Kvalita modelů a volba parametrů počítačové textové analýzy v programu COOA*

Václav Čepelák**

Fakulta sociálních věd, Univerzita Karlova v Praze

Model evaluation and parameter selection in COOA computer-assisted text analysis software

Abstract: *This article examines the reliability of statistical models that use visualization of word distances using computer-assisted text analysis. This study looks at the choice of parameters in the COOA - software for word co-occurrence analysis. The word co-occurrence analysis enables visualization of text structure through the exploration of the number of co-occurrences of words. The data visualization provided by a multi-dimensional scaling (MDS) procedure is susceptible to a particular form of error. The nonlinear relationship between words with significantly different frequencies lies at the root of this problem where words with higher frequencies are placed in the middle of a two-dimensional MDS map visualization. Words with lower frequency, on the other hand, are forced by the MDS estimator to the edge of the two-dimensional map and their estimated spatial positions are unstable. These two processes are potentially a major source of error in making inferences. One solution for reducing this source of error is to (a) reduce the number of words in a model or (b) increase of the number of model dimensions. This article, however, suggests that a detailed investigation of the word structure and a thorough analysis of the error sources and their meaningful interpretation may be a better solution.*

Key words: *computer-assisted text analysis, multidimensional scaling, similarity measures, contextual unit, data visualisation.*

Data a výzkum - SDA Info 2013, Vol. 7, No. 1: 7-27.

DOI: <http://dx.doi.org/10.13060/1802-8152.2013.7.1.1>

(c) Sociologický ústav AV ČR, v.v.í., Praha 2013.

* Tento článek vznikl v rámci řešení výzkumného projektu P404/10/0790, podpořeného Grantovou agenturou České republiky.

** Veškerou korespondenci směřujte na e-mail: v.cepelak@seznam.cz.

Úvod

Rozvoj analytické sociologie je do značné míry dán rozvojem informačních technologií, které umožňují popisovat data novým způsobem. Zároveň umožňují pracovat s novými typy dat, což je příklad analýzy sítí (social network analysis), pro niž předmětem analýzy nejsou informace o individuální jednotce, ale tzv. data relační, vypovídající o vztahu dvou či více jednotek [srov. Scott 2000: kap. 1].

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

Sociologie tradičně s texty pracuje s využitím dvou hlavních metodologických přístupů (paradigmat). První z nich tvoří Berelsonova kvantitativní obsahová analýza, která je systematickou analýzou textových dat vycházející z pozitivistické metodologie [Berelson 1952]. Druhé paradigma má kořeny v Gadamerově hermeneutice, která se zaměřuje na hledání alternativních významových rovin v textech s využitím principu hermeneutického kruhu [srov. Gadamer 2010 (1960)]. Konkrétním návodem k analytickému přístupu k textovým datům je pak grounded theory Gläsera a Strausse [Glaser, Strauss 1973] a její princip třístupňového kódování.

Trendem posledních dvaceti let je zejména zapojování počítačů do analýzy textů. Počítače umožňují zrychlení procesu analýzy a zvýšení její complexity, na druhé straně ale prohlubují otázku validity metod. Kvantitativní analýza textů přestává být založena výhradně na kvantifikaci předem stanovených schémat (proměnných), ale umožňuje explorativně odkrývat v textech předem neznámé struktury (významy).¹ Trendem je proto propojování obsahové analýzy s poznatky lingvistiky. Svou otevřeností se tak počítačová textová analýza přibližuje kvalitativní metodologii, ač stále pracuje s kvantifikací poznatků [Alexa 1997: 6; Baker 2006].

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 nechápe kvantitativní metodologii pouze jako přístup založený na kvantifikaci poznatků, ale tvrdí, 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].

Sblížení obou přístupů tak spočívá v tom, že explorace dat není omezena žádným a priori daným schématem, ale toto schéma je možné budovat vý-

1 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.

hradně na základě dat.² Ukázkou aplikace jedné z těchto metod je i následující text.

Cílem textu není metodu popisovat ani dávat návod k jejímu užití – tuto práci již odvedl Martin Hájek [2010]. Text může posloužit jako zdroj hlubšího pochopení principů zobrazování a interpretace dat při využívání počítačové textové analýzy, která není v sociologii běžně užívána, ale v určitých tematických oblastech má potenciál pro využití. Zároveň text může posloužit i jako úvaha nad využíváním mnohorozměrného škálování sociologických dat relační povahy, které s sebou nese určité limity – zejména z hlediska asymetrie vztahů mezi zkoumanými jednotkami.

V první části textu bude metoda stručně představena s odkazem na citovaný Hájkův článek. Dále se text zaměří na zobrazování spoluvýskytů slov v textu, a tedy struktury textu. Věnovat se bude jednotlivým parametrům, které ovlivňují strukturu zobrazení, a hodnocení kvality vizualizace. Poukáže na to, že chyby zobrazení nejsou dány pouze nepřesností metody, ale mají příčinu v povaze zobrazovaných vztahů mezi slovy – zejména mezi slovy s vysokou a slovy s nižší frekvencí výskytu.

Metoda zkoumání spoluvýskytu slov v textech

Pro zarámování celé analýzy je nutné alespoň ve stručnosti představit metodu, která byla pro analýzu použita. Jedná se o metodu počítačové textové analýzy (CATA – Computer Assisted Text Analysis) zkoumající spoluvýskyty slov v textu. K tomuto zkoumání je využit software COOA [COOA 2009], jehož autorem je Martin Hájek (více k metodě Hájek [2010]).

Metoda operuje s následujícími klíčovými pojmy:

1. *textový korpus* – soubor textů, na něž je metoda aplikována,
2. *lemma* – souhrn tvarů téhož klíčového slova, které je v textovém korpusu vyhledáváno,
3. *slovník* – seznam klíčových slov (lemmat), která jsou v textu zkoumána,
4. *kontextová jednotka* – úsek textu, v jehož rámci jsou počítány společné výskyty slov (lemmat) obsažených ve slovníku.

Při využití metody je třeba projít třemi fázemi. Nejprve je provedena *frekvenční analýza* výskytu slov v textu, která pomůže určit klíčová slova. Na

2 Tvrděním o sblížování kvalitativní a kvantitativní metodologie se nesnažíme relativizovat postavení kvalitativního výzkumu jako odnože výzkumu kvantitativního. Pouze se zde pokoušíme konstatovat, že mezi komplexním popisem individuálních případů na základě zcela otevřeného schématu (tj. kvalitativní metodologie) a induktivně-zobecňujícím popisem celku na základě a priori daného schématu (tj. kvantitativní metodologie) mohou existovat určité mezistupně. Z aplikace metody níže v tomto textu vyplývá, že typ informace využívá výhod otevřenosti metody a umožňuje vytvářet analytická schémata zcela na základě dat, na druhé straně dává odlišný typ informace, který je na vyšší úrovni obecnosti, zároveň však méně komplexní, než zjištění, která poskytuje kvalitativní analýza.

základě klíčových slov je sestaven *slovník*, se kterým výzkumník vstoupí do druhé fáze analýzy v programu COOA. Pro textové korpusy analyzované v tomto textu byla minimální frekvence lemmat okolo 100 výskytů.

V programu je třeba definovat *kontextovou jednotku* jako určitý počet slov, znaků, vět či odstavců. Program pak rozčlení text na jednotlivé úseky a v těchto úsecích sleduje, zda se zde společně vyskytne některá dvojice slov ze *slovníku*. Výstupem je pak frekvenční matice spoluvýskytů slov (*lemmat*) a zároveň matice podobnosti (resp. vzdálenosti) těchto slov. Podobnost dvojic slov je vypočtena na základě koeficientu podobnosti, který je rovněž zvolen jako parametr v programu COOA (nejčastěji používaným je koeficient Jaccardův).

Matice podobnosti je pak ve třetí fázi statisticky zpracována s využitím různých statistických procedur: faktorové analýzy, clusterové analýzy či mnohorozměrného škálování. Mnohorozměrné škálování primárně slouží k vizualizaci rozložení slov v textu a jejich interpretaci.

Volba slov do slovníku je klíčovou fází analýzy, která determinuje povahu výsledků (více k této fázi Hájek [2010]). Pro správné využití metody je třeba ale rozumět i tomu, jak zvolit jednotlivé parametry. V následujícím textu se tedy detailně zaměříme na druhou fázi analýzy, tedy na vliv volby parametrů analýzy na její výsledky a celkovou interpretaci.

Metoda CATA je 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 příbězích³, který se zaměřuje na studium životních zkušeností různých typů aktérů v období vymezeném lety 1968–1989. Na projektu se vedle sociologů podílejí i lingvisté a sociolingvisté.

Celkově byly v rámci projektu provedeny rozhovory se čtyřmi skupinami aktérů: 1. disidenty, 2. komunistickými funkcionáři, 3. dělníky, 4. inteligencí. Tabulka 1 shrnuje základní charakteristiky provedených rozhovorů. V následujícím textu srovnáváme výsledky analýzy za první dvě skupiny.

Reliabilita – vizualizace dat

Statistické zpracování matice spoluvýskytů je krokem metody, která umožňuje data smysluplně interpretovat. Nese však s sebou určitou míru zjednodušení, která vychází z toho, že se snažíme strukturu sémantického prostoru textových korpusů reprezentovat jako vzdálenosti v dvojrozměrném (případně trojrozměrném) eukleidovském prostoru, který umožní nahlédnout na strukturu původního mnohorozměrného prostoru. Vzniká zde několik problémů a otázek, které ovlivňují uspořádání této struktury, a tedy její interpretaci.

Obrázek 1 je ukázkou výstupu analýzy. Jedná se o vizualizaci vzdáleností lemmat v textovém korpusu, velikost bodu je ilustrací frekvence výskytu dané

3 GAP404/10/0790 – Instituce v životních příbězích. Víceúrovňová srovnávací analýza biografických vyprávění tří skupin aktérů české společnosti 2. poloviny 20. století (2010–2012, GAO/GA).

Tabulka 1. Vlastnosti analyzovaných textových korpusů

	Celkový počet narátorů	Průměrná délka rozhovoru (počet slov)	Průměrná délka odpovědi ¹	Průměrný věk narátorů	Podíl mužů
Disidenti	66	19 224	482	60,6	85 %
Funkcionáři	32	22 086	252	66,3	97 %
Dělníci	56	12 989	126	64,9	54 %
Inteligence	56	20 481	267	64,2	64 %

Zdroj: Projekt Instituce v životních příbězích

Poznámka: Průměrná délka odpovědi ukazuje, jak dlouhé odpovědi dávají nárátoři na otázky tazatelů. Jsou-li odpovědi delší, ukazuje to, že nárátoři dokážou sami vést vyprávění a nepotřebují příliš mnoho otázek. Krátké odpovědi naopak ukazují, že nárátoři nedokážou o tématu sami vyprávět a potřebují větší podporu tazatele.

Obrázek 1. Rozložení bodů v grafu MDS podle frekvence (disidenti)

Zdroj: Projekt Instituce v životních příbězích.

Legenda: Vizualizace (MDS) umísťuje slova s vyšší relativní frekvencí spoluvýskytu blíže k sobě a slova s nižší relativní frekvencí spoluvýskytu dále od sebe. Velikost bodu je dána frekvencí výskytu daného slova v textovém korpusu.

ho slova. Tento graf je třeba interpretovat se znalostí toho, s jakou logikou jsou zde slova/lemmata uspořádána.

Pro vizualizaci je použito mnohorozměrné škálování, konkrétně procedura PROXSCAL v programu SPSS. Mnohorozměrné škálování⁴ je podle Norušis [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 podobnosti (tj. *původní vzdálenosti* – pozn. autora) těchto bodů“ [Borg, Groenen 2005: 170].

Data ve formátu matice podobnosti slov získáme z programu COOA [COOA 2009] ve výše naznačené druhé fázi analýzy. V programu COOA je třeba zvolit i dva základní parametry: *koeficient podobnosti* a *velikost kontextové jednotky*.

Koeficient podobnosti

Koeficient podobnosti je variantou vzdálenosti či podobnosti používanou pro dvojice binárních proměnných.⁵ Poměřuje počet společných výskytů dvou prvků – v našem případě slov v kontextové jednotce – s celkovým počtem výskytů obou těchto prvků – slov v textu. Výpočet koeficientu podobnosti naznačuje Tabulka 2.

Chen, Härdle a Unwin [Chen, Härdle, Unwin 2008: 318] uvádějí 17 různých koeficientů podobnosti. Pro analýzu výskytu slov je vhodných pouze 8 z nich, které nepracují s proměnnou *d*.⁶ Tyto koeficienty, shrnuté v Tabulce 3, se liší 1. *rychlostí růstu a ne/linearitou funkce* a 2. *oborem hodnot*, a tedy mírou, s jakou zohledňují *nerovnoměrnost frekvence výskytu* dvou slov, v textovém souboru.

Koeficienty tak můžeme rozdělit na lineárně rostoucí (Simpson, Czekanowski-Sørensen-Dice, Braun-Blanquet, Kulczynski, Ochiai) a nelineárně rostoucí (Jaccard, Sokal-Sneath-Annenberg, Mountford).⁷ Druhou vlastností je různý obor hodnot koeficientů, který je patrný zejména u slov s různou

4 Anglicky multidimensional scaling, v literatuře je často používána zkratka MDS.

5 Pro kardinální data je nejčastěji používána eukleidovská vzdálenost či jiné varianty vzdálenosti, případně podobnosti. Jako míra podobnosti může sloužit i korelační koeficient.

6 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 díky vysoké míře společného ne-výskytu byly umístěny velmi blízko sebe.

7 Pro zde prezentovaný typ analýzy se jako jednoznačně nevhodný ukazuje Mountfordův koeficient, který roste velmi rychle až u velmi vysokého počtu spoluvýskytů.

Tabulka 2. Modelová kontingenční tabulka pro výpočet koeficientů vzdálenosti

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

Zdroj: Převzato z Chen, Härdle, Unwin [2008: 318], upraveno autorem.

Tabulka 3. Vzorce výpočtu jednotlivých koeficientů vzdálenosti

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

Zdroj: Převzato z Chen, Härdle, Unwin [2008: 318].

frekvencí výskytu. Tato vlastnost koeficientů je klíčová, protože zásadním způsobem působí na organizaci grafu.

Důsledky pro uspořádání grafů a jejich interpretaci

Volba koeficientu do jisté míry závisí na tom, jak chápeme vztah dvou slov s odlišnou frekvencí výskytu. Vztah těchto dvou slov je – na rozdíl od vzdálenosti v grafu – asymetrický. Jako příklad si vezmeme slova ČLOVĚK (6529 výskytů v korpusu disidentů) a CÍRKEV (269 výskytů v korpusu disidentů). Maximální

počet společných výskytů těchto dvou slov je 269 – otázka je, zda tato maximální frekvence výskytu ukazuje na silnou, nebo na slabou vazbu mezi těmito dvěma slovy. Z pohledu slova CÍRKEV jistě silnou – význam a použití slova CÍRKEV je silně svázáno se slovem ČLOVĚK, které se vždy vyskytuje v jeho kontextu. Pro slovo ČLOVĚK je ale naopak slovo CÍRKEV jen okrajovou součástí kontextu.

První vztah – tedy vztah méně frekventovaného slova k více frekventovanému – zohledňuje Simpsonův koeficient, který vztahuje počet společných výskytů k celkovému počtu výskytů méně frekventovaného slova.⁸ Druhý vztah pak reprezentuje koeficient Braun-Blanqueův.⁹ Ostatní koeficienty (zejm. Kulczyński, Ochiai a Jaccard) pak hledají střední hodnotu mezi oběma vztahy, a proto je považujeme za nejvhodnější. Nejčastěji užívaný a v literatuře uváděný koeficient pro tento typ analýzy je koeficient Jaccardův [Hájek 2010; Borg, Groenen 2005: 127; Mohammad, Hirst 2005].

Tyto skutečnosti mají pro logiku a interpretaci grafů dvojí význam. Jejich důsledkem je to, co je viditelné na Obrázku 1, a sice že frekvence slova je zásadní pro uspořádání grafu.¹⁰ Méně frekventovaná slova jsou vytlačována na okraj, neboť mají slabý vztah ke slovům více frekventovaným (při užití Braun-Blanqueova koeficientu) nebo proto, že více frekventovaná slova mají vazbu na více různých slov (při užití Simpsonova koeficientu).

Vztah mezi frekvencí slova a jeho vzdáleností od středu však není definiční.¹¹ Obvykle jsou tyto jevy způsobeny gramatickými vazbami mezi slovy (např. často užívané sousloví člen strany posouvá v korpusu disidentů slovo ČLEN blízko středu díky silné vazbě na slovo STRANA) nebo kontextovou vázaností slova (slovo RODINA má v korpusu disidentů vysokou frekvenci, ale je silně kontextově vázáno na vyprávění o dětství a rodině).

Druhým důsledkem asymetrie vztahu mezi slovy je nutná existence vyšší chybovosti způsobené tím, že blízkost dvou slov v grafu nemusí být způsobena pouze častým vzájemným společným výskytem, ale také společnou vazbou na třetí, více frekventované slovo.¹²

8 V tomto případě by tedy Simpsonův koeficient byl roven 1.

9 Braun-Blanqueův koeficient by byl roven 0,04.

10 Výskyt slov v textu podle frekvence je nerovnoměrný, na což poukazují tzv. Zipfovy zákony [podle Manning, Schütze 1999: 24 an.]. George Kingsley Zipf zkoumal ve 30. letech statistické rozložení slov v textu a všiml si určitých pravidelností. Jednou z nich – a pro nás nejdůležitější – je zákon distribuce slov podle jejich frekvence, podle něhož je součin frekvence výskytu slov v textu a jejich pořadí mezi všemi slovy podle frekvence přibližně konstantní. To jinými slovy znamená, že v náhodně vybraném textu se vyskytuje několik velmi frekventovaných slov a velké množství slov málo frekventovaných.

11 Tento vztah lze vyjádřit Spearmanovým korelačním koeficientem, který je pro analyzovaný slovník v korpusu disidentů roven -0,87 a pro slovník korpusu funkcionářů -0,88.

12 Tento jev lze ilustrovat následujícím příkladem. Slovo ČLOVĚK v učebnici vývoje lidského druhu má tendenci vytvářet silné vazby na přívlastky ZRUČNÝ, VZPŘÍMENÝ nebo

Obrázek 2. Naznačení modulární interpretace (disidenti)

Zdroj: Projekt Instituce v životních příbězích.

Legenda: Naznačení modulární interpretace (oblasti vyprávění)

- A. Jádro vyprávění
- B. Vyprávění o soukromém a rodinném životě
- C. Vyprávění o životě v opozici
- D. Vyprávění o politice a ideologii

To vede k nutnosti používat při čtení grafu tzv. *modulární interpretaci* [srov. Borg, Groenen 2005: 81]. V tomto způsobu uspořádání grafu rozdělíme prostor na kruhové výseče, v nichž mají body určité podobné vlastnosti. Tato logika vychází z výše uvedeného: slova se do grafu primárně řadí podle frekvence, ale další uspořádání již probíhá podle významových vazeb.

ROZUMNÝ. Jelikož však učebnice velmi pravděpodobně bude uspořádána do kapitol, kde se každá kapitola 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.

Modulární interpretaci ukazuje Obrázek 2. Zde sémantický prostor vyprávění disidentů rozdělujeme na jádro, které tvoří základní rámec vyprávění a obsahuje slova s nejvyšší frekvencí, popř. slova, která jsou propojena s různými kontexty. U okraje pak nacházíme slova s nižší frekvencí, která rozdělujeme do tří oblastí (výsečí): 1. soukromý a rodinný život, 2. život v opozici, 3. opoziční ideologie. Tyto oblasti pak reflektuje i uspořádání jádra vyprávění, kdy slova ČLEN a KSČ náleží blíže k oblasti soukromé, STÁTNÍ BEZPEČNOST k oblasti života v opozici a CHARTA k oblasti opoziční ideologie.

Určení hranic jádra a kruhových výsečí je v tomto případě čistě analytické a je součástí interpretace dat. Struktura sémantického prostoru obvykle nevytváří jasně vymezené podoblasti – znamenalo by to, že narátoři využívají zcela odlišnou slovní zásobu pro vyprávění o jednotlivých událostech/oblastech vyprávění a tyto části striktně odděluje, což se ovšem v praxi obvykle neděje. Určení hranic proto není založeno na výpočetní proceduře, která hranice oblastí stanoví objektivně. Výzkumník zde využívá toho, že mnohorozměrné škálování umožňuje nahlédnout na sémantický prostor komplexně a sledovat vztahy mezi jednotlivými součástmi vyprávění: těmi, které se vyskytují blízko, a těmi, které jsou naopak od sebe vzdáleny.

Dodejme, že s maticí vzdáleností lze pracovat i s užitím jiných statistických procedur, které umožňují akcentovat jiné vztahy mezi slovy. Pro stanovení interpretačních os textového korpusu lze užít faktorovou analýzu. Výstupem takové analýzy může být určení několika důležitých dimenzí vyprávění, které tvoří klíčové sémantické opozice.¹³

Stejně tak lze jednotlivé oblasti vyprávění – jak bylo naznačeno – stanovit i početně s pomocí hierarchické clusterové analýzy. Hierarchický graf pro korpus disidentů je možné najít v příloze. Obecně platí, že výsledek clusterové analýzy se nemusí plně shodovat se zobrazením pomocí mnohorozměrného škálování, neboť clusterová analýza ignoruje komplexitu vztahů mezi slovy včetně společných vazeb na třetí slovo, které jsou při zobrazení mnohorozměrným škálováním zohledňovány.¹⁴ Porovnání výsledků těchto tří statistických procedur může umožnit komplexnější pochopení a interpretaci dat.

13 Není zde prostor pro hlubší představení užití faktorové analýzy na matici kolokací výskytu slov v textu. Jako ilustrativní příklad můžeme uvést základní dimenze vyprávění, které identifikovala faktorová analýza v korpusu disidentů. Počet dimenzí byl arbitrárně omezen na pět pro účely prezentace. Jsou to: 1. soukromý – veřejný (rodina proti ostatním částem vyprávění), 2. oficiální – neoficiální (život v ilegality proti životu legálnímu), 3. politika – kultura, 4. bezpečí – nebezpečí, 5. před revolucí – po revoluci. Tyto dimenze základní binární opozice charakterizují text a jsou spojeny s (ne)výskytem určitých slov.

14 Hierarchická clusterová analýza řeší odlišným způsobem problém slov, která se vyskytují ve více kontextech (přičítá slova do jedné nejbližší skupiny), než mnohorozměrné škálování (které zohledňuje pozici slova i vůči ostatním kontextům). Tak například v clusterové analýze má slovo HAVEL blízko slovu DIVADLO, i když mnohorozměrné škálování jej umísťuje zejména do kontextu opoziční a porevoluční politiky.

Hodnocení kvality modelu

Zobrazení mnohorozměrného prostoru do prostoru méněrozměrného, které je podstatou mnohorozměrného škálování, s sebou nese určitou míru zkreslení.¹⁵ Tuto míru je třeba zhodnotit a promítnout do interpretace výsledné vizualizace dat. Míra zkreslení je vyjádřena tzv. hrubým stresem, vypočteným podle vzorce:

$$\sigma_r = \sum_{(ij)} [f(p_{ij}) - d_{ij}(X)]^2 \quad (5.1)$$

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:

$$\sigma_1^2 = \frac{\sum [f(p_{ij}) - d_{ij}(X)]^2}{\sum d_{ij}(X)^2} \quad (5.2)$$

Z této míry je pak dále odvozen Kruskalův Stress-1 jako její odmocnina. Tato míra bývá nejčastěji používána jako kritérium pro hodnocení kvality modelu. V literatuře bývají 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 by se měl pohybovat kolem 0,05. Borg a Groenen [Borg, Groenen 2005: 54] však nepovažují za vhodné užívat Kruskalova kritéria jako jediné měřítko kvality modelu. Autoři 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 měř 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 různé typy dat. Druhým Borgovým a Groenenovým doporučením je zabývat se stresem pouze orientačně a spíše zkoumat, zda se má zobrazená struktura tendenci v různých zobrazeních proměňovat, či je naopak stabilní.

Volba nejlepšího modelu by měla být závislá na vývoji velikost stresu se zvyšováním počtu dimenzí. To shrnuje Graf 1 (viz následující stranu). Borg a Groenen [2005: 48] doporučují vybírat to řešení, které se nachází ve zlomu

15 Hledání zobrazení mnohorozměrného prostoru je iterativní procedura, kdy se po stanovení určité úvodní konfigurace, která probíhá buď náhodně, či určitými výpočetními metodami, toto proložení iterativně zlepšuje do chvíle, kdy je nalezeno optimum podle určitých kritérií (maximální počet iterací, minimální snížení stresu oproti předchozí iteraci).

Graf 1. Hodnoty Stresu-1 pro řešení s různou dimenzionalitou (Jaccardův koeficient, kontextová jednotka 100 slov)

Zdroj: Projekt Instituce v životních příbězích.

křivky klesající hodnoty stresu v závislosti na počtu dimenzí. To by v našem případě mělo být řešení třídimenzionální.

Důvodem, proč volíme řešení dvoudimenzionální, je limitovaná možnost zobrazení, která značně ztěžuje interpretaci. Řešením situace by mohlo být vyřazení nestabilních slov z analýzy s vědomím určité redukce informací o struktuře vyprávění, případně analýza trojrozměrného zobrazení, která je jistě proveditelná, ale velice obtížné prezentabilní v odborném textu.

Analýza stability struktury zobrazení

Volbu dvoudimenzionálního řešení s vědomím větší chybovosti zobrazení dále podpoříme analýzou stability struktury různých zobrazení. Analýza zohledňuje stabilitu zobrazení s týmiž parametry, ale i stabilitu z hlediska změny základních parametrů, tj. koeficientu podobnosti (Jaccardův, Ochiaiův a Kulczynského koeficient)¹⁶ a kontextové jednotky (50, 100 a 150 slov).¹⁷

¹⁶ 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.

¹⁷ Přístupy k volbě délky kontextové jednotky shrnuje Hájek [2010 : 27]. Důležitý je fakt, že krátké kontextové jednotky (v řádu jednotek slov) jsou vhodné spíše pro sledování obecně jazykových jevů, pro tematickou analýzu jsou pak vhodnější delší kontextové jednotky. Preferovaná varianta 100 slov vychází z těchto úvah a také z předpokladu, že 100 slov přibližně odpovídá délce jednoho kratšího odstavce, a tedy délce smysluplné vý-

Rozbor proměn konfigurací bodů tedy byl proveden pro 9 skupin datových matic. Na každou z nich byla aplikována analýza mnohorozměrného škálování (PROXSCAL v programu SPSS) celkem pětkrát. Počátek analýzy byl stanoven metodou náhodných počátků v počtu 1000.¹⁸ Výsledkem tak bylo 5 různých „dobrých“ konfigurací pro každou skupinu.

V těchto 45 výstupech byly dále analyzovány posuny konfigurací jako 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. Velikost hodnoty průměrné vzdálenosti bodu od svého centroidu je mírou tendence bodu měnit svou pozici v grafu. Tato míra by měla být vztažena k měřítku grafu, které kvůli přehlednosti není ve vizualizacích uvedeno, ale pohybuje se přibližně od -1 do $+1$ v obou dimenzích.

Graf 2 (viz následující stranu) zobrazuje výsledky této analýzy pro konfiguraci disidentů. Každý z bodů má tři vlastnosti. Vodorovná osa zobrazuje míru variability bodu způsobenou změnou velikosti kontextové jednotky (50, 100, 150 slov).¹⁹ Svislá osa pak zobrazuje míru variability bodu způsobenou změnou koeficientu podobnosti (Jaccard, Ochiai, Kulczynski). Plocha bodu pak určuje míru nevysvětlené variability daného bodu, tj. variabilitu při stejně zvolených parametrech (u vybraných bodů uvedena i číslem).

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 0,2. 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 převážet jedna z těchto vazeb, která umístí slovo do jednoho z jeho kontextů.

Graf 3 (viz stranu 21) ukazuje analogickou situaci pro korpus funkcionářů. Vliv změn kontextové jednotky je v tomto korpusu u některých slov vyšší (např. u slova HAVEL se pohybuje v průměru kolem 0,6). Souvisí to s menší velikostí korpusu funkcionářů, která významně zkreslila graf pro kontextovou jednotku

povědi. Ostatní hodnoty kontextové jednotky zde byly zkoumány pro zaznamenání změn struktury způsobených změnou tohoto parametru. Dalším aspektům volby délky kontextové jednotky se budeme věnovat v diskuzi.

18 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).

19 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 variant výstupů. Míra určuje, zda daný bod má soustavnou tendenci měnit svou pozici s volbou jiné velikosti kontextové jednotky. Analogicky je vypočtena táž míra pro zobrazení s různým koeficientem podobnosti.

Graf 2. Variabilita zobrazení bodů při změně kontextové jednotky, míry podobnosti a nevysvětlená variabilita (disidenti)

Zdroj: Projekt Instituce v životních příbězích.

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

Zdroj: Projekt Institute v životních příbězích.

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

Zdroj: Projekt Instituce v životních příbězích.

Legenda: Vliv změny kontextové jednotky

→ (100 → 150 slov)

→ (100 → 50 slov)

50 slov. Zde se zvyšuje počet dvojic slov s nulovým počtem spoluvýskytů, což komplikuje možnost zobrazení.

Konkrétní realizaci této nestability ukazuje Obrázek 3. Vidíme, že některá slova se pohybují velmi výrazně. Důvody těchto výrazných pohybů lze však obvykle najít ve dvojí roli daného slova. Je to patrné u slova CÍRKEV, které se vyskytuje ve dvou dominantních kontextech: v kontextu polistopadové politiky a v kontextu dětství a dospívání narátorů. Podobně slovo SVOBODA se táhne obdobím disidentského boje i obdobím porevolučním, tedy obdobím jejího dosažení.

Obrázek 4 pak ukazuje proměny struktury grafu při změně kontextové jednotky ze 100 na 150 slov. Proměna struktury při snížení počtu slov v kontextové jednotce na 50 není naznačena, neboť je natolik zásadní, že by její zachycení ve vizualizaci bylo zcela nepřehledné. U korpusu funkcionářů se tak ukazuje, že volba kontextové jednotky o délce 50 slov zcela mění strukturu, která při zvýšení kontextové jednotky ze 100 na 150 slov už zůstává zachována.

Obrázek 4. Zobrazení změn pozic bodů se změnou kontextové jednotky (funkcionáři)

Zdroj: Projekt Instituce v životních příbězích.

Legenda: Vliv změny kontextové jednotky

→ (100 → 150 slov)

I v korpusu funkcionářů se vyskytují některá velmi nestabilní slova. Ukazuje se například, že slovo VOLBA se nachází v politickoideologické oblasti vyprávění, kde však hraje dvojí roli. Je spojeno jednak s volbami jako klíčovou institucí porevolučního společenského a politického vývoje, druhou rolí tohoto slova je pak spojení s vnitřním ideologickým životem komunistické strany.

Podobně zajímavá je oscilace slov DIVADLO a AUTO mezi dvěma póly soukromé a rodinné oblasti vyprávění, mezi rolemi těchto dvou slov v popisu mládí a v popisu rodinného života v souvislosti s funkcionářskou kariérou.

Nacházíme zde ale také příklad deformace struktury spojené s dvojí rolí slova STUDENT. Toto slovo se vyskytuje ve dvojím kontextu. První kontext, vyprávění o studentském životě narátora, je do značné míry potlačen druhým kontextem, rolí studentů v sametové revoluci. Vysoká vzdálenost slov STUDENT a VYSOKÁ ŠKOLA naznačuje, že komunističtí funkcionáři o sobě příliš často nehovořili jako o studentech, což souvisí s tím, že studovali stranické školy často při práci, a neprošli tak studentským obdobím života v tom smyslu, jaký je mnohem více patrný u disidentů.

Poměrně zajímavá situace nastává u slova MATKA, které se s prodloužením kontextové jednotky výrazně posouvá směrem do středu před slovo OTEC. Tento posun nemůže být dán frekvencí daného slova, která je nutně pořád stejná. Ukazuje se tak, že slovo se může dostat do středu grafu i díky silnější vazbě na jiná, důležitá slova.

Diskuze a závěr

Cílem tohoto textu bylo diskutovat kritéria kvality zobrazení výstupů mnoho-rozměrného škálování při používání počítačové textové analýzy v programu COOA. Sekundárním cílem pak bylo zhodnotit, jaký vliv na kvalitu modelu má volba parametrů (koeficientu podobnosti a velikosti kontextové jednotky).

Analýza ukazuje, že pro postižení struktury vyprávění touto metodou lze jen těžko stanovit objektivní kritéria kvality modelu a volby parametrů. Je spíše třeba s těmito kritérii experimentovat a analyzovat změny zobrazení, které samy o sobě přispívají k pochopení dané struktury.

Zásadním zdrojem chyby zobrazení je problém asymetrie vztahu mezi více frekventovaným slovem a slovem méně frekventovaným, který vede k nutnosti interpretovat graf nikoli jako mapu, kde vzájemně si blízká slova mají společné vlastnosti, ale jako oblasti (kruhové výseče), ve kterých se nacházejí slova charakterizující oblasti vyprávění.

Modely prezentované v textu neodpovídají obvykle užívaným kritériím kvality (tj. Kruskalovu kritériu, kdy stres ve výši 0,2 je hranice přijatelného modelu). Výzkumník má v tu chvíli několik možností, jak chybu redukovat: 1. vyřadit problematická slova a tím redukovat informaci v modelu obsaženou, 2. zvýšit dimenzionalitu modelu, což ovšem vede ke snížení možností snadné prezentace v odborném textu.

Třetí cestou je analyzovat zdroje této chyby a v interpretaci je zohlednit. Otázka je, zda je samotná struktura celkově stabilní a jen malá část prvků má tendenci výrazně měnit svou pozici, nebo je struktura celkově nestabilní. První případ je pro interpretaci použitelný. U nestabilních prvků je pak třeba rozhodnout, zda tuto nestabilitu dokážeme smysluplně interpretovat (jako např. u dvojího kontextu slova CÍRKEV ve vyprávění disidentů), či zda je takové povahy, že slovo z analýzy raději vyloučíme. Analýza ukázala, že v některých případech může znalost chování slova v různých zobrazeních sloužit jako interpretační nástroj i přesto, že je zdrojem chyby.

Druhá otázka, ke které je třeba se vyjádřit, je volba parametrů výpočtu matice vzdáleností v programu COOA. Analýza ukazuje, že volba koeficientu podobnosti by se měla omezit na koeficient Jaccardův, Kulczynského a Ochiaiův, které se pokoušejí eliminovat asymetrii vztahu mezi slovy s různou frekvencí. Rozdíly mezi těmito koeficienty nejsou pro výslednou konfiguraci již zásadně významné.

Volba kontextové jednotky již ve výsledné konfiguraci hraje větší roli. Hájek [2010: 27] poukazuje na to, že velikost a povaha kontextové jednotky závisí na

typu textu, způsobu jeho vytvoření a uspořádání. Zde prezentovaná analýza ukazuje, že optimální kontextová jednotka závisí i na frekvenci analyzovaných slov. Pro slova s nižší frekvencí výskytu (okolo 100) je krátká kontextová jednotka (okolo 50 slov) nevhodná, protože tato slova nejsou dostatečně pevně ve struktuře zakotvena a mají tendenci být vychylována náhodnými vazbami. Příliš velká kontextová jednotka na druhou stranu vede k tomu, že významová struktura může být méně zřetelná.²⁰

Z výše uvedeného plyne, že počítačovou textovou analýzu nelze aplikovat mechanicky podle předem daných pravidel. Konstrukce slovníku i volba parametrů významně závisí na povaze analyzovaných dat, na tom, zda jde o text psaný či přepis mluveného slova, na charakteristikách autora či mluvčího, případně na povaze komunikační situace a částečně také na velikosti textového korpusu a frekvenci výskytu analyzovaných lemmat.

Tyto závěry ukazují, že otevřenost metody znamená na jednu stranu větší šíři možností pro práci s daty, vede však k také nutnosti detailněji data poznat a parametry analýzy volit adekvátně tomuto poznání. Je třeba analýzu opakovat a korigovat ji úpravou slovníku a volbou délky kontextové jednotky. Otevřená metoda tak klade větší nárok na komplexní práci s daty a vyžaduje jednotlivé fáze výzkumu průběžně opakovat a upravovat parametry pro dosažení komplexního obrazu o analyzovaných datech.

VÁCLAV ČEPELÁK je doktorandem na Katedře sociologie Institutu sociologických studií Fakulty sociálních věd Univerzity Karlovy, kde se zabývá aplikací textových analýz v sociologii. Konkrétně se zaměřuje na studium ekonomického jednání jednotlivců a domácností a jeho chápání ve veřejném a mediálním diskurzu. Mimo studium pracuje v marketingovém výzkumu.

Literatura

- Alexa, Melina. 1997. „Computer-assisted text analysis in the social sciences.“ ZUMA Arbeitsbericht 97/07. Mannheim: Zuma.
- Baker, Paul. 2006. *Using corpora in discourse analysis*. New York: Continuum.
- Berelson, Bernard. 1952. *Content Analysis in Communication Research*. Glencoe, Illinois: The Free Press.
- Borg, Ingwer, Patrick J. Groenen. 2005. *Modern multidimensional scaling: theory and applications*. 2nd ed. New York: Springer.
- Chen, Chun-houh; Wolfgang Härdle; Antony Unwin. 2008. *Handbook of data visualization*. Berlin: Springer.
- COOA. Co-occurrence Analysis Software [počítačový software]. 2009. Praha: Fakulta sociálních věd UK [cit. 3. 11. 2012]. Dostupné z: http://publication.fsv.cuni.cz/attachments/471_setup_COOA.exe.

20 Je vidět, že délka kontextové jednotky by si zasloužila detailnější analýzu, zaměřenou na změny struktury zobrazení při významnějším zvětšování kontextové jednotky. Taková analýza je však nad rámec tohoto textu.

- Cox, Trevor F., Michael A. Cox. 2001. *Multidimensional scaling*. 2nd ed. Boca Raton: Chapman.
- Gadamer, Hans-Georg. 2010. *Pravda a metoda I: nárys filosofické hermeneutiky*. Vyd. 1. Praha: Triáda
- Glaser, Barney G., Anselm L. Strauss. 1973. *The discovery of grounded theory: strategies for qualitative research*. 1st pbk. ed. Chicago: Aldine Pub.
- Hájek, Martin. 2010. „Počítačová textová analýza metodou sledování spoluvýskytů slov.“ *Data a výzkum-SDA info*, 4(1): 19–37.
- Hebák, Petr. 2005. *Vícerozměrné statistické metody*. Vyd. 1. Praha: Informatorium.
- Manning, Christopher D., Heinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge: MIT Press.
- Mohammad, Saif, Graeme Hirst. 2005. *Distributional Measures as Proxies for Semantic Relatedness* [online]. Toronto : University of Toronto [cit. 3. 11. 2012]. Dostupné z: <ftp://ftp.cs.toronto.edu/pub/gh/Mohammad+Hirst-2005.pdf>.
- Norušis, Marija J. 2005. *SPSS 14.0 advanced statistical procedures companion*. Upper Saddle River : Prentice Hall : SPSS.
- Scott, John P. 2000. *Social Network Analysis: a handbook*. Thousand Oaks: SAGE Publications.

Příloha: Hierarchická clusterová analýza matice vzdáleností disidentů