UNIVERZITA KARLOVA V PRAZE

FAKULTA SOCIÁLNÍCH VĚD

Institut sociologických studií

Katedra sociologie

Václav Čepelák

Analýza biografických vyprávění pamětníků s užitím počítačové textové analýzy

Diplomová práce

Praha 2012

Autor práce: **Bc. Václav Čepelák**
Vedoucí práce: **Mgr. Martin Hájek, PhD.**

Rok obhajoby: **2012**

Bibliografický záznam

ČEPELÁK, Václav. *Analýza biografických vyprávění pamětníků s užitím počítačové textové analýzy.* Praha, 2012. 93 s. Diplomová práce (Mgr.) Univerzita Karlova, Fakulta sociálních věd, Institut sociologických studií. Katedra sociologie. Vedoucí diplomové práce Mgr. Martin Hájek, PhD.

Abstrakt

Vedle dotazníkových šetření představují texty důležitý zdroj dat pro sociologický výzkum již od počátků jeho rozvoje. Metody analýzy textů v sociologii zahrnují dva základní vývojové proudy: první představuje kvantitativní obsahová analýza Bernarda Berelsona, druhý hermeneutická analýza Hanse-Georga Gadamera. V posledních dvaceti letech jsou pak oba tyto metodologické proudy ovlivněny rozvojem informačních technologií. Předkládaná práce se zabývá jednou z metod počítačové textové analýzy (CATA), která stojí na pomezí obou těchto metodologických proudů, metodu sledování spoluvýskytu slov. Práce představuje tuto metodu v kontextu ostatních metod analýzy textu a zmiňuje se i o inspiračních zdrojích dalšího rozvoje těchto metod, o korpusové lingvistice a text miningu. Ve druhé části pak rozebírá jednotlivé kroky analýzy spoluvýskytů slov v textu: sestavení textového korpusu, sestavení slovníku, výpočet datové matice a vizualizace vzdáleností slov s užitím metody mnohorozměrného škálování. Metoda je dále aplikována na konkrétní data, dva textové korpusy sestavené z přepisů biografických interview s aktéry československé normalizace, s disidenty a komunistickými funkcionáři. U těchto korpusů je posouzena kvalita modelů v závislosti na volbě parametrů (koeficient vzdálenosti, velikost kontextové jednotky). Následně jsou tyto modely interpretovány. Pro posouzení validity jsou tyto interpretace dále konfrontovány s výsledky kvalitativní hermeneutické analýzy provedené na stejných datech.

Abstract

Besides the social survey data, texts have been an important source of sociological data since the beginning of the development of sociological methodology. Text analysis methods contain two main branches of development: Bernard Berelson's content analysis and Hans-Georg Gadamer's hermeneutic analysis. Both these methodological branches have been influenced by the development of information technologies in the last twenty years. The thesis presented here deals with one of the methods of computer text analysis (CATA), which stands on the border between these two methodological streams, a method of analyzing words' collocations in texts. The thesis presents the method in the context of other methods of text analysis, and mentions sources of inspiration for further development of these methods - corpus linguistics and text mining. The second part discusses the different steps of words' collocation analysis: building a text corpus, dictionary compilation, calculation of data matrix and visualisation of words' distances using multidimensional scaling (MDS). The method is also applied to a specific data, two text corpora compiled from transcripts of biographical interviews with actors of Czechoslovak normalization - with dissidents and Communist functionaries. Quality of the models is assessed, depending on the choice of parameters (distance coefficient, size of the context unit). Subsequently, these models are interpreted. To assess the validity, the interpretations are also confronted with the results of qualitative hermeneutic analysis performed on the same data.

Klíčová slova

metodologie, počítačová textová analýza, mnohorozměrné škálování, biografická analýza, normalizace, disidenti, komunističtí funkcionáři

Keywords

methodology, computer-assisted text analysis, multidimensional scaling, biographical analysis, normalization, dissidents, Communist functionaries

Rozsah práce: 166 685 znaků bez abstraktů a příloh

Prohlášení

1. Prohlašuji, že jsem předkládanou práci zpracoval/a samostatně a použil/a jen uvedené prameny a literaturu.
2. Prohlašuji, že práce nebyla využita k získání jiného titulu.
3. Souhlasím s tím, aby práce byla zpřístupněna pro studijní a výzkumné účely.

V Praze dne 17. 5. 2012 Václav Čepelák

Poděkování

Na tomto místě bych rád poděkoval Mgr. Martinu Hájkovi, PhD. za péči, kterou věnoval vedení této diplomové práce, za poskytnutí veškerých materiálů a podkladů (včetně softwaru) potřebných pro její zpracování a za cenné rady a konzultace, které mi v průběhu jejího zpracovávání poskytl.

Projekt diplomové práce

Autor: Bc. Václav Čepelák (v.cepelak@seznam.cz)
Vedoucí práce: Mgr. Martin Hájek, PhD.
Konzultant: PhDr. Ing. Petr Soukup

Téma diplomové práce

ANALÝZA BIOGRAFICKÝCH VYPRÁVĚNÍ PAMĚTNÍKŮ S UŽITÍM POČÍTAČOVÉ TEXTOVÉ ANALÝZY

Vymezení výzkumného problému a metoda zkoumání

Tématem diplomové práce bude využití počítačové textové analýzy (CATA) na přepisy biografických vyprávění tří skupin aktérů: disidentů, komunistických funkcionářů a „obyčejných lidí“. Rozhovory se zaměřují na život aktérů v Československu před rokem 1989. Autor zamýšlí v práci zejména demonstrovat využití metody počítačové textové analýzy v sociologii a nastínit specifika, možnosti a meze tohoto přístupu. To bude doplněno analýzou tří zmíněných textových korpusů.

Výsledkem analýzy bude identifikace klíčových slov konceptů a analýza jejich blízkosti skrze zkoumání frekvence spoluvýskytů těchto slov v rámci stanovené kontextové jednotky. Data budou následně zobrazena a analyzována pomocí metody mnohorozměrného škálování. Datová matice bude dále analyzována pomocí metody analýzy sítí (social network analysis), která umožňuje dané koncepty popsat pomocí dalších charakteristik.

V rámci analýzy budou nejprve určena klíčová slova s nejvyšší frekvencí výskytu (tj. frekvenční analýza, s užitím softwaru TextStat). Výběr slov bude zčásti arbitrární, neboť je třeba, aby do analýzy vstoupila pouze slova, která sama o sobě nesou význam. Bude vytvořen slovník pro každý korpus, kdy budou sloučeny různé tvary slov (např. pády) pod jeden koncept. Zároveň zde budou odděleny synonymní výrazy (např. stát jako podstatné jméno a stát jako sloveso).

Následně bude využit software COOA, který zjistí frekvence spoluvýskytů konceptů v rámci kontextové jednotky (např. odstavec). Výstupem bude datová matice, která bude dále statisticky zkoumána s užitím mnohorozměrného škálování (SPSS), případně analýzy sítí (např. UCINET).

Dalším cílem analýzy je pokusit se roztřídit koncepty do skupiny (např. aktéři vs. instituce). Tyto skupiny budou stanoveny na základě výsledků úvodní frekvenční analýzy všech tří korpusů. Skupiny je pak možné analyzovat buď společně (jeden výstup pro všechny koncepty s grafickým odlišením skupin), nebo odděleně (různé výstupy pro jednotlivé skupiny konceptů).

Cílem této analýzy je odhalit klíčové koncepty, které tvoří diskurs vyprávění o komunistickém režimu, a identifikovat rozdíly v uspořádání těchto konceptů u jednotlivých skupin aktérů. Metody síťové analýzy mají pomoci lépe odhadnout postavení těchto konceptů v rámci „diskursivních sítí“ a poukázat na centralitu postavení konceptů, tendenci vyskytovat se v různých kontextech či vytvářet významové clustery.

Současný stav poznání

Počítačová textová analýza je v současnosti doménou zejména korpusové lingvistiky a programování internetových vyhledávačů. V sociologii a jiných společenských vědách je tato analýza obvykle využívána v jednodušších formách a koncepty jsou zde často kódovány manuálně (např. v programu Atlas-ti). V poslední době využil pokročilejších a automatizovaných variant CATA Martin Hájek, vedoucí této práce, v rámci projektu Instituce v životních příbězích. Jeho data má v úmyslu analyzovat i tato diplomová práce. K tomu účelu Martin Hájek vyvinul i program COOA, který umožňuje tento způsob analýzy.

Předpokládaná struktura práce

1. Úvod
2. Metodologie počítačové textové analýzy
    i. definice
    ii. typy počítačové analýzy
    iii. oblasti využití
    iv. možnosti a meze využití metody v sociologii
3. Využití metody v praxi
    i. definice problému
    ii. sběr dat a jejich popis
4. Analytická část
5. Diskuze výsledků
6. Závěr

Literatura:

ALEXA, Mellina. Computer-assisted text analysis methodology in the social sciences. ZUMA –Arbeitsbericht 97/07. Mannheim: ZUMA, 1997.

COX, Trevor F.; COX, Micheal A. A. *Multidimensional Scaling*. Boca Raton : Chapman & Hall/CRC, 2001. xi, 308 s. ISBN 1-58488-094-5.

HÁJEK, M. Proměny dimenze soukromého a veřejného v biografických vyprávěních pamětníků. In: Sborník z konference „1989-2009: Společnost. Dějiny. Politika“ A. Gjuričová (ed.). Praha, 2009.

HÁJEK, Martin. Počítačová textová analýza metodou sledování spoluvýskytů slov. *Data a výzkum - SDA Info.* 2010, 4, 1, s. 19-37.

Krippendorff, K. Content analysis: An introduction to its methodology. Thousand Oaks, CA: Sage, 2004.

MOHR, John W. 1998. Measuring Meaning Structures. Annual Review of Sociology, 1998, 24, s. 345–70.

NORUŠIS, Marija J. *SPSS 14.0 advanced statistical procedures companion*. Upper Saddle River : Prentice Hall : SPSS, 2005. xiii, 366 s. ISBN 0-13-174700-2.

POPPING, Roel. Computer-assisted Text Analysis. London : Sage, 2000. x, 229 s. ISBN 0-7619-5378-7.

ROBERTS, Carl W. *Text Analysis for the Social Sciences : Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah : Erlbaum, 1997. ix, 316 s. ISBN 0-8058-1734-4.

VANĚK, Miroslav. *Obyčejní lidé--?! : pohled do života tzv. mlčící většiny : životopisná vyprávění příslušníků dělnických profesí a inteligence*. Praha : Academia, 2009. 1304 s. ISBN 978-80-200-1791-8.

WEST, Mark D. *Theory, Method, and Practice in Computer Content Analysis* . Westport, Connecticut : Ablex Publishing, 2001. 199 s. Dostupné z WWW: <http://www.questia.com/PM.qst?a=o&d=102154188>. ISBN 1-56750-502-3.

Souhlas konzultanta se spoluprací

Konzultant: Mgr. Martin Hájek, PhD.

Datum: Podpis:

_____ _____

Obsah

Seznam grafů

Seznam obrázků

Seznam tabulek

Úvod

Mnozí vědci upozorňují na trend ve vědě, jehož výsledkem bude specializovaná datová věda.[1] Empirický výzkum se v čase proměňuje. Dříve byly produkce a zpracování vědeckých dat velmi náročnou fází výzkumu a to determinovalo i povahu výsledků. Nové vědecké objevy jsou umožněny tím, že se tyto dvě fáze zjednodušují, a to zejména díky počítačům. Vědec v různých oborech tak má k dispozici velké množství dat, které může snáze zpracovat.

Budoucnost vědy je v rozvoji nových způsobů, jak zpracovávat data a získávat z nich nové informace. Z těchto poznatků bude moci těžit i empirická sociologie a získávat o společnosti (resp. společnostech) stále detailnější poznatky. Datový vědec budoucnosti nebude již pravděpodobně stát před otázkou, jak data získat, ale spíše se bude ptát, o čem data vypovídají a co v nich lze objevit, jakým způsobem je zpracovat. Zatímco zpracování dat bude umožněno datovému specialistovi a bude vyžadovat poznatky z oblasti statistiky a informačních technologií, otázky o povaze dat a jejich vztahu k realitě budou naopak otázkami výsostně sociologickými.

Jako data pro sociologické analýzy mohou sloužit i texty, které společnost různými kanály (skrze masová média, internet, byrokratický aparát atp.) produkuje. Texty lze popisovat a analyzovat různými způsoby.

V sociologii – stejně jako v jiných humanitních a společenských vědách – dochází v polovině 20. století k jazykovému obratu.[2] Zkoumání společenských institucí, lidského jednání, kultury a dalších základních sociologických kategorií, je podmíněno jejich

[1] Debata na téma budoucnosti statistiky probíhá zejména v médiích. Zejména lze podtrhnout článek Stevea Lohra v New York Times s názvem *For Today's Graduate, Just One Word: Statistics* (Lohr, 2009). Rozvoj datové vědy je spojován zejména s rozvojem internetu a jako vizionář je v článku citován chief executive Googlu Hal Varian. O datové vědě hovoří Bell, Hey a Szalay (2009) jako o „čtvrtém paradigmatu", tedy vedle teorie, experimentu, počítačové simulace jako o čtvrtém způsobu, jak získávat vědecké poznatky o světě kolem nás.

[2] Obrat k jazyku má své počátky v analytické filozofii, jejímž základním principem je „poznání, že pro řešení filozofických otázek je zcela zásadní analyzovat jazyk, v kterém si tyto otázky klademe a v němž se na ně má odpovídat" (Machová a Švehlová, 2001 : 68). Jazyk a komunikace jsou centrálním pojmem tohoto vědeckého paradigmatu. Jak dále uvádí Machová a Švehlová (2001 : 70): „Každá komunikace je zprostředkována (a) znakově verbálním a neverbálním kódem, (b) sociální strukturou, zvl. sociálními normami a rolemi, (c) cílovou hodnotou; tyto momenty se prolínají a podmiňují". Právě prolnutí znakového kódu jazyka se sociální strukturou je to, na co je zde poukázáno.
Jednotlivé sociální skupiny, třídy, kategorie, organizace či instituce užívají specifický jazyk, zároveň však jsou tyto entity vymezovány a reprodukovány v jazyce. Zkoumání jazyka jako média reprodukce společenských fenoménů má centrální postavení ve fenomenologické sociologii nebo v kritické analýze diskursu. Diskursivní analýza se zaměřuje zejména na vztah mezi jazykem a mocí. V lingvistice (zejm. v pragmatice jako jedné z disciplín lingvistiky) je prolnutí jazyka a společnosti vyjádřeno pojmem komunikační situace, který je chápán jednak v užším významu jako bezprostřední okolnosti a kontexty konkrétní promluvy, ale i v širším významu jako vliv kultury a společenských norem na jazyk a řeč. Více ke studiu jazyka v sociologii srov. Petrusek (1993 : 69 an.).

jazykovým vyjádřením. Sociologové se v rámci tohoto paradigmatu snaží zkoumat jazyk, v němž toto vyjádření probíhá, zkoumat strukturu popisu světa jednotlivců a rekonstruovat jejich realitu v závislosti na jazyce, který užívají. Texty - ať už v podobě autentických dokumentů či přepisů verbálních projevů - zde slouží jako základní analytický materiál.

Tradičně tato snaha o rekonstrukci struktury narativů či textů probíhá s pomocí kvalitativní (hermeneutické) analýzy či pomocí kvantitativní obsahové analýzy. V současnosti je možné budovat stále větší a specializovanější textové korpusy vytvořené různými způsoby (blogy na internetu, internetové diskuze, digitalizované novinové a časopisecké články), mnohdy zasahující až do soukromí jednotlivců (databáze emailů, profily na sociálních sítích, jejichž výsledky jsou prodávány pro účely cílení marketingové komunikace). Práce s velkými textovými korpusy tak znesnadňuje klasickou hermeneutickou analýzu, kdy není v silách výzkumníka přečíst a interpretovat tak velké množství dat. Zároveň se však díky digitální podobě těchto textů otevírají i jiné možnosti zpracování dat a jejich interpretace. A zde je prostor pro počítačovou textovou analýzu (CATA – Computer Assisted Text Analysis).

Jak tvrdí Alexa (1997 : 4): „S využitím textové analýzy má sociální vědec v ruce nástroj, který mu umožňuje *popisovat, klasifikovat, interpretovat* nebo *činit závěry* o společenských normách, hodnotách, chování či strukturách na základě „skutečných“ dat, tj. přirozeně vytvořených textových dat, reprezentativních stejně jako relevantních vzhledem k určitému kontextu (či určitým kontextům) situací, které jsou zkoumány.“[3]

Tato diplomová práce má za cíl představit jeden z přístupů počítačové textové analýzy, *metodu zkoumání spoluvýskytu slov* v textu, a to jednak teoreticky, v kontextu jiných podobných přístupů, a jednak prakticky, skrze využití této metody na konkrétních datech. Cílem práce je tak čtenáře, který podobnou metodu zamýšlí využít, seznámit s šíří podobných metod, se silnými a slabými místy této konkrétní metody, ale zejména mu dát praktický návod, jak metodu aplikovat a jak její výsledky interpretovat.

V první kapitole práce je čtenář seznámen s různými přístupy k analýze textů: s korpusovou lingvistikou, text miningem, kvantitativní obsahovou analýzou a hermeneutickou analýzou. Všechny tyto metody představují důležité inspirační zdroje pro počítačovou textovou analýzu představovanou v této práci. Speciální důraz pak je v této kapitole kladen na užití počítačů v obsahových analýzách, což je případ i zde prezentovaného analytického přístupu.

[3] Kurzíva je převzata z originálního textu.

V druhé kapitole jsou představeny jednotlivé kroky metody obecně. Tato část obsahuje definici klíčových pojmů a vysvětlení fungování jednotlivých kroků metody od sestavení textového korpusu, přes sestavení slovníku k vizualizaci dat s využitím mnohorozměrného škálování.

Ve třetí kapitole jsou na konkrétních datech hlouběji rozebrány jednotlivé kroky metody představené v první kapitole. Budou zde představena analyzovaná data a zhodnocen vlivy volby jednotlivých parametrů analýzy na její výsledek. Na základě toho je zhodnocena reliabilita analýzy.

Ve čtvrté kapitole budou výsledky analýzy interpretovány. Čtenář tak získá představu o tom, jakou povahu výsledků metoda přináší. Výsledky analýzy pak budou srovnány s výsledky hermeneutické analýzy provedené na stejných datech. Účelem tohoto srovnání je demonstrovat validitu tohoto analytického přístupu.

1 Přehled metod analýzy textů

Analýza textových dat se v sociologii uplatňuje od počátku budování empirického sociologického výzkumu. Jako významný milník v této oblasti je možné uvést práci Thomase a Znanieckeho *Polský sedlák v Evropě a Americe* (Thomas a Znaniecki, 1996 [1918-1920]), kde byla jako jedna z metod zkoumání uplatněna analýza osobních dokumentů. Studie měla velmi silné teoretické zázemí. Užitá metodologie pak měla odhalit individuální a sociální stránky lidského jednání a chápání sebe sama.[4] Později se však setkala s kritikou z pozice tehdy převládající pozitivistické metodologie (Blumer, 1979 [1939]).[5]

Systematickou analýzu textových dat, vycházející z pozitivistické metodologie, zavedl v 50. letech do sociologie (resp. metodologie společenských věd) Bernard Berelson (1952) v podobě metody kvantitativní obsahové analýzy.[6] V 60. letech v souvislosti s rozvojem interpretativního paradigmatu dochází k rozvoji alternativních metodologických přístupů se zaměřením na studium interakce. (srov. Petrusek, 1993 : 133 an.)[7]

Obě paradigmatické větve i nadále užívají vlastní metodologii analytického zpracování textů. Trendem posledních dvaceti let je využití počítačů. Užití počítače pro textovou (resp. obsahovou) analýzu lze na různých úrovních a v rámci různých metodologií. Kvalitativní metodologie využívá počítače k budování otevřeného kódu a analytických kategorií, tedy jako nástroj hermeneutické analýzy, který vychází z interpretace dat výzkumníkem a není formalizován. Program ATLAS.ti (2010), který je příkladem počítačového nástroje pro tento typ analýzy. Mezi funkce tohoto programu však patří i možnosti kvantifikace a vizualizace struktur.

Rozvoj počítačů mění užití počítačové textové analýzy. Počítače nepomáhají pouze rychleji klasifikovat texty do předem stanovených schémat (proměnných), jak je tomu u klasické obsahové analýzy, ale zároveň umožňují nové druhy zkoumání, zaměřené zejména

[4] Jak autoři tvrdí: „*Příčinou sociálního či individuálního fenoménu není nikdy jiný sociální či individuální fenomén sám o sobě, ale vždy kombinace sociálního a individuálního fenoménu.*“ (Thomas a Znaniecki, 1996 [1918-1920] : 44, kurzíva původní)

[5] Blumer si pokládá otázku, nakolik mohou osobní dokumenty sloužit jako vědecký nástroj zkoumání. Tvrdí, že tento nástroj neodpovídá žádnému ze čtyř následujících kritérií: (1) reprezentativita, (2) adekvátnost, (3) reliabilita, (4) možnost s pomocí dat validizovat teoretické interpretace (Blumer, 1979 : xxviii – xxix).

[6] V anglicky psané literatuře se pro tuto metodu užívá pojem *content analysis*, tedy *obsahová analýza* bez přívlastku *kvantitativní*.

[7] Konkrétním metodologickým příspěvkem k analýze textů je metoda zakotvené teorie Glasera a Strausse (1973) a tři stupně kódovacího procesu, které navrhuje.

na odkrývání významů v textech.[8] Trendem je proto propojování obsahové analýzy s poznatky lingvistiky (Alexa, 1997 : 6).

Alexa (1997 : 10) dále poukazuje na to, že rozlišení kvalitativní a kvantitativní metodologie není v případě textových analýz triviální. Podle Alexy se v rámci textové analýzy spíše jedná o kontinuální než o dichotomické rozlišení těchto dvou metodologických větví. Alexa uzavírá tuto úvahu tvrzením, že [...] zdůrazňování kvantitativního aspektu analýzy se více vztahuje k faktu, že obsahová analýza (content analysis) patří k empirické výzkumné tradici, a tím odkazuje k průkaznému, systematickému, objektivnímu a na datech založenému přístupu, spíše než pouze k počítání [frekvencí] slov.“ (ibid)

Tuto distinkci, na kterou Alexa upozorňuje, lze vztáhnout na rozlišení mezi explorační a konfirmační kvantitativní analýzou. Zatímco konfirmační přístup směřuje k potvrzení či vyvrácení předem stanovených hypotéz na základě statistického testování (induktivní statistiky), přístup explorační se pokouší v datech odhalit určité – více či méně komplexní – pravidelnosti či vzorce. Tyto dva odlišné přístupy využívají i odlišných statistických procedur (např. explorační a konfirmační faktorová analýza).

V rozvoji textových a obsahových analýz lze tedy identifikovat trend stírání rozdílů mezi kvalitativním a kvantitativním přístupem s rozvojem mezních postupů, který je umožněn pokračující komputerizací celého oboru. Vstupují sem rovněž poznatky lingvistiky a informačních technologií.

V následujících odstavcích se pokusíme vymezit metodologii počítačové textové analýzy ve vztahu k ostatním zmíněným metodám. Nejprve se budeme zabývat *korpusovou lingvistikou* a *text miningem*, tedy inspiračními zdroji z jiných disciplín. Poté se zaměříme na sociologické metody zabývající se textovými daty, tedy kvantitativní obsahovou analýzu a jednotlivé přístupy využívající počítače. Na závěr se zaměříme na kvalitativní metody, zejm. hermeneutickou analýzu. Celá tato kapitola poslouží čtenáři jako ilustrace šíře metod umožňujících zpracovat textová data a jejich metodologické zázemí.[9]

[8] Pro tento přístup se v informačních technologiích užívá termín *information retrieval*. Informační technologie se v práci s automatizovanými způsoby extrakce významů z textu dostaly již velmi daleko. Projevuje se to například v rozvoji vyhledávacích algoritmů pro webové vyhledávače.

[9] Dodejme, že práce si neklade ambice udělat úplný výčet těchto metod. Literatury k tomuto tématu je dostupné velké množství a dotýká se velkého množství oborů, počínaje informačními technologiemi, korpusovou a komputační lingvistikou a konče sociologií, historií či kulturálními studii. Komplexní výčet by tak vyžadoval jednak velmi mnoho prostoru a jednak velmi širokou erudici autora.

1.1 Korpusová lingvistika

Předmětem lingvistiky je jazyk, což je velmi složitý systém a tato složitost je reflektována v složitém dělení disciplín lingvistiky (Černý, 1998 : 72-73). Korpusová lingvistika se specificky zaměřuje na jazyk v jeho přirozeném užívání (Baker, 2006). Pracuje s textovými korpusy, které ve své obecné (nikoliv specializované) formě mají reprezentativně zastupovat jazyk dané země v užité podobě.

Textový korpus je velký soubor digitalizovaných textů (případně přepisů mluveného slova), který reprezentuje způsoby užívání jazyka. Povaha korpusů je vždy determinována způsobem jejich sestavení. Korpusy jsou zkoumány z hlediska struktury (synchronní přístup) a dynamiky (diachronní přístup) jazykového systému. Korpusová lingvistika tak představuje induktivní metodu ke zkoumání užívání a proměn jazyka ve třech rovinách. Tradičně se lingvistika dělí na syntax (jazyková pravidla), lexikologii (slovní zásobu) a pragmatiku (řečovou praxi, realizaci pravidel a slovní zásoby) a korpusoví lingvisté přistupují k jazyku jako k živému systému, který se neustále proměňuje ve všech těchto rovinách.

Jak ukazuje Bakerova (2006) práce, analýza textových korpusů může být přínosnou metodou i pro sociologii a ostatní společenské vědy. Korpusová lingvistika poskytuje metody pro zkoumání jazyka také v jeho společenském kontextu, což předurčuje analýzu textových korpusů jako důležitou metodu pro sociolingvistiku.

Svůj význam ve studiu společenského kontextu užívání jazyka pak mají některé typy specializovaných korpusů.[10] Za specializované korpusy můžeme považovat i databáze mediálních textů, které vytvářejí agentury poskytující monitoring médií. Specifickým textovým korpusem je pak web jako celek,[11] sociální média apod. Pro analýzu jazyka internetu slouží například WebCorp (2012), což je nástroj pracující na podobném principu jako webové vyhledávače, jeho výstupem je však výskyt slova spolu s předem stanoveným kontextem.

Korpusová lingvistika poskytuje celou řadu metodologických přístupů ke zkoumání jazyka, ale je determinována objektem svého zájmu. Zatímco pro sociologii je jazyk médiem,

[10] Ústav českého národního korpusu (Ústav, 2012) archivuje různé druhy textových korpusů. Vedle obecných korpusů (které lze rozdělit na referenční a nereferenční) zde nacházíme i specializované typy korpusů, a to zejména z hlediska způsobu užití jazyka (psaný a mluvený) a z hlediska žánru (publicistický korpus, korespondence). V mluveném korpusu nacházíme například Pražský a Brněnský mluvený korpus či korpus vyučovacích hodin. Právě tyto typy korpusu využívané specifickou společenskou skupinou či kategorií mohou být využity pro výzkum v sociologii, antropologii i jiných příbuzných vědách.

[11] Viz Biewer, Hundt a Nesselhauf (2007).

skrze nějž se reprodukuje (či konstruuje) společenská realita,[12] pro lingvisty je objektem zájmu jazyk jako takový. V rovině procesuální je rozdíl v kontextu, který lingvistu a sociologa zajímá. Lingvisté obvykle zkoumají proměny užívání slov a dodržování gramatických pravidel a zaměřují se na bezprostřední kontext slov. Sociologa naopak zajímá práce s těmito kontexty a zaměřuje se na širší vztahy mezi určitou skupinou slov. Někdy mohou mít oba přístupy podobný zájem.

Na základě analýzy výskytů slova KOMUNISTA můžeme objevit určité významové nuance tohoto slova vzhledem ke kontextu užití, které mohou být zajímavé pro lingvistu i pro sociologa. Zatímco lingvistu zajímá význam slova a jeho gramatické charakteristiky, sociolog se zaměřuje na roli slova KOMUNISTA ve vyprávění aktérů a při konstrukci reality minulého režimu. Jak ale ukazuje Baker (2006), metodologie korpusové lingvistiky je použitelná i pro více sociologicky, resp. diskursivně zaměřené analýzy.

V rámci české sociologie či ostatních humanitních a společenských věd se projevuje zřetelná inspirace lingvistickými disciplínami. Ta je přirozená zejména v oblasti kvalitativní metodologie, pro niž je jazyk základním dorozumívacím kódem a reflexe jeho užívání je důležitým metodologickým nástrojem. V tomto ohledu může česká lingvistika navazovat na tradici Pražského lingvistického kroužku, jehož členové kladli důraz na společenskou a kulturní dimenzi jazyka zejména v užívání pojmu řečová kultura (viz Kraus, 2010 : 131 an.).

V sociologické oblasti lze vliv lingvistiky dokumentovat dvěma monotematickými čísly Sociologického časopisu pod edičním vedením Jiřího Nekvapila (Sociologický časopis, 2002; Sociologický časopis, 2006). Co se týče využívání studia korpusů s využitím statistických metod, v českém sociologickém výzkumu nacházíme práce využívající metodologii, o níž pojednává tato práce (Bayer et al., 2009; Hájek, 2004; Hájek, Kabele, Vojtíšková, 2006; Hájek, Bayer, 2007; Hájek, 2009; k metodologii pak Hájek, 2010). Mezi pracemi korpusových lingvistů lze najít i některé práce, které lze považovat za sociologicky relevantní. Mezi ně patří například Slovník komunistické totality (2010). Tato práce tvoří základ výzkumu totalitního jazyka na multidisciplinární bázi.[13]

[12] S výjimkou specifických disciplín jako je sociolingvistika či sociologie jazyka.

[13] Dodejme, že Slovník komunistické totality představuje čistě statistický popis určitých textových korpusů bez jejich hlubšího rozboru. Je to příležitost pro mediální vědce či sociology zabývající se propagandou. Součástí publikace je i CD s analyzovaným korpusem.

1.2 Text mining

Vedle metodologie korpusové lingvistiky je třeba zmínit se ještě o jednom přístupu, který je alternativou tzv. data miningu, užívanou specificky pro textová data. Nejedná se v pravém slova o obor či disciplínu, ale spíše o multidisciplinární metodologické pole užívané pro identifikaci smysluplných pravidelně se vyskytujících vzorců v datech, konkrétně tedy v datech textových.

Zatímco data mining je záležitostí statistiky a informatiky, u text miningu sem vstupují ještě poznatky lingvistiky. Výsledky text miningu mohou být dále využívány v *komputační lingvistice* (jinak též *natural language processing*), která se pokouší vytvořit takový počítačový systém, který by rozuměl lidské řeči a byl schopen i odpovídat.

Text mining je tedy systém, v jehož rámci jsou v surových textových datech odhalovány pravidelnosti. Tento systém zahrnuje transformaci dat, výpočty a vizualizaci výsledků. Pracuje obvykle na úrovni pojmů či konceptů. Zatímco pojmy (terms) jsou slova či skupiny slov se stejným významem, koncepty vznikají na základě propojení pojmů do určitých struktur, které bývají nazývány ontologie. Soustavy konceptů (ontologie) vznikají v rámci určité domény (oboru či objektu zájmu). Při budování ontologií se navíc často pracuje s předchozí znalostí o doméně (knowledge base), kdy výzkumník sestavuje strukturu konceptů na základ svých znalostí o daném problému.[14]

Feldman a Sanger (2007 : 19) hovoří o distribuci, frekvenci a asociaci jako o hlavních analytických operacích, které systémy text miningu provádějí. Podobně Baker (2006) naznačuje několik přístupů k popisu textů s pomocí korpusové lingvistiky, mezi něž řadí zkoumání frekvence a distribuce slov v textech, konkordance (tj. užívání slov v určitých kontextech s pomocí metody KWIC), kolokace (systematický popis okolí slov) či klíčových slov.

Zaměříme-li se pouze na metody užívané korpusovými lingvisty, budeme limitováni zaměřením na jazyk jako objekt zkoumání. U text miningu nacházíme nepřeberné množství různých přístupů ke statistické analýze textů a také vizualizaci výsledků, naopak zde zase postrádáme jakékoliv zakotvení v teorii. Přesto oba přístupy pro sociologickou metodologii představují významný inspirační zdroj. Metodologie počítačové textové analýzy představovaná v této práci pracuje do velké míry se základními poznatky obou těchto

[14] Výzkumník například ví, že gynekolog, chirurg a oftalmolog jsou druhy lékaře, což ovšem automatický systém nemůže vědět. Na základě znalostí vztahů mezi těmito konceptu může být vybudována ontologie lékařské profese v rámci domény nemocnice. Systém pak s těmito koncepty může pracovat poučeněji.

přístupů, a proto zde existuje potenciál pro rozvoj sociologické metodologie v interakci s těmito přístupy.

1.3 Kvantitativní obsahová analýza a užití počítačů

Autorem kvantitativní obsahové analýzy je Bernard Berelson, který ji definoval jako „výzkumnou techniku pro objektivní, systematický a kvantitativní popis manifestního obsahu komunikace“ (Berelson, 1952 : 74). Podle Krippendorffa (2004: 18) je obsahová analýza „výzkumná technika, která umožňuje činit na základě studia textů (či jiných nositelů významu) replikovatelné a validní závěry o kontextu jejich užití.“ Z obou definic jsou vidět základní vlastnosti této metody. Je to její *replikovatelnost* a *nezávislost na osobě výzkumníka*. Při opakovaném užití stejné metody na táž data dospěje jakýkoliv výzkumník k týmž výsledkům. Další, ne zcela zřejmou vlastností kvantitativní obsahové analýzy je *zobecnitelnost* závěrů.[15]

Krippendorff rovněž sestavil model obsahové analýzy, který popisuje její epistemologický základ. Analyzovaná data (texty) jsou v tomto případě vytržena z kontextu svého užití a zkoumána sama o sobě, tj. jsou interpretována pouze výzkumníkem. Kontext je tak v tomto případě dán výzkumníkovým věděním, s nímž k textům přistupuje. K datům je přistupováno na základě analytických konstruktů, které jsou operacionalizovány do určitého klasifikačního systému. Na základě výsledků zkoumání jsou pak činěny určité logické úsudky, které mohou být na základě validní evidence zobecněny (Krippendorff, 2004: 30 an.).

Užití počítačů v obsahové analýze

Stále se rozvíjející informační technologie postupně proměňují i obsahovou analýzu.[16] Počítače celou analýzu zrychlují a zvyšují její reliabilitu, neboť počítače pracují spolehlivě a rychle. Neřeší však - a spíše zesilují - problém validity, neboť nedokážou textům v pravém slova smyslu rozumět. S užitím počítačů tak hrozí odcizení výzkumníka od jeho dat a je třeba dbát na validizaci výsledků počítačových analýz. Krippendorff upozorňuje, že automatická analýza by měla být podpořena i analýzou skrze čtení původních textů. Počítače dokážou

[15] Podotýkáme, že tato vlastnost se týká pouze konfirmačního přístupu založeného na statistické indukci. Explorativní přístup se primárně zabývá (obvykle komplexnějším) popisem statistických dat bez ambice zobecňování.

[16] Krippendorff využití počítačů v obsahové analýze věnuje speciální kapitolu (Krippendorff, 2004 : 257-312).

v rámci obsahové analýzy zpracovat jen určité mezikroky, jako jsou např. vyhledávání, kódování, řazení, výpočty. Hovoří se proto o Computer Aided (resp. assisted) Text Analysis (počítačem *podpořená* textová analýza).[17] (Krippendorff : 2004, 260-261)

Roli počítačů v obsahové analýze lze rozdělit do několika skupin:

1. Popis typických slov či slovních spojení pomocí výčtu, řazení, počtu a krostabulace
2. Popis textových jednotek (jednotlivých textů) skrze výskyt určitých slov či slovních spojení
3. Komputační obsahová analýza, tj. modelování kontextu užití textů na základě určité teorie významu
4. Interaktivní hermeneutický přístup, tj. tvorba kódovacích schémat v průběhu čtení textů

Podrobný přehled druhů a možností užití počítačové textové analýzy (CATA) poskytuje Alexa (1997 : 4). Tato autorka poukazuje také na fakt, že pojmy obsahová analýza (content analysis) a textová analýza (textual analysis) bývají používány jako synonyma, zejména ve společenských vědách. Hlavní rozlišení, které Alexa (1997 : 8) uvádí, je rozlišení na manuální a automatické přístupy. Zatímco manuální přístupy využívají ručního kódování konceptů, automatické přístupy pracují s výskytem slov a slovních spojení. Výhodou automatických přístupů je rychlost a možnost zpracování velkého množství textů, výhodou manuálních přístupů je přesnost. U automatických přístupů hrozí chyby spojené například s homonymií slov.[18] Pro automatické kódování jsou proto vhodné sémanticky omezené texty, u nichž tyto víceznačnosti nejsou tak časté. Zatímco autoři jako Kathleen Carley a Roel Popping jsou zastánci ručního kódování a tvorby konceptů s pomocí počítačů, Krippendorff (2007) více reflektuje trend využívání automatizovaných přístupů zkoumajících výskyt a vzdálenosti určitých slov.

Alexa (1997 : 14) dále rozlišuje přístupy apriorní (na základě předem stanovených schémat) a induktivní (automatické generování kategorií na základě frekvence výskytu).

Popping (2006) dělí počítačové textové analýzy na tématickou, sémantickou a síťovou. Tématická analýza textu - jako nejjednodušší způsob textové analýzy - se zaměřuje na četnost

[17] V této práci je užíván zkrácený překlad *počítačová textová analýza*, který nevystihuje zcela tuto roli počítače v analýze. Rovněž je v práci využívána anglická zkratka CATA v souladu s literaturou k tomuto tématu.
[18] *Homonymie* je jazykový jev, kdy stejně znějící slova nesou různé významy.

výskytu (nebo spoluvýskytu) konceptů. Texty jsou zasazeny do konceptuálního rámce a vzniká sada proměnných, jejichž hodnoty jsou stanoveny pro každou kódovací jednotku.

Sémantický přístup k analýze textu se nezaměřuje na pojmy, ale spíše na vztahy mezi nimi a jejich význam. Tento přístup se zaměřuje na otázky týkající se "specifických vztahů mezi pojmy používanými v různých sociálních kontextech" (Popping, 2006: 28). V procesu kódování se používá tzv. sémantické gramatiky. Jedná se o model, na jehož základě jsou dané vztahy kódovány. Popping (2006: 28-29) uvádí několik druhů těchto sémantických gramatik. Jedním z druhů je například vzor Agent – Pozice – Akce - Objekt, kde agent a objekt jsou dva subjekty zapojené do vztahu, pozice vyjadřuje povahu tohoto vztahu a akce je určitá aktivita agenta směrem k objektu.

Základem sémantického přístupu je tak udělování významů vztahům mezi pojmy. Podobně však pracuje i síťová analýza. Ta na základě sémantického přístupu vytváří sítě a pojmů a vztahů mezi nimi. Takto definována je síťová analýza poměrně složitá. Základním vztahem mezi koncepty je blízkost či spoluvýskyt.[19] Carley (1997) uvádí pro další popis vztahů mezi koncepty 4 charakteristiky: směr, síla, pozitivita/negativita a význam.

Krippedorff (2007), který se zaměřuje na automatizované přístupy, uvádí pět typů počítačové analýzy, které se liší podle teorie významu, na jejímž základě fungují. Jsou to:

- Slovníkově kódovací přístup (coding/dictionary approach)
- Statisticky asociační přístup (statistical association approach)
- Sémanticky síťový přístup (semantic network approach)
- Memetický přístup (memetic approach)
- Interaktivně hermeneutický přístup

Slovníkově kódovací přístup je charakteristický zkoumáním výskytu slov v textech a jejich spojováním do společných konceptů. Jednou z aplikací je tzv. lemmatizace, tj. spojování různých tvarů téhož slova do jedné skupiny. V abstraktnější rovině tato analýza probíhá tvorbou analytického slovníku, který určitým slovům přiděluje určité analytické dimenze. Následně je skrze výskyt určitých typů slov textu daný text statisticky popsán podle těchto dimenzí. Přístupy se liší v tom, zda dovolují zařadit určitá slova jako indikátor pouze jedné či více dimenzí.

[19] Angličtina užívá těžko přeložitelný výraz *adjacency*.

Příkladem tohoto přístupu je program LIWC (Linguistic Inquiry and Word Count). LIWC (Pennebaker, Booth a Francis, 2007) pracuje se slovníky, které umožňují hodnotit texty podle stupně manifestace negativních a pozitivních emocí, vzájemného odkazování atp. Z frekvence výskytu slov indikujících pozitivní emoce pak usuzuje na pozitivní emocionalitu daného textu.[20]

Statisticky asociační přístup se zaměřuje na zkoumání textů bez nutnosti předchozích úprav. Tento přístup obecně předpokládá, že statistické vztahy mezi výskytem slov a jejich kontextem mohou definovat jejich význam. Tento přístup se kryje s text miningem a můžeme do této skupiny přiřadit i metodu prezentovanou v této práci. Tyto přístupy se mohou lišit druhy užitých statistických analýz a šíří uvažovaného kontextu: mohou se zaměřovat na okolí jednotlivých slov či na texty jako celky a jejich strukturu.

Sémanticky síťový přístup vychází ze dvou inspiračních zdrojů: z analýzy sociálních sítí (Wasserman a Faust, 1994) a sémantických přístupů (Popping, 2006; Carley, 1997). Zkoumá uspořádání vazeb mezi koncepty do kognitivních map a význam vzájemných vztahů těchto konceptů (paradigmatických i syntagmatických). Příkladem integrace sémantických a síťových přístupů je Event Structure Analysis. Jak uvádí Corsaro a Heise (1990), autoři tohoto analytického přístupu, Event Structure Analysis slouží k zarámování kvalitativních dat (konkrétně z etnografického výzkumu) do empiricky podložených logických struktur s pomocí analýzy, která je systematická a uniformní. Jak sám autor uvádí, počítač zde neslouží k usnadnění práce, ale k větší preciznosti analýzy. Analytik s využitím této analýzy nejprve rekonstruuje jednotlivé události v rámci určitého příběhu či pozorování. Jednotlivé události popisuje z hlediska aktérů, cílů, využívaných zdrojů atp.

Memetický přístup se zaměřuje na vývoj významů určitých slov či konceptů v čase a na odkazování na svá předchozí užití v jiných textech. Pojem *mem* pochází od genetika Richarda Dawkinse, který jím označuje prvky lidské kultury („řetězce informací“), které zajišťují její reprodukci, podobně jako geny zajišťují reprodukci biologickou (Dawkins, 1998). Memy – stejně jako geny – se musí reprodukovat, aby zůstaly součástí kultury. Teorie memů – memetika – ovlivnila i rozvoj metodologie CATA. V rámci memetického přístupu se textová analýza zaměřuje na dynamickou stránku významů obsažených v textech.

Krippendorff uvádí jako příklad užití memetického přístupu práci Bestovu (1996). Best hovoří o „korpusové ekologii“ (corporal ecology), čímž zdůrazňuje přirovnání textových korpusů k živým (dynamickým) organismům. Využívá zkoumání kolokací slov

[20] Tohoto přístupu užívá například práce Mihalcey a Pulmana (2009).

v internetových diskusních příspěvcích, čímž se snaží identifikovat sebereprodukující se textové jednotky (tj. memy).

Každý z přibližně 10 000 diskusních příspěvků, které jsou součástí textového korpusu, je převeden na vektor slov, která obsahuje. Následně jsou pomocí analýzy hlavních komponent identifikována slova, která mají tendenci se vyskytovat společně ve stejném příspěvku. Následně se Best pokouší data různými postupy statisticky popsat a zpracovat.

Posledním typem počítačové analýzy podle Krippendorffa je *interaktivní hermeneutický přístup*. Ten využívá hermeneutických metod k vytvoření konceptů, které lze pak zkoumat statisticky. Metoda vytvořená na malém souboru textů pak může být automatizována a aplikována na větší soubory. Tím, že je metoda hermeneutická, vyhýbá se výzkumník interpretačním chybám. Jistá míra formalizace zase umožňuje činit zobecnění. Postup je iterativní: formalizovaná analýza vyvolává další impulzy ke změně konceptualizace a vrací výzkumníka k hermeneutické analýze. Cílem je uspokojivé pochopení textu. Podpora počítače je v tomto přístupu využita třemi způsoby:

- Manipulace s textem a jeho organizace,
- Vnesení více systematicity do výzkumníkova čtení,
- Lepší odlišení určitých distinkcí v textu.

Svým pátým typem analýzy vstupuje Krippendorff do oblasti kvalitativních textových analýz, kterým se bude věnovat následující podkapitola. Celkově však Krippendorffovo třídění ukazuje, jak odlišně chápané statistické (resp. komputační) vyjádření významu v textu může vést k zajímavým postupům. V případě přístupu CATA užitého v této práci zkoumáme společné výskyty slov, která vzájemně definují své významy a identifikují určité významové shluky a oblasti v textu. Jak ale ukazuje Krippendorff, lze na význam pohlížet i jinak a získat jinou povahu výsledků analýzy.

1.4 Kvalitativní (hermeneutická) analýza

Zatímco cílem kvantitativní metodologie je popis, kvalitativní metodologie směřuje k porozumění.[21] Jak tvrdí Čermák (2002 : 11) : „Implikace pro výzkumnou praxi je evidentní - kvalitativní výzkumníci studují jevy ve svých přirozených podmínkách a pokoušejí se jim dát smysl nebo je interpretovat v termínech významů, které jim lidé dávají.“ Jak dále Čermák poukazuje, rozdíl mezi kvalitativní a kvantitativní metodologií nelze redukovat pouze na rozdíl v počtu studovaných jednotek a počtu zkoumaných proměnných. Výzkumník na poli narativního výzkumu se vzdává objektivity a reprezentativnosti metody, zároveň však získává možnost studovat objekt komplexně. Jako další rozdíl mezi kvalitativním a kvantitativním výzkumem se uvádí konstrukce kategorií zkoumání, která je u kvantitativního výzkumu apriorní – obvykle vychází z teorie, zatímco u kvalitativního aposteriorní.[22] (Hendl, 2008 : 44 an.)

Kvalitativní výzkum velmi často pracuje s textovými daty. Mnohé přístupy, např. biografická, narativní, konverzační či diskurzivní analýza, se snaží získávat poznatky o společenské realitě skrze to, co lidé o této realitě vypovídají, jakým způsobem o ní vypovídají a jak se chovají (resp. jak jednají). Tyto poznatky se snaží získat v přímé interakci s aktérem (rozhovor, pozorování) či studiem objektů (např. dokumentů), které nějakým způsobem vypovídají o jedincových postojích, názorech, jednání, každodennosti atp. Vyjmeme-li kvalitativní studium manifestních projevů lidského jednání (např. skrze pozorování), pak základním prostředkem zkoumání společenské reality pro kvalitativního výzkumníka je jazyk (případně jiný znakový systém). A právě text je materiálním vyjádřením jazyka a umožňuje tak jeho systematické studium. Díky tomu se analýza textů stává rozšířenou technikou kvalitativní metodologie.[23]

---

[21] Tím se drží metodologické tradice společenských věd začínající u Wilhelma Diltheye a Maxe Webera (a jeho rozumějící sociologie).

[22] V tomto ohledu bývají metody často kombinovány (tzv. mixed-mode survey). Kvalitativní metodologie může sloužit například jako nástroj ke konstrukci kategorií pro kvantitativní výzkum nebo jako nástroj pro prohloubení porozumění výsledkům kvantitativního výzkumu. Možnostmi kombinace kvalitativních a kvantitativních metod se detailněji zabývá práce Evy Veisové (2009).

[23] Blíže se texty a textualitou v sociologii zabývá Zdeněk Konopásek (1996a; 1996b; 1997). Konopásek se na textualitu a její roli ve společenských vědách dívá ze dvou stran. Z hlediska teorie (Konopásek, 1996a) poukazuje na metaforu společnosti jako textu. Text je na jednu stranu produktem určité komunikační situace, resp. autorské intence, na druhou stranu má materiální formu, která jej z této situace vytrhuje a činí jej faktem. Tento postřeh na jednu stranu pomáhá osvětlit problém dilematu struktury a jednání v sociologické teorii – a dodejme, že s metodologickou stránkou studia textů v sociologii nemá tento Konopáskův článek mnoho společného - , na druhou stranu však je problém jednání a struktury přenositelný na samotné texty. Text je na jednu stranu individuální realizací pravidel a intencí, na druhou stranu má i své strukturní charakteristiky. Jedná se například o studium paradigmatických a syntagmatických vztahů či odkazování textů na jiné texty (tzv.

Text jako základní objekt analýzy je typický zejména pro hermeneutiku, resp. hermeneutickou analýzu, kterou můžeme považovat za první kvalitativní metodu, která reagovala na dominanci pozitivismu ve společenských vědách.[24] Princip této analýzy je vyjádřen hermeneutickým kruhem: výzkumník přistupuje k textu, který interpretuje, s určitým předporozuměním, které v interakci s textem konfrontuje a v této interakci postupně text interpretuje a buduje porozumění.

Jak uvádí Hendl (1998), hermeneutická metoda se vyznačuje pluralitou přístupů. Přesto lze stanovit základní principy textové hermeneutické analýzy. Analytik si stanoví problém či otázku, na níž hledá odpovědi v textu. Stanovuje relevantní místa v textu a utváří první interpretaci. Tuto interpretaci konfrontuje s dalšími částmi textu a s dalšími texty a postupně ji zpřesňuje a upravuje. Součástí interpretace je i vysvětlení („Proč se tak děje?“). Ačkoliv se tedy snaží analytik dobrat porozumění skrze pečlivé studium nejen samotného textu, ale i ostatních relevantních textů, vždy svou analýzu opírá o jednotlivé texty a jejich individuální vlastnosti a významy v nich obsažené. To je významná odlišnost od přístupu počítačové textové analýzy, která se pokouší rekonstruovat významy na úrovni struktury textů.

Hendl (2008 : 63-100) vedle hermeneutiky uvádí stručný přehled dalších jednotlivých teorií, které dále ovlivňovaly vývoj a podobu kvalitativního výzkumu[25] a v návaznosti na to se blíže věnuje jednotlivým metodologickým přístupům (ibid : 101-141). Velmi přesně popsaný postup analýzy nachází zejména u metody zakotvené teorie Glasera a Strausse (1973). Tato metoda využívá kódování jako klíčového postupu. V rámci kódování přisuzuje jednotlivé úseky textu určitým konceptům – v této fázi se jedná o tzv. otevřené kódování. V dalším kroku hledá vztahy mezi těmito koncepty, z nichž utváří kategorie – zde se jedná o kódování axiální. V poslední fázi pak výzkumní své výsledky integruje do zakotvené teorie - provádí selektivní kódování.

V procesu kódování můžeme vidět určitou paralelu se zkoumáním výskytu jednotlivých slov v metodě CATA užité v této práci. Tato metoda pracuje s předpokladem, že výskyt určitého slova (či více slov) zastupuje určitý kontext – či koncept, řečeno slovy zakotvené teorie. Kategorizace konceptů je prováděna zkoumáním frekvence spoluvýskytů jednotlivých slov. Rozdíl je samozřejmě v přesnosti a významové definovanosti konceptů a logických

---

intertextualitu). Studium textů na úrovni individuálních realizací a na úrovni strukturálních charakteristik lze označit za dva komplementární přístupy k jejich popisu.

[24] Viz Hendl (2008 : 70).

[25] Řadí sem fenomenologii, pragmatismus, symbolický interakcionismus, konstruktivismus a realismus. Pro tuto práci není zásadní se těmito teoriemi zabývat, spíše nám půjde o zachycení vývoje metodologie, který byl rozvojem těchto teorií ovlivněn.

vztahů mezi nimi. Nespornou výhodou metody zakotvené teorie je hloubka analýzy a přístup k získání nových a nečekaných poznatků. Pro vybudování celkového pohledu na data a jednotlivé kategorie však musí výzkumník vyvinout velké úsilí, zatímco CATA to dokáže o poznání rychleji a zejména dokáže zpracovat větší množství textů.

1.4.1 Narativní biografická metoda

Na tomto místě je důležité se zabývat také narativní biografická metodou, neboť tato metoda byla užita pro sběr dat analyzovaných v této práci s pomocí CATA. Při využití této metody jsou data získána od respondentů (resp. narátorů či informátorů) v životopisném narativním interview. Metoda vychází z předpokladu, že příběh (narativ) je základním prostředkem udělování významu objektům žitého světa daného individua a zároveň zajišťuje koherenci a vzájemné propojení těchto významů. Důležitou vlastností významů je, že jsou vyjednávány v procesu interakce s ostatními jedinci. Fischer-Rosenthal a Rosenthalová (2001) považují vyprávění za vhodný prostředek sdílení vlastních zkušeností. Sociální zkušenost je podle nich utvářena neustálým interpretačním procesem, její vlastností je proměnlivost v závislosti na komunikační situaci.

V tomto ohledu je důležité minimalizovat intervenci výzkumníka do budování těchto významů. Schütze (1999) uvádí jako důležitý aspekt metody narativního interview snahu povzbudit výzkumníka k vyprávění a minimalizovat nutnost pokládat mu další otázky. Tyto otázky by dále měly být pokládány tak, aby do rozhovoru nevnášely nové tematizace a spíše rozvinuly boční linie vyprávění s nerealizovaným vyprávěcím potenciálem.

1.4.2 Reflexivita v kvalitativní metodologii

Důležitým aspektem kvalitativní metodologie je otázka reflexivity (Konopásek, 1997). Vzhledem k tomu, že utváření významů žitého světa vzniká v interakci, je osoba výzkumníka zároveň aktérem tohoto utváření významů. On sám interpretuje žitý svět druhých lidí vzhledem ke svému vlastnímu žitému světu.

Fischer-Rosenthal a Rosenthalová (2001) upozorňují na problematičnost vztahu mezi realitou a textem. Jeden z bodů kritiky tohoto vztahu je spojen s modelem strukturalismu. Podle autorů se jedná o to, [...] zda lze kontinuitu sociální interakce přiměřeně vysvětlit

pomocí pojmu struktury a jeho polárních forem (hloubková a povrchová struktura, manifestní a latentní smysl atd.).“ Odmítána bývá představa skrytě působící generativní struktury, která má být odkryta pomocí vědeckých metod. Hlavním argumentem je právě situační vázanost komunikace, v níž je daná struktura utvářena.

Druhým bodem kritiky je pro Fischer-Rosenthala a Rosenthalovou otázka kartezianismu v sociálních vědách, tj. rozlišování sociálního světa na kognitivní a objektivní sféru. Toto rozlišení je pouze analytické a je třeba si vždy uvědomovat, že k objektivní sféře – pokud nějaká taková vůbec existuje[26] – se dostáváme skrze komunikační situaci.

Celkově lze vývoj kvalitativní metodologie charakterizovat jako odklon od naivního realismu k zohledňování dalších vlivů na prezentaci biografických narací. Konopásek (1996b) rozlišuje trojí orientaci biografického výzkumu reprezentovanou třemi částmi slova „auto-bio-grafie“ („Já“, „život“, „psaní“).[27] Tyto tři orientace vymezují nejen tři složky „reality“ života, ale zároveň i tři perspektivy, z nichž můžeme dané tři složky pojímat – vzniká tak 9 kombinací. Z těchto kombinací si výzkumník – i s ohledem na výzkumnou tradici, k níž se odvolává – vybírá ty, které jsou pro něj relevantní.[28]

CATA jako metoda z větší části formalizovaná se musí vyrovnávat s kritikou z pozic kvalitativní metodologie – zejm. s ohledem na reflexivitu výzkumné metodologie. Jedná se hlavně o nebezpečí „odcizení výzkumníka od jeho dat“, které bývá zmiňováno ve spojitosti s užitím počítačů v kvalitativním výzkumu. Souvisí to i s kritikou Fisher-Rosenthala a Rosenthalové (2001), zda lze ze studia strukturních aspektů textu vyvodit poznatky o realitě. Souvisí to však i s tím, že počítače činí analýzu méně transparentní, zejm. je-li založena výhradně na datech a nikoli na předem validizovaném schématu.

Vztahu mezi kvalitativní a kvantitativní obsahovou analýzou se věnuje i Alexa (1997 : 10), která řadí CATA na kontinuum „kvalitativní – kvantitativní“ spíše ke kvalitativní výzkumné tradici. I u takto zdánlivě objektivizované metody, jakou je CATA, tedy není možné zapomínat na její kvalitativní aspekty, zejm. na reflexivitu interpretace výsledků. Tyto strukturní aspekty vyprávění je třeba vždy vztahovat k vyprávění samému a k zamýšlenému významu v tomto vyprávění obsaženému.

---

[26] Kritika kartezianismu vychází z toho, zda existuje nějaká situace prožitku, či zda tato situace není utvářena až v komunikaci.

[27] „Já“ odkazuje k definici sebe sama, „život“ k pojetí vlastní životní dráhy a „psaní“ k aktuální komunikační situaci a k textu jako takovému.

[28] Naivní realismus je podle Konopáska reprezentován perspektivou *života*, tzn. vše, o čem jedinec vypráví, lze interpretovat s ohledem na život, který žije.

Výhoda CATA je v tom, že vedle čtenářského pohledu na biografická data umožňuje pohled komplexnější. To, že vypravěči soustavně spojují – či naopak oddělují - určitá témata odhalíme díky CATA mnohem jednodušeji a rychleji než s pomocí hermeneutické analýzy. Zároveň se tímto analytickým přístupem dostáváme „za významy“ obsažené v jednotlivých vyprávěních, na vyšší úroveň analýzy.

1.5 Shrnutí

V první kapitole byly představeny jednotlivé metody analýzy textů v sociologii a příbuzných vědách a jejich inspirační zdroje. Korpusová lingvistika a text mining posloužily jako disciplíny, které dlouhodobě pracují s textovými daty. Obě disciplíny mají význam zejména praktický: korpusová lingvistika se primárně zabývá studiem jazyka, text mining je čistě metodologická disciplína bez teoretického zakotvení z hlediska objektu svého studia.

Sociologie studuje texty (textová data) v rámci kvalitativní a kvantitativní metodologie. Kvantitativní metodologie se věnuje studiu textů pouze okrajově, a sice v podobě kvantitativní obsahové analýzy. Pro kvalitativní metodologii patří naopak texty mezi základní zdroje dat.

Trendem posledních dvaceti let je zejména zapojování počítačů do analýzy textů. Počítače umožňují zrychlení procesu analýzy a zvýšení její komplexity, na druhé straně ale prohlubují otázku validity metod.

Ukazuje se také, že rozlišení na kvalitativní a kvantitativní analýzu přestává být s užitím počítačů striktně dichotomické, ale stává se spíše kontinuálním, tj. dává vzniknout přístupům na pomezí obou metodologií. A právě metoda, která bude v této práci dále prezentována, leží na tomto pomezí, blíže metodologii kvalitativní. Musí se tedy vyrovnávat s požadavky obou výzkumných tradic, na jedné straně s otázkou *replikovatelnosti* a *nezávislosti výsledků na osobě výzkumníka*, na druhé straně hlavně s otázkou *reflexivity* interpretace výsledků.

2 CATA – metoda sledování spoluvýskytů slov

Předchozí kapitola měla představit metodu CATA v souvislosti s ostatními metodami analýzy textů, což posloužilo jako základ pro podrobnější rozbor této metody. Jak bylo rovněž naznačeno v předchozí kapitole, lze rozlišit více různých přístupů v rámci počítačových textových analýz. Nyní se zaměříme na konkrétní metodu zkoumání vzdáleností slov v textových korpusech skrze *frekvenci společných výskytů v kontextových jednotkách*.

Jak již bylo naznačeno, lze v rámci Krippendorfova dělení zařadit tuto metodu do skupiny statisticky asociačních přístupů, kde primárním cílem je odhalit statistické vztahy mezi výskyty jednotlivých slov. V rámci analýzy textových dat je však třeba projít několika kroky, z nichž zvláště přípravná fáze analýzy je časově poměrně náročná.

Využití CATA v této práci má charakter explaračního přístupu, což vychází z toho, že analyzuje velké množství textových dat, ovšem jen těžko lze tato data poměřovat kritériem reprezentativity. Nejde nám tedy o usuzování určitých vlastností na určitou cílovou populaci skrze výběr, ale o komplexnější popis daného souboru textů. Metoda CATA v tomto smyslu slouží primárně k vizualizaci dat, skrze niž můžeme usuzovat na jejich strukturu.[29] Vizualizace nám je přijatelným zjednodušením komplexní struktury dat při zachování možnosti interpretovat data jako celek a nikoli jako určitý partikulární výsek reality.[30]

S pomocí metody CATA se pokoušíme zjistit, která slova mají větší tendenci se vyskytovat společně. Celý text je v této analýze rozdělen do úseků a v každém úseku je vypočítán společný výskyt dvojic slov. K vizualizaci je pak využito mnohorozměrné škálování, tedy statistická metoda, která umožňuje vícerozměrný prostor zredukovat do dvou dimenzí.

V následující kapitole bude nejprve ve stručnosti představen celý proces analýzy s využitím této metody. V další části pak budou klíčové fáze rozebrány detailněji. Jedná se o problém extrakce významu spojený s komputační teorií významu a sestavení slovníku jako

[29] Vědomi si výše citované Alexiny poznámky můžeme řadit počítačovou textovou analýzu prezentovanou v této práci blíže ke kvalitativní analýze, i když ne zcela. Pracujeme zde s kvalitativními daty, i když vzhledem k jejich velkému množství se je nepokoušíme popsat jejich hloubkovým studiem, ale určitým zjednodušujícím popisem typickým spíše pro analýzu kvantitativní. Na rozdíl od kvantitativní analýzy však klasifikační schéma nevychází z teorie, ale je budováno přímo z dat, jak uvidíme dále. Tento přístup je typický spíše pro kvalitativní metodologii.

[30] V tom spočívá jedna z nevýhod kvalitativní metodologie a zejm. grounded theory, že výzkumník se pokouší konstruovat určitou analytickou strukturu či analytický model textových dat bez možnosti pohlédnout na data jako na celek.

klíčovou část celé analýzy. Zbytek kapitoly pak bude věnován mnohorozměrnému škálování, tedy statistické proceduře, která je využívána pro vizualizaci dat.

2.1 Shrnutí postupu analýzy

Sestavení textového korpusu

Postup analýzy je následující. Prvním krokem je sestavení analyzovaného textového korpusu. Textová data mohou být různé povahy: novinové články, přepisy hloubkových rozhovorů, databáze emailů, výsledky internetových vyhledávání atd. Metoda CATA popisovaná v této práci je vhodná pro analýzu tematicky specializovaných korpusů, tj. např. na vyprávění či novinové články pojednávající o určitém tématu. Jejím cílem je zaměřit se na celek tohoto vyprávění, identifikovat klíčová témata a jejich celkové uspořádání v textech.

Výhodou CATA je, že dokáže analyzovat velké textové korpusy. V určitých případech však je třeba texty vybírat a v těchto případech je nutné řídit se kritériem reprezentativity. Příkladem může být výběr novinových článků s určitým obsahem z databáze výstřižkové služby (např. Newton Media). V takových případech je třeba dbát na metodologickou správnost výběrových kritérií.[31]

Sestavení slovníku

Máme-li sestavený textový korpus, je nejprve třeba vytvořit slovník konceptů, které budou do analýzy vstupovat. Slovník tvoří seznam konceptů (lemmat), jejichž rozložení v textu je analyzováno skrze zkoumání počtu vzájemných společných výskytů v textu. Každý koncept je tvořen seznamem slov, která pod daný koncept patří. Obvykle je koncept tvořen různými tvary (pády) téhož lemmatu, tj. základním tvar slova zahrnující ostatní jeho tvary – hovoří se proto o lemmatizaci.

Volba klíčových slov není zcela formalizována, záleží do jisté míry na volbě výzkumníka. Hlavními kritérii jsou: 1. frekvence výskytu v rámci textového korpusu, 2. jejich vztah k předmětu zkoumání a 3. jejich významová jednoznačnost.

[31] Tomuto tématu se dále věnovat nebudeme vzhledem k tomu, že v této práci je pro analýzu použit nevýběrový textový korpus. Pro více informací viz Krippendorff (2004 :111 an.), Baker (2006 :25 an.) nebo Alexa (1997 : 14 an.).

V některých případech je třeba slova dále upravovat. Týká se to zejména synonym a homonym. U homonymních výrazů oddělit daná slova v jednotlivých významech, jako je tomu např. u slov STÁT a STRANA.[32]

Výpočet vzdáleností

Ve druhé fázi jsou na základě slovníku vypočteny vzdálenosti klíčových slov v textovém korpusu, a sice na základě frekvence jejich spoluvýskytu. K výpočtu je využit program COOA (2009), který nejprve textový korpus rozdělí na předem definované kontextové jednotky, a následně v těchto jednotkách načítá počty společných výskytů dvojic slov. Kontextová jednotka je opět dána volbou výzkumníka. Program pak postupuje tak, že vybírá jednotlivé úseky textu čítající daný počet slov a v každém z nich vyhledá dvojice slov, které se zde vyskytují společně. Výsledkem je pak datová matice počtu společných výskytů, která je dále transformována na matici vzdáleností s užitím Jaccardova koeficientu.

Nastavení parametrů textové analýzy pak určuje výzkumníkův pohled na data. Tento proces lze přirovnat k pohledu do mikroskopu, kdy výzkumník rovněž volí kontrastní látku (klíčová slova) a snaží se mikroskop co nejlépe zaostřit (volba kontextové jednotky).

Vizualizace výsledků

Ve třetí fázi je pak tato matice s pomocí procedury mnohorozměrného škálování převedena do podoby dvojrozměrného grafu reprezentujícího vzdálenosti slov. V této fázi již výzkumník pouze volí parametry této statistické metody, jako je zejména volba míry nepřesnosti zobrazení (tzv. stresu).

Dodejme, že celý proces analýzy je obvykle iterativní. Výzkumník těmito fázemi prochází opakovaně, snaží se výsledky interpretovat a dále upravuje a čistí slovník. V průběhu tvorby slovníku (ale i při interpretaci) je třeba používat metodu KWIC (Key Words in Context), která pomáhá nacházet významové nuance u jednotlivých slov či lépe interpretovat blízkost některých slov.

Výstupem je pak graf, kde jednotlivá slova jsou reprezentanty klíčových bodů vyprávění (textu). Z grafu se pak pokoušíme zjistit, která témata se často vyskytují společně a tvoří

[32] Tyto postupy jsou často specifické při užívání různých jazyků. Při aplikaci metody CATA v češtině je tato fáze složitější vzhledem ke komplikovaným pravidlům skloňování a časování.

tematický celek. Tato úvaha však není triviální a stojí na určité komputační teorii významu (Krippendorff, 2004 : 309 an.), tedy určité představě o tom, jak lze významy slov (či konceptů) reprezentovat s pomocí počítače a statistických asociací.

2.2 Validita a reliabilita metody

Validita a reliabilita této metody se projevuje ve dvou aspektech. Validita je dána aspektem sémantickým, tj. otázkou, zda metoda nějak reprezentuje realitu textu a přináší smysluplně interpretovatelné výsledky. Otázkou validity například je, jakým způsobem by měla být interpretována blízkost slov či naopak jejich vysoká vzdálenost.

S tím dále souvisí otázka reliability, tj. otázka, zda grafické zobrazení dat tato data spolehlivě reprezentuje, zda blízkost v grafu reprezentuje blízkost slov ve skutečnosti, zda je výpočet vzdálenosti slov vhodně zvolen atp.

Vidíme tedy, že při reprezentaci složitých textových dat dochází k dvojí redukci: redukci sémantické (spojené s výběrem slov a užitím matice vzdáleností jako indikátoru jejich významu) a redukci matematické (spojené s volbou metody zobrazení těchto vzdáleností).

Sémantická redukce je spojena s hodnocením validity metody. To lze učinit pouze konfrontací výsledků získaných touto metodou s výsledky jiných metod (viz kap. 4). Zhodnocením matematických vlastností modelu, zejm. míru nepřesnosti zobrazení spojenou s volbou parametrů a koeficientů vzdálenosti (viz kap. 3.3 a 3.4).

2.3 Komputační teorie významu a extrakce významu

Krippendorff (2004 : 309 an.) poukazuje na to, že každý přístup k obsahové analýze je zakotven v určité teorii významu, a každá z technik počítačové analýzy je tak závislá na své teoretickém a konceptuálním rámci.

Jak Krippendorff (2004 : 309) tvrdí: „Velice často jsou obsahové analýzy konceptualizovány (dodejme, že naivně) v intencích metod tradičních behaviorálních věd určených k analyzování dat ne-textové povahy: jako měřící nástroj pro generování dat přístupných inferenční statistice, testování hypotéz a modelování příčin a důsledků.“ Krippendorff tak poukazuje na fakt, že statistický popis textu nutně nemusí vypovídat nic o

jeho významu. Je třeba k analýze textových dat přistoupit s určitým teoretickým základem, který umožní skrze počítačový popis textu rekonstruovat jeho význam.

Podle Krippendorffa mají počítačové analýzy tendenci pracovat na základě těchto behavioristických předpokladů, tj. zkoumat texty pouze statisticky bez zohlednění jejich významů. Komputační teorie významu má za cíl „[rozšířit] teoretizování o individuálních kognitivních schopnostech typického čtenáře na různé způsoby využívání textů různými komunitami, o veřejné procesy, které činí určité atributy textu signifikantními, a o sociální procesy, v nichž sídlí sociální instituce“ (ibid). Součástí teorie by měly být i určité analytické konstrukty.

Podle Krippendorffa jsou komputační teorie, vztažené k jeho jednotlivým typům analýzy, stále neúplné a vyžadují značné prohloubení. Je nad rámec této práce věnovat se hlouběji komputační teorii významu. Jedna z dimenzí tohoto problému je však pro pochopení fungování metody CATA velmi důležitá, a sice otázka *extrakce významu*.

Problému extrakce významu se věnují French a Labiouse (2002) ve svém příspěvku *Four Problems with Extracting Human Semantics from Large Text Corpora*, kde identifikují čtyři základní problémové body ve vztahu mezi výstupy počítačové analýzy a významu analyzovaného textu. Jedná se o: 1. vnitřní deformovatelnost sémantického prostoru, 2. detekci abstraktních struktur, 3. možnost zahrnutí sémantické informace, 4. atomizaci slov. Tyto čtyři problémy odkazují k představě, že význam slova lze popsat jeho lokací v sémantickém prostoru.

První problém – *vnitřní deformovatelnost sémantického prostoru* – odkazuje na kontextovou vázanost uspořádání slov. Slova nemají v sémantickém prostoru pevné místo, ale jejich pozice vždy závisí na kontextu zkoumání. Strukturní uspořádání slov a jejich významů není pevné, ale mění se s různými způsoby jejich užívání. Slovo putuje prostorem v souvislosti s tím, jak variuje jeho zamýšlený význam. Problém je v tom, že techniky zkoumající spoluvýskyt slov identifikují spíše určitou průměrnou pozici slov v sémantickém prostoru. To ovšem ne zcela reflektuje kontextové významové variace.

Na tento problém reaguje CATA požadavkem významově jednoznačných slov vstupujících do analýzy. Tato slova umožňují jasněji vymezit kontext svého užití. Oddělování homonymních slov rovněž reaguje na tento požadavek.

Druhý problém – *problém detekce abstraktních struktur* – popisují French a Labiouse jako problém interpretace kolokace dvou slov. Autoři naznačují, že vztahy mezi spolu se vyskytujícími se slovy mohou být různé povahy. Vznikají tak určité *abstraktní struktury*,

které determinují povahu vztahů mezi slovy. Častý spoluvýskyt dvou slov může indikovat určitou strukturní podobnost, ale tento vztah může být i jiné povahy.[33] Pro techniky zkoumání spoluvýskytů jsou tyto struktury těžko zachytitelné.

Tento problém ukazuje na omezenou možnost využití metody CATA. Na jejím základě můžeme definovat určité oblasti textu (vyprávění), ale jen těžko můžeme hlouběji interpretovat roli jednotlivých slov v textu vzhledem k neznalosti povahy vztahů mezi jednotlivými slovy.[34]

Třetím problémem je *zahrnutí sémantické informace* do modelu. Autoři poukazují na rozdíly mezi lidským čtením a automatickým čtením počítačů. Vztahy mezi slovy jsou ovlivňovány znalostí širšího kulturního kontextu, který lze jen velmi těžko zahrnout do počítačového zpracování. Jako příklad autoři uvádějí znalost o existenci napětí mezi Izraelci a Palestinci, nebo znalost, že otec je vždy muž. Význam slov je budován na základě této znalosti, kterou ovšem počítače nemají.

Zahrnutí sémantické informace do modelu v CATA probíhá skrze výzkumníkovu interpretaci. Ačkoliv se tento problém týká spíše omezenosti počítačového zpracování informací, odkazuje i k problému sociologické metodologie, že výzkumníkova interpretace je vždy vázána jeho kulturním a dobovým kontextem.

Čtvrtý problém - *atomizace slov* – vychází z problematičnosti vnímání slov jako základních kamenů textu nesoucích význam. Programy zkoumající spoluvýskyt tak zanedbávají vliv nižší roviny významu, tj. suprasegmentálních prvků (tón, přízvuk, intonace, frázování atp.) či užití různých tvarů téhož slova.

Ačkoliv autoři French a Labiouse vycházejí z jiného způsobů využití statistické analýzy zkoumání spoluvýskytů slov než tato práce, jejich čtyři problémy jsou pro CATA inspirativní poznámkou (snad s výjimkou problému atomizace slov, který se až příliš zaměřuje na nižší rovinu zkoumání). Dobře dokumentují hranice této metody při zachycování významů obsažených v textu.

Metoda zkoumání spoluvýskytu slov v textu, jak je prezentována v této práci, se nepokouší rekonstruovat význam jednotlivých slov na základě jejich umístění v sémantické

[33] Autoři se zaměřují zejména na zkoumání malých kontextových jednotek a v tomto ohledu si všímají dvou základních vztahů mezi slovy: atribučních a relačních. U atribučních vztahů je význam jednoho slova budován skrze vnější podobnost s jiným slovem, zatímco u relačních vztahů je podobnost dána ve vztahu ke třetímu objektu. Zatímco atribuční vztahy jsou snadno detekovatelné, u relačních vztahů, kontextově podmíněných už to tak jednoduché není.

[34] Takovou analýzu si lze představit, ale byla by velmi náročná. Vyžadovala by buď manuálně vyhledat všechny spoluvýskyty a zhodnotit jejich povahu, nebo analyzovat výskyt dalších slov v blízkosti společně se vyskytující dvojice.

síti. Tím se odlišuje od pojetí, které naznačují French a Labiouse. Zatímco tedy tito autoři uvažují o malých kontextových jednotkách (v řádku jednotek slov), v této práce je uvažováno o širších kontextových jednotkách (v řádu desítek slov), které odpovídají délce smysluplné souvislé odpovědi na otázku, tj. přibližně jednomu odstavci textu.

Metoda tak nezamýšlí zachytit podrobnou strukturu textu, ale spíše rekonstruovat uspořádání jednotlivých kontextů v celém korpusu. Kontext může analytik podrobněji poznat s užitím metody KWIC. Při užití metody nás tedy nezajímají vztahy mezi slovy, které generují kontext, ale spíše vztahy mezi kontexty. Naše úvahy pak směřují k vyšší („makro-“) rovině textové struktury.

Tyto vztahy mezi kontexty jsou však vyjádřeny pouze svou blízkostí či vzdáleností v rámci textu. Metoda nedokáže říct nic o významu jednotlivého kontextu v celém vyprávění, o tom, jakým způsobem o tom narátoři uvažují a vymezují-li se vůči daným tématům negativně či pozitivně. S absencí sémantické informace, jak o ní hovoří French a Labiouse, je proto třeba počítat. Zároveň je třeba dávat pozor na interpretaci kolokace. Nelze s jistotou tvrdit, že kolokace určitých slov automaticky znamená logickou vazbu mezi nimi.[35] Chceme-li něco takového tvrdit, musíme to podpořit hlubším rozborem vztahu těchto dvou slov s pomocí jiných metod, např. KWIC.

Asi nejdůležitějším problémem je problém vnitřní deformovatelnosti sémantického prostoru. Poukazuje na to, že volba slov determinuje to, jak je struktura textu ve výstupech metody CATA zobrazena. Interpretace výsledků metody tak musí být prováděna pouze s ohledem na uzavřenou množinu slov, která daný graf reprezentuje.[36]

Z toho vyplývá, že uspořádání slov v grafu se může proměňovat podle toho, jak vybereme klíčová slova do slovníku. Slova v rámci grafu nemají pevné místo, jejich místo je dáno vztahy k ostatním slovům.

Důležitým momentem teorie významu pro nás je pak její propojení s biografickou metodou výzkumu. Sociologové využívající biografickou metodu si uvědomují, že význam narativu je vždy zakotven v komunikační situaci. Při interpretaci výstupů metody CATA je

[35] Viz problém *detekce abstraktních struktur* u Frenche a Labiouse.

[36] Představme si například, že bychom s pomocí metody analyzovali kuchařku, ale pro analýzu bychom zvolili pouze slova reprezentující zeleninu. Získali bychom klasifikaci druhů zeleniny podle toho, které druhy zeleniny se vyskytují ve stejných receptech. Zahrneme-li do analýzy i slova označující druhy masa, získáme zcela novou klasifikaci: maso, které se podává bez zeleniny, maso, které se podává s určitými druhy zeleniny a zelenina, která se podává bez masa. V prvním případě nedokážeme vůbec popsat tu část textu, ty recepty, kde se žádná zelenina nevyskytuje. Zároveň mohou mezi jednotlivými prvky nové vazby. Bude-li kuchařka obsahovat mnoho receptů na vepřové s dušenou zeleninou (mrkví, hráškem, kukuřicí), budou tyto druhy zeleniny v grafu blízko slova VEPŘOVÉ. I když se tyto druhy zeleniny společně v žádném receptu nevyskytují, přiblíží se k sobě díky vazbě na společné slovo VEPŘOVÉ. To jsou ony relační vztahy mezi slovy, o nichž hovoří French a Labiouse.

třeba si být vědom synekdochičnosti jazykové komunikace (viz Kraus, 2008 : 41 an.) ve vztahu ke komunikační situaci. Jak tvrdí Kraus, „podstata synekdochického přístupu ke komunikaci je založena na faktu, že rozsah našeho vědění (i když leckdy jen vědění racionálně nepodloženého, tedy přesněji mínění) je vždy větší než to, co jsme schopni vyjádřit.“ (Kraus, 2008 : 41)

Tato synekdochičnost je podmíněna společnou encyklopedií podavatele a příjemce sdělení (resp. podavatelovým předpokladem o tom, co příjemce ví), komunikační kompetencí obou a znalostí kontextu komunikační situace. Podavatel sdělení, tedy vypravěč, sděluje to, co považuje v dané situaci za relevantní. Vychází zde z encyklopedie adresáta, kdy předpokládá, že určité informace jsou adresátovi známé a není třeba mu je objasňovat. Shoda na významu slov jako nutném předpokladu úspěšné komunikace je vyjádřena ve Wittgensteinově konceptu řečové hry (Wittgenstein, 1993 : 171). V tomto ohledu je pro podavatele sdělení důležitá zpětná vazba příjemce, který tím dává podavateli sdělení najevo, co je třeba vysvětlit a čemu naopak rozumí. U biografické metody by měl výzkumník tuto zpětnou vazbu minimalizovat natolik, že nechává narátora, aby určoval relevanci témat. V tomto požadavku je obsažen předpoklad, že ve vyprávění tímto způsobem utvářeném je nejlépe reflektována životní zkušenost narátora.

Z těchto metodologických předpokladů vychází i metoda CATA zejména ve fázi konstrukce slovníku. Volba slov zahrnutých do slovníku by měla být v první řadě provedena na základě jejich frekvence. Zastoupení jednotlivých témat a jejich uspořádání v rámci textu by mělo reflektovat životní zkušenost narátorů.

U analýzy biografického vyprávění s pomocí metody CATA nám tedy jde o to identifikovat určité skryté uspořádání, které vypravěči (soustavně) volí, aby do něj vložili zamýšlený význam. Z reality komunistického režimu v našem případě vybírají určitá místa, která považují pro posluchače za relevantní a nesamozřejmá, a sestavují je takovým způsobem, aby vyjádřili svůj vztah k tomu, co tehdy dělali. CATA umožňuje toto uspořádání zobrazit s limity, které dobře dokumentují French a Labiouse (2002).

2.4 Sestavování slovníku

Má-li analytik sestaven textový korpus, který bude analyzovat, pak dalším krokem analýzy je sestavení slovníku. Analytik si vždy musí uvědomit, že sémantický prostor, který

se CATA snaží reprezentovat, je dynamický a proměnlivý. Tento prostor analytik zachycuje v určitém časovém bodě, v určité komunikační situaci[37] a zároveň v určitém společenském a kulturním kontextu. Sémantický prostor je vymezován i volbou konkrétních parametrů analytického nástroje.[38] V těchto různých rámcích je třeba o výsledcích přemýšlet a interpretovat je.

Jak uvádí Alexa (1997 : 16 an.), v rámci počítačových textových analýz lze rozlišit dva přístupy. První přístup je apriorní, kdy analytik buduje slovník před sebráním dat na základě svého zájmu, zkušenosti a chápání a vědomostí o kontextu celého zkoumaného tématu. Tímto způsobem vytvořený slovník výzkumník dále pretestuje a validizuje na testovacích datech a následně používá na různých datech, která tímto srovnává. Výsledkem je standardizovaný slovník, který měří určité vlastnosti textu.[39]

Tato práce však využívá přístupu aposteriorního, založeného na datech. Zde je slovník vytvořen na základě frekvenční analýzy výskytu slov v textovém korpusu. Nejedná se o zcela formální proceduru, nýbrž je do velké míry dáno volbou výzkumníka. Jak bylo ukázáno v předchozí kapitole, volba slov determinuje i výsledek analýzy. Slova, která výzkumník zvolí, budou tvořit rámec celé analýzy. Při volbě slov hraje roli několik kritérií:

1. Frekvence výskytu v rámci textového korpusu
2. Významová jednoznačnost slova
3. Vztah slova k tématu výzkumu
4. Použitelnost slova k interpretaci

Výzkumník při sestavování slovníku postupuje tak, že nejprve provede frekvenční analýzu textového korpusu. K této analýze lze použít různé textové programy, např. TextStat (2012). Výskyt slov v textu podle frekvence lze vyjádřit tzv. Zipfovými zákony (viz Manning a Schütze, 1999 : 24 an.).

George Kingsley Zipf ve 30. letech zkoumal statistické rozložení slov v textu a všiml si určitých pravidelností, které shrnul do svých zákonů. Jedním z těchto zákonů – a pro nás nejdůležitější – je zákon distribuce slov podle jejich frekvence, který lze shrnout do následující rovnice.

[37] To se týká okolností toho, jak text vznikl, kdo je jeho autorem, komu je určen.
[38] Konkrétně se jedná o volbu slov, která budou zahrnuta do analýzy, volbu statistické metody a jejích konkrétních parametrů (indikátorů ad.) a volbu zobrazení. Tomu se blíže budeme věnovat později.
[39] Tento přístup by odpovídal Krippendorffovu slovníkově-kódovacímu přístupu.

$$F * r = k \quad (2.1)$$

Zákon říká, že součin frekvence výskytu slov v textu a jejich pořadí mezi všemi slovy podle frekvence je přibližně konstantní. To jinými slovy znamená, že v náhodně vybraném textu se vyskytuje několik velmi frekventovaných slov a velké množství slov málo frekventovaných.[40]

Tento Zipfův zákon je nejznámější, Manning a Schütze (1999 : 27 an.) však uvádějí i další Zipfovy zákony týkající se frekvencí slov v textových korpusech. Zipf například identifikoval korelaci mezi počtem významů slova a odmocninou z jeho frekvence. Pro nás zajímavým Zipfovým poznatkem je, že více frekventovaná slova mají tendenci se v textu vyskytovat v „trsech“, tj. kumulovat své jednotlivé výskyty. Zipf dále identifikoval inverzní vztah mezi frekvencí výskytu slova a jeho délkou.

Zipfovy zákony, zejména první z nich, napoví, jaký bude výsledek frekvenční analýzy textového korpusu. Frekvenční analýza odhalí jako nejfrekventovanější slova spojky a zájmena, která ovšem pro analýzu nemají smysl. Je třeba vybírat slova, která sama o sobě nesou význam. Výběr takových slov zčásti závisí na tématu výzkumu. Obecně se jedná o slova, která mohou odkazovat k nějakému širšímu kontextu. Doporučujeme se zaměřit pouze na podstatná jména, i když pro určitá témata mohou být důležitá i slovesa či přídavná jména.

Ze seznamu je pak třeba vyloučit slova, která mají velkou tendenci reprezentovat tzv. deixe. Kraus (2008 : 68) uvádí, že „významy výrazů užitých v komunikačních aktech se dotvářejí podle konkrétních situací.“ Deixe jsou pak podle Krause takové výrazy, které „na tuto situaci (na její místo, čas, účastníky) bezprostředně ukazují a odkazují“ (ibid). Například slovo HODINA obvykle nenese samo o sobě význam, neboť obvykle utváří časovou lokalizaci děje, o němž se hovoří. Tyto výrazy nejsou pro analýzu vhodné, neboť se vyskytují v mnoha různých kontextech.

Slovník je klíčovou determinantou výsledků celé analýzy. Volba charakteristických slov determinuje obraz, který vznikne. V této fázi má výzkumník největší možnost ovlivnit kvalitu výsledků. V následující výpočetní fázi již lze pouze volit parametry zobrazení dat.

[40] Je třeba upozornit, že Zipfovy zákony jsou empirické generalizace, spíše než v pravém slova smyslu zákony. Je však zajímavé si všimnout, že tento Zipfův zákon přibližně odpovídá Paretovu principu - resp. rozdělení (20% typů slov odpovídá 80% jejich výskytů). Jak Manning a Schütze (1999 : 25) dále upozorňují, tento první Zipfův zákon platí pro slova přibližně v pořadí 10 – 10 000 (u velkých textových korpusů). Nelze jej tedy přesně uplatnit na prvních 10 slov a na nejméně frekventovaná slova v textu.

2.5 Mnohorozměrné škálování

Výpočetní fáze výzkumu umožňuje převést data získaná metodou CATA do podoby grafu, který je možné snáze interpretovat. Graf reprezentuje blízkost a vzdálenost jednotlivých slov podle míry jejich společného výskytu v jednotlivých úsecích textu. Vizualizace této struktury textu je provedena s pomocí metody mnohorozměrného škálování.

Mnohorozměrné škálování[41] je podle Norušis (2005 : 288) statistická metoda navržená pro zkoumání dat vyjadřujících stupeň vzájemné rozdílnosti (dissimilarity data) či podobnosti (similarity data) proměnných. V rámci modelu mnohorozměrného škálování jsou jednotlivé proměnné reprezentovány jako body ve vícerozměrném prostoru. Cílem procedury mnohorozměrného škálování je „najít takovou méněrozměrnou konfiguraci bodů, kde vzdálenosti mezi body co nejlépe reprezentují míru nepodobnosti [tj. původní vzdálenosti] těchto bodů“ (Borg & Groenen, 2005 : 170). Cílem je minimalizovat chybu reprezentace, která je definována jako rozdíl mezi vzdáleností a nepodobností.

Mnohorozměrné škálování umožňuje vizualizaci dat tak, že bližší body jsou si podobné, zatímco vzdálené body jsou rozdílné. Procedura pracuje odlišně s daty podobností, která uspořádá data v prostoru tak, že vyšší podobnost poskládá data blíže k sobě, zatímco nepodobnost určuje vzdálenost mezi daty.

Metodu lze klasifikovat do několika typů. Lze rozlišit škálování jednoduché (na základě jedné datové matice) a replikované (více datových matic). V rámci eukleidovského modelu lze stanovit váhu jednotlivých dimenzí (vážený a nevážený eukleidovský model). Pracovat lze se symetrickou maticí (vzdálenost z bodu A do bodu B je táž jako vzdálenost z bodu B do bodu A) či asymetrickou (obě vzdálenosti se liší). Data mohou do analýzy vstupovat jako matice vzdáleností či jako matice proměnných.

Důležité je rozlišení na metrickou variantu užívanou pro kardinální (intervalová či poměrová) data a nemetrickou variantu užívanou pro data ordinální. Jak uvádějí Borg a Groenen (2005 : 199), ve společenských vědách bývá obvykle využívána ordinální varianta mnohorozměrného škálování, která pracuje s pořadím vzdáleností, s tzv. disparitami.

S ordinální variantou mnohorozměrného škálování pracuje i CATA. Důvod je ten, že spoluvýskyty slov v textu nelze považovat v pravém slova smyslu za vzdálenosti. Vzdálenosti

[41] Anglicky *Multidimensional scaling*, v literatuře je často používána zkratka MDS.

zde reprezentují míru podobnosti rozmístění dvou slov v textu. Pro takový typ dat doporučují Borg a Groenen (2005 : 30) užívat ordinální variantu.

Eukleidovský model prokládá mnohorozměrnými daty rovinu či trojrozměrný prostor, čímž tato data zjednodušuje a zkresluje skutečné vzdálenosti. Ideální proložení roviny je v případě nemetrické varianty iterativní proces, který pro svůj výpočet vyžaduje určité parametry. Přesnost proložení roviny mnohorozměrným prostorem lze stanovit s pomocí několika měr, jako je s-stres, Kruskalův stres či čtvercový korelační koeficient (neboli koeficient determinace). Tyto míry poměřují reálné vzdálenosti v mnohorozměrném prostoru se vzdálenostmi v rovině a vyjadřují jejich (ne)podobnost. Procedura PROXSCAL v SPSS pracuje s hrubým stresem, který je vypočten podle následujícího vzorce (Borg & Groenen, 2005 : 42):

$$\sigma_r = \sum_{(i,j)} [f(p_{ij}) - d_{ij}(X)]^2 \tag{2.2}$$

Hrubý stres je součet čtverců rozdílů mezi mírou nepodobnosti dvou prvků a vzdáleností zobrazených bodů, tj. mezi původní vzdáleností v mnohorozměrném prostoru a novou vzdáleností v prostoru méněrozměrném. Tento hrubý stres je pak dále normalizován, tj. vydělen čtvercem vzdálenosti zobrazených bodů:

$$\sigma_1^2 = \frac{\Sigma [f(p_{ij}) - d_{ij}(X)]^2}{\Sigma d_{ij}(X)^2} \tag{2.3}$$

Z této míry je pak dále odvozen Kruskallův Stress-1 jako její odmocnina.

Minimalizace funkce stres – jako základní algoritmus mnohorozměrného škálování - je umožněna procedurou iterativní majorizace (ibid : 178). Jedná se o modifikaci výpočtu derivace pro vektory a matice. Základním principem této procedury je nalezení jednodušší funkce, která v základních bodech kopíruje průběh původní funkce stres. Tato majorizace může být provedena v různých bodech, přičemž procedura hledá takový bod, kdy se nová funkce nejméně liší od funkce původní.

2.5.1 Práce se stresem

Pro hodnocení kvality modelu bývají v literatuře nejčastěji uváděna Kruskalova kritéria velikosti stresu (viz Hebák, 2005; Cox & Cox, 2001; Borg & Groenen, 2005). Podle Kruskala by stres neměl překročit hodnotu 0,2 a v ideálním případě se pohybovat kolem 0,05. Zobrazení prezentovaná v této práci však pracují se stresy mnohem vyššími.

Borg a Groenen (2005 : 54) uvádějí, že stres jako míra kvality zobrazení mnohorozměrných dat je závislý na několika aspektech, a sice zejména na: 1. počtu proměnných v modelu, 2. dimenzionalitě modelu, 3. druhu a rozsahu chyby měr podobnosti, 4. typu originální konfigurace, která má být zobrazena a 5. počtu chybějících pozorování v datech.

Podle Borga a Groenena proto lze jen těžko usuzovat na optimální a přijatelnou velikost stresu. V tomto ohledu dávají Borg a Groenen dvě základní doporučení. Jednak odkazují na studie datových simulací, které pracují s náhodně generovanými daty a hodnotami stresu pro tyto typy dat.

Autoři konkrétně citují Spenceovu studii (1979), která na základě zkoumání náhodně generovaných dat pro různé počty proměnných uvádí následující vzorec pro výpočet stresu pro m-dimenzionální zobrazení matice n prvků. Hodnota stresu pro konkrétní zobrazení by měla být významně nižší než hodnota náhodně generovaného stresu.

$$\sigma_1 = 0{,}001 * (-542{,}25 + 33{,}8*m - 2{,}54*n - 307{,}26*\ln(m)\ 558*\sqrt{\ln(n)} \qquad (2.4)$$

Na základě tohoto vzorce však lze jen těžko usuzovat na statistickou významnost rozdílu naměřené hodnoty stresu od této simulované hodnoty. Druhým Borgovým doporučením je proto se Stresem zabývat pouze orientačně a spíše zkoumat, zda se zobrazená struktura má tendenci v různých zobrazeních proměňovat či je naopak stabilní.

2.5.2 Interpretace grafů mnohorozměrného škálování

Borg a Groenen (2005 : 81) uvádějí tři hlavní principy interpretace výsledků mnohorozměrného škálování. Prvním principem je sledování tvarů a uspořádání bodů v prostoru. Jedná se o skupiny bodů, které lokálně tvoří lze reprezentovat jako křivky či jednoduché útvary v 2D zobrazení. Tyto útvary pak lze interpretovat jako vyjádření určitého (kvazi)funkčního vztahu mezi body.

Druhým principem je tzv. regionální interpretace. V tomto případě rozdělíme prostor na regiony, které sdružují prvky s podobnými vlastnostmi. V grafu lze identifikovat clustery, tj. shluky bodů umístěných ve vzájemné blízkosti oddělené prázdnými prostory. Tento přístup je nejvýznamněji používán i při interpretaci grafů kolokací slov, kdy určité regiony reprezentují „oblasti“ vyprávění.

Třetím principem je dimenzionální interpretace, kdy je zkoumáno celkové uspořádání bodů v prostoru a hledány určité principy, podle nichž se body řadí do jednotlivých dimenzí.

Tyto tři přístupy jsou podobné třem mnohorozměrným statistickým metodám: mnohonásobné regresní analýze, clusterové analýze a faktorové analýze.

Ve spojitosti s regionální interpretací Borg a Groenen (ibid : 99) uvádějí tři typy základních prostorových konfigurací[42] v grafech mnohorozměrného škálování: *axiální, modulární* a *polární*. Tyto typy uspořádání se vztahují k povaze zkoumaných dat. *Axiální konfigurace* organizuje jednotlivé skupiny proměnných do vzájemně oddělených oblastí – „pruhů“ – oddělených osami. *Modulární konfigurace* je dána vzdáleností od středu a jednotlivé skupiny tvoří kruhy a mezikruží. *Polární konfigurace* rozděluje prostor na oblasti oddělené osami, které se setkávají ve středu grafu, tj. kruhové výseče. Tento model uspořádání je použit jako interpretační nástroj v této práci pro grafy kolokací slov.

Konfiguraci grafů lze zkoumat apriorně i aposteriorně: lze zkoumat, zda mají jednotlivé body podobných vlastností tendenci shlukovat se do některé z konfigurací, lze však také na základě očekávání určité konfigurace zkoumat shlukování bodů. V této práci je uplatňován tento aposteriorní přístup, kdy preferovaná polární konfigurace je dána logikou použitých koeficientů vzdálenosti a statistickými vlastnostmi výskytu slov v textu.

Zároveň je třeba při interpretaci vnímat základní odlišnost mezi regiony, clustery a faktory. U regionů jsou rozhodující substantivní vlastnosti prvků a jejich odlišitelnost od ostatních prvků. Cluster je definován výhradně na základě vzájemných vztahů mezi prvky. Zatímco u regionů je důležité vnímat kontinuitu, u clusterů je důležitější stanovení jasných hranic a volných prostorů odlišujících clustery.

Faktory se od obou významně odlišují svým důrazem na polaritu prostorové (makro)struktury. Faktory jsou v MDS zobrazeny jako osy zobrazení. Díváme se pak, zda logiku uspořádání regionů je možno interpretovat jako působení určitých skrytých faktorů.

[42] Autoři (ibid : 87 ad.) spojují regionální interpretaci výstupů s tzv. fasetovou teorií (*facet theory*). Tato teorie poskytuje podklad pro vytvoření klasifikačního systému, kde jeden prvek může patřit do více tříd (na rozdíl od taxonomie).

2.5.3 Metrika vzdálenosti mezi slovy

Mnohorozměrné škálování pracuje s daty vyjadřujících míru odlišnosti hodnot proměnných (dissimilarity data). Tyto míry lze převést na vzdálenosti v mnohorozměrném prostoru. Pro metrickou variantu mnohorozměrného škálování je nejčastěji využívána Eukleidovská vzdálenost, kterou lze vypočítat pomocí vzorce:

$$\delta_{rs} = \sqrt{\Sigma_i (x_{ri} - x_{si})^2} \tag{2.5}$$

Chen, Härdle a Unwin (2008 : 317) dále uvádí 12 různých dalších metrik či modifikací Eukleidovské vzdálenosti, mezi jinými například korelační koeficient. Tím lze vyjádřit podobnost mnohorozměrného škálování s ostatními mnohorozměrnými metodami, např. faktorovou či clusterovou analýzou. Od těchto metod se mnohorozměrné škálování odlišuje zejména svým primárním důrazem na vizualizaci dat.

Speciální koeficienty jsou využívání pro určení vzdálenosti binárních dat. Tyto koeficienty vycházejí z počtu (resp. podílu) společných výskytů dvou hodnot proměnné. Výpočet těchto koeficientů vychází ze čtyřpolní kontingenční tabulky. Tabulka 1 ukazuje tuto kontingenční tabulku a poslouží dále jako podklad pro vzorce výpočtu jednotlivých koeficientů.

		Výskyt slova Y		
		1 = ano	0 = ne	celkem
Výskyt slova X	1 = ano	a	b	a+b
	0 = ne	b	d	c+d
	celkem	a+c	b+d	a+b+c+d

Tabulka 1: Modelová kontingenční tabulka pro výpočet koeficientů vzdálenosti (převzato z Chen, Härdle a Unwin, 2008 :318) – upraveno autorem.

Koeficient	Vzorec výpočtu
Simpson	$s_{rs} = \frac{a}{\min\{(a+b), (a+c)\}}$
Kulczynski (2)	$s_{rs} = \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right)$
Ochiai	$s_{rs} = \frac{a}{\sqrt{(a+b)(a+c)}}$
Czekanowski-Sørensen-Dice	$s_{rs} = \frac{2a}{2a+b+c}$
Braun-Blanque	$s_{rs} = \frac{a}{\max\{(a+b), (a+c)\}}$
Jaccard	$s_{rs} = \frac{a}{a+b+c}$
Sokal-Sneath-Annenberg	$s_{rs} = \frac{a}{a+2(b+c)}$
Mountford	$s_{rs} = \frac{2a}{a(b+c)+2b}$

Tabulka 2: Vzorce výpočtu jednotlivých koeficientů vzdáleností (převzato z Chen, Härdle a Unwin, 2008 :318).

Chen uvádí celkem 17 koeficientů vzdáleností. V následujícím textu se budeme zabývat pouze 8 z nich, které pro svůj výpočet nepotřebují proměnnou d.[43] Jak uvádějí Borg a Groenen (2005 : 128), tyto koeficienty nejsou vhodné pro vzácně se vyskytující případy, což se týká i výskytu slov v textu. Autoři to zdůvodňují tím, že vzácně se vyskytující případy by

[43] Proměnnou d ve výpočtu nevyužívá první Kulczynského koeficient $s_{rs} = a / (b + c)$, který ovšem má jiný obor hodnot než ostatní koeficienty – od 0 do nekonečna, čímž znemožňuje snadné porovnání jednotlivých koeficientů.

díky vysoké míře společného ne-výskytu byly umístěny velmi blízko sebe. V tabulce 2 tedy uvádíme seznam srovnávaných koeficientů a vzorce jejich výpočtu.

Jednotlivé koeficienty vzdáleností vycházející ze čtyřpolní kontingenční tabulky se liší 1. *rychlostí růstu a ne/linearitou funkce* a 2. mírou, s jakou zohledňují *nerovnoměrnost frekvence výskytu* dvou slov v textovém souboru. Výběr koeficientu tak do určité míry ovlivňuje výsledný graf mnohorozměrného škálování. Použitelnost těchto koeficientů pro analýzu kolokací bude detailně rozebrána při využití na konkrétních datech. Nejčastěji užívaným koeficientem pro tento typ dat je však koeficient Jaccardův (Hájek, 2010; Borg & Groenen, 2005 : 127; Mohammad & Hirst, 2005).

2.6 Shrnutí

Cílem druhé kapitoly bylo v obecných konturách představit průběh celé analýzy, tj. 1. sestavení korpusu, 2. sestavení slovníku a jeho úpravy, 3. výpočet datové matice, 4. vizualizace dat s užitím mnohorozměrného škálování, 5. interpretace dat.

Klíčové body celého postupu, které vyžadují hlubší rozbor, byly představeny podrobněji. Otázka validity je spojena s logikou, s jakou je význam extrahován pomocí statistického popisu textového souboru. Problém extrakce významu se ukazuje jako zásadní v tom smyslu, že činí volbu klíčových slov určující pro povahu celé analýzy. Jak ukazuje text Frenche a Labiouse (2002), analýza spoluvýskytů slov má mnohá omezení, přičemž asi nejdůležitější omezení je dáno *vnitřní deformovatelností sémantického prostoru*, tj. vázanosti pozice slova v sémantickém prostoru na kontext zkoumání.

Otázka reliability je pak spojena s fází vizualizace dat s užitím mnohorozměrného škálování. Tato metoda umožňuje mnohorozměrný prostor vazeb mezi slovy redukovat na prostor dvourozměrný s co nejmenší chybou zobrazení. Vzhledem k tomu, že neměříme vzdálenosti mezi slovy, ale míru jejich vazeb, používáme *ordinální variantu* mnohorozměrného škálování. Souvisí to i s tím, že vzdálenost je zastoupena koeficienty spoluvýskytu, jejichž výpočet vychází ze čtyřpolní kontingenční tabulky vzájemného společného (ne)výskytu.

Obecné formulace z druhé kapitoly budou rozvedeny v kapitole třetí, kde budou použity na konkrétních datech. Bude předvedena jejich interpretace a detailní zhodnocení modelu.

3 Analýza korpusů biografických vyprávění disidentů a komunistických funkcionářů

Následující kapitola má ukázat aplikaci metody CATA na konkrétních datech a dále tak čtenáři ozřejmit její fungování. Cílem této kapitoly je jednak poskytnout návod, jak s výstupy získanými s pomocí této metody nakládat. Zároveň bude zhodnocena reliabilita této metody.

V kapitole budou nejprve popsána data, na něž je metoda aplikována. Následně bude detailně rozebrán proces analýzy – sestavení slovníků, volba kontextové jednotky, volba koeficientu vzdálenosti a zhodnocení kvality vizualizace dat vytvořené s využitím metody mnohorozměrného škálování.

3.1 Analyzovaná data

Metoda CATA je v této části práce aplikována na dva textové korpusy. Jedná se o přepisy biografických vyprávění aktérů reálného socialismu v Československu získané v rámci interdisciplinárního grantového projektu Instituce v životních příbězích,[44] který se zaměřoval na studium životních zkušenosti různých typů aktérů v období vymezeném lety 1968-1989. Na projektu se vedle sociologů podílejí i lingvisté a sociolingvisté. Cílem projektu je „zjistit institucionální podmíněnost biografických vyprávění pomocí inovativní víceúrovňové metodologie a zasadit tato zjištění do kontextu institucionální analýzy moderní společnosti.“ (Projekt, 2011)

Rozhovory probíhaly zčásti biografickou metodou, kdy narátoři byli požádáni o vyprávění svého životního příběhu. V druhé části pak rozhovor probíhal formou polostrukturovaného interview, kdy byly narátorům pokládány doplňující otázky ohledně jejich životní dráhy.

Celkově byly v rámci projektu provedeny rozhovory se čtyřmi skupinami aktérů: 1. disidenti, 2. komunističtí funkcionáři, 3. dělníci, 4. inteligence. Tabulka 3 shrnuje základní charakteristiky provedených rozhovorů.

[44] GAP404/10/0790 - Instituce v životních příbězích. Víceúrovňová srovnávací analýza biografických vyprávění tří skupin aktérů české společnosti 2. poloviny 20. století (2010-2012, GA0/GA)

	Celkový počet narátorů	Průměrná délka rozhovoru (počet slov)	Průměrná délka odpovědi[45]	Průměrný věk narátorů	Podíl mužů
Disidenti	66	19224	482	60,6	85%
Funkcionáři	32	22086	252	66,3	97%
Dělníci	56	12989	126	64,9	54%
Inteligence	56	20481	267	64,2	64%

Tabulka 3: Vlastnosti analyzovaných textových korpusů

3.2 Slovníky

Slovníky pro analýzu obou korpusů byly konstruovány *aposteriorním způsobem*, tj. na základě dat. Při konstrukci slovníku byla primárně zohledňována kritéria frekvence slov a významová jednoznačnost. Nejprve byla v původních hrubých textech každého z korpusů provedena frekvenční analýza v programu TextStat (2012). Bylo rozhodnuto, že do analýzy vstoupí pouze podstatná jména. Ta zastupují objekty, které považujeme za hlavní nositele významu a tvůrce kontextu.[46]

Na základě výše uvedených podmínek dále výzkumník vybírá širší skupinu slov. Doporučujeme si stanovit určitou minimální frekvenci, která oddělí potenciální slova, která do analýzy vstoupí. Frekvenční analýza je schopna odhalit nikoli výskyt konceptů, ale jejich jednotlivých tvarů slov. S tím je třeba v této fázi tvorby slovníku počítat a stanovit širší skupinu slov.

Další fází je totiž *lemmatizace slov*,[47] tzn. sloučení různých tvarů téhož slova pod jeden koncept. V této fázi se pořadí výskytu slov může proměňovat. Pro každé slovo v prvním výběru je třeba stanovit všechny tvary (pády), které se v textu vyskytují. Slovník pak výzkumník vytváří v textovém editoru, kde jednotlivé tvary vynese do téhož řádku a oddělí je čárkou. Lze rovněž použít regulárních výrazů, tj. zkrácených slov s hvězdičkou. Například všechny tvary slova ČLÁNEK lze shrnout do jednoho řádku jako „*článek, článk**,

[45] Průměrná délka odpovědi ukazuje, jak dlouhé odpovědi dávají narátoři na otázky tazatelů. Jsou-li odpovědi delší, ukazuje to, že narátoři dokážou sami vést vyprávění a nepotřebují příliš mnoho otázek. Krátké odpovědi naopak ukazují, že narátoři nedokážou o tématu sami vyprávět a více potřebují podporu tazatele.
[46] Toto rozhodnutí je čistě arbitrární a bylo provedeno ad hoc vzhledem k tématu analýzy.
[47] KONCEPTY, resp. LEMMATA jsou v textu označována velkými písmeny. *Jednotlivá slova* jsou označována kurzívou.

článcích“.[48] Při užití regulárního výrazu však je vždy třeba zkontrolovat, zda se v textu nevyskytne slovo, které pod lemma ČLÁNEK nepatří. To nám umožní opět TextStat. Zároveň je třeba dávat pozor, aby každý regulární výraz byl jednoznačný, tj. aby žádné slovo v textu nebylo možné reprezentovat dvěma slovníkovými slovy.

Po vytvoření první široké verze slovníku zjistíme *frekvenci jednotlivých lemmat*. To již učiníme s užitím programu COOA (2009). Pořadí slov podle frekvence se nyní může mírně měnit, protože některá slova mohou mít více různých tvarů (ČLOVĚK: *člověk, člověka, člověku, člověče, člověkem, lidé, lidí, lidem, lidi, lidech, lidmi*), zatímco jiná mají tvarů méně (ŘEŠENÍ: *řešení, řešením, řešeními, řešeních*) či jen jeden (zkratky).

Dále je třeba u nejfrekventovanějších lemmat *zkontrolovat jejich kontext* s užitím procedury KWIC (Key Words in Context), která zobrazí všechny výskyty zadaného slova spolu s okolními slovy. Pro tuto proceduru užijeme rovněž program TextStat.[49] U slov kontrolujeme, zda se v textu nevyskytují ve více významech – v takovém případě je třeba tyto významy oddělit. U některých slov je to předem jasné. Například slovo STÁT může být podstatné jméno i sloveso, což jsou homononyma. Oba tyto výskyty nemají žádný významový vztah a je třeba je oddělit. To učiníme tak, že projdeme každý výskyt slova a slovesu stát v textu předřadíme nějakou předem stanovenou značku (např. _stát). Tato procedura je časově poměrně náročná, ale je nutná.[50]

Poměrně často se také stává, že některé významové nuance si výzkumník neuvědomí, je proto třeba provést kontrolu kontextů u všech slov. Analytik tím zároveň získá přesnější představu o roli slova v textu, což mu pomůže při interpretaci.

U některých slov naopak existují synonymní výrazy, které je možné slučovat pod jeden výraz. To však doporučujeme dělat jen velmi opatrně a v odůvodněných případech. Není například problematické slučovat slova SESTRA a SÉGRA pod koncept SESTRA.[51] Méně snadné je pak sloučit například jednotlivé druhy škol, o kterých narátoři vyprávějí pod

[48] Užití regulárního výrazu je vhodné hlavně proto, že urychlí zpracování textu a výpočet vzdáleností v programu COOA (2009). Při tvorbě slovníku je dále třeba myslet na to, aby slova byla disjunktní. To se týká zejména používání regulárních výrazů, jako např. „komunist*“, kde hvězdička nahrazuje libovolný počet znaků. Tyto výrazy doporučujeme používat zejména proto, že zrychlují výpočet výsledné matice v programu COOA. Je třeba však dávat pozor, aby pod takto definovaný koncept nebylo omylem zahrnuto i jiné slovo a aby se takto definovaný koncept nepřekrýval s jinými koncepty.

[49] V programu TextStat se daná funkce nazývá *Concordances*.

[50] Proceduru lze někdy urychlit hromadným nahrazením určitých ustálených výskytů. To je dobře vidět u slova STRANA, kde je třeba oddělit výskyt slova ve významu „politické uskupení“ od ostatních výskytů. Toto slovo se často vyskytuje v ustálených spojeních, jako např. „na jedné straně“, „na jednu stranu“. Samozřejmě je třeba dát pozor, aby se náhodou námi požadovaný význam nevyskytl ve formě podobné tomuto ustálenému spojení.

[51] V našem případě nezkoumáme, zda se kontext a potažmo význam slova *sestra* nějak liší od slova *ségra*. Jde nám o pozici rodinných příslušníků v rámci celého vyprávění.

společné koncepty. Komunističtí funkcionáři například prošli různými druhy terciárních vzdělávacích institucí (např. VUML), které byly pro účely analýzy zahrnuty pod jeden koncept. Podobným problémem pak bylo v rámci výskytů slova ŽENA oddělení výskytů tohoto slova ve významu MANŽELKA a zahrnutí pod tento koncept.

Máme-li sestavený a vyčištěný širší slovník, vybereme *několik nejfrekventovanějších slov*, s nimiž vstoupíme do analýzy. Stanovení počtu slov, která do analýzy vstoupí, je dáno několika kritérii. První kritérium je přehlednost výstupů. Příliš velké množství slov by graf učinilo nepřehledným.

Druhým kritériem je minimální frekvence výskytu slova v textu, která závisí i na velikosti textového korpusu. U velkých textových korpusů je vzhledem k platnosti Zipfových zákonů zaručeno, že i méně frekventovaná slova budou mít dostatečnou frekvenci, aby mohla vstoupit do analýzy. Doporučujeme, aby minimální frekvence lemmatu ve slovníku byla okolo 100. Méně frekventovaná slova mají přílišnou tendenci tvořit náhodné vazby, které lze jen těžko interpretovat strukturně.

V této práci vstupuje do analýzy celkem 50 slov, což je hranice daná zkušeností s využíváním této metody a možností interpretace celé vizualizace (Hájek, 2010 : 24-25).

Zároveň doporučujeme zkontrolovat rozložení výskytu lemmatu v textu, např. s pomocí programu AntConc (Anthony, 2011). Získáme představu o tom, zda je dané slovo rozloženo po celém textu, nebo je kumulováno do jedné jeho části. V našem případě, kdy pracujeme s vyprávěními různých aktérů jako s jedním textovým souborem, se může stát, že jeden narátor může svým specifickým vyprávěním ovlivnit celkový obraz dané skupiny.[52] Je pak na zvážení výzkumníka, zda dané slovo v analýze ponechá s vědomím této anomálie, či jej z analýzy vyřadí.

3.2.1 Slovník pro korpus disidentů

Na základě frekvenční analýzy korpusu vyprávění disidentů bylo vybráno 270 slov, která byla potenciálními kandidáty pro vstup do analýzy. Z těchto 270 slov byla vytvořena první verze slovníku.

Mezi homonymní výrazy, které bylo třeba oddělit, patřilo zejména zmíněné slovo STÁT. Zároveň v některých případech byly pod jeden koncept zahrnuty synonymní výrazy.

[52] Příkladem je slovo CÍRKEV v textovém korpusu disidentů, kdy toto slovo významně používala pouze subskupina katolických disidentů, zatímco v ostatních vyprávěních bylo marginální.

Příkladem je slovo RODIČ, kam byly zahrnuty výrazy označující matku (*máma, maminka, matka*) i otce (*otec, táta, tatínek*) a jiné synonymní výrazy (*naši* ve významu rodiče). Vedle zmíněného konceptu RODIČ bylo podobně nakládáno i s koncepty MANŽEL a SOUROZENEC. V těchto případech byla pod jeden koncept zahrnuta slova *manžel/manželka* a *bratr/sestra*. Zároveň sem byly zahrnuty různé varianty označení rodinných příslušníků, jako např. *žena, muž,*[53] *ségra, brácha*.

Specificky bylo pracováno se slovy označujícími VĚZNĚ a VĚZENÍ. Zde byla pod jeden koncept zahrnuta různá označení: *mukl, trestanec, kriminálník, zavřenej, lágr, loch, basa, kriminál*.

Jednou z důležitých součástí bylo vzdělání a škola. Zde bylo v obou korpusech odděleno základní a středoškolské vzdělání od vzdělání vysokoškolského. V některých případech samostatně použitého slova *škola* bylo třeba z širšího kontextu odvodit, zda narátor hovoří o škole střední či vysoké.

Postupně bylo vybráno 50 slov, která byla použita pro první analýzu. Výsledný graf byl zhodnocen z hlediska možností interpretace a vhodnosti výběru slov. Na základě vizualizace dat byla vybrána některá problémová slova, která nebyla snadno interpretovatelná či naopak nesla více různých významů. Na základě první vizualizace tedy byl slovník dále dočišťován. Tento proces byl několikrát zopakován.

Výsledkem pak byl slovník pro korpus disidentů, který je uveden v Příloze 1, včetně frekvence každého lemmatu. Slovník obsahuje centrální lemma ČLOVĚK (6529 výskytů) a další velmi frekventované slovo RODIČ (2257). Od třetího slova v pořadí (PRÁCE, 1104) již frekvence dalších slov klesá pozvolna.

3.2.2 Slovník pro korpus funkcionářů

Stejným způsobem pak vznikl i slovník pro korpus funkcionářů, uvedený v Příloze 2. Specifikem korpusu funkcionářů bylo zahrnutí institucí vzdělávající komunistické funkcionáře, jako byla Večerní univerzita marxismu-leninismu (tzv. VUML), pod koncept VYSOKÁ ŠKOLA. Dalším důležitým funkcionářským slovem byl VÝBOR. Slovo *výbor* bylo velmi často součástí zkratky, jako např. *ÚV*, *MNV*, *KNV*. Podobně bylo naloženo

[53] Zde bylo samozřejmě třeba oddělit ty výskyty slova *žena/muž*, kde se nejednalo o označení partnera, ale o označení člověka obecně.

s pojmem *svaz*, který byl zahrnut pod koncept ČSM, jednalo-li se o Svaz mládeže (podobně jako ČSM či SSM), resp. pod koncept SSSR, jednalo-li se o Sovětský svaz.

Další speciální úpravou bylo vytvoření dvou konceptů pro slovo *práce*. Toto slovo se ve vyprávění funkcionářů vyskytovalo ve dvou základních distinktivních významech: 1. povolání, činnost zajišťující obživu (PRÁCE), 2. *politická práce*, příp. *stranická práce, práce strany* jako speciální výraz pro činnost vyplývající z plnění funkce ve stranické hierarchii, resp. činnost celé politické instituce (POLITICKÁ PRÁCE).

Frekvenční analýza slovníku funkcionářů (rovněž obsaženého v Příloze 2) ukazuje podobné rozložení frekvencí jako v případě slovníku disidentů: tři lemmata s výrazně vyšší frekvencí (ČLOVĚK, 3574 výskytů; VÝBOR, 2475; STRANA, 1945) a zbytek lemmat s postupně klesající frekvencí (FUNKCE 890; VYSOKÁ ŠKOLA, 875; atd.).

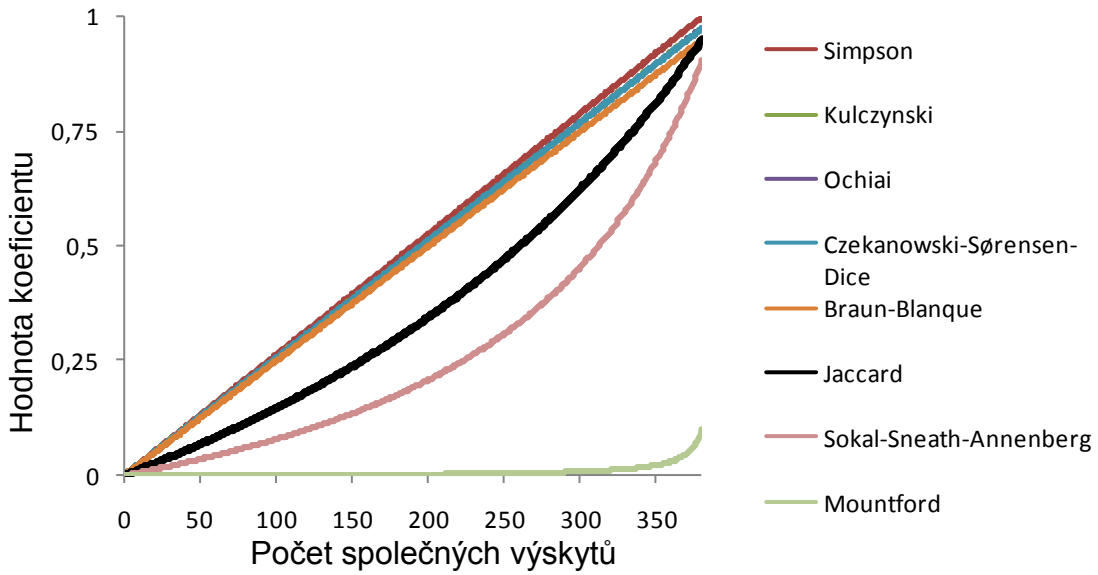
3.3 Volba koeficientu vzdálenosti

Sestavení slovníku je nejdůležitější částí analýzy, tato fáze nejvíce ovlivňuje povahu výsledků. Je však třeba ještě rozhodnout o dalších parametrech, které mohou částečně ovlivnit povahu výsledků analýzy. Jedním z nich je volba koeficientu vzdálenosti.

Jednotlivé koeficienty vzdálenosti a jejich logika již byly představeny výše. Nyní si na příkladu vazby mezi dvojicemi slov ukážeme možné rozdíly v užití jednotlivých koeficientů a různé možnosti interpretace jejich výsledků. Zároveň tím posoudíme jejich vhodnost pro užití analýze kolokací.

V Grafu 1 je zobrazen vývoj hodnoty jednotlivých koeficientů s rostoucím počtem společných výskytů v textovém souboru. Jedná se o konkrétní slova vybraná z textového korpusu disidentů, POLITIKA a ČLEN. Celkový výskyt slova POLITIKA v souboru byl 380, slovo ČLEN se vyskytlo 400 krát. Dodejme, že skutečný počet spoluvýskytů v kontextové jednotce 100 slov byl 8.

Jak je vidět z Grafu 1, pro slova s podobným výskytem nabývá většina koeficientů (Simpsonův, Kulczynského, Ochiaiův, Czekanowski-Sørensen-Diceův a Braun-Blanqueův) podobných hodnot. Výpočet těchto koeficientů je v zásadě podobný: vztahují počet společných výskytů k celkovému počtu výskytů jednoho ze dvou slov, jejich aritmetického či geometrického průměru.



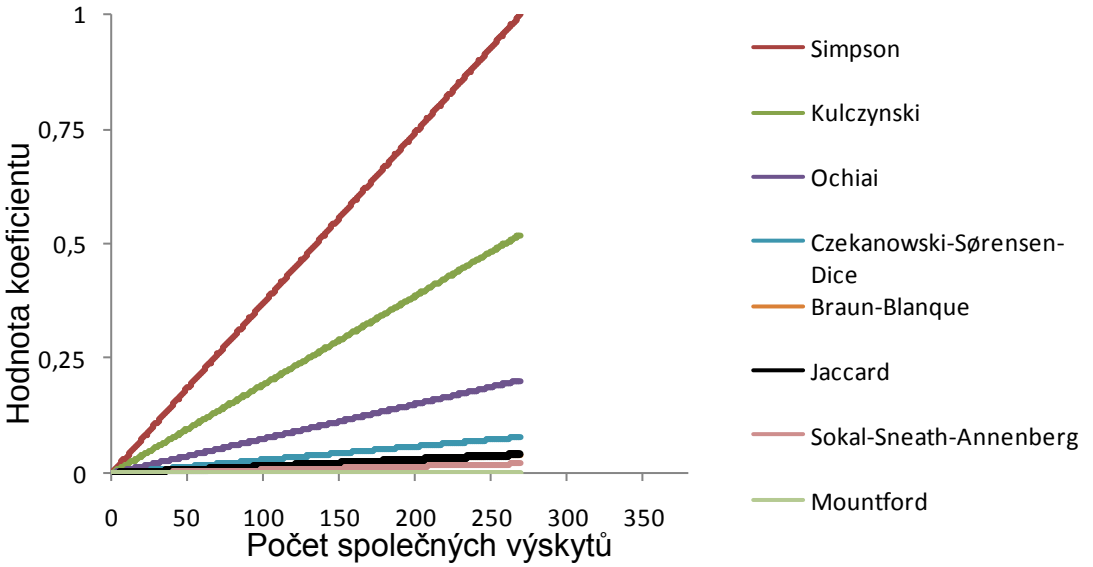
Graf 1: Hodnoty koeficientů vzdáleností v závislosti na počtu spoluvýskytů slov s podobnou frekvencí výskytu v korpusu

Od této skupiny koeficientů se liší Jaccardův, Sokal-Sneath-Annenbergův a Montfordův koeficient. Vyjádříme-li tyto koeficienty jako funkce, pak v případě těchto tří koeficientů se nejedná o funkce lineární, nýbrž funkce lineární lomené. Grafem těchto funkcí tedy není přímka, nýbrž hyperbola (resp. její část, vzhledem k definičnímu oboru). Znamená to, že tyto koeficienty pomaleji rostou u malých výskytů a rychleji naopak v případě výskytů větších.

Již nyní vidíme, že Mountfordův koeficient je jednoznačně nepoužitelný, neboť roste velmi pomalu a vyšších hodnot nabývá u velmi vysokého počtu spoluvýskytů. Zároveň nedokáže zohlednit rozdílný počet výskytů dvou slov.

Zvolíme dále dvě slova s velmi odlišnou frekvencí a porovnáme koeficienty také v jejich případě. Jedná se o slova ČLOVĚK (6529 výskytů)a CÍRKEV (269 výskytů). Výsledky jednotlivých koeficientů pro různé počty spoluvýskytů ukazuje Graf 2.

Z Grafu 2 vidíme, že s výjimkou Simpsonova koeficientu, který vždy nabývá hodnot od 0 do 1, všechny ostatní koeficienty znemožňují přiblížení slov s různou frekvencí. Znamená to, že při užití těchto ostatních koeficientů budou mít slova s nejvyšší frekvencí tendenci být umístěna ve středu grafu, neboť od ostatních méně frekventovaných slov budou vždy relativně vzdálena. Slova s malou frekvencí budou vytlačována na okraj a budou mít silnější vazbu s některými z dalších méně frekventovaných slov.



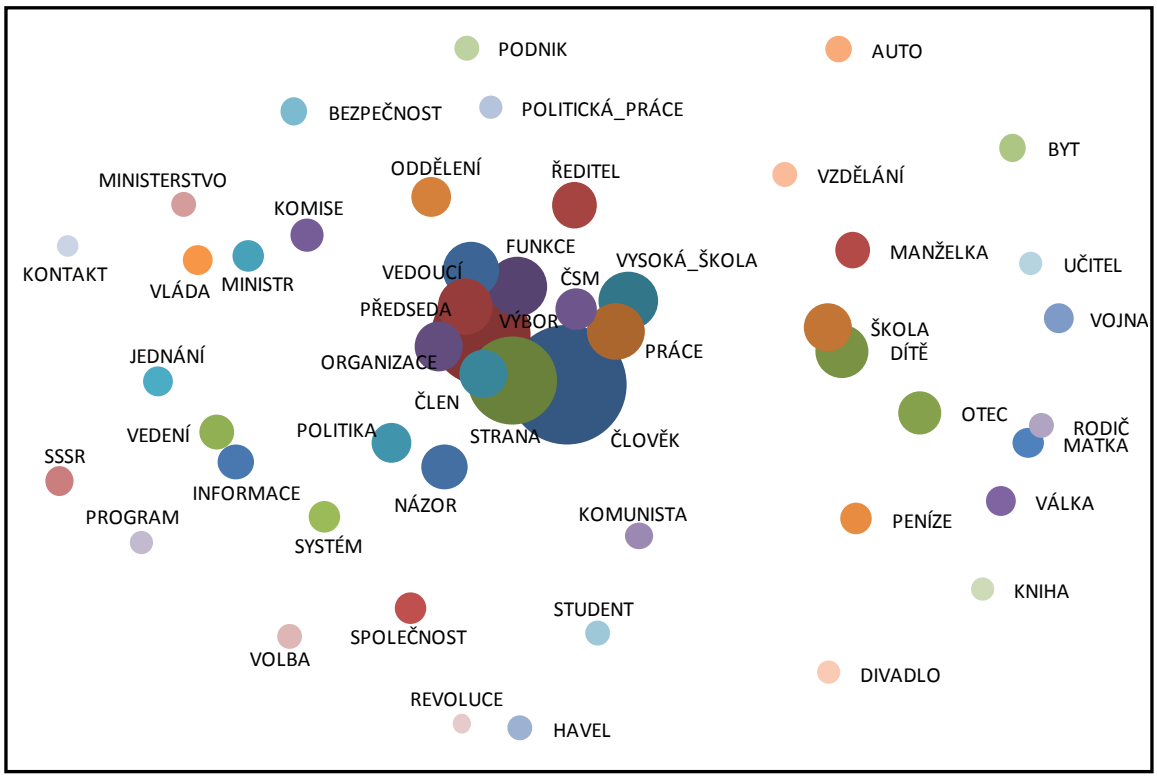
Graf 2: Hodnoty koeficientů vzdáleností v závislosti na počtu spoluvýskytů slov s různou frekvencí výskytu v korpusu

Podívejme se nyní na výslednou vizualizaci vztahů mezi slovy. Jedná se o vizualizaci kolokací pro korpus funkcionářů. Souřadnice bodů jsou výstupem z procedury PROXSCAL[54] v programu SPSS. Jako míra vzdálenosti zde byl užit Jaccardův koeficient. Velikost bodu pak vyjadřuje frekvenci daného bodu: ČLOVĚK má nejvyšší frekvenci (3574) a REVOLUCE nejnižší (93).

Odhlédněme nyní od interpretace daného rozložení slov a sledujme organizaci slov v závislosti na frekvenci na Obrázku 1.[55] Vidíme, že nejfrekventovanější slovo (resp. lemma) ČLOVĚK stojí ve středu grafu. Toto slovo bylo do grafu zařazeno záměrně, neboť působí jako přirozený střed grafu. Další slova v pořadí frekvence jsou VÝBOR (2475), STRANA (1945), FUNKCE (890), VYSOKÁ ŠKOLA (875), PRÁCE (822). Všechna tato slova se nacházejí v blízkosti středu grafu. Vidíme však, že již deváté slovo v pořadí – DÍTĚ – se nachází mimo střed grafu, neboť tematicky nesouvisí se slovy uprostřed a naopak slovo ČSM (reprezentující Svaz mládeže a jeho různá označení), které je až šestnácté v pořadí je v těsné blízkosti středu.

[54] V SPSS jsou implementovány dvě procedury mnohorozměrného škálování: ALSCAL a PROXSCAL. PROXSCAL je novější varianta a bývá preferována před ALSCAL. ALSCAL umožňuje pouze klasické mnohorozměrné škálování pracující s jednou symetrickou maticí vzdáleností. PROXSCAL je v tomto směru operativnější a umožňuje použít i pokročilých modifikací mnohorozměrného škálování. Zároveň nabízí více možností identifikace počáteční konfigurace, zejm. metodu více náhodných startů užívanou i v prezentované analýze.

[55] Frekvence je na Obrázku 1 a také na Obrázku 2 reprezentována velikostí bodu.



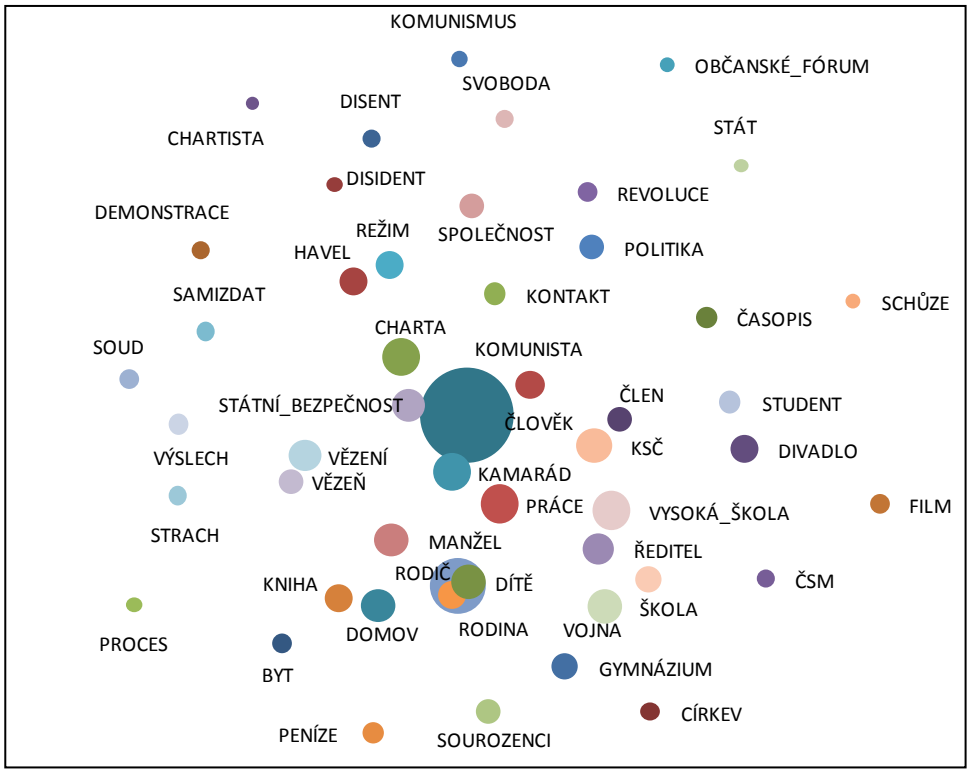
Obrázek 1: Rozložení bodů v grafu MDS podle frekvence (funkcionáři)

Je tedy vidět, že frekvence slova působí jako důležitý organizační princip[56] rozložení slov v grafu. Tento organizační princip bude silnější u pomaleji rostoucích koeficientů v Grafu 2 (zejm. Jaccardův a Sokal-Sneath-Annenbergův koeficient).

Výsledný graf bude zorganizován podle poněkud odlišného organizačního principu při užití Simpsonova koeficientu. I v tomto případě však budou mít více frekventovaná slova tendenci být umístěna ve středu grafu. Nebude to způsobeno jejich vysokou frekvencí, ale tendenci mít vazbu na více různých slov. V zásadě lze ale říci, že centrální postavení slova je dáno podobnou „přitažlivou“ či „odpudivou“ silou všech ostatních slov, zatímco slova na okraji jsou umístěna slova, která jsou „přitahována“ či „odpuzována“ selektivně pouze některými slovy.

Vzhledem k volbě ordinální varianty mnohorozměrného škálování však nehraje volba koeficientu tak zásadní roli. Jednotlivé koeficienty dávají velmi podobné výsledky, což bude ještě dále rozebráno níže.

[56] Vztah mezi frekvencí slova a vzdáleností od středu grafu můžeme vyjádřit Spearmanovým korelačním koeficientem, který je pro tento graf roven -0,87. To znamená, že pořadí frekvence je téměř inverzní pořadí vzdálenosti od středu. Podobných hodnot dosahují i ostatní koeficienty (Ochiaiův, Kulczynského a Simpsonův) a totéž platí i pro korpus disidentů. Je tedy vidět, že frekvence slova velmi silně ovlivňuje jeho pozici v grafu.



Obrázek 2: Rozložení bodů v grafu MDS podle frekvence (disidenti)

Frekvence slova jako hlavní organizační princip vizualizace mnohorozměrných dat zde platí jen do jisté míry. Z vizualizace korpusu disidentů (Obrázek 2) je vidět, že RODIČ jako druhé nejfrekventovanější slovo se nenachází blízko středu grafu, ale naopak stojí mimo. U funkcionářů (Obrázek 1) je střed velmi silně obsazen nejfrekventovanějšími slovy, ale objevují se zde i některá méně frekventovaná, např. ČLEN nebo ČSM.

Vztah mezi frekvencí slova a jeho vzdáleností od středu tak sice existuje, ale není definiční. Tento vztah lze vyjádřit Spearmanovým korelačním koeficientem,[57] který je pro graf disidentů roven -0,87 a pro graf funkcionářů -0,88. Ukazuje se, že Jaccardův koeficient, který znemožňuje slovům s rozdílnou frekvencí vytvořit si k sobě vzájemně silnou vazbu, působí jako hlavní strukturující princip konfigurace. I přesto ale některá slova tato omezení dokážou prolomit. Je to vidět na slovu RODIČ v grafu disidentů, kde se toto velmi frekventované slovo vzdálilo od středu díky specifickým vazbám na jiná slova týkající se rodiny. Podobně v grafu funkcionářů slovo ČSM (tedy Svaz mládeže) se i přes nízkou frekvenci dostalo do jádra.

[57] Vzhledem k tomu, že je využívána nemetrická (tj. ordinální) varianta mnohorozměrného škálování, je třeba vztah mezi frekvencí a vzdáleností také pojímat ordinálně. Proto je zde jako měřítko zvolen Spearmanův korelační koeficient.

3.4 Volba kontextové jednotky

Druhým parametrem, který je třeba stanovit, je velikost kontextové jednotky, tj. délka úseku textu, v níž jsou počítány společné výskyty dvojic slov. Program COOA postupuje tak, že text rozdělí na kontextové jednotky dané délky a zkoumá, v kolika těchto jednotkách se každé slovo (lemma) definované ve slovníku vyskytlo a kolikrát se vyskytlo společně s jiným slovem (lemmatem).

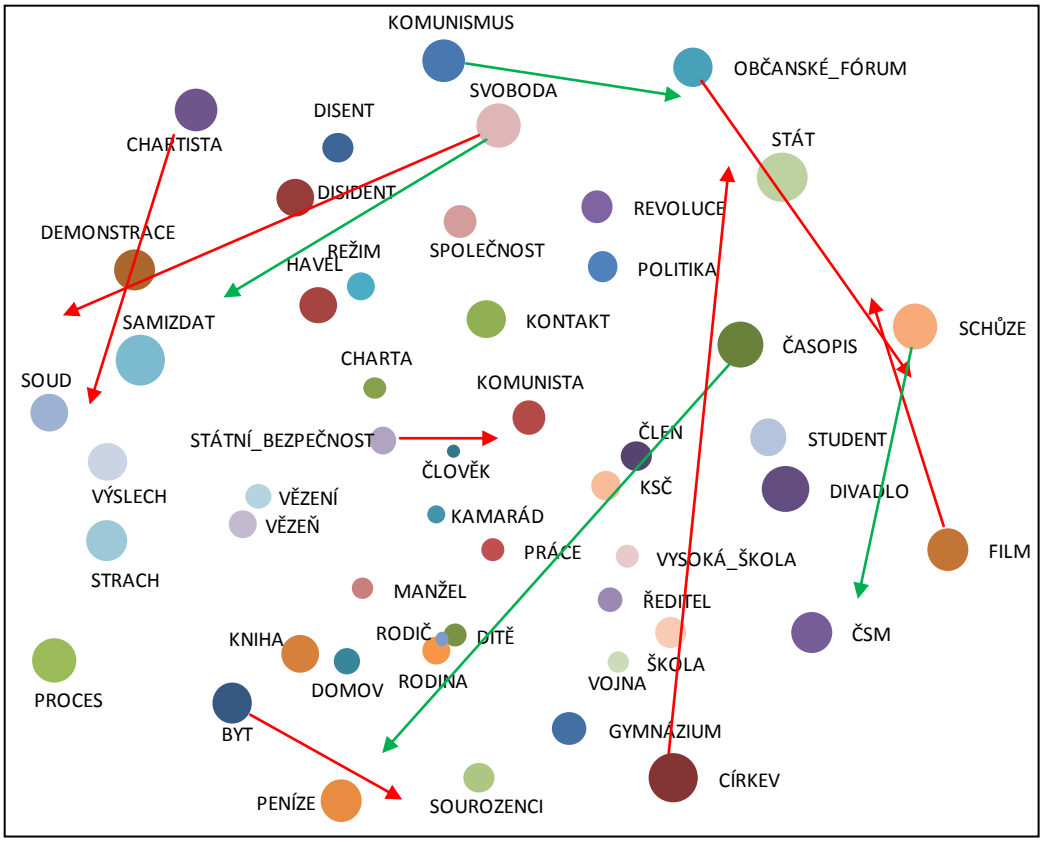
COOA nabízí různé možnosti definice kontextové jednotky. Ta může být vymezena počtem znaků, slov, vět či odstavců případně jiným definovaným dělícím znakem. Definice kontextové jednotky by měla vycházet ze zaměření výzkumu a výzkumné otázky. Obecně platí, že pro zkoumání lingvistických aspektů textů (sémantiky, syntaxe) je třeba operovat s malými kontextovými jednotkami v řádu jednotek slov.

Zajímá-li nás – jako v tomto případě – tematické uspořádání textu, volíme delší kontextové jednotky, v řádu desítek až stovek slov. Zdůvodnitelnou volbou je společný výskyt dvou slov v jednom odstavci. V případě dodržení jazykově stylistického úzu tvoří věty v jednom odstavci logický celek. Použití odstavce jako kontextové jednotky však může mít i nevýhody, neboť může záviset na autorovi textu, zda má tendenci volit spíše krátké či dlouhé odstavce. Textový korpus tvořený více autory tak může vnést do dat drobná zkreslení.

Z podobných důvodů je využití odstavce jako kontextové jednotky problematické i v případě zde prezentovaných dat. Psaný text a transkript mluvené řeči mají odlišné řazení. Výpovědi narátorů nelze přirozeně členit do odstavců. Odstavec zde tak tvoří souvislou odpověď na tazatelovu otázku, která je v určitých částech tvořena souvislým tokem textu, v jiných částech jsou odpovědi kratší. Volíme zde proto kontextovou jednotku definovanou počtem slov.

Ukažme si nyní na korpusu disidentů, jak působí změna kontextové jednotky na změny v grafu mnohorozměrného škálování. Obrázek 3 ukazuje rozložení slov při užití Jaccardova koeficientu a kontextové jednotky 100 slov.[58] Červené šipky pak ukazují významné (strukturní)[59] posuny pozice slova v uspořádání při změně kontextové jednotky na 50 slov. Zelená šipka označuje totéž pro změnu kontextové jednotky na 150 slov.

[58] Velikost bodu na Obrázku 3 a také na Obrázku 4 nyní reprezentuje velikost stresu pro daný bod, tj. míru nepřesnosti jeho zobrazení.
[59] Jedná se o takovou změnu pozice, která posouvá slovo do okolí jiných slov, než tomu bylo původně.

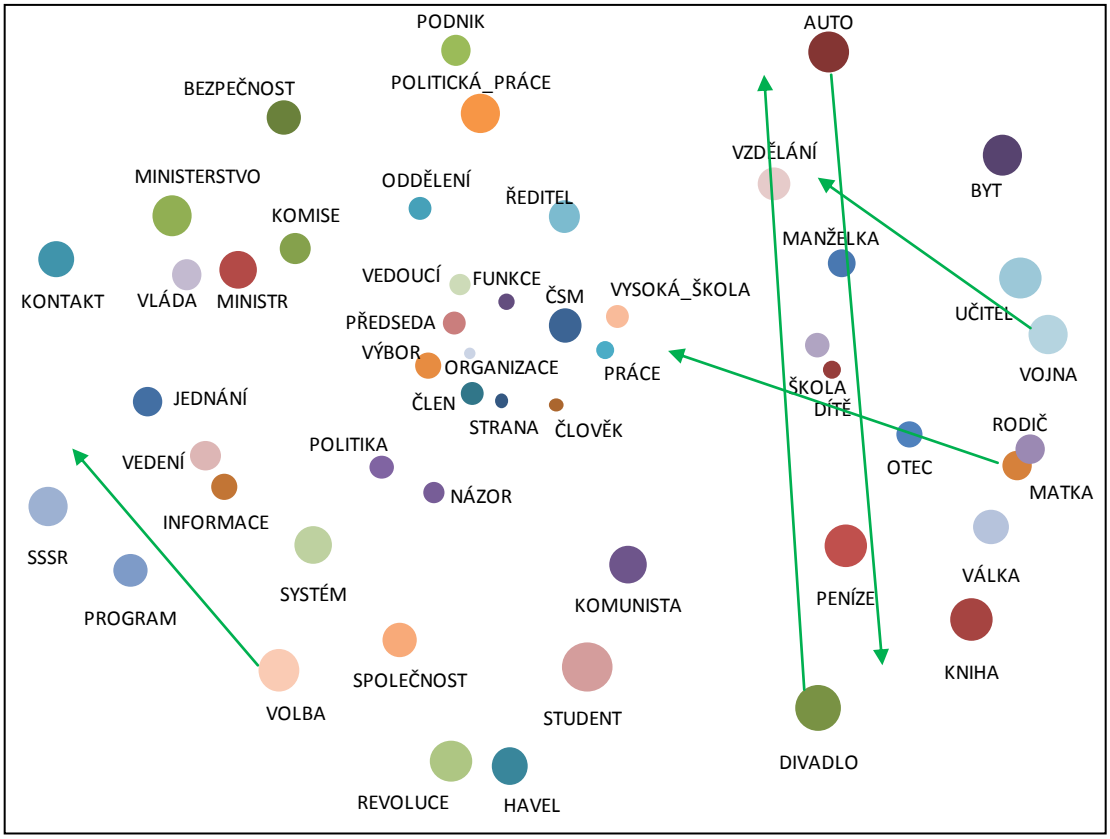


Obrázek 3: Zobrazení změn pozic bodů se změnou kontextové jednotky (disidenti)

Obrázek slouží jako ilustrace posunů, které se dějí při změnách kontextových jednotek. Zásadním zjištěním je, že základní grafu zůstává v podstatě stabilní. Je vidět, že obecně vyšší tendenci k pohybu mají méně frekventovaná slova na okraji (tj. slova s vyšším stresem). Tato slova vytvářejí náhodné vazby, které se zvýrazňují při menších kontextových jednotkách.

Jsou však slova, která zaznamenávají velmi výrazné posuny. Jedná se zejména o slovo CÍRKEV, které při zmenšení kontextové jednotky na 50 slov radikálně změnilo svůj kontext z oblasti vyprávění o dětství (okolí slov GYMNÁZIUM, VOJNA, ŠKOLA) do oblasti vyprávění o porevoluční politice (okolí slov STÁT a REVOLUCE). To je způsobeno jednak poměrně malou frekvencí slova a jednak jeho zastoupením ve dvou těchto kontextech. Při rozšíření kontextové jednotky pak může mít jiný z těchto kontextů tendenci převážit a ovlivnit pozici slova. Specifikem slova CÍRKEV v korpusu disidentů je jeho nerovnoměrné rozložení v textu, což indikuje, že bylo užíváno pouze určitou skupinou narátorů (spojených s katolickým disentem).

Podobně se chová i slovo ČASOPIS, které se při rozšíření kontextové jednotky přesouvá z kontextu politického (okolí slov SCHŮZE, POLITIKA) do kontextu rodinného (okolí slov RODINA, DOMOV).



Obrázek 4: Zobrazení změn pozic bodů se změnou kontextové jednotky ze 100 na 150 slov (funkcionáři)

Poněkud jinak se při změnách kontextových jednotek chová korpus funkcionářů (Obrázek 4). Při zmenšení kontextové jednotky na 50 slov se struktura grafu zcela mění.[60] Problém je v příliš malé frekvenci společných výskytů, která zdůrazňuje náhodné vazby mezi slovy.

Při rozšíření kontextové jednotky ze 100 na 150 slov již struktura grafu zůstává poměrně stabilní. Významné posuny nacházíme pouze na pravé straně grafu, která se týká vyprávění o soukromém a rodinném životě. Důležité je, že všechny významné posuny probíhají v rámci této oblasti vyprávění.

Při určování kontextové jednotky je tedy třeba zkoumat zejména pohyby jednotlivých slov v grafu a celkovou stabilitu struktury. Je vidět, že zatímco u vizualizace korpusu disidentů je struktura vyprávění poměrně stabilní i pro malé kontextové jednotky, v grafu funkcionářů se struktura na úrovni kontextové jednotky 50 slov smazává.

[60] Z tohoto důvodu chybí v grafu červené šipky, protože by graf ztratil svou přehlednost. Grafický výstup pro kontextovou jednotku 50 slov je součástí přílohy (viz Příloha 3).

Problémem malých kontextových jednotek je větší zdůraznění náhodných vazeb. V malé kontextové jednotce je v globálu identifikováno méně společných výskytů a jeden náhodný společný výskyt může významně změnit pořadí grafu. Problémem příliš velkých kontextových jednotek pak je zdůraznění vazeb relativně vzdálených slov. V takových případech se i slova, která spolu souvisí jen málo, mohou dostat do vzájemné blízkosti pouze díky svému vyššímu výskytu. Struktura se tak rovněž smazává.

Je tedy třeba hledat optimum, podobně jako probíhá ostření v mikroskopu, zkoumat pohyby slov a sledovat celkovou strukturu a její stabilitu.

3.5 Hodnocení kvality vizualizace kolokací

Kvalita vizualizace je v mnohorozměrném škálování poměřována s pomocí míry stresu, tj. rozdílu mezi mírou nepodobnosti v původním mnohorozměrném prostoru a vzdáleností ve vizualizovaném prostoru dvojrozměrném.

Tabulka 4 a Tabulka 5 ukazují tyto hodnoty pro různé kombinace parametrů koeficientu vzdálenosti a velikosti kontextové jednotky.[61]

		Kontextová jednotka		
		50 slov	100 slov	150 slov
Koeficient vzdálenosti	Jaccard	0,30990	0,29876	0,29337
	Ochiai	0,31929	0,30890	0,30307
	Kulczynski	0,32426	0,31518	0,33013

Tabulka 4: Kruskallův stres-1 pro jednotlivé kombinace parametrů (disidenti)

		Kontextová jednotka		
		50 slov	100 slov	150 slov
Koeficient vzdálenosti	Jaccard	0,29889	0,27998	0,27421
	Ochiai	0,31192	0,28797	0,28818
	Kulczynski	0,31187	0,28810	0,30234

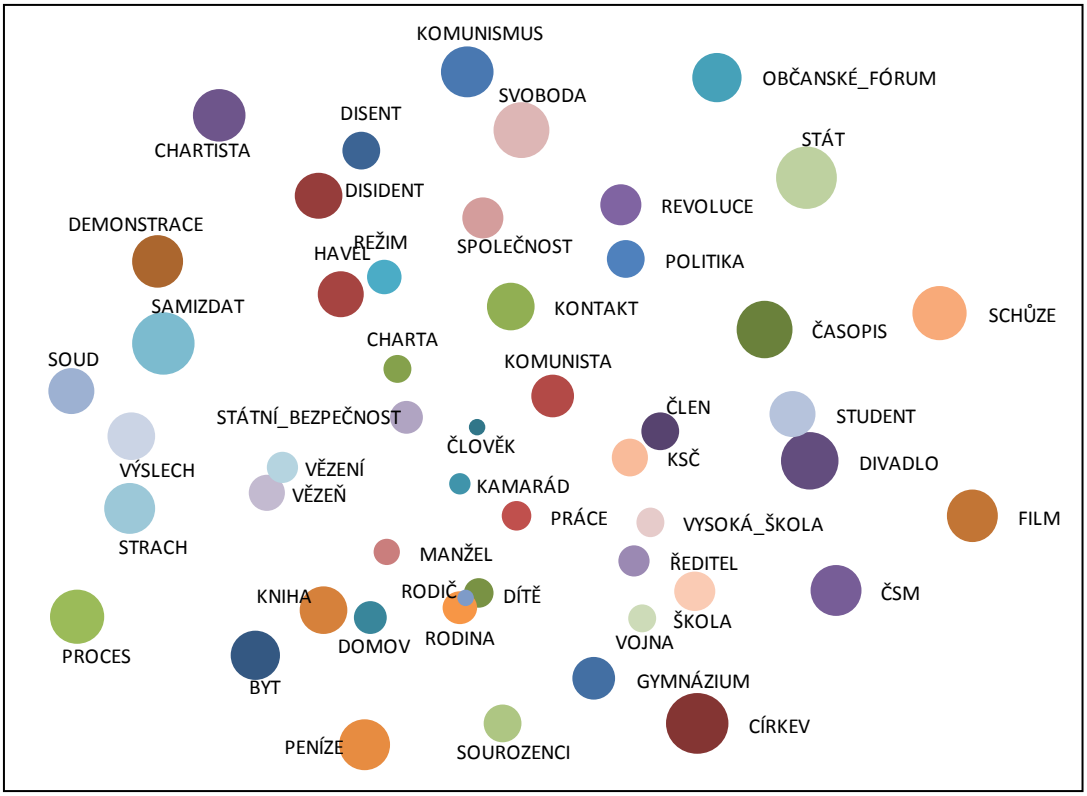
Tabulka 5: Kruskallův stres-1 pro jednotlivé kombinace parametrů (funkcionáři)

[61] Jedná se o průměrný stres vypočtený z pěti různých modelů vytvořených pro každou kombinaci koeficientu vzdálenosti a velikosti kontextové jednotky.

Hodnoty ukazují, že při užití Jaccardova koeficientu má model při stejné kontextové jednotce nižší stres. Zároveň stres klesá s rostoucí kontextovou jednotkou. To je způsobeno nižším výskytem nulových společných výskytů v matici, které působí problémy při zařazení slov v grafu.

V absolutních číslech je Kruskallův stres poměrně vysoký, ovšem dle doporučení Borga a Groenena (2005) je při vyšším počtu proměnných a malém počtu dimenzí vysoký stres logický. Je třeba se zaměřit na stabilitu zobrazení a rozložení stresu v grafu vizualizace.

Obrázek 5 a Obrázek 6 ukazují příspěvek každého bodu grafu k celkovému stresu, tzn. čím větší kruh, tím vyšší stres daného bodu.[62] Je vidět, že logika rozložení stresu je inverzní k rozložení frekvence slov. Zatímco slova ve středu grafu mají většinou nízký stres, tzn. jejich zobrazení je poměrně spolehlivé, méně frekventovaná slova na okraji grafu mají stres relativně vysoký. Z vizualizace modelu pro korpus funkcionářů je však vidět, že ve středu se nacházejí i tematicky vázaná slova s vyšším stresem (např. ČSM).

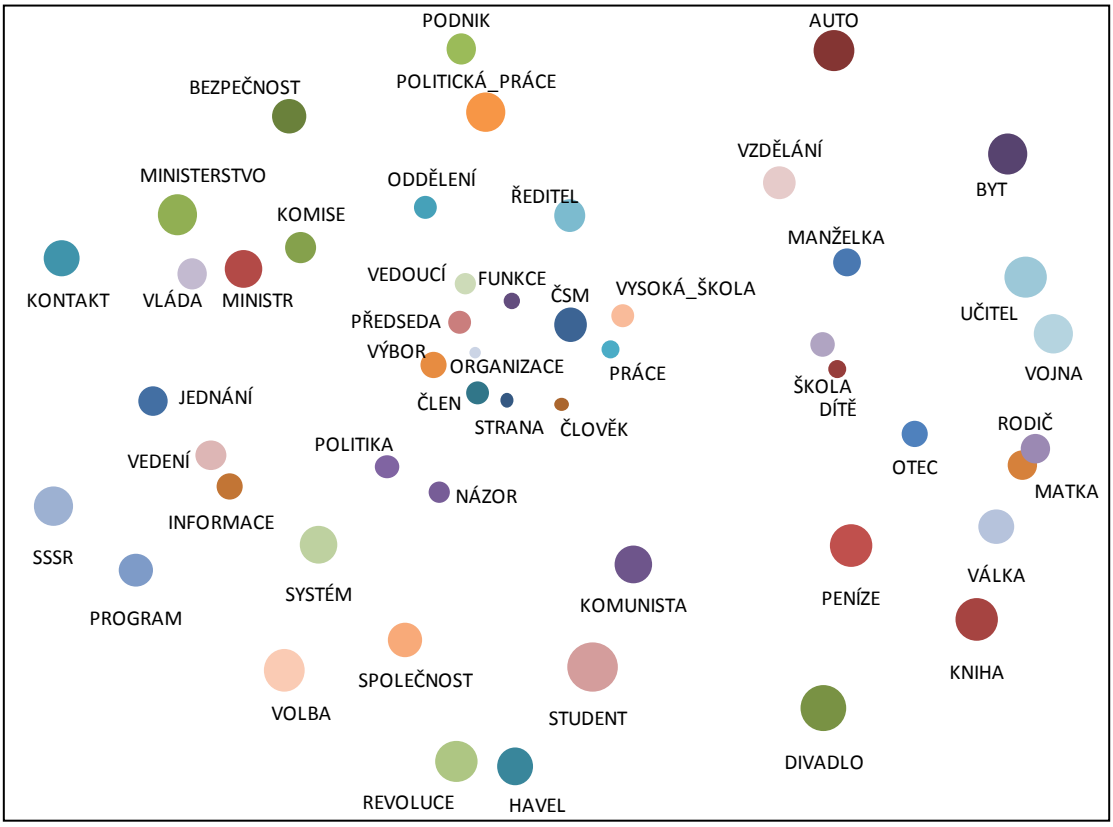


Obrázek 5: Rozložení stresu v konfiguraci MDS (disidenti)

[62] Základem procedury PROXSCAL v SPSS je minimalizace normalizovaného hrubého stresu (tj. druhé mocniny Kruskallova stresu-1), která je průměrnou hodnotou stresů pro jednotlivé body. V modelu pro korpus disidentů má nejvyšší normalizovaný hrubý stres slovo CÍRKEV (0,177) a nejnižší slovo ČLOVĚK (0,0129).

Analýza dále ukazuje, že stres je závislý na frekvenci. Spearmanův koeficient porovnávající pořadí frekvence bodů s pořadím velikosti stresu má hodnotu -0,83 pro graf disidentů i pro graf funkcionářů.

Můžeme vidět tři hlavní příčiny tohoto rozložení stresu. Asi nejdůležitější příčinou je zmíněná existence nulových vazeb mezi méně frekventovanými slovy. Druhou příčinou je společná vazba dvou vzájemně poměrně vzdálených slov na společné třetí slovo. To lze identifikovat například u trojice slov RODIČ, DÍTĚ, RODINA v grafu disidentů. Vzájemná vazba slov DÍTĚ a RODINA je slabší než vazba obou slov na slovo RODIČ. Tato nerovnoměrnost pramení hlavně z toho, že slovo RODIČ má velmi vysokou frekvenci výskytu.[63]



Obrázek 6: Rozložení stresu v konfiguraci MDS (funkcionáři)

Třetí příčinou může být vazba jednoho slova na více vzájemně si vzdálených slov. To nastává v případě, kdy se dané slovo vyskytuje ve dvou oddělených kontextech. Program pak

[63] Tento jev lze ilustrovat následujícím příkladem. Slovo ČLOVĚK v učebnici vývoje lidského druhu má tendenci vytvářet silné vazby na přívlastky ZRUČNÝ, VZPŘÍMENÝ nebo ROZUMNÝ. Jelikož však učebnice velmi pravděpodobně bude uspořádána do kapitol, kde každá kapitola se bude věnovat podrobně jednomu druhu, vyskytne se velmi málo kolokací jednotlivých přívlastků. Tím se zvýší jejich stres. Hlavní příčinou je nerovnoměrnost výskytu jednotlivých slov: zatímco ČLOVĚK se vyskytuje velmi často, jednotlivé přívlastky mají mnohem nižší frekvenci.

umístí slovo v grafu víceméně náhodně k jednomu i druhému kontextu. Tento typ slova lze identifikovat opakováním procedury mnohorozměrného škálování a sledování pohybů slova v rámci grafu.

Při zkoumání míry stresu jsme se dále na doporučení Borga a Groenena (2005 : 54) zabývali proměnami konfigurace bodů s opakováním analýzy a změnami jejích základních parametrů: s užitím různé míry podobnosti (Jaccardův, Ochiaiův a Kulczynského koeficient)[64] a s užitím různé kontextové jednotky pro výpočet kolokací (50, 100 a 150 slov).[65] Provedená analýza je pouze explorativní, výsledky nebyly statisticky testovány.

Rozbor proměn konfigurací bodů tedy byl proveden pro 9 skupin datových matic. Na každou z devíti datových matic byla aplikována analýza mnohorozměrného škálování (PROXSCAL v programu SPSS) celkem pětkrát. Počátek analýzy byl stanoven metodou náhodných počátků v počtu 1000.[66] Výsledkem tak bylo 5 různých „dobrých" konfigurací pro každou skupinu.

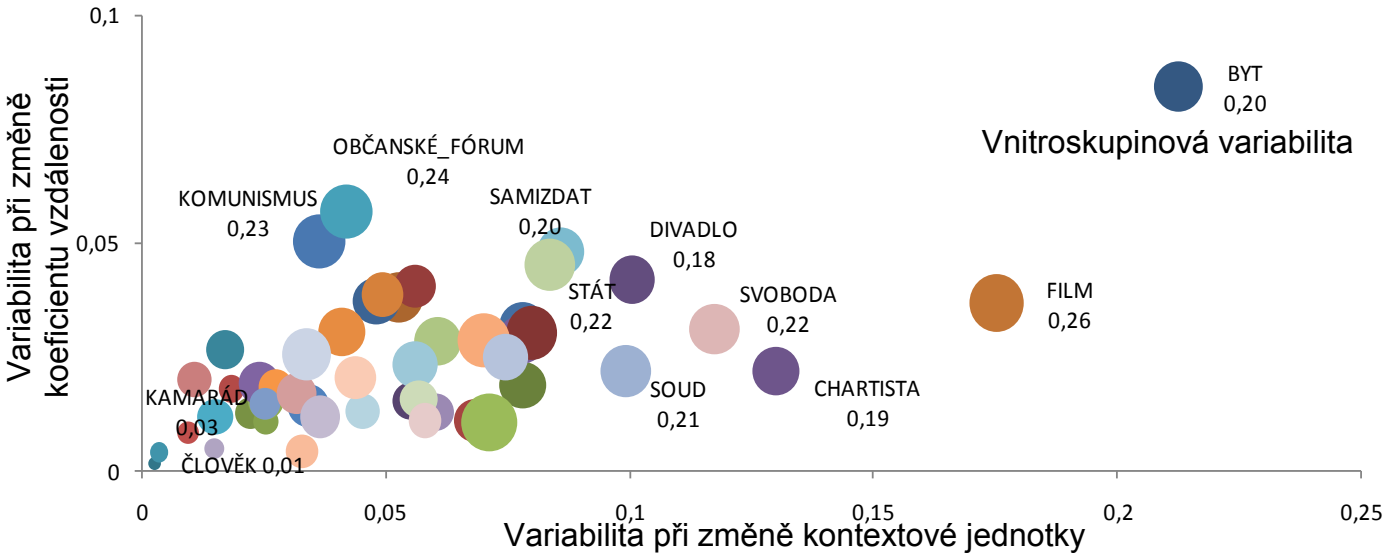
Procedura mnohorozměrného škálování postupuje tak, že nejprve položí dva body s nejvyšší vzdáleností, zvolí pozici třetího bodu na základě vzdálenosti od dvou předchozí a následně přidává postupně na základě vzdálenosti od ostatních již zanesených bodů. Vzhledem k tomuto faktu jsou všechna řešení symetrická podle vodorovné a/nebo svislé osy. Nebyla tedy prováděna náročnější prokrustovská analýza, která je schopna stanovit takové postavení dvou grafů, kdy čtverce vzdáleností mezi body jsou minimální.[67]

Při analýze posunů konfigurací byly zkoumány průměrné vzdálenosti od průměrné konfigurace daného bodu, tzv. centroidu. Ten byl stanoven jako průměr jednotlivých hodnot každé ze dvou dimenzí daného bodu. Průměrná vzdálenost od centroidu byla zvolena jako míra tendence bodu měnit svou pozici v grafu.

Graf 3 zobrazuje výsledky této analýzy pro konfiguraci disidentů. Každý z bodů má tři vlastnosti. Vodorovná osa zobrazuje míru variability bodu způsobenou změnou velikosti

[64] Předběžná analýza ukázala, že tyto tři koeficienty podobnosti produkují velmi podobné konfigurace. Významně se lišil zejména koeficient Simpsonův, který jsme proto do analýzy již nezahrnovali, neboť vyžaduje odlišnou interpretaci vztahů mezi body a není pro analýzu, jak je využívána v této práci, příliš vhodný.
[65] Preferovaná varianta 100 slov vychází z předchozí zkušenosti a z předpokladu, že 100 slov přibližně odpovídá délce smysluplné výpovědi. Ostatní délky kontextové jednotky zde byly zkoumány pro zaznamenání změn.
[66] Tento typ analýzy stanoví náhodnou úvodní konfiguraci bodů, kterou dalšími iteracemi zpřesňuje, dokud nejsou naplněna kritéria. Tento postup program opakuje 1000krát a vybírá tu konfiguraci, která má nejnižší míru normalizovaného hrubého stresu (viz výše).
[67] K prokrustovské analýze více viz Borg a Groenen (2005 : 429 an.)

kontextové jednotky (50, 100, 150 slov).[68] Svislá osa pak zobrazuje míru variability bodu způsobenou změnou míry podobnosti (Jaccard, Ochiai, Kulczynski).[69] Plocha bodu pak určuje míru nevysvětlené variability daného bodu.[70]



Graf 3: Variabilita zobrazení bodů při změně kontextové jednotky, míry podobnosti a nevysvětlená variabilita (disidenti)

Z grafu je na první pohled vidět, že vliv volby jedné ze tří měr podobnosti na výslednou konfiguraci bodů není příliš velký. V grafu, kde se souřadnice bodů na obou osách pohybují přibližně od -1 do +1, je posun způsobený volbou jiného koeficientu roven maximálně 0,1. Naopak volba jiné kontextové jednotky již způsobuje výraznější posuny některých slov. Jedná se o slova s nízkou frekvencí nacházející se na okraji grafu. Důvody pro tyto posuny již byly naznačeny: při nízké frekvenci mají vazby na slova ve větší míře náhodný charakter a nejsou pevně zakotveny v určitých kontextech. Při vizualizaci má pak tendenci převážit jedna z těchto vazeb, která umístí slovo do jednoho z jeho kontextů.

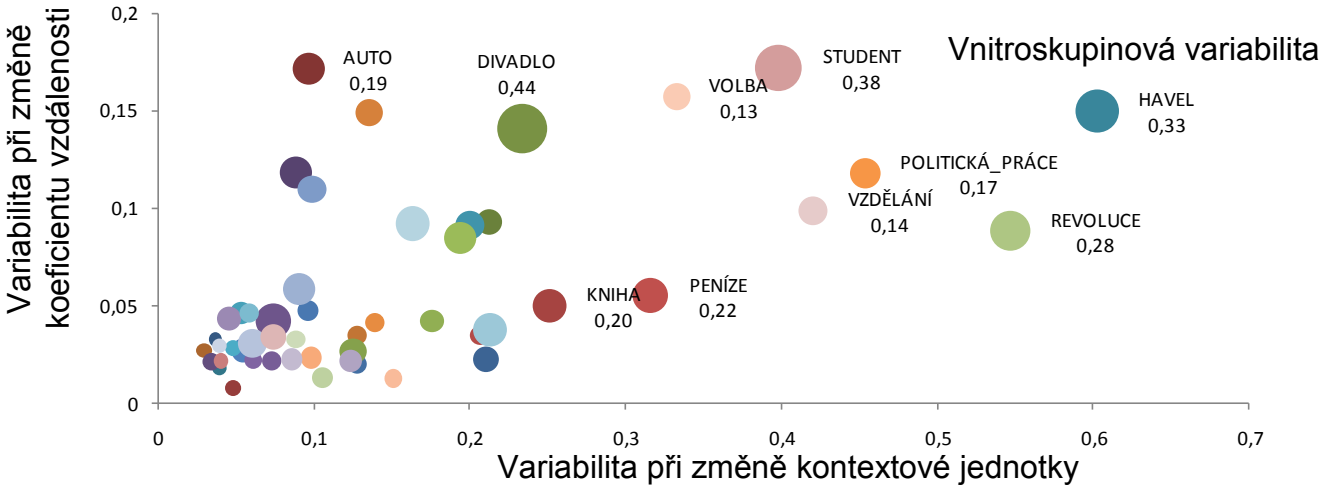
Graf 4 ukazuje analogickou situaci pro korpus disidentů. Vliv změn kontextové jednotky a volby koeficientu vzdálenosti je vyšší stejně jako nevysvětlená variabilita některých slov. I přesto, že model pro korpus funkcionářů má mírně nižší Kruskallův stres, tendence bodů

[68] Jedná se o průměrnou vzdálenost průměrné pozice (centroidu) každé ze tří velikostí kontextové jednotky (50, 100, 150 slov) od průměrné pozice bodů spočtené ze všech 45 měření. Míra určuje, zda daný bod má soustavnou tendenci měnit svou pozici s volbou jiné velikosti kontextové jednotky.

[69] Tato míra je vypočtena analogicky předchozí. Pro každou ze tří užitých měr podobnosti je vypočtena průměrná konfigurace a je zjišťována průměrná vzdálenost od středu těchto tří centroidů.

[70] Ta je vypočtena jako průměrná vzdálenost bodu v každé ze 45 konfigurací od centroidu své skupiny. Pro skupinu pěti měření pro matici Jaccardových vzdáleností s kontextovou jednotkou 50 slov byl vypočten centroid a průměrná vzdálenost každého měření od tohoto centroidu. Totéž bylo provedeno pro 8 skupin. Dále byl vypočten průměr z těchto průměrů, který vyjadřuje míru nevysvětlené variability.

putovat grafem v závislosti je vyšší. Souvisí to s tvarem vizualizace: funkcionáři mají jasně oddělený střed grafu, což proceduře mnohorozměrného škálování usnadňuje lokaci těchto bodů a zvyšuje její přesnost. Zároveň to souvisí s menší velikostí korpusu funkcionářů, která významně zkreslila graf pro kontextovou jednotku 50 slov.



Graf 4: Variabilita zobrazení bodů při změně kontextové jednotky, míry podobnosti a nevysvětlená variabilita (funkcionáři)

3.6 Shrnutí

Třetí kapitola ukázala logiku, podle níž jsou matice vzdáleností - v konkrétním případě dvou textových korpusů analyzovaných v této práci - převedeny do vizuální podoby. Zároveň bylo podrobně analyzováno, zda je tato vizualizace reliabilní a jaký vliv má výběr parametrů (kontextová jednotka, koeficient vzdálenosti) na výslednou organizaci bodů ve vizualizaci.

Bylo ukázáno, že jedním z klíčových aspektů ovlivňujících vizualizaci matice spoluvýskytu slov v textu je nerovnoměrné rozdělení frekvence těchto slov. Vzniká otázka, jak definovat vzdálenost mezi dvěma slovy s různou frekvencí, neboť vztah mezi takovými slovy je asymetrický. I když se první, méně frekventované slovo vyskytuje výhradně v kontextu druhého, vysoce frekventovaného slova, těžko se mu může přiblížit, protože pro toto druhé slovo je ono první naopak pouze marginální součástí jeho kontextu. Tato asymetrie pak působí problémy při vizualizaci struktury uspořádání slov v textu.

Při analýze chování jednotlivých koeficientů vzdálenosti byly identifikovány tři až čtyři koeficienty potenciálně vhodné pro užití při výpočtu míry podobnosti pozice dvou slov ve struktuře textu. Jedná se o Jaccardův, Kulczynského a Ochiaiův koeficient, které – jak se

ukázalo při detailním rozboru – pro data zde prezentovaná dávají velmi podobné výsledky. Jako mezní případ je zmiňován koeficient Simpsonův.

Všechny tyto koeficienty však mají velmi podobnou logiku rozložení konfigurace bodů při vizualizaci matice. Tato logika pak determinuje i interpretaci této konfigurace. Ve všech případech se uprostřed vizualizace nacházejí slova s nejvyšší frekvencí výskytu. Tato slova zároveň mají nejmenší stres, tj. nejmenší míru nepřesnosti zobrazení. Na okraji grafu pak nacházíme méně frekventovaná slova s vyšší nepřesností zobrazení. Některá z těchto slov mají tendenci se při opakování procedury mnohorozměrného škálování posouvat. Důvodem je to, že často mají slabé vazby na slova různá slova a při vizualizaci mají tendenci se přiklonit k jedné z nich podle aktuální konstelace.

Druhým parametrem je volba velikosti kontextové jednotky. Obecně se ukazuje, že volba větší kontextové jednotky je spolehlivější, ovšem příliš široké kontextové jednotky započítávají i náhodné spoluvýskyty slov, které nemají významový vztah. V obou případech volíme kontextovou jednotku 100 slov, která se oproti volbě kontextové jednotky 150 slov již ukazuje jako poměrně stabilní.

Poměrně vysoká míra stresu je v případě zde prezentovaných analýz způsobena velkým počtem proměnných v modelu a malým počtem dimenzí. Problém je však také v tom, že přístup, který je zde prezentován, se pokouší převést vzdálenosti mezi slovy v textu převést na vzdálenosti v eukleidovském dvourozměrném prostoru. Logika distribuce slov v sémantickém prostoru, i vzhledem k asymetrii jejich vzdáleností, však přesné zobrazení v eukleidovském prostoru neumožňuje.

Je důležité si uvědomit, že vizualizace struktury do podoby vizualizace slouží pouze jako interpretační nástroj. Stejně jako se čtenář mapy učí v mapách číst a chápat, že vzdálenost mezi Prahou a Brnem v mapě má určitý vztah k realitě, musí se analytik užívající vizualizaci struktury textu naučit chápat, jaký vztah má vzdálenost mezi slovem ČLOVĚK a slovem KOMUNISTA k textu, který je analyzován. K tomu poslouží následující kapitola.

4 Interpretace výstupů textové analýzy

V předchozí kapitole byl představen celý postup metody CATA na konkrétních datech: na korpusech sestavených z přepisů biografických vyprávění dvou skupin aktérů československé normalizace, *disidentů* a *komunistických funkcionářů*. Pro obě skupiny byl představen průběh sestavování slovníků, představena vizualizace kolokací slov v textu, posouzen vliv volby koeficientu vzdálenosti a velikosti kontextové jednotky na tuto vizualizaci a na základě toho posouzena míra nepřesnosti této vizualizace. Celkově tedy byla posouzena reliabilita výsledků této metody.

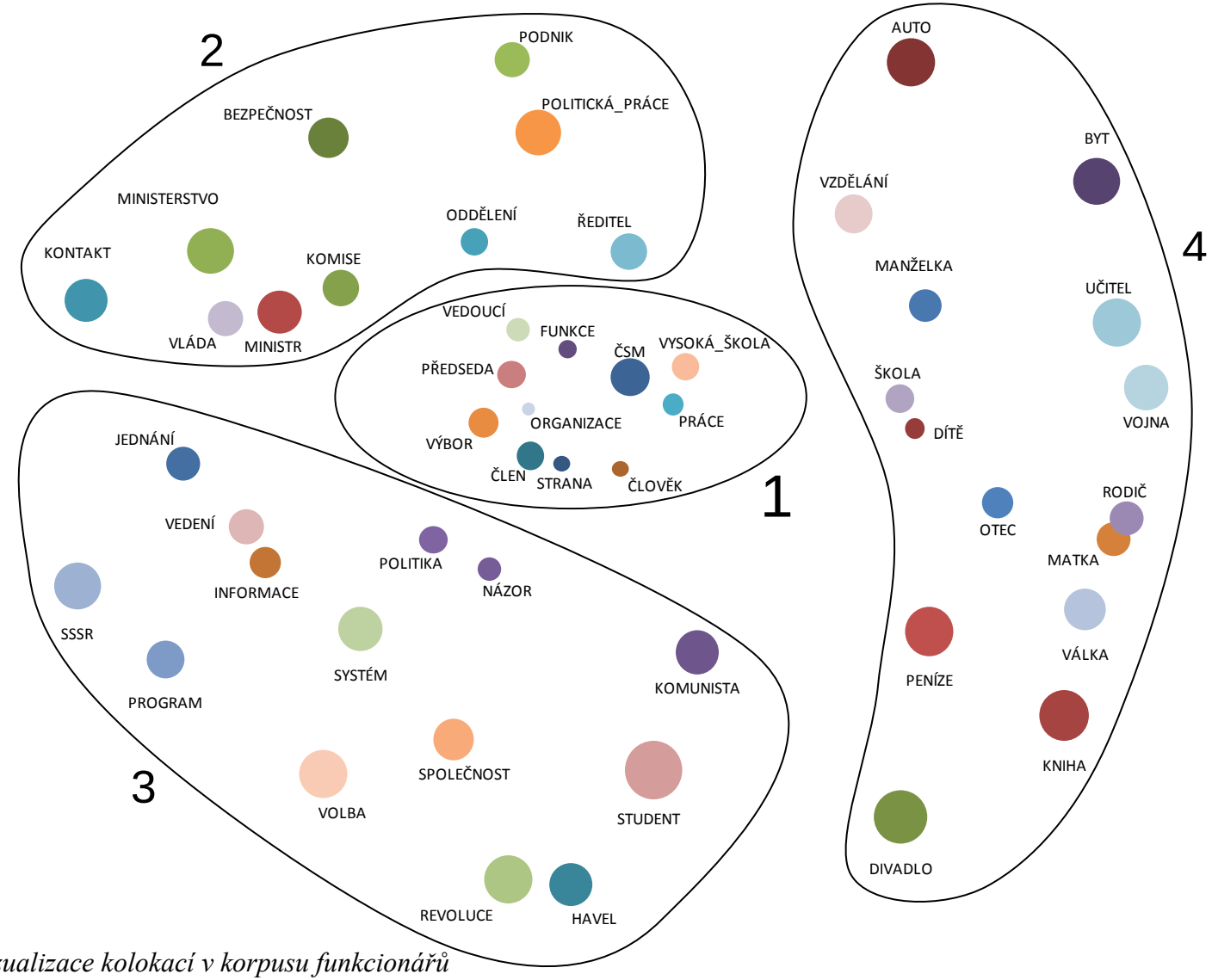
V této kapitole budou představené vizualizace dat dále interpretovány. Zaměříme se vedle toho na validitu dat, kterou se pokusíme prokázat porovnáním této původní interpretace s interpretací výsledků klasických metod analýzy biografických vyprávění s pomocí hermeneutické analýzy.

Pro analýzu byla zvolena procedura s užitím Jaccardova koeficientu a kontextovou jednotkou 100 slov. Jak ukázal rozbor v předchozí kapitole, volba koeficientu nemá na výslednou konfiguraci významný vliv. Drželi jsme se proto doporučeními z literatury (Hájek, 2010; Borg & Groenen, 2005 : 127; Mohammad & Hirst, 2005) a zvolili Jaccardův koeficient. Kontextová jednotka 100 slov byla zvolena na základě zjištění, že rozdíl mezi jednotkou 100 a 150 slov byl u obou korpusů minimální.

Nejprve se v této kapitole budeme zabývat samostatným rozborem vizualizací pro oba korpusy: nejprve pro korpus funkcionářů, poté pro korpus disidentů. Poté se zaměříme na porovnání výsledků této metody s výsledky analýzy Petry Schindler-Wisten *Rodinné prostředí příslušníků politických elit a disentu* (Schindler-Wisten in Vaněk et al., 2006). Analýza je součástí sborníku *Mocní? A bezmocní?* (Vaněk et al., 2006). Sborník, redigovaný pracovníky Centra orální historie Ústavu pro soudobé dějiny AV ČR, obsahuje orálně-historické analýzy provedené na stejných datech, která analyzujeme zde.

4.1 Rozbor vizualizace korpusu funkcionářů

Obrázek 7 ukazuje na větším prostoru vizualizaci kolokací pro korpus funkcionářů. Jedná se o stejné zobrazení jako na Obrázku 6 (kap. 3.5), pouze pro přehlednost zvětšené a s vyznačenými klíčovými oblastmi vyprávění.



Obrázek 7: Vizualizace kolokací v korpusu funkcionářů

Z vyznačení grafu vyplývá, že zde využíváme tzv. polární interpretací mnohorozměrného škálování (viz kap. 2.5.2). Sledujeme oblasti grafu vycházející od středu a zkoumáme jejich společné vlastnosti.[71] Vzhledem k tomuto způsobu interpretace je dobré se dívat i na uspořádání jádra, tj. která slova z jádra přiléhají ke které oblasti. Zčásti se zde opíráme také o interpretaci polární, která umožňuje sledovat rozložení bodů v rámci skupin (clusterů). Pro srovnání byla na stejné datové matici provedena i clusterová analýza, jejíž výstup je součástí přílohy.

Body ve vizualizaci funkcionářů se rozpadají na čtyři viditelné skupiny slov. Skupina (1) tvoří *jádro vyprávění*, jednotný rámec, který rámuje další části vyprávění,[72] skupiny (2), (3) a (4). Tyto skupiny slov jsou vzájemně relativně odděleny.

Hustě zaplněný střed grafu ukazuje na vysokou míru strukturovanosti vyprávění životních příběhů komunistických funkcionářů. U funkcionářů mají příběhy jasně daný rámec vyprávění, jasně danou strukturu utvářenou kolem angažmá těchto aktérů v komunistické straně. Vyskytují se zde slova jako STRANA, VÝBOR, ČLEN, PŘEDSEDA, FUNKCE, PRÁCE.[73]

Vysoká zaplněnost jádra vyprávění ukazuje na jeho vysokou strukturovanost. Tato vysoká strukturovanost se přizpůsobuje strukturovanosti a hierarchizaci státního a stranického aparátu. Odkazy ke struktuře stranické aparátu tvoří těžiště celého vyprávění. Clusterová analýza do jádra nezahrnuje VYSOKOU ŠKOLU[74] jako důležitý nástroj výchovy stranických funkcionářů.

Rozložení slov v jádru odpovídá přiléhajícím skupinám. Je vidět, že část jádra týkající se práce ve straně se kloní blízko skupiny slov (2), která se týká tématu *stranického života*. Tento oddíl v podstatě doplňuje onu právě popsanou část jádra. V této skupině slov můžeme najít dva póly: 1. pól *vysoké politiky* charakterizovaný slovy VLÁDA, MINISTR,

[71] Jak bylo uvedeno, tato interpretace odpovídá užití clusterové analýzy. Obsahem přílohy je i stromový graf, který je výstupem hierarchické clusterové analýzy. Ta dává pro táž data velmi podobné (i když ne zcela identické) výsledky. Při další interpretaci budeme dále porovnávat výsledky obou metod.

[72] Ve stromové struktuře funkcionářů (viz Příloha 4) je rovněž patrné brzké oddělení jádra od zbytku slov.

[73] I v clusterové analýze se jádro diskursu odděluje jako první a dále se rozpadá na 2 skupiny (A1, A2). Ve skupině A1 jsou zastoupena slova jako STRANA, VÝBOR, ČLEN, PŘEDSEDA, ORGANIZACE a týká se organizace strany, druhá část jádra A2 zastoupena slovy FUNKCE, PRÁCE, VEDOUCÍ a ODDĚLENÍ se týká respondentovy práce ve straně.

[74] Vznikla oddělením školy vysoké od ostatních škol při tvorbě slovníku. Pod koncept vysoké školy nebyly zahrnuty pouze klasické výskyty sousloví vysoká škola, případně univerzita, fakulta, výška apod., ale i různé stranické školy, jako byla Večerní univerzita marxismu-leninismu, vyšší stranické školy atp.

MINISTERSTVO, 2. pól *lokální politiky* charakterizovaný slovy POLITICKÁ PRÁCE,[75] PODNIK, ODDĚLENÍ. [76]

Odděleně od skupiny (2) stojí skupina slov (3), která se týká tématu *politiky a ideologie*. Oddělenost těchto dvou částí vyprávění je poměrně zajímavá. Ukazuje se, že funkcionáři ve svých vyprávěních oddělují svůj život ve straně a vysvětlení svého angažmá v ní – jak lze také chápat tuto skupinu slov.

Zároveň je zde patrné, že vedení strany je zde odděleno od běžného stranického života. Tuto součást vyprávění lze opět rozdělit na dva póly, kdy napravo stojí ideologické otázky komunistického režimu (VEDENÍ, JEDNÁNÍ, PROGRAM, POLITIKA, SSSR) a nalevo nacházíme vyústění v podobě sametové revoluce, ke které se zde funkcionáři vztahují (REVOLUCE, HAVEL, SPOLEČNOST, VOLBA[77]).

Toto rozdělení na dva póly je způsobeno rozdíly mezi vysokými a nižšími stranickými funkcionáři. Oddělení těchto dvou skupin koresponduje se dvěma póly stranického života. Z blízkosti krajních pólů *vyprávění o stranickém životě* a vyprávění o *ideologii a politice* můžeme usuzovat, že u nich je práce ve straně a ideologie silněji spojena, než tomu bylo u nižších funkcionářů.[78] Vysocí funkcionáři rovněž nepovažují svůj osobní a rodinný život za tolik relevantní a hovoří spíše o své práci v aparátu strany.

V centru celé oblasti stojí slovo POLITIKA, které se vztahuje jak předrevoluční, tak k revoluční a porevoluční fázi vyprávění. Důležité je, že v předrevoluční fázi se nevyskytují slova jako NÁZOR či SPOLEČNOST, což ukazuje na jistou míru odcizenosti předrevoluční politiky. Osobní ideologickou konfrontaci požadavků společnosti s vlastními názory a s názory strany prezentují funkcionáři až v revoluční oblasti vyprávění.

Je zároveň zajímavé, že až v této oblasti vyprávění se objevuje slovo KOMUNISTA. Označování sebe sama a své referenční skupin tímto slovem se v předchozích fázích téměř nevyskytuje. Zapojuje se až ve fázi, kdy se člověk musí vyrovnávat se změnou politických poměrů.

---

[75] Tento koncept a jeho vytvoření jsou vysvětleny v kapitole 3.2.2.

[76] Clusterová analýza v tomto ohledu kopíruje skupinu 2 jen částečně, ale základ této skupiny se překrývá se skupinou D, která reflektuje i rozdělení na 2 póly. Slova KONTAKT a BEZPEČNOST však řadí do široké skupiny E, která se týká rodinného života, a POLITICKOU PRÁCI dává do souvislosti se skupinou F, která se týká vzdělání.

[77] Resp. *volby*, které pod tento koncept taky zahrnujeme.

[78] Clusterová analýza ukazuje, že zatímco ideologická oblast vyprávění se z větší části kryje se skupinou C, revoluční pól je podskupinou skupiny E, tedy každodenního života. Ukazuje se tak, že revoluční události roku 1989 byly předělem, který proměnil nejen ideologické uvažování respondentů, ale i jejich každodenní a rodinný život.

Zároveň, a to zejména u nižších funkcionářů, zde slovo REVOLUCE označuje také časové období, které bylo důležitým mezníkem v životě aktérů (něco se stalo „před revolucí“, něco „po revoluci“).

Těžiště tohoto rodinného života pak představuje oblast vyprávění (4). Tato oblast vyprávění věnující se *soukromému a rodinnému životu* zaujímá velkou část prostoru grafu mnohorozměrného škálování a ukazuje, že tvoří důležitou – a do jisté míry oddělenou – součást celého vyprávění.[79] Centrum této oblasti tvoří rodina, a to ať už rodina, z níž aktér pochází (OTEC, MATKA, RODIČE), tak rodina, kterou aktér založil (MANŽELKA, DÍTĚ, ŠKOLA).

V dolní části této oblasti stojí rodinné aspirace: AUTO, BYT, VZDĚLÁNÍ. Přiléhají k lokálnímu pólu diskursu stranického života, což ukazuje k tomu, že naplnění těchto aspirací (vlastnictví auta, vzdělání) je do jisté míry spojeno se stranickou kariérou. Na druhém pólu rodinného života pak stojí kulturní soukromý život narátorů. Obě tyto krajní oblasti se však mají tendenci pohybovat v rámci rodinné oblasti podle aktuální konstelace. Týká se to slov AUTO a DIVADLO. Tyto pohyby můžeme vysvětlit vazbou na slovo REVOLUCE v jeho časově lokalizačním významu.[80]

Jen zdánlivě s těmito aspiracemi nejsou spojeny PENÍZE. Ty stojí spíše na kulturní straně této oblasti. Clusterová analýza naznačuje blízkost slova PENÍZE k AUTU a BYTU. PENÍZE však souvisejí také s dětstvím aktérů, kteří často pocházeli z chudších rodin a vyrůstali za války, kdy právě majetek hraje důležitou srovnávací roli.

Důležité je také si všimnout, že vyprávění o soukromém a rodinném životě – celá oblast vyprávění o rodině, ŠKOLA, ale i VYSOKÁ ŠKOLA - stojí zcela v opozici k ideologickým částem biografie funkcionářů. Je tedy vidět, že proces vzdělávání a výchovy ve vyprávění aktérů není spojen se získáváním ideologického názoru. Ideologie byla u komunistických funkcionářů internalizována a tvoří specifickou oblast vyprávění. Proces utváření této ideologie v životě jednotlivců zde tedy absentuje – jedná se o něco hotového a neměnného.

[79] V clusterové analýze reprezentována skupinou E představující široký soukromý život, skupinou F představující vzdělání jako součást životní dráhy a také dráhy potomků respondentů, a specifickou skupinou B zabývající se rodinou, z níž respondenti pocházejí.

[80] Před revolucí získali funkcionáři nové auto, Škodu Favorit.

4.2 Rozbor vizualizace korpusu disidentů

Obrázek 8 ukazuje vizualizaci kolokací slov pro korpus disidentů. Stejně jako v předchozím případě se jedná o zvětšenou verzi Obrázku 5 (kap. 3.5). Při pohledu na vizualizaci pro korpus disidentů nelze jednoduše identifikovat jednotlivé části vyprávění, jako tomu bylo v případě funkcionářů.[81] To může mít několik vysvětlení. Jednu příčinu můžeme hledat v rozdílné zkušenosti, která ovlivňuje celé vyprávění. Zatímco funkcionářská zkušenost i jejich životní příběh jsou si v zásadě velmi podobné, disidentská vyprávění jsou individuálně více odlišná, co se týče životní zkušenosti i volby lexika.

Pro tento klíčový rozdíl však můžeme zvolit i vyšší rovinu interpretace. Tuto různou úroveň strukturovanosti životní zkušenosti dvou protikladných skupin lze promítnout i do různé úrovně strukturovanosti žitého světa narátorů. Žitý svět komunistických funkcionářů byl silně zakotven v institucionální struktuře normalizační společnosti, disidenti stáli téměř zcela mimo tuto strukturu – setkávali se pouze s jejím represivním aparátem. Tomu obě skupiny přizpůsobují svůj jazyk, do nějž se tato institucionální struktura promítá.

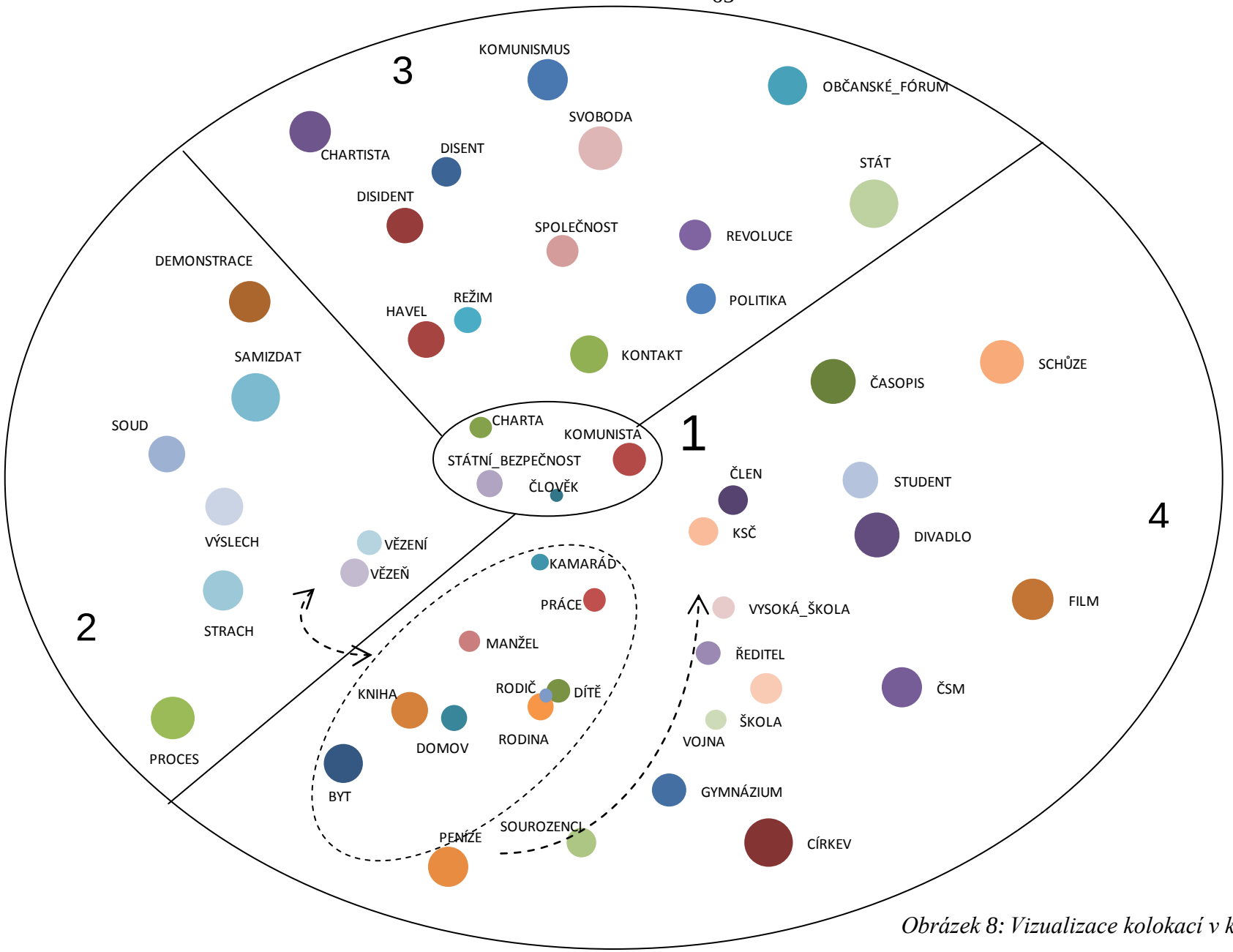
Jak již bylo naznačeno, diskurs disidentů můžeme rozdělit na podobné oblasti, jako tomu bylo u komunistických funkcionářů. Tato podobnost vychází opět ze způsobu, jakým byl rozhovor veden. Nacházíme zde opět *jádro* (1), které ovšem obsahuje méně slov, než tomu bylo v předchozím případě,[82] a dále oblasti (2), (3) a (4).

Důležitými pojmy – či společnými jmenovateli jednotlivých vyprávění – jsou CHARTA, STÁTNÍ BEZPEČNOST a KOMUNISTA/KOMUNISTÉ. Jádro tak charakterizuje celkovou strukturu vyprávění jako vnější systému. V jádru nenacházíme základní kameny systému (STRANU, VÝBORY, FUNKCE atd.) jako u komunistických funkcionářů, ale systém zde vystupuje spíše amorfně či holisticky (reprezentován jednak konceptem KOMUNISTA a také konceptem REŽIM).[83]

[81] Tutéž informaci nám dává i hierarchická clusterová analýza, která ukazuje, že zatímco disidenti tvoří větší a méně soudržné skupiny slov, od kterých se postupně oddělují dvojice či trojice významově spojených konceptů (např. člen a KSČ), diskurs funkcionářů tvoří více soudržných skupin textu.

[82] Na základě stromového grafu hierarchické clusterové analýzy provedené na korpusu disidentů (viz Příloha 5) lze tvrdit, že jádro vyprávění zcela chybí. Jednotlivé body blízko středu grafu totiž netvoří homogenní cluster, jako tomu bylo v případě korpusu funkcionářů. Ukazuje to na to, že vazby mezi slovy zahrnutými do jádra nejsou tak silné. Jejich příslušnost do jádra je dána jejich důležitostí v rámci celého vyprávění.

[83] Toto setkání se systémem v rovině ideologické reprezentuje skupina E v clusterové analýze.



Obrázek 8: Vizualizace kolokací v korpusu disidentů

Naopak je zde důležité setkání s represivní složkou systému, které je charakteristické pro oblast vyprávění (2), kterou zde nazveme *vyprávění o životě v opozici*. Tato oblast však na rozdíl od funkcionářů zabírá méně prostoru a je velmi málo strukturována. Redukuje se v podstatě na vězeňskou a represivní zkušenost, ale netvoří integrální součást vyprávění.[84]

Vyprávění o životě v opozici pak volně přechází do oblasti označené číslem (3), týkající se *opoziční ideologie*. Nacházíme zde podobné rozlišení mezi „životem" a „ideologií" jako v případě komunistických funkcionářů. Zatímco ve vyprávěních komunistických funkcionářů spolu tyto dvě stránky politického života příliš nesouvisely, u disidentů jsou obě tyto oblasti více provázány.

Stejně jako u funkcionářů je zde REVOLUCE přirozeným vyústěním, které stojí na opačném pólu než ideologie disidentská. Zajímavé v tomto srovnání je odlišné postavení konceptu POLITIKA, který je u funkcionářů spojen s pólem předrevolučním, zatímco disidenti jej spojují s pólem revolučním a porevolučním.

Pro disidenty jsou centrálním bodem této části vyprávění slova REŽIM a SPOLEČNOST. Chápou období normalizace jako období zvláštní, což podtrhují právě častým používáním slova REŽIM. Ve svém vyprávění příliš nepersonalizují ty, kdo vládnou, naopak volí optiku SPOLEČNOSTI, tedy optiku ovládaných. Slovo POLITIKA pak používají až v souvislosti s porevolučními událostmi. Své předchozí angažmá nechápou jako v pravém slova smyslu politiku.

Poslední oblastí vyprávění, která tvoří téměř polovinu celého grafu, je oblast č. 4 – *vyprávění o rodině a soukromém a rodinném životě*.[85] Osu této části tvoří řada slov (naznačená šipkou), která označuje vyprávění o mládí a dospívání disidentů. Vidíme, že tento vývoj je velmi zřetelný od studia na střední škole či gymnáziu přes vojnu a vysokou školu. Na konci této řady pak stojí konfrontace disidentů s členstvím v KSČ, které bylo důležitým prvkem jejich vysokoškolského studia. Mezi disidenty se vyskytuje mnoho bývalých členů KSČ. Členství v KSČ patřilo i do vysokoškolského kolektivu.

K VYSOKÉ ŠKOLE přiléhá i kulturní oblast této části vyprávění, která ovšem má rovněž blízko k členství v KSČ. Toto období je typické první kulturně politickou angažovaností, která měla pro vývoj disidentů určující vývoj.

[84] V clusterové analýze je tato oblast reprezentována skupinou D.

[85] Rodinná oblast je reprezentována v grafu clusterové analýzy skupinou C, oblast kulturní pak velkou a dále členitou skupinou F. Kulturní oblast je opět oddělena v souvislosti s její rolí v disidentském životě s vyústěním v revolučních událostech roku 1989. Rodinná oblast je od těchto událostí do jisté míry oddělena. Vidíme, že je blíže skupině D, tedy represivní disidentské zkušenosti, než skupinám E a F, reprezentujícím „konstruktivní" stránku disidentského života.

Do jisté míry oddělenou část vyprávění pak tvoří oblast rodiny, a to jak rodiny, ve které narátor vyrostl, tak rodiny, již založil. Vyprávění o rodině má blízko vyprávění o perzekucích prožívaných narátorem.

Tato blízkost rodiny k represivní stránce disidentského života tvoří základní odlišnost od vyprávění funkcionářů. Zatímco tedy funkcionáři ve svých vyprávěních oddělovali rodinu od svého politického života, u disidentů toto nenacházíme. Represe amorfního komunistického systému byla totální, zasahovala i RODINU, DOMOV a PRÁCI. Velmi zajímavou odlišnost v této oblasti zároveň nacházíme ve výrazně centrální pozici konceptu KAMARÁD, který u komunistických funkcionářů zcela chybí. Je patrné, že zatímco uvnitř systému dominují ve vyprávění formální vztahy, v disentu naopak tyto formální vztahy téměř zcela chybí. Vně systému zde tedy vzniká určitý neformální systém vztahů na systému nezávislý, který zároveň vstupuje i do represivních zásahů tohoto systému do světa disentu.

4.3 Konfrontace hermeneutické a počítačové analýzy biografických rozhovorů

V následující části se pokusíme výsledky získané rozborem výsledků počítačové analýzy konfrontovat s hermeneutickou analýzou rodinného prostředí funkcionářů a disidentů. Upozorníme na odlišnosti povahy otázek, které nám mohou obě metody zodpovědět, a zároveň se zamyslíme, zda jejich vzájemná interakce může sloužit k získání nových informací. Zároveň konfrontace obou metod má napomoci validizaci výsledků počítačové textové analýzy. Tato konfrontace má pomoci čelit námitce, že výstupy počítačové analýzy jsou určité umělé konstrukty, které nemají žádný smysluplný vztah k textům, z nichž vyšly.[86]

Výše citovaná analýza Petry Schindler-Wisten *Rodinné prostředí příslušníků politických elit a disentu* (Schindler-Wisten in Vaněk et al., 2006) si na začátku pokládá výzkumnou otázku, „jak rodinné prostředí působilo na konkrétní narátory, jaké vazby a vztahy v jejich rodinách fungovaly“ (Schindler-Wisten in Vaněk et al., 2006 : 205). Autorka tuto otázku zkoumá ve dvou rovinách: v rovině vlivu rodičů narátorů (tj. orientační rodiny) na jejich názory a přesvědčení a v rovině vlivu politické činnosti narátora na chod jím založené rodiny.

[86] U zastánců klasické kvalitativní metodologie může užívání počítačů v analýze vzbuzovat určitou nedůvěru. Kelle (1997) „hovoří o odcizování výzkumníka od jeho dat“. Je tedy třeba prokázat, že formalizovaný přístup, který se klasické kvalitativní analýze příčí, může být pro analýzu přínosný a hlavně s ní není v rozporu.

Už z této pozice je vidět, že přístup klasické kvalitativní analýzy biografických vyprávění je odlišný od počítačové analýzy. Velký soubor kvalitativních dat pro výzkumníka představuje soubor velkého množství informací, které je třeba analyzovat. Zatímco v klasickém přístupu si výzkumník nejprve pokládá otázku, která vymezí rovinu, v níž výzkumník bude texty analyzovat, čímž se analýza stane proveditelnou, počítačová analýza se naopak text snaží zkoumat jako celek ve své bohatosti, přičemž tuto bohatost redukuje zjednodušením analýzy na výběr klíčových slov. Pohledem na tuto makro-strukturu textu pak výzkumník jednak získává informaci o postavení jednotlivých oblastí v rámci celku vyprávění, dále však může generovat hypotézy, jejichž hlubší prozkoumání může být provedeno s pomocí kvalitativní analýzy. Zjednodušeně můžeme tvrdit, že zatímco počítačová analýza hledá strukturu jednotlivých bodů vyprávění, kvalitativní analýza zkoumá vztahy mezi vybranými body.

Aplikujeme-li tyto závěry konkrétně na náš případ, pak analýza Petry Schindler-Wisten zkoumá vztahy mezi rovinou vyprávění vztahující se k rodině a ideologicko-politickou rovinou vyprávění. Jak ukazuje počítačová analýza, můžeme vztah mezi těmito dvěma rovinami rozdělit na dvě části. Jednak je to postavení vyprávění o rodině (4) vůči jádru (1) a jednak vůči ideologickému vyprávění (3) a vyprávění o straně, resp. o disidentském životě (2).

V následujícím oddílu textu přednesu 5 konkrétních závěrů analýzy Petry Schindler-Wisten a pokusím se je konfrontovat s výsledky zde prezentované analýzy. Je třeba podotknout, že obě analýzy mají svá specifika. Kvalitativní analýza Petry Schindler-Wisten pracuje pouze s částí rozhovorů vzhledem k množství textů, které bylo v autorčiných možnostech zpracovat. Počítačová analýza nemůže být zároveň konfrontována s některými autorčinými závěry, které vycházejí z třídění podle pohlaví a věku (generace) narátorů, které nebyly v počítačové analýze brány v potaz.

Závěr 1: Orientační rodina ve vyprávěních funkcionářů a disidentů

Co se týče postavení orientační rodiny ve vyprávění disidentů i funkcionářů, lze analýzu Petry Schindler-Wisten shrnout do tvrzení, že narátoři ve vyprávění věnovali poměrně málo prostoru své orientační rodině (tj. rodině, do níž se narodili), neboť to nepovažovali za relevantní.

Jak autorka dále uvádí, vztahují se narátoři ke své orientační rodině různými způsoby. Téma „dětství a rodina“ bylo prvním tematickým okruhem, kterým byl rozhovor zpravidla zahajován (Schindler-Wisten in Vaněk et al., 2006 : 207), takže se ho všichni narátoři nějakým způsobem dotkli. Nepovažovali to však za relevantní téma pro jejich vyprávění. „Někteří narátoři se omezili

na stručný výtah ze svého rodinného zázemí, poskytli jakési resumé o svém původu a prostředí, v němž vyrůstali“ (Schindler-Wisten in Vaněk et al., 2006 : 207-208).[87] Jiní narátoři věnovali své orientační rodině více prostoru, ale jak uvádí autorka: „U některých [funkcionářů] můžeme pozorovat jistou snahu oddálit zevrubným vypravováním o dětství a vztahu ke všem členům široké rodiny další, pro ně nepříjemná témata, kterým se snažili vyhnout“ (Schindler-Wisten in Vaněk et al., 2006 : 208). To ukazuje na nízkou relevanci tématu.

Při pohledu na výstupy počítačové textové analýzy u komunistických funkcionářů nacházíme zřetelněji oddělenou orientační rodinu od rodiny, již založili. Důležitým společně se vyskytujícím se slovem je VÁLKA, která zde má význam jednak časové lokalizace a jednak důležité zkušenosti, která ovlivňovala dětství mnoha narátorů.

Jak již bylo uvedeno, funkcionáři věnují soukromému a rodinnému životu více prostoru než disidenti a zřetelněji jej oddělují od života politického. Vidíme ovšem, že slova MATKA a RODIČ se vyskytují na okraji grafu a mají menší důležitost než MANŽELKA odkazující k rodině, již narátor založil. Naopak slovo OTEC se vyskytuje častěji – je blíže středu grafu. Slovo DÍTĚ odkazuje k oběma rodinám, zastupuje dětství samotného narátora, ale i jeho potomky.

Orientační rodina disidentů vystupuje v poněkud odlišné roli v jejich vyprávění. Rodinný život v rámci vyprávění o soukromí narátorů zaujímá poměrně úzký prostor a zřetelně přiléhá k represivní zkušenosti s komunistickým režimem. Je zde zřetelné, že narátoři rodině příliš prostoru nevěnují a spíše se v rámci soukromého diskursu zabývají svým osobním životem, mládím a cestou, kterou se do disentu dostali a kde rodina nehrála centrální roli.

Obecně tedy lze tvrdit, že jak funkcionáři, tak disidenti tématu orientační rodiny příliš prostoru nevěnují. Zatímco hermeneutická analýza ukazuje na tendenci pracovat s tímto tématem spíše jako se zdrojem umožňujícím zmenšit prostor pro vyprávění o méně příjemných tématech (zejm. u komunistických funkcionářů), počítačová analýza ukazuje, že v rámci celku vyprávění hrály větší důležitost samotné politické aktivity narátorů (spojené s rodinou, již založili) než životní dráha, která narátory k jejich disidentství přivedla. V rámci vyprávění o této životní dráze pak hraje orientační rodina jen malou roli.

[87] Z analýzy je patrné, ač to není explicitně řečeno, že méně prostoru rodinným tématům věnovali zejména vysocí funkcionáři a nejvíce angažovaní disidenti, pro něž toto téma zkrátka nebylo relevantní. Jako ilustrace je zde citován Miroslav Štěpán, který své dětství shrnul do dvou vět (viz -Wisten in Vaněk et al., 2006 : 208).

Závěr 2: Politický vliv rodičů na funkcionáře

Podle analýzy Petry Schindler-Wisten není pravidlem, že by orientační rodina funkcionáře automaticky determinovala k vykonávání funkce. Důležitou roli přisuzují sociálnímu a politickému prostředí, v němž vyrůstali.

Autorka se zaměřuje na roli, kterou politické směřování rodičů hrálo při krystalizaci politického smýšlení a funkcionářské kariéry. Nepřistupuje však k vyprávění narátorů příliš kriticky a nechápe tento ideologický vliv rodičů jako zpětnou racionalizaci jejich politického angažmá.

Jak uvádí: „Kromě výchovy, kterou jim rodiče poskytli, považuji za velmi důležitý element společenského a politického dozrávání prostředí, v němž vyrůstali. Často zmiňují skromné životní podmínky, které v nich vyvolaly a zakořenily silné sociální cítění“ (Schindler-Wisten in Vaněk et al., 2006 : 219). Ve vzorku se však vyskytli i narátoři, jejichž rodiče nebyli politicky angažovaní v komunistické straně a kteří prezentují své politické angažmá jako „souběh různých životních okolností“ s primárně pragmatickou motivací.

V tomto ohledu je důležitější role otců narátorů, kteří byli jako muži přirozeně více politicky angažováni než ženy. Vzhledem k zaměření výzkumu je pro narátory politické zaměření rodičů relevantní informací, o které hovoří „téměř automaticky“ (Schindler-Wisten in Vaněk et al., 2006 : 219).

Počítačová analýza způsobem, jakým je užívána v této práci, neumožňuje snadné třídění podle vlastností narátorů. Proto není v intencích této práce odlišit narátory s rodiči politicky angažovanými a neangažovanými a zkoumat jejich vyprávění odděleně. Narativní praxe je zkoumána jako celek.

Z tohoto ohledu je ve vyprávění funkcionářů vidět oddělenost (resp. nízká míra spoluvýskytu) orientační rodiny narátorů od ideologie a politického života. V tomto ohledu je relevantní vztah orientační rodiny k jádru vyprávění, tj. k funkcionářskému angažmá narátorů. Je vidět, že role otce v utváření tohoto angažmá má větší důležitost než role matky. Slovo OTEC hrálo v celém vyprávění větší roli: nachází se blíže středu, má vyšší frekvenci. Počítačová analýza nám ukazuje, že spojení, které Petra Schindler-Wisten zdůrazňuje, tedy vztah angažmá rodičů a dětí, nepřesahuje svou relevancí vztah rámce vyprávění a jedné z epizod.

Naopak se skrze spojení slov RODIČE, MATKA, OTEC a VÁLKA, PENÍZE ukazuje, že sociální prostředí bylo relevantním tématem ve spojení s identifikační rodinou. Zatímco slovo VÁLKA zde slouží spíše jako časová lokalizace, slovo PENÍZE v souvislosti orientační rodinou

ukazuje, že sociální a ekonomické postavení této rodiny bylo pro komunistické funkcionáře relevantním tématem vyprávění.

Závěr 3: Politický vliv rodičů na disidenty

Co se týče identifikační rodiny disidentů, její vliv na politický vývoj narátorů je mnohem různorodější – najdeme zde disidenty vyrůstající v disidentské, politicky neutrální, ale i komunisticky orientované rodině. Vedle rodiny hrají významnou roli v „politické socializaci" disidentů jejich vrstevníci, četba knih a poslech zahraničního rozhlasu.

Jak tvrdí Petra Schindler-Wisten: „U disidentů bylo rodinné prostředí z hlediska politických názorů poněkud pestřejší než u komunistických funkcionářů. Ve větší míře tu pozorujeme odklon od názoru rodičů a v několika případech sledujeme úplný názorový střet, který dokonce vyústí v rozchod s rodiči" (Schindler-Wisten in Vaněk et al., 2006 : 222). Autorka identifikuje tři skupiny narátorů. Do první skupiny náleží narátoři s disidentskou rodinnou tradicí a sem lze zařadit i katolické rodiny. V druhé skupině najdeme disidenty, „kteří vyrůstali v neutrálním prostředí, kde rodiče nedávali své politické názory najevo. Mohli mít pro to dva důvody: buď se politikou vůbec nezabývali, byla jim lhostejná, anebo se nechtěli vyjadřovat před dětmi." (Schindler-Wisten in Vaněk et al., 2006 : 223). Výjimečně pak najdeme ve vzorku i narátory, kteří se negativně vymezili vůči komunistickému přesvědčení svých rodičů. Tyto dvě skupiny pak zdůrazňují roli vrstevníků, četby knih a poslechu zahraničního rozhlasu, což ovlivnilo jejich další politickou dráhu.

Jak již bylo uvedeno, počítačová analýza ukazuje spojení vyprávění o rodině s represivní zkušeností narátorů. Komunistické represe zasahují rodinu, již založili, i rodinu identifikační. Podíváme-li se na část oblasti (4) umístěnou v blízkosti oblasti (2), vidíme, že slova jako RODIČ, RODINA, DOMOV, která se nacházejí v blízkosti zmiňované represivní zkušenosti, dále ústí do dalšího vyprávění, které už není tolik blízké represivním složkám. Je charakterizováno slovy jako ŠKOLA, CÍRKEV, VYSOKÁ ŠKOLA, ČSM. V rámci vyprávění o rodině pak hraje důležitou roli slovo KAMARÁD. Ukazuje se tak poměrně stabilní struktura vyprávění o dětství a škole, kde svou roli hraje i církev jako činitel politické socializace. Rodinné a přátelské vztahy formované politikou jsou pak zasaženy i komunistickými represemi. Lokace slova KNIHA v blízkosti rodiny by mohlo ukazovat zároveň na důležitou roli četby knih v politické socializaci.

Konfrontace s členstvím v KSČ představuje vyústění průběhu celého dospívání narátorů, zejména se vztahuje k období vysokoškolského studia. Jak víme, rodinný život je spojen

s ideologickou rovinou politického uvažování disidentů více, než tomu bylo u funkcionářů. Přesto těžiště vyprávění spojuje rodinu s represemi režimu.

Závěr 4: Vliv politické kariéry funkcionářů na rodinu, již založili

Funkcionáři prezentují svou funkci tak, že oběti rodinnému životu výrazně převažují nad benefity, které jim funkce přinesla. Rodinný život podle této prezentace trpěl nedostatkem času a odloučením od rodiny spojeným se studijními pobyty v Moskvě.

Důležitým momentem vlivu stranické funkce na rodinný život je zdůrazňování negativních vlivů. „Z rozhovorů s komunistickými funkcionáři vyplývá, že často museli své funkci obětovat mnoho času. O děti a o domácnost se většinou staraly jejich manželky samy. Narátoři totiž jezdili na různá školení, stranické schůze se často protáhly do pozdních hodin nebo se konaly o víkendu“ (Schindler-Wisten in Vaněk et al., 2006 : 226). Důležitým momentem citelně zasahujícím rodinu byl zahraniční studijní pobyt na stranické vysoké škole v Moskvě. „V několika rozhovorech se vyskytují zmínky o tom, že manželky měly často značné výhrady proti manželovu odchodu na studia“ (ibid). To u některých funkcionářů působilo rodinné problémy někdy vedoucí až k rozvodu, ale většinou byl rodinný život stranické funkci podřízen a příkazy byly přijímány automaticky.[88] „Manželky funkcionářů zastávaly stejné politické přesvědčení jako manželé, a třebaže nebyly aktivní činností manželů přímo nadšeny, vždy se je snažily podporovat“ (ibid : 228).

Počítačová analýza u funkcionářů identifikuje silně zaplněné jádro, které rámuje celé vyprávění. Tento rámec je tvořen slovy, která reflektují jejich členství a pozici ve straně a cestu, jak se k ní dostali. Když se podíváme na strukturu tohoto jádra a na to, jak slova v jádru přiléhají jednotlivým okrajovým sektorům (2-4), vidíme, že hlavní propojení mezi rodinou a jádrem tvoří slova PRÁCE, ČSM[89] a VYSOKÁ ŠKOLA. To je v souladu s tím, že stranický život zasahuje do rodinného života skrze časovou náročnost práce a stranického studia.

I když autorka tvrdí, že benefity funkcionářského života jsou ve vyprávěních zamlčovány, počítačová analýza přesto identifikuje diskursivní uspořádání, kdy slova zastupují tyto benefity se mají tendenci vyskytovat v souvislosti s rodinou natátorů. Jedná se o slova BYT, AUTO, VZDĚLÁNÍ. Jejich pozice v blízkosti každodenního stranického a pracovního života zároveň ukazuje, že je zde určitá tendence hovořit o těchto tématech v souvislosti se stranickým životem.

[88] Jak ukazuje citace jednoho z narátorů k tomuto tématu: „Tehdá disciplína byla disciplína“ (ibid : 226).
[89] Angažmá v ČSM jako předstupeň stranické kariéry má spíše návaznost na identifikační rodinu a na děti narátorů.

Závěr 5: Vliv politické kariéry disidentů na rodinu, již založili

Podle analýzy Petry Schindler-Wisten byl rodinný život disidentů významně zasažen zejména represemi komunistického režimu. Přirozenou snahou disidentů bylo odstínit rodinu od bezpečnostních a existenčních rizik spojených s vykonáváním disidentské činnosti.

Jak uvádí Petra Schindler-Wisten, rodinná shoda na politických postojích byla nutným předpokladem fungování rodiny disidenta. „Všichni narátoři v našem souboru hodnotí svůj politický postoj ve větší či alespoň minimální míře jako stejný či podobný s politickým přesvědčením manžela, resp. manželky. V opačném případě by pravděpodobně jejich soužití nebylo vůbec možné“ (ibid : 229). Rodina disidentů byla zasažena jednak existenčními potížemi spojenými se ztíženým přístupem k lépe placené práci a vzdělání dětí. To vedlo rodiny k dělbě rolí, kdy zatímco manžel byl politicky činný, manželka pouze zajišťovala (ekonomicky i jinak) chod rodiny, i když se s manželovými politickými názory ztotožňovala. „[...] z vyprávění našich narátorů cítíme, že si byli vědomi určitého nebezpečí, které aktivní opoziční činnost přinášela, a nechtěli proto svoji rodinu do těchto záležitostí zatahovat, mnohdy se doma o svých aktivitách ani konkrétně nezmiňovali“ (ibid).

Přesto komunistické represe v podobě sledování, propouštění z práce a šikany dětí ve škole rodinu nutně zasahovaly. „Neustálé sledování ze strany StB, domovní prohlídky, výslechy atd., to vše bylo u odpůrců komunistického režimu na denním pořádku a často se citlivě dotýkalo členů jejich rodiny“ (ibid : 230). Rodinná podpora pak byla důležitá pro disidenta ve vazbě či ve vězení. „Když byl nějaký disident vzat do vazby nebo později uvězněn, byl všemožně podporován manželkou (resp. manželem), např. povzbudivými dopisy, balíčky, žádostmi o propuštění, účastí na soudních procesech apod.“ (ibid).

Komunistické represe na druhou stranu posilovaly rodinnou soudržnost disidentských rodin a rovněž podporovaly občanskou angažovanost dětí, které se často nějakým způsobem disidentského života rovněž účastnily.

Co se týče vlivu disidentského života na rodinu, počítačová analýza ukazuje zřetelný vliv komunistických represí na rodinný život, vyjádřený blízkostí skupiny slov MANŽEL, RODINA, DOMOV, KAMARÁD, PRÁCE a dalších[90] ke slovům jako VĚZEŇ, VĚZENÍ, STRACH, VÝSLECH a v centru grafu stojící STÁTNÍ BEZPEČNOSTI. Toto uspořádání slov je zřetelným vyjádřením postavení rodiny ve vyprávění disidentů v souvislosti s represemi komunistického

[90] Na Obrázku 8 zvýrazněn přerušovanou čárou.

režimu. Vidíme, že těmito represemi byly zasaženy jak rodinné a přátelské vztahy, tak ekonomická situace rodiny, vyjádřená PRACÍ. Věznění rodinu jednak zasahovalo existenčně a bezpečnostně, na druhou stranu byla pro vězněného důležitá podpora jeho rodiny a přátel. Právě toto propojení rodinných a přátelských vztahů v disidentském prostředí je určitým specifikem, na které můžeme na základě počítačové analýzy poukázat.

4.4 Shrnutí

Čtvrtá kapitola představila praktické užití metody kvantitativní textové analýzy. Na dvou textových korpusech, vyprávěních komunistických funkcionářů a disidentů, byly identifikovány rozdíly v narativní praxi obou těchto skupin, které reflektují odlišné chápání žitého světa normalizačního období u obou skupin.

Prvním zásadním rozdílem byla zaplněnost středu grafu. Vysoce zaplněný střed v grafu funkcionářů ukazuje na vysokou strukturovanost jejich životní zkušenosti v normalizačním období. Disidenti naopak tento jednotný rámec postrádají a chápou normalizační institucionální uspořádání jako vnější a nestrukturovanou entitu, která jejich život ovlivňuje skrze represe.

Druhou významnou odlišností je pozice soukromého a rodinného života ve vyprávění zejména v souvislosti s životem politickým. Tuto rovinu vyprávění reflektuje i analýza Petry Schindler-Wisten (in Vaněk et al., 2006). Zatímco funkcionáři svůj rodinný život ve vyprávěních významně oddělují od života politického, pro disidenty jsou rodina a politika významně propojeny zejména skrze represivní zkušenost s režimem.

Třetím významným rozdílem je odlišné chápání politiky obou skupin. Disidenti spojují slovo POLITIKA pouze s revoluční a porevoluční částí jejich vyprávění. Ve středu politicko-ideologické oblasti vyprávění stojí slova REŽIM a SPOLEČNOST. Funkcionáři slovo POLITIKA používají častěji a ve vyprávění jak předrevolučním, tak porevolučním. Zajímavé však je, že slova jako NÁZOR či SPOLEČNOST se vyskytují spíše v souvislosti s vyprávěním porevolučním.

Toto odlišné jazykové pojetí reflektuje odlišnou životní zkušenost spojenou s tím, co je chápáno jako normální a běžné a co naopak jako nenormální. Slovo REŽIM užívané disidenty reflektuje fakt, že se jedná o něco vnějšího, co nastavuje pravidla pro žitý svět. Naopak funkcionáři tento žitý svět přijímají jako zcela normální a vůbec jej neproblematizují. To se děje až v revoluční fázi vyprávění, kdy častěji užívají slovo KOMUNISTA, které reflektuje problematizaci jejich identity v době, kdy se postavení komunistické strany zcela převrací.

Závěr

Cílem této diplomové práce bylo představit metodu kvantitativní textové analýzy v kontextu jiných metod analýzy textů v sociologii a ostatních vědách. Metoda byla představena v jejích jednotlivých krocích a zhodnocena z hlediska validity a reliability svých výsledků.

Co se týče reliability, ukázalo se, že i přes poměrně vysoké hodnoty stresu (jako míry nepřesnosti zobrazení matice spoluvýskytů v dvojrozměrném prostoru) je rozložení slov ve vizualizaci dat poměrně stabilní. Má tendenci měnit se pouze u některých méně frekventovaných slov, která se vyskytují ve více možných kontextech.

Tato slova v další analýze ovšem nebyla vyloučena jako nevhodná. I tento typ slov – s vědomím jejich slabé vazby na své okolí – může být pro interpretaci použitelný. Je však třeba znát jeho chování v rámci grafu a s touto znalostí k interpretaci přistupovat.

Z hlediska validity bylo ukázáno, že interpretace výsledků metody CATA bylo možné propojit s výsledky klasické hermeneutické analýzy. Ukázalo se však také, že metoda přináší oproti klasické hermeneutické analýze poněkud jiný typ výsledků, což je dáno odlišnou rovinou zkoumání aplikovanou u obou přístupů.

Situace je podobná používání kombinovaných metodologických přístupů kvalitativní a kvantitativní analýzy. Ideálem společného užití obou těchto přístupů je komplementarita, kdy výsledky obou metod nejsou ve vzájemném rozporu a vzájemně se doplňují a vytvářejí celistvější obraz zkoumaného problému. Interakce mezi kvantitativními a kvalitativními metodami může probíhat ve všech čtyřech fázích výzkumu: ve fázi konceptualizace, fázi metodologické, fázi analytické a fázi formulace závěrů (Veisová, 2009 : 26).

Podobné společné užití obou výzkumných metodologických přístupů je žádoucí i u analýzy textových dat. Potřeba je o to větší, že zatímco u kvantitativního dotazníkového šetření získává výzkumník data na základě vlastní aktivní intervence (stanovení otázek, na které výzkum odpoví) a již předem očekává, jaké výsledky může touto metodou získat a co znamenají, u kvantitativní textové analýzy se výzkumníkova intervence omezuje na sestavení textového korpusu a výběr slov do slovníku.

Alespoň rámcová znalost textů, které jsou analyzovány, je pro práci s metodou potřebná. Nelze opírat pouze o frekvenční analýzu výskytu jednotlivých slov. Každá interpretace výsledků metody vyžaduje alespoň částečné předporozumění. Chce-li výzkumník interpretovat vztahy mezi

slovy, musí znát význam, který v textu hrají, a kontexty, v nichž se vyskytují.[91] Vychází to z logiky kvantitativní textové analýzy, na kterou upozorňuje Laver et al. (2003). Autoři tvrdí o svém přístupu k počítačové textové analýze, že „[...] se odlišuje od ‚tradičních' technik textové obsahové analýzy tím, že nakládá s texty nikoli jako s diskursy, které je třeba číst, interpretovat a porozumět jim [...], ale jako soubor dat o slovech obsahující informaci o pozici, kterou autor vůči předem daným dimenzím [tématu analýzy] zaujímá" (ibid : 312). Tento způsob nakládání se slovy v sobě nutně předpokládá, že slovům a jejich roli v textu bude výzkumník rozumět.

Odlišná rovina zkoumání interpretace metody CATA oproti hermeneutické analýze je dána zkoumáním frekvence společných výskytů daných slov v textu. Metoda tak otevírá pohled na data, který si čtenář běžným čtením textu ne vždy uvědomuje.

Silné stránky metody tkví v její schopnosti poměrně rychle zpracovat velké množství textů. Díky tomu analytik získá komplexní představu o analyzovaných datech. V kombinaci s dalšími metodami je pak možné zpřesňovat svou představu o datech a nacházet nové vztahy mezi pojmy. Hermeneutická analýza spojená s užitím metody KWIC (Key Words in Context) pak může sloužit k dalšímu rozboru vztahů mezi jednotlivými klíčovými slovy a koncepty. Například vysoká vzdálenost může jednak ukazovat, že vztah mezi určitými analytickými pojmy je slabý, ale také to může znamenat, že narátoři mají vysokou tendenci tento vztah zamlčovat. Distinkci mezi těmito dvěma interpretacemi můžeme odhalit skrze čtení textů a jejich výklad.

Existují také určité směry, kterými lze metodu dále rozvíjet. Týká se to například zavedení časového hlediska do analýzy a zkoumání proměn konfigurací klíčových slov v diachronních textových korpusech.

Další směr rozvoje metody se týká efektivnějšího posouzení kvality modelu na základě kritérií statistické významnosti. V této práci byla tato zkoumání činěna pouze explorativně bez porovnání s určitou cílovou hodnotou.

To souvisí i s hledáním indikátorů pro posouzení vhodnosti slova pro zahrnutí do modelu. Ty by měly již v přípravné fázi výzkumu pomoci výzkumníkovi posoudit, zda dané slovo je potenciálně vhodné pro zahrnutí do modelu. Toto rozhodnutí je dosud činěno výzkumníkem spíše arbitrárně, jediným empirickým indikátorem je frekvence výskytu slova. Lze si však představit existenci určitých dalších indikátorů, které budou měřit vhodnost slova pro analýzu na základě (ne)rovnoměrnosti jeho výskytu v textovém korpusu, na základě počtu různých slov vyskytujících se v jeho okolí apod.

[91] Určitou představu o významu slov zahrnutých v modelu výzkumník získává průběžně při sestavování slovníku. Opírá se o metodu KWIC (Key Words in Context), s jejíž pomocí kontroluje, zda slovo zahrnuté do slovníku se v textu nevyskytuje v neočekávaných významech. Určitou znalost tak výzkumník získává v této přípravné fázi.

Třetím takovým možným vylepšením je efektivnější způsob nalezení vhodné velikosti kontextové jednotky na základě vlastností textu, tedy již v přípravné fázi výzkumu. I takové rozhodnutí by bylo možné činit na základě určitých statistických vlastností textů.

Je samozřejmé, že metoda má vedle uvedených silných stránek také několik omezení. Ta jsou dána jednak tím, že matematické a statistické vlastnosti výskytu a kolokací slov v textech zcela neodpovídají běžným veličinám, které zkoumáme při analýze proměnných v empirickém výzkumu. To se projevuje například tím, že vzdálenost dvou slov v grafu nelze interpretovat stejně jako vzdálenost např. v geografické mapě.

Druhý problém představuje fakt, že široká dostupnost velkého množství textů, která umožňuje nový způsob jejich zkoumání, ještě neznamená, že v těchto textech objevíme nějaké nečekané souvislosti. Ukazuje se to například na textech mediálních, kdy rostoucí počet kanálů pro šíření mediálních sdělení a rostoucí objem mediálních obsahů jimi šířených neimplikuje rostoucí rozmanitost těchto obsahů.

To, že budeme získávat další a další výpovědi o společenské realitě stále novými kanály, ještě automaticky neznamená, že obraz reality obsažený v těchto výpovědích bude komplexnější a hlavně koherentní.

Metoda zkoumání textů skrze jejich kvantitativní popis nám tak může umožnit pokládat si nové otázky o společenské realitě, ale to ještě neznamená, že tím získáme nové, smysluplné a validní odpovědi.

Vedle pokládání nových otázek metoda umožňuje analyzovat nové typy dat: velké textové korpusy, jakým je například Český národní korpus či velké soubory mediálních textů. Tato data představují soubor výpovědí o společenské realitě, který lze poměřovat kritérii reprezentativity. Zároveň však tento analytický přístup není striktně pozitivistický, ale zaměřuje se na užívání jazyka a utváření výpovědí o subjektivně prožívané realitě na základě sociálně konstruovaných pojmů.

Metoda CATA užívající zkoumání kolokací slov v textu, jak byla prezentována v této práci, je stále poměrně jednoduchá metoda zabývající se jednoduchými vztahy mezi slovy. Jazyk je však mnohem složitější symbolický systém. Stále komplexnějšímu popisu užívání jazyka a jeho zpracování se velmi intenzivně věnuje počítačová věda a korpusová lingvistika. V rozvoji obou těchto disciplín existuje velký inspirační potenciál také pro další rozvoj kvantitativních textových analýz.

Summary

The diploma thesis deals with the access to computer-assisted text analysis (CATA) which examines the collocation of words in large text corpora and possibilities of its use in sociology and other social sciences.

This analytical approach is compared with other approaches to text analysis. Apart from corpus linguistics and text mining, the main emphasis is put on the quantitative and qualitative analysis of texts in social sciences. When quantitative analysis is concerned, the text focuses on the use of computers. Qualitative analysis is presented as a method that works with *understanding* and *reflexivity* as the key methodological approaches to the analysis of texts. As shown in the work of Alexa (1997), one cannot strictly speak of qualitative and quantitative analysis as two separate methodologies, but rather as a "qualitative - quantitative" continuum. The method presented here can be placed somewhere in the middle of this continuum.

The second chapter is devoted to the CATA general procedure and presents analysis of its individual steps: preparation of a text corpus, building dictionary, distance calculation and visualization of results. The resulting model depends on the choice of other parameters: the coefficient of distance between words and the size of the context units. The choice of these parameters depends on the investigator's decision and should be made with regard to the interpretability of the model.

The third chapter shows the use of procedures introduced in the second chapter on two text corpora, the biographical narratives of Czechoslovak normalization actors: dissidents and Communist functionaries. The process of compilation of dictionaries based on analysis of the frequency of words' occurrence in the text corpora is introduced and evaluation of the quality of the resulting models (depending on the chosen parameters) is made. The model using *Jaccard's coefficient* and the *100 words* context unit is chosen to be interpreted.

This interpretation is introduced in the fourth chapter of the thesis. The analysis identifies three fundamental differences in narrative practice of both groups, which reflect different understanding of the lived world of the normalization in both groups.

The first major difference is the high number of words placed in the centre of the functionaries' visualization, which shows the high rate of structuredness of their life experience in the normalization period. Dissidents on the contrary lack this kind of a common framework and understand the normalization's institutional arrangements as an external unstructured entity that affects their lives through the repressive authorities.

The position of the private and the family life in the narrative (especially in relation to the political life) is another important difference. While functionaries in their family life narratives significantly separate political life and family, in dissidents' narratives, policy and family are significantly linked mainly through experience with the repressive regime.

The third important difference is the different understanding and usage of the word POLICY in both groups. Dissidents associate POLICY with the post-revolutionary and revolutionary part of their narrative. Their political-ideological narrative is mostly structured by the words REGIME and SOCIETY. Functionaries use the word POLICY when speaking about both the pre-revolutionary and the post-revolutionary phase. It is interesting that words like OPINION and SOCIETY occur more often in the post-revolutionary narrative.

This different language usage reflects different life experience connected with what is seen as normal and common and what is seen as abnormal. The word REGIME used by dissidents reflects the fact that it is an external entity that sets the rules for the lived world. On the contrary, the functionaries accept the lived world as normal and do not question it.

In the interpretation process, we can see an advantage of CATA over the hermeneutic analysis as the possibility to quickly process large text corpora. On the higher level of analysis of texts, the model reveals relationships between the key words that are hardly detectable in the normal reading. At the same time, however, the analysis requires a prior understanding of the analyzed texts and knowledge of the analyzed words' meaning.

Použitá literatura:

ALEXA, Melina. Computer-assisted text analysis in the social sciences. *ZUMA Arbeitsbericht*. 1997, č. 7.

BAKER, Paul. *Using corpora in discourse analysis*. New York: Continuum, c2006, 198 s. ISBN 08-264-7725-9.

BAYER, Ivo, Jitka KOLÁŘOVÁ, Marta KOLÁŘOVÁ a Martin VÁVRA. *Zobrazování nerovností a hodnotová poselství v časopisech pro děti a mládež na příkladu časopisu Bravo*. Praha: Sociologický ústav AV ČR, 2009, 127 s. ISBN 978-80-7330-172-9.

BELL, Gordon, Tony HEY a Alex SZALAY. Beyond the Data Deluge. *Science*. 2009, č. 323, s. 1297-1298. ISSN 0036-8075. Dostupné z:

<http://www.cloudinnovation.com.au/Bell_Hey%20_Szalay_Science_March_2009.pdf>

BERELSON, Bernard. *Content Analysis in Communication Research*. Glencoe, Illinois: The Free Press, 1952. 220 p.

BEST, Michael Lloyd. *An Ecology of the Net: Message Morphology and the Evolution in Net News*. Cambridge, MA, May 10, 1996. Dostupné z: http://mit.dspace.org/bitstream/handle/1721.1/61089/35522664.pdf?sequence=1. Master degree thesis. MIT. Vedoucí práce Kenneth B. Haase Jr.

BIEWER, Carolin, Marianne HUNDT a Nadja NESSELHAUF. Corpus linguistics and the web. New York: Rodopi, 2007, 305 s. *Language and computers, no. 59*. ISBN 90-420-2128-4.

BLUMER, Herbert. *Critiques of research in the social sciences: an appraisal of Thomas and Znaniecki's The Polish peasant in Europe and America*. New Brunswick, N.J.: Transaction Books, c1979, 210 s. Bulletin (Social Science Research Council (U.S.)), 44. ISBN 08-785-5694-X.

BORG, Ingwer a Patrick J. GROENEN. *Modern multidimensional scaling: theory and applications*. 2nd ed. New York: Springer, c2005, 614 s. ISBN 03-872-5150-2.

CARLEY, Kathleen M. Network Textual Analysis. ROBERTS, Carl W. *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. New York: Routledge, 1997, s. 79-100. ISBN 0-8058-1734-4.

CHEN, Chun-houh, Wolfgang HÄRDLE a Antony UNWIN. *Handbook of data visualization*. Berlin: Springer, c2008, 936 s. ISBN 35-403-3036-4.

CORSARO, William A. a David R. HEISE. Event Structure Models from Etnographic Data. Clifford C. Clogg. Cambridge, MA: Basil Blackwell, 1990. *Sociological methodology, vol. 30*. Dostupné z: http://www.indiana.edu/~socpsy/papers/ethnog/ethnography.htm

COX, Trevor F. a Michael A. COX. *Multidimensional scaling*. 2nd ed. Boca Raton: Chapman, c2001, 308 s. ISBN 15-848-8094-5.

ČERMÁK, Ivo. Myslet narativně (kvalitativní výzkum „on the road“). In I. Čermák, M.. Miovský (Ed). *Sborník z konference Kvalitativní výzkum ve vědách o člověku na prahu třetího tisíciletí*. Brno: Psychologický ústav AV ČR, Nakladatelství Albert, 2002, s. 11-25. Dostupné z: <http://www.rorschach.cz/?p=189>.

ČERNÝ, Jiří. *Úvod do studia jazyka*. 2. vyd. Olomouc: Rubico, 2008, 248 s. ISBN 978-807-3460-938.

DAWKINS, Richard. *Sobecký gen*. Vyd. 1. Překlad Vojtěch Kopský. Praha: Mladá fronta, 1998, 319 s. ISBN 80-204-0730-8.

FELDMAN, Ronen a James SANGER. *The text mining handbook: advanced approaches in analyzing data*. New York: Cambridge University Press, 2007, 410 s. ISBN 05-218-3657-3.

FISCHER-ROSENTHAL, Wolfram a Gabriele ROSENTHAL. Analýza narativně-biografických rozhovorů. *Biograf: časopis pro biografickou a reflexivní sociologii*. Přeložila Denisa Palečková. 2001, č. 24. ISSN 1211-5770. Dostupné z: http://www.biograf.org/clanky/clanek.php?clanek=v2402

FRENCH, Robert M. a Christophe LABIOUSE. Four Problems with Extracting Human Semantics from Large Text Corpora. In: *Proceedings of the 24th Annual Conference of the Cognitive Science Society*. 2002. Dostupné z:

<http://lead.u-bourgogne.fr/people/french/CogSci2002.french_labiouse.pdf>

GLASER, Barney G. a STRAUSS, Anselm L. *The discovery of grounded theory: strategies for qualitative research*. 1st pbk. ed. Chicago: Aldine Pub, 1973. ISBN 978-020-2302-607.

HÁJEK, Martin. Počítačová textová analýza metodou sledování spoluvýskytů slov. *Data a výzkum-SDA info*. 2010, roč. 4, č. 1, s. 19-37. ISSN 1802-8152.

HÁJEK, Martin. Pojem normality v tisku a běžném hovoru: vývoj, sémantika a sociologické aspekty. In: KABELE, Jiří, Martin POTŮČEK, Irena PRÁZOVÁ a Arnošt VESELÝ (eds.). *Rozvoj české společnosti v Evropské unii: (příspěvky z konference konané ve dnech 21.-23.10.2004)*. Vyd. 1. Praha: Matfyzpress, 2004. ISBN 80-86732-35-5. Dostupné z: <http://instituty.fsv.cuni.cz/~hajek/temp/hajek_2004_normalita.pdf>

HÁJEK, Martin. Proměny dimenze soukromého a veřejného v biografických vyprávěních pamětníků. In: GJURIČOVÁ, Adéla (ed.). Sborník z konference „1989-2009: Společnost. Dějiny. Politika“ [online]. Praha, 2009 [cit. 2012-05-15]. Dostupné z:

<http://www.boell.cz/downloads/hbs-studie-Hajek.pdf>

HÁJEK, Martin a Ivo BAYER. Diskurzivní stabilita „ne/spravedlivého“: kvantitativní obsahová analýza českých deníků z let 1996–2006. In: Pražské sociálně vědní studie. Praha: FSV UK, 2007. Sociologická řada, SOC-010. ISSN 1801-5999. Dostupné z: <http://publication.fsv.cuni.cz/attachments/236_Hajek,%20Bayer.pdf>

HÁJEK, Martin, Jiří KABELE a Kateřina VOJTÍŠKOVÁ. Zázemí' a 'bojiště' v usilování o spravedlnost: textová analýza odborářské, feministické a lidskoprávní mediální komunikace. *Sociologický časopis*. Praha: Sociologický ústav AV ČR, 2006, roč. 42, č. 2, s. 269-290. ISSN 0038-0288.

HEBÁK, Petr. *Vícerozměrné statistické metody*. Vyd. 1. Praha: Informatorium, 2005. ISBN 80-733-3039-3.

HENDL, Jan. Hermeneutika: metodologické poznámky. In: HOGENOVÁ, Anna. *Hermeneutika sportu*. 1. vyd. Praha: Karolinum, 1998. ISBN 80-7184-744-5.

HENDL, Jan. *Kvalitativní výzkum: základní teorie, metody a aplikace*. 2., aktualiz. vyd. Praha: Portál, 2008, 407 s. ISBN 978-80-7367-485-4 (VáZ.).

KELLE, Udo. Theory Building in Qualitative Research and Computer Programs for the Management of Textual Data. *Sociological Research Online* [online]. 1997, roč. 2, č. 2 [cit. 2012-05-03]. Dostupné z: http://www.socresonline.org.uk/2/2/1

KONOPÁSEK, Zdeněk. Text a textualita v sociálních vědách: Část druhá - metodologické motivace. *Biograf: časopis pro biografickou a reflexivní sociologii* [online]. 1996b, č. 8 [cit. 2012-02-21]. ISSN 1211-5770. Dostupné z: http://www.biograf.org/clanky/clanek.php?clanek=802

KONOPÁSEK, Zdeněk. Text a textualita v sociálních vědách: Část první - teoretické motivace. *Biograf: časopis pro biografickou a reflexivní sociologii* [online]. 1996a, č. 7 [cit. 2012-02-21]. ISSN 1211-5770. Dostupné z: http://www.biograf.org/clanky/clanek.php?clanek=703

KONOPÁSEK, Zdeněk. Text a textualita v sociálních vědách: Část třetí - refelxivní impuls. *Biograf: časopis pro biografickou a reflexivní sociologii* [online]. 1997, č. 9 [cit. 2012-02-21]. ISSN 1211-5770. Dostupné z: http://www.biograf.org/clanky/clanek.php?clanek=902

KRAUS, Jiří. *Jazyk v proměnách komunikačních médií*. Vyd. 1. Praha: Karolinum, 172 s. Učební texty Univerzity Karlovy v Praze, 15. ISBN 978-802-4615-783.

KRIPPENDORFF, Klaus. *Content analysis: an introduction to its methodology*. 2nd ed. Thousand Oaks, Calif.: Sage, 2004. ISBN 07-619-1545-1.

LAVER, Michael, Kenneth BENOIT a John GARRY. Extracting Policy Positions from Political Texts Using Words as Data. *The American political science review*. 2003, roč. 97, č. 2, s. 311-331. ISSN 0003-0554.

LOHR, Steve. For Today's Graduate, Just One Word: Statistics. In: *New York Times* [online]. 5. srpna 2009 [cit. 2012-02-19]. Dostupné z:

<http://www.nytimes.com/2009/08/06/technology/06stats.html>

MACHOVÁ, Svatava a Milena ŠVEHLOVÁ. *Sémantika a pragmatická lingvistika*. Praha: Univezita Karlova, Pedagogická fakulta, 2001, 159 s. ISBN 80-729-0061-7.

MANNING, Christopher D a Heinrich SCHÜTZE. *Foundations of statistical natural language processing*. Cambridge: MIT Press, c1999, 680 s. ISBN 02-621-3360-1.

MIHALCEA, Radu a Stephen PULMAN. Linguistic Ethnography: Identifying Dominant Word Classes in Text. In: GELBUKH, Alexander. *Computational linguistics and intelligent text processing: 10th International Conference, CICLing 2009, Mexico City, Mexico, March 1 - 7, 2009, proceedings*. Berlin: Springer, 2009, s. 594-602. ISBN 3-642-00381-8 ISSN 0302-9743.

MOHAMMAD, Saif a Graeme HIRST. *Distributional Measures as Proxies for Semantic Relatedness*. Toronto : University of Toronto, 2005. Dostupné z:

<ftp://ftp.cs.toronto.edu/pub/gh/Mohammad+Hirst-2005.pdf>

NORUŠIS, Marija J. *SPSS 14.0 advanced statistical procedures companion*. Upper Saddle River : Prentice Hall : SPSS, 2005. xiii, 366 s. ISBN 0-13-174700-2.

PATEL, Malti, John BULLINARIA a Joseph P. LEVY. Extracting Semantic Representations from Large Text Corpora. In: *Proceedings of the Fourth Neural Computation and Psychology Workshop*. London: Springer-Verlag, 1997, s. 199-212.

PETRUSEK, Miloslav. *Teorie a metoda v moderní sociologii*. Vyd. 1. Praha: Univerzita Karlova, 1993, 204 s. ISBN 80-706-6799-0.

POPPING, Roel. *Computer-assisted text analysis: an introduction to its methodology*. 2nd ed. Thousand Oaks, Calif.: Sage Publications, 2000, 229 s. ISBN 07-619-5379-5.

Projekt GAP404/10/0790 - Instituce v životních příbězích: Víceúrovňová srovnávací analýza biografických vyprávění tří skupin aktérů české společnosti 2. poloviny 20. století (2010-2012, GA0/GA). *Informační systém výzkumu, experimentálního vývoje a inovací: Výzkum, vývoj a inovace podporované z veřejných prostředků ČR* [online]. 2011, 30. 3. 2012 [cit. 2012-04-21]. Dostupné z: <http://www.isvav.cz/projectDetail.do?rowId=GAP404%2F10%2F0790>

SCHINDLER-WISTEN, Petra. Rodinné prostředí příslušníků politických elit a disentu. In VANĚK, Miroslav, et al. *Mocní? A bezmocní?*. Vyd. 1. Praha : Prostor, 2006. s. 411.

SCHÜTZE, Fritz. Narativní interview ve studiích interakčního pole. *Biograf: časopis pro biografickou a reflexivní sociologii*. Přeložila Petra Pavlištíková. 2001, č. 24. ISSN 1211-5770. Dostupné z: http://www.biograf.org/clanky/clanek.php?clanek=2003

Slovník komunistické totality. Vyd. 1. Editor František Čermák, Václav Cvrček, Věra Schmiedtová. Praha: Nakladatelství Lidové noviny, 2010, 302 s. Korpusová lexikografie, sv. 3. ISBN 978-807-4220-609.

Sociologický časopis. Praha: Sociologický ústav AV ČR, 2002, roč. 38, č. 4. ISSN 0038-0288.

Sociologický časopis. Praha: Sociologický ústav AV ČR, 2006, roč. 42, č. 2. ISSN 0038-0288.

SPENCE, Ian. A simple approximation for random rankings stress values. *Multivariate Behavioral Research*. 1979, roč. 14, č. 3, s. 355-365.

THOMAS, William Isaac, Florian ZNANIECKI a Eli ZARETSKY. *The Polish peasant in Europe and America: a classic work in immigration history*. Urbana: University of Illinois Press, c1996, 127 s. ISBN 02-520-6484-4.

ÚSTAV ČESKÉHO NÁRODNÍHO KORPUSU. *Český národní korpus* [online]. Praha, 2012 [cit. 2012-04-28]. Dostupné z: <http://ucnk.ff.cuni.cz>

VANĚK, Miroslav. *Mocní? a bezmocní?: politické elity a disent v období tzv. normalizace : interpretační studie životopisných interview*. Vyd. 1. Praha: Prostor, 2006, 411 s. ISBN 978-807-2601-615.

VEISOVÁ, Eva. *Možnosti a důsledky kombinace metod v sociologickém výzkumu se zřetelem na metody focus groups a internetového výzkumu*. Praha, 2009. Dizertační práce. Fakulta sociálních věd Univerzity Karlovy. Vedoucí práce Hynek Jeřábek.

WASSERMAN, Stanley a Katherine FAUST. *Social network analysis: methods and applications*. New York: Cambridge University Press, 1994, 825 s. ISBN 05-213-8707-8.

WITTGENSTEIN, Ludwig. *Filosofická zkoumání*. Vyd. 1. Praha: Filozofický ústav AV ČR, 1993, 294 s. Základní filosofické texty, sv. 2. ISBN 80-700-7040-4.

Software a online nástroje pro analýzu textů:

ANTHONY, Lawrence. *AntConc (Version 3.2.2)* [počítačový software]. Tokyo : Waseda University, 2011. Dostupné z: <http://www.antlab.sci.waseda.ac.jp/>

Atlas.ti (verze 6.2) [počítačový software]. Berlin : ATLAS.ti Scientific Software Development GmbH, 2010. Dostupné z: <http://www.atlasti.com/demo.html>

COOA. *Co-occurrence Analysis Software* [počítačový software]. Praha: Fakulta sociálních věd UK, 2009. Dostupné z: <http://publication.fsv.cuni.cz/attachments/471_setup_COOA.exe>

PENNEBAKER, James W., BOOTH, Roger J. a FRANCIS, Martha E. *LIWC: Linguistic Inquiry and Word Count*. Austin, Texas : Pennebaker Conglomerates Inc., 2007. Dostupné z: <http://www.liwc.net/index.php>

TextStat (Verze 2.9). Berlin : Freie Universität Berlin, 2012. Dostupné z: <http://neon.niederlandistik.fu-berlin.de/static/textstat/TextSTAT-2.9.zip>

WebCorp: *The Web as Corpus* [online]. Birmingham: Research and Development Unit for English Studies, Birmingham City University, 2012 [cit. 2012-04-28]. Dostupné z: <http://www.webcorp.org.uk/live/>