebo věznice) a v nízkých frekvencích skutečně dochází k synonymní volbě slov (Faltýnek, 2020). Toto zjištění nás motivovalo k ověření předpokladů vztahu osobnostních charakteristik k jazykovému chování, a zaměřili jsme se na to v následující studii.

Třetí studie se taktéž zaobírala rozhovory o zdraví, nicméně zde jsme nehodnotili okolí superhapaxů kvalitativně, ale využili jsme programu LIWC k analýze sentimentu. Proto bylo

zahrnulo všech pět participantů a nikoliv pouze jednotlivec. Jelikož bylo naším záměrem porovnat nejen souvislost jazykového chování s osobností autora, ale také ověřit, jaký je rozdíl mezi povrchovým a hlubokým sentimentem, rozhodli jsme se pro analýzu jednak celkového rozsahu textu, jednak okolí v rozsahu kolem 40 slov, samozřejmě pro každého autora zvlášť. Důležitým bodem je nutnost překladu do angličtiny, jelikož LIWC zatím nemá vytvořený slovník pro češtinu. Ač došlo ke zběžné kontrole překladu, vzhledem k přítomnosti hovorových, nespisovných výrazů, nebyl vždy překlad přesně proveden. Participantům byl administrován také inventář NEO-FFI, jelikož je reliabilním a dostupným sebe-posuzujícím nástrojem, který se zaměřuje na pět dimenzí osobnosti. Z výsledků vyplynulo, že v případě korelace okolí superhapaxů v LIWC se tyto dvě metody setkávají ve více kategoriích, než když analyzujeme celý rozsah textu. Vzhledem k tomu, že počet participantů nebyl velký, rozhodli jsme se pro striktnější přístup a brali pouze korelace vyšší než .75 na obě strany. Byť se naše výsledky plně neshodují s proběhlými studiemi, nepředpokládáme, že by měly být zatraceny. K nesouladu mohlo dojít z důvodu striktního přijímání / nepřijímání velikosti korelace, jelikož ostatní výzkumníci ve svých pracech takto přísní nebyli. Zároveň ani výsledky zmíněných autorů se mezi sebou plně neshodují. Dalším rozdílem může být, že jsme neporovnávali rozsah celého textu autorů, ale pouze okolí, což se u jiných výzkumníků neobjevilo a náš výzkum tak přináší nové světlo do této oblasti. Jeví se, že největší míra korelací LIWC se vyskytuje u dimenze Otevřenost vůči zkušenostem. Ostatní dimenze jsou poté téměř vyrovnané. Soulad či nesoulad s výsledky jsme již uvedli v dílčí diskuzi příslušné kapitole (6) a proto se zde nebudeme opakovat.

Limity a doporučení pro další analýzy v této souvislosti vidíme následovně: je nutné provést další testy, které se zaměří na délku analyzovaného textu v LIWC ve vztahu k osobnostním inventářům. Také by bylo přínosné korelace ověřit na původně psaném textu a ne přepisu rozhovorů, což by mohlo přinést potvrzení nalezených výsledků nebo naopak nová zjištění. Nejdůležitějším faktorem je provést analýzu na větším počtu participantů, aby bylo možné provést test p-hodnoty. V takovém případě je pak možnost korelovat více faktorů, například pohlaví, jehož vliv zahrnuli předchozí výzkumy a byly objeveny rozdíly (a nejen kvůli tomu, ale nýbrž proto, že normy pro pohlaví používá i NEO-FFI a liší se od norem pro celkový soubor). V případě rozhovorů vnímáme ještě jednu nástrahu, ke které se zde ještě vyjádříme, a to sice přesah okolí přes otázky výzkumníka, jelikož superhapaxy se mohou vyskytovat i na začátcích nebo koncích výpovědí participantů. S tím je potřeba také pracovat a dbát na patřičné hlídání okolí, které se do analýzy dostává.

Byť se může zdát, že k tomuto přístupu máme velké výhrady, opak je pravdou. V této metodě korelace nízkých frekvencí slov s osobností jedince vidíme velký přínos a shodu, a právě proto se snažíme nasměrovat další výzkum směrem, jenž by prokázal její validitu v co největší možné míře a dokázal konzistentnost jazyka s tím, jaký autor je.

Naše čtvrtá analýza se zabývala posouzením povrchového a hlubokého sentimentu v autorském textu. Jednalo se o přepis ze soudního jednání s E. Holmes a dalšími, ze kterého byly vybrány právě její výpovědi. Zkoumali jsme zde, jak se na grafu vykreslí hodnoty sentimentu získané z celého textu v porovnání s okolím. Předpokladem je, že hluboký sentiment má na rozdíl od povrchového tendenci zdůrazňovat extrémy, a to ať už se týká emocí nebo jiné kategorie. Tento efekt jsme skutečně objevili, a pokud se podíváme na křivky, skutečně vidíme, že hodnoty okolí kopírují vesměs směr spojnice celého textu, nicméně jsou o poznání více extrémní. S tímto jevem můžeme dále pracovat a interpretovat ho. Je možné se následně vracet do textu a zkoumat pomocí close readingu, co přesně se v oblastech výkyvu událo, což bychom určitě doporučovali do dalších výzkumů. Tedy, pomocí této studie jsme dokázali, že hluboký sentiment vykazuje silnější hodnoty než povrchový a jeho další využití má potenciál. Další velkou výhodou je práce s menším rozsahem textu, není třeba zkoumat dlouhé texty, které si člověk do jisté míry může vědomě hlídat, ale k analýze nám stačí jeho nevědomá hloubka v podobě nízko frekventovaného lexika.

Naše poslední případová studie se zabývala fenoménem primingu slov. Pro jeho otestování byly sebrány dva korpusy od dvou autorek. Jeden z nich obsahoval úvahy a motivační dopisy a byl kratší (cca 15 000 slov). Druhý zahrnoval práce do školy a přepisy monologu a mluvených zpráv a z důvodu nekonzistence textů byl delší (cca 23 000 slov). Jelikož jednotlivé texty měly různé délky, rozhodli jsme se pro jejich sloučení a strojové rozdělení, abychom předešli potenciálnímu riziku, že mezi vzorky bude nepřiměřený rozdíl v počtu slov, dokonce v řádu tisíců. Nicméně i při slučování textů jsme dbali na tematické řazení, abychom tuto soudržnost příliš nenarušili. Vzhledem k tomu, že podobná analýza (význam vlivu superhapaxů na emoce) dosud neproběhla, slouží tato studie jako výchozí bod pro následující výzkumy, a proto výsledky neporovnáváme s předchozími zjištěními, ale detailně popisujeme různá poznání.

Pro vytvoření dotazníku, kterým jsme priming testovali, jsme vybrali superhapaxy, u nichž analýza LIWC prokázala vysoké hodnoty v kategoriích, které značily emocionalitu.

Nicméně jsme se nespolehli pouze na strojovou analýzu, ale prošli okolí ručně a zvolili slova, která se lidskému oku zdála v souladu s výsledky softwaru. Tento postup se zřejmě osvědčil, jelikož obě autorky volily svá citlivá slova oproti ostatním. Vytvořené věty byly nejprve otestovány na kontrolní skupině participantů, aby se předešlo zkreslení z důvodu nepřiměřeného sestavení větných celků.

Z této studie můžeme vytěžit cenné informace co se týká potřebné délky textu pro zjištění sentimentu. I když jsme použili relativně nízký počet slov (15 000) – relativní z toho důvodu, že předchozí zjištění udávala minimální rozsah vzorku 6250 slov (Faltýnek et al., 2020) a tedy bychom tohoto čísla při rozdělení do tří částí nedosáhli – dostali jsme výsledky, které se ukázaly jako platné. Také se nepotvrdila naše skepse ohledně smíchání různých typů textů (viz autorka Dana). Musíme však podotknout, že sentiment zde nebyl tak jednoznačný, jako v případě Dity. Další naší obavou v případě Dany byl přepis mluvených zpráv, jelikož se jednalo o krátké úseky, které od sebe v některých případech byly odděleny odpověďmi druhé strany nebo časem. Nicméně zřejmě ani z toho důvodu se vliv superhapaxů nezměnil. I když jsme zde používali program LIWC, který se osvědčil, z některých předchozích studií vyplývá, že pro hodnocení sentimentu krátkých úseků, jakými jsou zprávy nebo příspěvky na sociálních sítích, jsou vhodnější přístupy strojového učení (Naive Bayes, Maximum Entropy a SVM) a klasifikační metody (SVM a Multinomial Naive Bayes) než metody LIWC založené na lexikálním přístupu (Tausczik, Pennebaker, 2010; Bermingham, Smeaton, 2010).

K této studii lze závěrem dodat, že poskytuje významné zjištění v oblasti primování pomocí superhapaxů. Opět doporučujeme ověřit výsledky na větším množství participantů a uchopit problematiku i z jiné perspektivy. Dále by bylo vhodné najít spodní práh počtu slov a ověřit, zda se nějak liší u různých druhů textu (korespondence, volný proud myšlenek, přepis monologu apod.). Důležitým zjištěním bylo i prokázání stability sentimentu superhapaxů v čase. Potenciál v této oblasti je výrazný a hovoří i pro praktické využití.

Souhrnem do diskuze předkládáme i jisté otázky a povšimnutí, která se nám v průběhu všech analýz vyskytla a o kterých zde ještě nepadla zmínka. První, již zmiňovanou, je velikost okolí v rozhovorech nebo konverzacích. Vzhledem k tomu, že v těchto případech vnímáme narušení druhé osoby svým proslovem, nedá se určit, zda je možné okolí navázat z předchozí odpovědi či nikoliv. V naší práci jsme se ve většině případů přiklonili k variantě nezahrnutí okolí, které přesáhlo jednu výpověď autora. Další otázkou je, jak je to v případě, že se autor do textu vrací a upravuje ho, vpisuje do něj, popřípadě se k nedopsanému vrací po čase. Nicméně tady nám zázemí může poskytnout poslední provedená analýza, v níž se potvrdilo, že superhapaxy by měly být na čase nezávislé a úpravy od autora by též neměly být na škodu. Jako jeden z nejzajímavějších postřehů vnímáme shodu v superhapaxech u různých autorů. Zde to byla hlavně „doba“ v různých pádech, která se jako superhapax objevila ve více analýzách (od různých autorů). Obecně se mezi superhapaxy objevuje spoustu slov, které nějak souvisí s označením času („rok“ „tejdnů“ atd.). Tento efekt by do budoucna mohl být prozkoumán hlouběji.

Samozřejmě i tato práce má své limity. Tím největším, který je však neovlivnitelný, je nedostatek předchozích studií, se kterými by se zjištěné výsledky daly ověřit. Například hluboký sentiment nebyl doposud nikým jiným dříve (před Faltýnkem, Benešovou a Kučerou, 2022) popsán a stejný osud se týká i superhapaxů. Tím pádem je právě na těchto prvních studiích, aby prorazily cestu novému směru, který se dále může ověřovat a vytvářet tak validní data.

10 ZÁVĚR

Tato práce si kladla za cíl prozkoumat novou metodu získávání hlubokého autorského sentimentu z více úhlů pohledu, jelikož je vzhledem k jejím počátkům prozatím tato oblast neprobádaná. K získání hlubokého sentimentu bylo využito práce s tzv. „superhapaxy“ což je jeden z nově objevených typů nízko frekventovaného lexika, který má potenciál určit nejen autora daného textu, ale také jeho osobnostní charakteristiky, postoje a nálady, které do textu ze sebe přenáší.

Do této chvíle byl zkoumán pouze sentiment celého rozsahu textu, který jsme zde nazvali jako povrchový. Tento typ sentimentu se zabývá zjištěním, jak moc je daný text emočně laděn a jeho praktické využití je například v marketingu nebo politické scéně. K těmto analýzám se používá nejrůznějších softwarových metod, které jsou vhodné pro různé velikosti i tematicky odlišné texty. V naší práci přinášíme nový termín – hluboký sentiment – jenž se zakládá nikoliv na práci s celým textem, ale pouze s nízko frekventovaným lexikem. Toto lexikum má svá specifika a jedná se o speciální případ hapax legomen, kdy je na ně kladena podmínka opakování po určitých úsecích, tedy je nutné mít od daného autora více textů (popřípadě jeho dostatečně dlouhý text rozdělit do několika vzorků), získat z nich hapaxy a tyto hapaxy pak porovnat a najít ty, které se ve vzorcích opakují. Tato citlivá slova pak vykreslují postavu autora a jsou nositeli jeho osobnosti.

Z první studie lze závěrem konstatovat, že přístup průniku superhapaxů může poskytnout užitečný vhled do nízko frekventovaného lexika z jiného úhlu pohledu, než který přináší jejich zkoumání v jednotlivých analýzách různých řezů. Výsledky analýzy ukázaly, že na nízkých frekvencích autor nevědomě užívá určitá slova (často autosémantika) a že v některých případech může být možné spojit výsledky této analýzy s reálnými situacemi. Nicméně je důležité brát tato zjištění s rezervou a vyvarovat se jejich přílišné interpretaci, zejména vzhledem k tomu, že problematika superhapaxů není doposud plně prozkoumána.

Ze druhé případové studie, která byla zaměřená na analýzu přepisu mluveného slova, který přináší autentičnost, ale zároveň nástrahy, vyplynulo, že i tato forma textu přináší důležitá zjištění a dokonce i na nižších rozsazích vzorků. Při analýze transkriptu je však důležité brát v úvahu nespisovné výrazy a fonetické změny řeči a rozhodnout se, zda přepsat text do spisovné formy nebo zahrnout tyto změny všude rovnoměrně. Jelikož strojový přepis

ještě není na dokonalé úrovni, jsou často kladeny vysoké nároky na výzkumníka. Zjištění z této studie naznačilo, že superhapaxy přitahují určité jazykové chování (opakování jiných slov) a v nízkých frekvencích skutečně dochází k synonymní volbě slov. Tyto závěry vedly k ověření předpokladů vztahu osobnostních charakteristik k jazykovému chování v následující studii.

Následující analýza se zaměřila na zkoumání souvislosti jazykového chování (respektive hlubokého sentimentu textu – tzn. okolí superhapaxu v rozsahu zhruba 20 slov před ním a stejný počet za ním) pomocí programu LIWC a osobnostních charakteristik, které vyplynuly ze sebe-posuzujícího inventáře NEO-FFI. Přestože překlady z češtiny do angličtiny byly zhruba kontrolovány, přesnost mohla být ovlivněna hovorovými nebo nestandardními výrazy. Korelace zmíněných dvou metod, odhalila vyšší míru shody mezi sentimentem okolí a osobnostními rysy než při analýze celého textu. Při vyhodnocování výsledků jsme zvolili striktní přístup spočívající v tom, že byly brány pouze korelace vyšší než 0,75 na obou stranách, což však mohlo způsobit rozpory s předchozími studiemi, které byly ve svých metodách méně přísné. Přesto studie vrhá nové světlo na vztah mezi užíváním jazyka a osobnostními rysy, zejména pokud jde o otevřenost vůči zkušenosti. Mezi omezení a doporučení pro další výzkum patří testování korelace na větším počtu účastníků, aby se získaly spolehlivější výsledky, analýza původního psaného textu namísto přepisů a zohlednění dalších faktorů, jako je pohlaví.

V předposlední analýze jsme zkoumali grafické zobrazení povrchového a hlubokého sentimentu, čímž se tato studie lišila od předchozích. Z výsledků je patrné, že hluboký sentiment má lepší vypovídající hodnotu než dosud užívaný povrchový díky tomu, že vykresluje motivy autora na nižších frekvencích. Z těchto analýz se následně může čerpat do dalších analýz čtení textu, jelikož dostaneme přesné údaje o tom, na jaký úsek se zaměřit nejvíce.

Poslední zmíněná případová studie se zaměřila na fenomén slovního primingu. Byly shromážděny dva korpusy od dvou autorek, jeden obsahoval reflexe a motivační dopisy a druhý školní práce, přepisy monologů či mluvených zprávy z konverzace. Tyto dva korpusy byly spojeny a strojově rozděleny, aby se předešlo potenciálním rozdílům v počtu slov. Studie měla za cíl analyzovat vliv superhapaxů na emoce a byla vytvořena dotazníková metoda, která testovala primingový efekt těchto slov. Sentiment superhapaxů byl identifikován pomocí softwarové analýzy LIWC a následného manuálního hodnocení, aby

se předešlo zkreslení dat nepatřičným strojovým zařazením slov a výsledné věty byly testovány na kontrolní skupině, aby se zamezilo nezáměrnému obecně vnímanému kladnému/zápornému hodnocení celků. Studie poskytla cenné poznatky o potřebné délce textu pro určení sentimentu a výsledky byly platné i přesto, že byl relativně nízký počet slov s ohledem na předchozí studie, které udávaly jiný potřebný počet slov. Smíchání různých typů textů u jedné z autorek nemělo významný vliv na superhapaxy, i když v některých případech byl sentiment méně jasný.

Byť má tato práce své limity, je jedna z prvních, která se problematikou hlubokého sentimentu v autorských textech zabývá a poskytuje tak náhled na dosud nepoznanou oblast. Naším cílem bylo z textu především vytáhnout samotného autora a poznat jeho osobnost skrze nízko frekventovaná slova, jež užívá nevědomě a opakovaně. Znát někoho, totiž znamená *tušit všechna jeho opakování*.

LITERATURA

Abbasi, A., Javed, A. R., Iqbal, F., Jalil, Z., Gadekallu, T. R., & Kryvinska, N. (2022). Authorship identification using ensemble learning. Scientific reports, 12(1), 9537.

Alsayat, A. (2022). Improving sentiment analysis for social media applications using an ensemble deep learning language model. Arabian Journal for Science and Engineering, 47(2), 2499-2511.

Baayen, H., Van Halteren, H., & Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. Literary and linguistic computing, 11(3), 121-132.

Baek, Y. M., & Ihm, J. (2021). Word Use as an Unobtrusive Predictor of Early Departure From Organizations. Journal of Language and Social Psychology, 40(2), 238-259.

Bermingham, A., Smeaton. A. F. (2010). Classifying Sentiment in Microblogs: Is Brevity an Advantage? In ACM International Conference on Information and Knowledge Management (CIKM).

Bettis, B. (2021). The Relationship Between Personality Traits, Coping Behaviors, and Language Use Among Individuals Diagnosed with Cancer (Doctoral dissertation, The Chicago School of Professional Psychology).

Biel, J. I., Tsiminaki, V., Dines, J., & Gatica-Perez, D. (2013, December9–13). Hi youtube!: Personality impressions and verbal content in social video [Conference session]. Proceedings of the 15th ACM on International Conference on Multimodal Interaction, Sydney, NSW, Australia.

Binongo, J. N. G. (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship at- tribution. Chance, 16(2): 9–17.

Borchers, C., Rosenberg, J., Gibbons, B., Burchfield, M. A., & Fischer, C. (2021). To Scale or Not to Scale: Comparing Popular Sentiment Analysis Dictionaries on Educational Twitter Data. In EDM.

Boyd, R. L., & Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. Journal of Language and Social Psychology, 40(1), 21-41.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. Austin, TX: University of Texas at Austin.

Cvrček, V. (2017): HAPAX. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny.

Dong, X. L., & De Melo, G. (2018). A helping hand: Transfer learning for deep sentiment analysis. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Faltýnek, D. (2020). It will certainly be found that some words are literally repeated: Horecký's hypersyntax (Celkom iste sa príde na to, že niektoré slová sa opakujů doslovne: k Horeckého hypersyntaxi). Journal of Linguistics (Jazykovedný časopis), 71(2): 185–196.

Faltýnek, D., Benešová, M., & Kučera, O. (2022). Low-frequency critical discourse analysis–methodological framework and manifestos' analysis (a new formal CDA paradigm).

Faltýnek, D., Kučera, O. (2022). Parasyntax jako struktura nízko frekventovaných částí textu Hapax legomenon prostředkem textové koheze.

Faltýnek, D., Matlach, V. (2021). Hapax remains: Regularity of low-frequency words in authorial texts. Digital Scholarship in the Humanities, 37(3), 693-715.

Faltýnek, D., Matlach, V., & Owsianková, H. (2020). Hapax legomena jako indikátor autorského stylu a formální znak koheze textu. Preprint.

Godsay, M. (2015). The process of sentiment analysis: a study. International Journal of Computer Applications, 126(7).

Gonçalves, P., Benevenuto, F., & Cha, M. (2013). Panas-t: A pychometric scale for measuring sentiments on twitter. arXiv preprint arXiv:1308.1857.

Hladká, Z., Novotná, R., Karlíková, H. (2017): Hapax legomenon. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny.

Hřebíčková, M. (2004). NEO–PI-R. NEO osobnostní inventář (podle NEO-PI-R Costy a McCraee). Praha: Testcentrum Hogrefe.

Hřebíčková, M., & Urbánek, T. (2001). Big five. NEO pětifaktorový osobnostní inventář. Praha: Testcentrum Hogrefe.

Huang, Q., Chen, R., Zheng, X., & Dong, Z. (2017, August). Deep sentiment representation based on CNN and LSTM. In 2017 international conference on green informatics (ICGI) (pp. 30-33). IEEE.

Chen, J., Qiu, L., & Ho, M. H. R. (2020). A meta-analysis of linguistic markers of extraversion: Positive emotion and social process words. Journal of Research in Personality, 89, 104035.

Jang, H. J., Sim, J., Lee, Y., & Kwon, O. (2013). Deep sentiment analysis: Mining the causality between personality-value-attitude for analyzing business ads in social media. Expert Systems with applications, 40(18), 7492-7503.

Jelodar, H., Wang, Y., Orji, R., & Huang, S. (2020). Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. IEEE Journal of Biomedical and Health Informatics, 24(10), 2733-2742.

Kateřina Veselovská (2017): POSTOJOVÁ ANALÝZA. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny.

Koutsoumpis, A., Oostrom, J. K., Holtrop, D., van Breda, W., Ghassemi, S., & de Vries, R. E. (2022). The kernel of truth in text-based personality assessment: A meta-analysis of the relations between the Big Five and the Linguistic Inquiry and Word Count (LIWC).Psychological Bulletin, 148(11-12), 843–868.

LIWC (n.d.). How it works. https://www.liwc.app/help/howitworks

Maks, I., & Vossen, P. (2012). A lexicon model for deep sentiment analysis and opinion mining applications. Decision support systems, 53(4), 680-688.

Martincová, O. (2017): OKAZIONALISMUS. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny.

Martins, R., Almeida, J. J., Henriques, P., & Novais, P. (2021). A sentiment analysis approach to improve authorship identification. Expert Systems, 38(5), e12469.

Mehl, M. R. (2006). The lay assessment of subclinical depression in daily life. Psychological Assessment, 18, 340–345.

Minaee, S., Azimi, E., & Abdolrashidi, A. (2019). Deep-sentiment: Sentiment analysis using ensemble of cnn and bi-lstm models. arXiv preprint arXiv:1904.04206.

Nadeau, D., Sabourin, C., De Koninck, J., Matwin, S., & Turney, P. D. (2006). Automatic dream sentiment analysis. In Proceedings of the Workshop on Computational Aesthetics at the Twenty-First National Conference on Artificial Intelligence.

Nichols, R. (n.d.). Introducing the Linguistic Inquiry and Word Count. The University of British Columbia: Centre for Human Evolution, Cognition and Culture – Quantitative Textual Analysis. https://hecc.ubc.ca/quantitative-textual-analysis/qta-practice/linguistic-inquiry-and-word-count/

Onan, A. (2018). Sentiment analysis on Twitter based on ensemble of psychological and linguistic feature sets. Balkan Journal of Electrical and Computer Engineering, 6(2), 69-77.

Pang, B., Lee, L., Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint cs/0205070.

Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015.Austin, TX: University of Texas at Austin.

Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you tweet: Personality expression and perception on Twitter. Journal of Research in Personality, 46, 710–718.

Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., & Benevenuto, F. (2016). Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Science, 5, 1-29.

Selner, M. (12. březen 2016 - 13. únor 2019). Autismus a Chradonnay. http://selner84.blogspot.com/search?updated-max=2016-04-23T15:06:00-07:00&max-results=7&start=91&by-date=false

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. PLOS ONE, 8(9), Article Article e73791.

Schwartz, H. A., Sap, M., Kern, M. L., Eichstaedt, J. C., Kapelner, A., Agrawal, M., ... & Ungar, L. H. (2016). Predicting individual well-being through the language of social media. In Biocomputing 2016: Proceedings of the Pacific Symposium (pp. 516-527).

Tabak, F. S., Evrim, V. (2016, October). Comparison of emotion lexicons. In 2016 HONET-ICT (pp. 154-158). IEEE.

Tausczik, Y.R. & Pennebaker, J.W. (2014). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29 (1), 24-54.

Zhang, B., & Provost, E. M. (2019). Automatic recognition of self-reported and perceived emotions. In Multimodal Behavior Analysis in the Wild (pp. 443-470). Academic Press.

Zhou, Z. G. (2022). Research on Sentiment Analysis Model of Short Text Based on Deep Learning. Scientific Programming.

PŘÍLOHY

Seznam příloh:

1. Abstrakt v českém jazyce
2. Abstrakt v anglickém jazyce
3. Dotazník – hodnocení vět

ABSTRAKT DIPLOMOVÉ PRÁCE

Název práce: Analýza hlubokého a povrchového sentimentu v autorských textech: případové studie

Autor práce: Bc. Libuše Kormaníková

Vedoucí práce: doc. Mgr. Dan Faltýnek, PhD.

Počet stran a znaků: 71, 115 317

Počet příloh: 3

Počet titulů použité literatury: 49

Abstrakt:

Autorský sentiment je do dnešní doby zkoumán pouze z jednoho pohledu, a to sice z práce s textem v jeho plném rozsahu. Nicméně nedávná zjištění naznačují, že autora a jeho osobnost můžeme lépe poznat spíše v hlubokých strukturách, které se vyskytují na úrovni nízko frekventovaného lexika, než na povrchu. Cílem této práce je představit novou metodu pro získání hlubokého autorského sentimentu z více perspektiv. K získání hlubokého sentimentu využíváme analýzy „superhapaxů“, což je nově objevený typ nízkofrekvenčního lexika, který má potenciál určit nejen autora daného textu, ale také jeho osobnostní charakteristiky, postoje a nálady sdělované v textu. Je založen na pravidelném opakování v autorských textech a v jeho okolí se vyskytuje zvláštní jazykové chování. Na případových studiích je představeno, jak se dá s tímto jevem zacházet a jaké skrývá možnosti.

Klíčová slova: autorský sentiment, autorství, hapax legomena, superhapax

ABSTRACT OF THESIS

Title: Deep sentiment and shallow sentiment analysis in authorial texts: case studies

Author: Bc. Libuše Kormaníková

Supervisor: doc. Mgr. Dan Faltýnek, PhD.

Number of pages and characters: 71, 115 317

Number of appendices: 3

Number of references: 49

Abstract:

This thesis addresses the current topic of deep sentiment. To date, authorial sentiment has been examined from only one perspective, namely, working with the text in its entirety. However, recent findings suggest that the author and his or her personality can be better recognized in the deep structures that occur at the level of low-frequency words, rather than on the surface. The goal of this thesis is to present a new method for extracting deep authorial sentiment from multiple perspectives. To extract deep sentiment, we use the analysis of "superhapaxes", a newly discovered type of low-frequency lexicon that has the potential to identify not only the author of a given text, but also his/her personal characteristics, attitudes and sentiments conveyed in the text. It is based on regular repetition in the author's texts and there is a particular linguistic behaviour around it. Case studies are used to show how this phenomenon can be handled and what possibilities it holds.

Key words: authorship, authorship sentiment, hapax legomena, superhapaxes

PŘÍLOHA Č. 3: DOTAZNÍK – HODNOCENÍ VĚT

Hodnocení vět

Přeji krásný den,
prosím o vyplnění následujících položek, které se týkají hodnocení emočního náboje vět.
Nad odpovědí dlouho nepřemýšlejte, odpovězte tak, jak to cítíte, respektive, jak bys nejspíše odpověděli Vy.
Odpovědi zde nejsou správné nebo špatné, jedná se pouze o zjištění vnímání vět.
Děkuji předem za vyplnění

* Označuje povinnou otázku

1. Která z následujících vět na tebe působí nejpozitivněji? *

   Označte jen jednu elipsu.

   - ◯ Chtěla bych ti něco sdělit
   - ◯ Chtěla bych ti něco říct
   - ◯ Chtěla bych ti něco povědět

2. Která z následujících vět na tebe působí nejvíc negativně? *

   Označte jen jednu elipsu.

   - ◯ A pak se událo,...
   - ◯ A tak se stalo,...
   - ◯ A jak už to bývá,....

3. Která z následujících vět na tebe působí nejpozitivněji? *

   Označte jen jednu elipsu.

   - ◯ Mohly jsme si vybrat
   - ◯ Měly jsme možnost si vybrat
   - ◯ Chtěly jsme si vybrat

4. Která z následujících vět na tebe působí nejvíc negativně? *

Označte jen jednu elipsu.

- ◯ Peníze, se kterými platil
- ◯ Kolo, se kterým šel do sklepa
- ◯ Taška, se kterou jel na nákup

5. Z následujícího výčtu slov vyber dvě, která tě zaujmou *

Zaškrtněte všechny platné možnosti.

- ☐ pět
- ☐ jenom
- ☐ mohly
- ☐ lavička
- ☐ poznat
- ☐ vědět
- ☐ pouze

6. poznámka: *

_____