TEMATICKÁ KONCENTRACE TEXTU V ČEŠTINĚ

Radek Čech

ÚSTAV FORMÁLNÍ
A APLIKOVANÉ LINGVISTIKY

STUDIES IN COMPUTATIONAL AND THEORETICAL LINGUISTICS

Radek Čech

TEMATICKÁ KONCENTRACE TEXTU V ČEŠTINĚ

Published by the Institute of Formal and Applied Linguistics
as the 15th publication in the series
Studies in Computational and Theoretical Linguistics.

Printed by Printo, spol. s r. o.

ISBN 978-80-88132-00-4

Obsah

Předmluva

Téma patří mezi základní vlastnosti textu. Samozřejmě existují texty, které žádné téma nemají – například některé básně Ch. Morgensterna či díla dadaistů –, ale u naprosté většiny textů se otázky typu „O čem ten článek je?“ nebo „O čem ten člověk mluvil?“ jeví jako adekvátní. Podobně adekvátní jsou většinou i otázky typu „Držel se autor daného tématu?“ nebo „Jak silnou roli má v textu probírané téma?“. Cílem této knihy je představit a důkladně prozkoumat jednu z kvantitativnělingvistických metod, jejímž prostřednictvím je možné na výše uvedené otázky odpovědět, a to způsobem, který v co největší míře eliminuje vliv subjektivity.

V této knize navazuji na tradici české kvantitativní textologie, která je spojena se jménem Marie Těšitelové a jejích spolupracovnic a spolupracovníků. Zejména se však opírám o teoretická a metodologická východiska toho směru kvantitativní lingvistiky, o jehož rozvoj se zasloužil především Gabriel Altmann a který dnes reprezentují badatelé sdružení v *Mezinárodní asociaci kvantitativní lingvistiky* (*International Quantitative Linguistics Association*). V českém prostředí se mohou čtenáři s tímto přístupem už více než 30 let seznamovat v pracích Luďka Hřebíčka, jenž je asi jediným představitelem tohoto směru v domácí kvantitativní textologii.

Kniha by nevznikla bez pomoci mých přátel a spolupracovníků. Především děkuji Gabrielu Altmannovi za všechny rady, inspirace, kritiku i za stovky (asi to už budou i tisíce) emailů, které jsme si za několik let vyměnili; Jánu Mačutkovi za trpělivost, kterou projevuje už několik let jako statistik s matematicky nevzdělaným lingvistou; Jaroslavu Davidovi a Janě Davidové Glogarové za pečlivé přečtení rukopisu a za všechny kritické připomínky a poznámky; Vladimíru Matlachovi za ochotu a rychlost, se kterou vyřešil všechny problémy související s automatickým zpracováním textů.

1
Úvod

Text, ať psaný či mluvený, je produkt lidského chování, který se vyznačuje jednak určitými pravidelnostmi (zejména těmi, které jsou způsobeny gramatikou daného jazyka), jednak obrovskou variabilitou: například existuje jak bezpočet cílů, proč píšeme či mluvíme (informování; vědomé lhaní; rozkazování; mluvení „o ničem“, jehož smyslem je sociální interakce; „zabíjení“ času atd.), tak i bezpočet způsobů, jak tyto cíle realizovat prostřednictvím přirozeného jazyka. Pokud bychom třeba dali dvěma lidem napsat jednu stranu textu na určité téma, je jen minimální pravděpodobnost, že se v obou textech vyskytnou identické věty, natož pak identické odstavce. Navzdory obrovské variabilitě možností se však v textech zároveň projevují i takové pravidelnosti, které nejsou způsobeny gramatikou a které lze interpretovat jako důsledek obecných principů, jež mají rozhodující vliv na charakter řečového chování; jde například o Zipfův princip nejmenšího úsilí (Zipf 1949) či samoregulaci v synergetickém modelu jazyka (Köhler 1986, 2005). Za všechny pravidelnosti tohoto druhu uveďme třeba známý vztah mezi frekvencí a délkou slova (čím je slovo frekventovanější, tím je kratší), mezi frekvencí a polysémií (čím je slovo frekventovanější, tím má více významů) či vztah mezi velikostí inventáře fonémů daného jazyka a průměrnou délkou slov (čím je inventář fonémů větší, tím jsou slova v průměru kratší). Důležité je přitom zejména to, že tyto pravidelnosti, jež mají stochastickou povahu, je možné nejen popsat, ale i matematicky modelovat a, v nejlepším případě, predikovat jejich chování v textu či jazyce.

Cílem této knihy je systematická analýza tzv. tematické koncentrace textu. Tato analýza je založena na následujících předpokladech:

(a) v různých textech se autor na dané téma či témata může zaměřovat s různou intenzitou;
(b) lze identifikovat jazykové jednotky, které je možné chápat jako nositele určitého tématu či témat;
(c) míru zaměření se na dané téma či témata lze detekovat analýzou frekvenčních charakteristik textu;
(d) míra zaměření se na dané téma či témata není náhodná, tj. přepokládá se její systematické chování vzhledem jak k jiným vlastnostem textu, tak k faktorům pragmatickým.

Předpoklad (a) je zřejmě nejméně problematický – asi každý má zkušenost jak s textem, kde se „přeskakuje od tématu k tématu“, tak s textem, ve kterém se autor důsledně daného tématu drží.

V případě předpokladu (b) se dostáváme k problematice definice jazykových jednotek. Vzhledem k tomu, že se jedná o problém netriviální, je mu níže věnována jedna celá kapitola (kap. 4). Ale už zde bych rád zdůraznil, že přístup prezentovaný v této knize obecně nepředpokládá, že by jazyková data (jakéhokoliv druhu) byla nějak dopředu „dána" a že by těmto datům (s větší či menší přesností) odpovídaly naše pojmy (např. pojem *slovo*). Jazykové jednotky jsou chápany jako naše konstrukce, jejichž prostřednictvím se snažíme „manipulovat" s realitou, v tomto případě za účelem analýzy tematických charakteristik textu.

K předpokladu (c) je třeba jasně říct, že jsem si dobře vědom omezení, které jeho přijetí s sebou nese. Jde zejména o to, že nepochybně existují situační kontexty, díky nimž se určité téma vůbec nemusí projevit ve frekvenčních charakteristikách sledovaných jednotek; dále například můžeme mít relativně dlouhý text o smrti, ale díky pestré paletě metaforických prostředků se samotné slovo ‚smrt' (v žádném tvaru) vůbec nemusí objevit, případně se objeví s minimální frekvencí – to, že účastníci komunikace dovedou odvodit hlavní téma textu, je zpravidla dáno znalostí významu metafor, ale i kontextu (v nejširším slova smyslu). V případě předpokladu (c) je zde prezentovaný přístup nepochybně redukcionistický, jako jsou ovšem redukcionistické analýzy jakéhokoliv druhu.

Nejproblematičtějším, ale zároveň teoreticky nejzajímavějším je předpoklad (d), který se stal hlavní motivací vzniku této knihy. Z hlediska poznání toho, jak funguje text, jsou totiž důležité nikoliv jednotlivé vlastnosti textu (těch je de facto nekonečně mnoho), ale vzájemné vztahy mezi nimi, případně vztahy mezi nimi a faktory pragmatickými. Vzhledem k tomu, že však doposud nebyla vytvořena žádná teorie textu – ve smyslu souboru tvrzení, z nichž je možné odvodit empiricky testovatelné hypotézy – je velmi obtížné predikovat, jak se bude tematická koncentrace projevovat vzhledem k jiným textovým charakteristikám. Na druhou stranu již byly vytvořeny modely některých vlastností textu (např. slovního bohatství), takže je možné se pokusit dedukovat, jaký bude vztah mezi tematickou koncentrací a těmito vlastnostmi, a následně tyto dedukce experimentálně ověřit. Podobně lze testovat rozdíly mezi pragmatickými vlivy, na základě kterých dochází k různým klasifikacím textů (např. textové typy, žánry atp.), a tematickou koncentrací.

Analýzy prezentované v této monografii navazují na výzkum týkající se problematiky tematické koncentrace textu, na kterém jsem se doposud podílel (srov. Čech et al. 2013a,b; Davidová Glogarová, Čech 2013; Davidová Glogarová et al. 2013; Čech 2014a; Čech et al. 2014a; Čech et al. 2015). Jedním z hlavních cílů této knihy je důkladně prozkoumat jednotlivé aspekty tematické koncentrace z hlediska metodologického (Část I). Proto se nejdříve zaměřuji na možné způsoby měření této vlastnosti (kap. 2 a 3) a na vliv volby jazykových jednotek na charakter jejího měření (kap. 4). V následující kapitole analyzuji vliv délky textu na jednotlivé způsoby měření tematické koncentrace (kap. 5). Délka textu je totiž faktorem, který se většina kvantitativních textových analýz obecně snaží eliminovat, protože tato vlastnost má často rozhodující vliv na hodnoty sledovaných indexů (type-token poměr, indexy slovního bohatství,

distribuce hapaxů legomenon atd.). Bez náležité znalosti vlivu délky textu na zvolený způsob měření hrozí možnost nenáležitých interpretací (srov. Čech 2015), proto se mu zde věnuji důkladně. V závěrečné kapitole (kap. 6) Části I je představen způsob měření a testování vývoje tematické koncentrace v textu. Sledování sekvenčního vývoje jakékoliv vlastnosti textu totiž přináší detailnější pohled na danou vlastnost – namísto jedné číselné hodnoty charakterizující text jako celek totiž získáváme uspořádanou množinu více hodnot. Může se tedy stát, že dva texty budou vykazovat stejnou celkovou tematickou koncentraci, ale v každém s ní bude „nakládáno" zcela odlišným způsobem.

V Části II se zaměřuji na vztah tematické koncentrace a dvou vlastností textu, které by s ní z teoretického hlediska měly souviset. Konkrétně jde o slovní bohatství (kap. 7) a distribuci klíčových slov (kap. 8). V případě slovního bohatství vycházím z předpokladu, že by mezi ním a tematickou koncentrací měl být inverzní vztah, tj. čím je větší slovní bohatství, tím by měla být menší tematická koncentrace (měřená daným způsobem) a vice versa. V případě klíčových slov budu sledovat, do jaké míry se obě metody shodují, či liší při určování slov, která reprezentují hlavní témata textu.

V závěrečné části knihy (Část III) budou prezentovány příklady konkrétního využití analýzy tematické koncentrace v textologii. Bude ukázán způsob, jak vytvořit a zkoumat tzv. asociativní tematickou strukturu textu (kap. 9) a jak za pomoci jednotlivých indexů tematické koncentrace klasifikovat texty a testovat rozdíly mezi nimi (kap 10).

Pro analýzu bylo použito 1168 textů různých žánrů, které byly zpracovány prostřednictvím softwaru QUITA (Kubát et al. 2014) (kap. 11). Číslovaný seznam textů a hodnoty jednotlivých indexů tematické koncentrace jsou uvedeny v Příloze 1. Při výběru textů jsem se snažil vybrat takový vzorek, v němž by byly jednak zastoupeny různé žánry, jednak texty o různé délce. V žádném případě jej nelze považovat za vzorek tzv. reprezentativní vzhledem k češtině obecně. Pro naplnění cílů této knihy (viz výše) jej však považuji za dostatečný.

Abych se vyhnul případným nedorozuměním, raději zde připojím terminologickou poznámku: v následujícím textu často používám pojem ‚slovo'. Tímto výrazem je označován buď slovní tvar, přičemž ten je vymezen grafikou, tj. jde o řetězec grafémů mezi mezerami, nebo lemma, tj. základní (slovníkový) tvar slova, který zastupuje všechny jeho další tvary (podrobněji viz kap. 4). Pokud používám tento výraz bez bližšího určení, může zastupovat jak slovní tvar, tak lemma – zpravidla se tak děje na místech, kde vysvětluji principy měření tematické koncentrace, u nichž je volba dané jednotky až druhotná. V ostatních případech používám pro rozlišení termíny ‚slovní tvar' a ‚lemma'.

I METODOLOGICKÉ ASPEKTY MĚŘENÍ TEMATICKÉ KONCETRACE TEXTU

2

Tematická koncentrace textu

Metodu měření tematické koncentrace lze zařadit k typům textových analýz, které jsou obecně označovány jako *obsahové analýzy*, srov. jejich přehled v Krippendorff (2013). Svým charakterem má také blízko ke kvantitativní analýze tzv. klíčových slov[1] (Stubbs 1996; Adolphs 2006; Scott, Tribble, 2006). Jak je však patrné již z názvu této metody, jejím primárním cílem není odhalit hlavní témata textu reprezentovaná danými jazykovými jednotkami (byť i to umožňuje), jako je tomu například u analýzy klíčových slov, ale postihnout to, do jaké míry se autor v textu na dané téma či témata zaměřuje celkově. Vychází se přitom z předpokladu, že míru zaměřenosti je možné kvantifikovat na základě frekvenčních charakteristik textu. Z obecnějšího pohledu se jedná o metodu, jejímž prostřednictvím se dá modelovat určitý aspekt řečového chování.

Metoda analýzy tematické koncentrace vznikla v rámci lingvistického směru, který je označován jako kvantitativní lingvistika (Köhler et al. 2005; Köhler, Altmann 2011). Směru, v němž při jazykových analýzách není rozhodující samotná kvantifikace (jak by se snad z jeho názvu mohlo zdát), ale důraz na možnost statistického testování hypotéz (srov. Čech et al. 2014a, kap. 2). Kvantifikace je v tomto ohledu tedy jen pouhým nástrojem, který takovéto testování umožňuje. Proto i metoda analýzy tematické koncentrace byla od počátku konstruována tak, aby ji bylo možné využít pro statistické testování rozdílů mezi texty, případně skupinami textů (viz níže). Tuto vlastnost je třeba zvlášť zdůraznit, protože používání statistických metod stále nepatří mezi standardy lingvistické práce, srov. slova S. Griese: „[...] it may appear surprising that statistical methods are not that widespread in linguistics. This is all the more surprising because such methods are very widespread in disciplines with similarly complex topics such as psychology, sociology, economics. To some degree, this situation is probably due to how linguistics has evolved over the past decades [...]" (2009, s. 4).

Metoda analýzy tematické koncentrace textu byla poprvé představena Popescem (2007), dále byla rozpracována Popescem et al. (2009a), Popescem a Altmannem (2011) a Čechem et al. (2013b, 2015). V rámci textologie byla aplikována Sanadou (2013), v literární teorii a historii Wilsonem (2009), Davidovou Glogarovou et al. (2013), Davidovou Glogarovou a Čechem (2013), v historické sémantice Čechem et al. (2013a), v analýze politických projevů Tuzzi et al. (2010) a Čechem (2014a) a v tzv. postojové analýze Veselovskou a Čechem (2014).

[1] Klíčová slova byla a jsou analyzována také prostřednictvím nekvantitativních metod, za všechny srov. Němec et al. (1980), Michálek (1981), Danaher (2010); celou problematiku přehledně shrnuje David (2014).

2.1 Metoda měření tematické koncentrace

Princip měření tematické koncentrace textu je založen na vlastnostech tzv. frekvenční struktury textu. Vezmeme-li jakýkoliv text a stanovíme-li libovolné jednotky (slabiky, slova, lemmata, slovní druhy, koreferenční jednotky atp.), můžeme jednoduše spočítat frekvenci těchto jednotek. Pokud jednotky v textu uspořádáme podle frekvence od nejvyšší hodnoty k nejnižší, získáme tzv. rankovou frekvenční distribuci sledovaných jednotek. Pro ilustraci sledujme rankovou frekvenční distribuci slovních tvarů v básni J. Skácela *Odvaha k tomu* (text č. 200 v Příloze), viz Tab. 2.1:

Odvaha k tomu

Lhal jsem a říkal, že tam mrtvý není.
Tak pozdě v noci – nebude nikdo na ty housle hrát.
A byl jsem zděšený a prázdný
jak v zimě sad.
A byl tam. Docela tam byl.
Tím rovným dílem ticha jsme to znali.
Dovedli v stráni ukřižovat les –
a vodu ukamenovali.

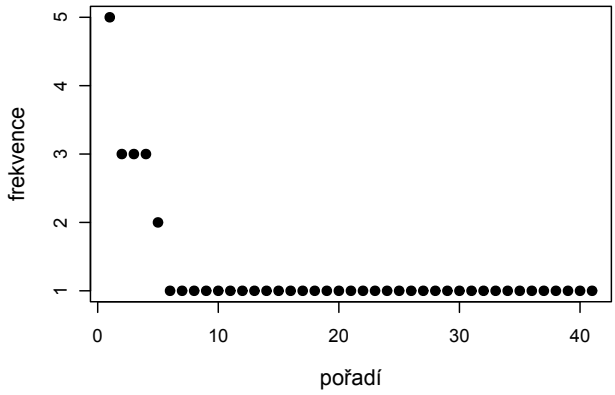
pořadí	slovní tvar	frekvence
1	*a*	5
2	*tam*	3
3	*byl*	3
4	*v*	3
5	*jsem*	2
6	*tím*	1
7	*prázdný*	1
8	*dílem*	1
9	*rovným*	1
(...)	(...)	(...)
41	*zděšený*	1

Tabulka 2.1: Ranková frekvenční distribuce slovních tvarů v básni J. Skácela *Odvaha k tomu* (text č. 200).

Vlastnosti rankových distribucí jsou systematicky zkoumány již téměř 80 let (jen namátkově: Zipf 1935, 1949; Mandelbrot 1953; Simon 1955; Rapoport 1982; Ferrer i Cancho a Solé 2001; Popescu et al. 2010, 2011), přičemž se ukazuje, že vlastnosti této distribuce reflektují základní mechanismy řídící řečové chování. Na druhou stranu v otázce

teoretického významu rankové frekvenční distribuce existují i hluboké spory (srov. Miller 1957; Miller et al. 1958; Miller, Chomsky 1963).

Sledujeme-li rankovou frekvenční distribuci téměř jakýchkoliv textů, zjistíme, že platí, že se v každém textu vyskytuje několik málo slov s relativně vysokou frekvencí a mnoho slov s frekvencí malou. Nejinak je tomu i v případě velmi krátkého textu, jímž je Skácelova báseň *Odvaha k tomu*, viz Obr. 2.1, který přehledně ilustruje tuto vlastnost textu – nejvyšší frekvenci ($f = 5$) má v textu jediný slovní tvar, zatímco frekvenci nejnižší ($f = 1$) 36 různých slovních tvarů.



Obrázek 2.1: Grafické znázornění rankové frekvenční distribuce slovních tvarů v básni J. Skácela *Odvaha k tomu* (viz Tab. 2.1).

Popescu (2007), inspirován tzv. Hirschovým indexem, který se používá pro hodnocení publikační činnosti vědců (Hirsch 2005), se pokusil aplikovat do analýzy rankové frekvenční distribuce tzv. pevný bod[2] (nazval jej h-bod), který by umožnil analyzovat frekvenční charakteristiky novým způsobem. Tento bod je definován jako místo, kde se pořadí slova rovná jeho frekvenci, tj.

$$h = f(h), \tag{2.1}$$

kde h je pořadí slova a $f(h)$ frekvence slova v daném pořadí. Pokud v dané rankové frekvenční distribuci nedojde k tomu, že $h = f(h)$, vypočítá se h-bod následujícím

[2] Pevný bod, ve smyslu, jak je užit zde, byl definován S. Banachem v r. 1922. V matematice má široké uplatnění, srov. Kirk a Sims (2001), Granas a Dugundji (2003).

způsobem:

$$h = \frac{f(i)j - f(j)i}{j - i + f(i) - f(j)}, \tag{2.2}$$

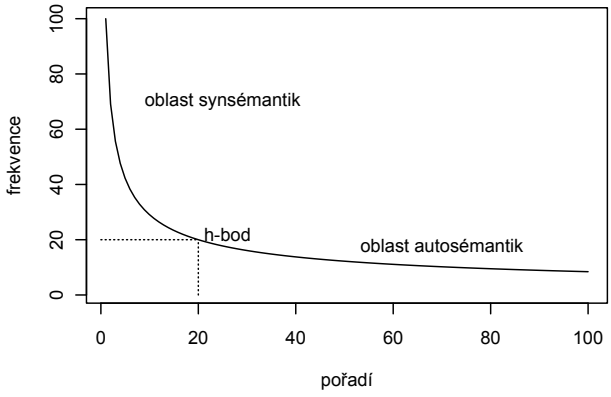
kde i a j jsou pořadí slov a $f(i)$ a $f(j)$ jsou jejich frekvence, přičemž $i < j$, kde i je největší takové číslo, pro které $i < f(i)$, a j je nejmenší takové číslo, pro které $j > f(j)$. V případě básně *Odvaha k tomu* je hodnota h rovna 3, protože třetí slovní tvar v pořadí má frekvenci $f = 3$ (viz Tab. 1). Vytvoříme-li však rankovou frekvenční distribuci slovních tvarů například z textu Bieblovy básně *S lodí jež dováží čaj a kávu* (text č. 11), viz Tab. 2, vidíme, že $h \neq f(h)$, proto použijeme vzorec (2.2) a dostáváme

$$h = \frac{4 \cdot 3 - 2 \cdot 2}{3 - 2 + 4 - 2} = 2,67.$$

pořadí	slovní tvar	frekvence
1	*a*	7
2	*s*	4
3	*na*	2
4	*za*	2
5	*loď*	2
6	*pojedem*	2
7	*kávu*	2
8	*lodí*	2
9	*jež*	2
10	*čaj*	2
11	*dováží*	2
12	*mořskou*	1
13	*přes*	1
14	*trávu*	1
(…)	(…)	(…)
66	*zvedá*	1

Tabulka 2.2: Ranková frekvenční distribuce slovních tvarů v básni K. Biebla *S lodí jež dováží čaj a kávu* (text č. 11).

Metoda h-bodu byla použita zejména pro analýzy tzv. geometrických vlastností frekvenční struktury textu (např. Popescu et al. 2009a, 2009b, 2012). Dále se ukázalo, že h-bod je možné vnímat jako hranici mezi slovy synsémantickými a autosémantickými v rankové frekvenční distribuci, viz Obr. 2.2. Jde samozřejmě jen o hranici přibližnou či neostrou, protože v oblasti autosémantik se mohou vyskytnout synsémantika a naopak.



Obrázek 2.2: h-bod oddělující dvě oblasti frekvenční distribuce slov; v grafu je hodnota h-bodu rovna 20, což znamená, že dvacáté nejfrekventovanější slovo v textu má frekvenci $f = 20$ (srov. Popescu et al 2009a, s. 17; Čech et al. 2014a, s. 15).

Vzhledem k funkci synsémantik v jazyce – jde o relační gramatické funkce předložek, spojek, částic a substituční gramatické funkce zájmen a číslovek – nejsou jejich frekvenční charakteristiky (tj. jde o slova s vysokou frekvencí) samozřejmě ničím překvapivým. Jejich vysoká frekvence se dá interpretovat jako důsledek vlivu gramatiky. Se slovy autosémantickými je to trochu složitější, protože jejich frekvence nezávisí na gramatice. Obecně mají autosémantika tendenci mít frekvenci nižší než synsémantika. Pokud se však v rankové frekvenční distribuci tato slova objeví v oblasti synsémantik (srov. Obr. 2.2), lze to považovat za jistý druh „anomálie“, která je odrazem specifické vlastnosti zkoumaného textu, konkrétně silné zaměřenosti autora na určité téma (či témata), reprezentované právě autosémantiky (či jinak vymezenými tzv. tematickými slovy, viz níže). Texty, u nichž se v oblasti synsémantik nevyskytuje žádné autosémantikum, proto definujme jako texty tematicky neutrální, texty, u nichž se v oblasti synsémantik autosémantikum vyskytne, jako texty tematicky koncentrované. Pro ilustraci bude porovnán text tematicky neutrální básně J. Skácela *Odvaha k tomu* (viz Tab. 2.1, ze které je patrné, že se nad h-bodem nenachází žádné autosémantikum) s básní *Smuténka* stejného autora (text č. 207):

Smuténka

To až se v září stmívá,
už bez sametu, drsně naholo,

po poli chodí smuténka
a zpívá,

smuténka chodí kolem hrud
šedých jak skřivani a zpívá,
(je příběh starší nežli já,
než moje smrt,
než smutek ze mne, odpusť)
zpívá si na poli smuténka
a chodí
po konopných cestách podzimu.

pořadí	slovní tvar	frekvence
1	*smuténka*	4
2	*a*	3
3	*chodí*	3
4	*zpívá*	3
5	*poli*	2
6	*po*	2
7	*než*	2
8	*starší*	1
9	*příběh*	1
(…)	(…)	(…)
39	*září*	1

Tabulka 2.3: Ranková frekvenční distribuce slovních tvarů v básni J. Skácela *Smuténka* (text č. 207).

Jak je vidět z Tab. 2.3, $\mathrm{h} = 3$, přičemž nejfrekventovanějším slovním tvarem je výraz ‚smuténka', který má frekvenci $\mathrm{f} = 4$, nachází se tedy v oblasti synsémantik a reflektuje tematické zaměření básně. Autosémantika vyskytující se nad h-bodem proto budu dále v textu označovat termínem *tematická slova*.

V případě tak krátkých textů, jako jsou obě citované básně, se může zdát výše uvedený postup zbytečně komplikovaný a „umělý" – stačí si přece oba texty přečíst a je jasné, že první báseň je tematicky nevyhraněná, zatímco v druhé dominuje právě výraz ‚smuténka'. U delších textů však často situace tak přehledná není. Navíc, jak ukazuje například praxe literární kritiky, mnohdy není možné mezi různými interpretátory textů nalézt interpersonální shodu, která by se týkala jak tématu, tak i míry

zaměřenosti daného textu. Hlavní výhodou existence h-bodu a celé kvantifikace tematické koncentrace je to, že umožňuje intersubjektivní hodnocení textu.

Jednoznačná definice jak pevného bodu v rankové frekvenční distribuci, tak i tematických slov (tj. jde o autosémantika nad h-bodem) dovoluje tematickou koncentraci textu kvantifikovat. Popescu et al. (2009a) ji kvantifikovali na základě dvou vlastností tematických slov:[3] (1) vzdálenosti mezi h-bodem a pořadím tematického slova v rankové frekvenční distribuci[4] a (2) frekvence tematického slova. Co se týká vzdálenosti (ad 1), je zřejmé, že čím nižší je pořadí slova, tím je vyšší jeho frekvence, a tudíž slovo s nízkým pořadím má na tematické koncentraci větší podíl než slovo s vyšším pořadím (podíl slova na tematické koncentraci textu budu dále označovat jako *tematickou váhu slova*). Samotná vzdálenost je definována jako

$$h - r', \tag{2.3}$$

kde r' je pořadí autosémantika nad h-bodem. Ilustrujme si tento jev na frekvenční distribuci novinového článku *V Beskydech blesk zapálil chatu, vítr lámal stromy* (text č. 267), viz Tab. 2.4.

pořadí	slovní tvar	frekvence
1	*v*	18
2	*a*	15
3	*na*	13
4	*hasiči*	10
5	*voda*	7
6	*do*	6
7	*i*	6
8	*zaplavila*	5
9	*ze*	4
(…)	(...)	(...)
340	*zdravotníkům*	1

Tabulka 2.4: Ranková frekvenční distribuce slovních tvarů v článku *V Beskydech blesk zapálil chatu, vítr lámal stromy* (text č. 267).

Z Tab. 2.4 je patrné, že h = 6 a že se nad h-bodem nacházejí dvě autosémantika (tj. tematická slova), konkrétně ‚hasiči' ($r' = 4$, $f = 10$) a ‚voda' ($r' = 5$, $f = 7$). U výrazu ‚hasiči' je

[3] Pro větší přehlednost a jednoduchost vysvětlení celého postupu výpočtu indexu určujícího míru tematické koncentrace budu tento postup ilustrovat na slovních tvarech; tento postup je však platný pro jakékoliv zvolené jednotky (více viz kap. 4).

[4] Níže vysvětluji, že je potřeba pracovat s průměrným pořadím. Tento fakt ale prozatím nechávám stranou, i když s ním implicitně pracuji i v ilustrativních příkladech. Více viz s. 21.

vzdálenost od h-bodu rovna hodnotě 2, u výrazu ‚voda' hodnotě 1. Dalším důležitým faktorem, který je třeba vzít v potaz, je frekvence (ad 2): je totiž zřejmé, že vzdálenosti různých slov od h-bodu mohou být v různých textech sice stejné, ale jejich frekvence se mohou výrazně lišit. Proto Popescu et al. (2009a) navrhují vzdálenost vynásobit právě frekvencí daného tematického slova, čímž činí analýzu přesnější:

$$(h - r') \cdot f(r'), \tag{2.4}$$

kde $f(r')$ je frekvence slova v daném pořadí. Celý problém ilustrujme na porovnání Tab. 2.4 (viz výše) a Tab. 2.5, v níž je ranková frekvenční distribuce novinového článku *Kouření v restauracích by mohlo být zakázáno od ledna 2016* (text č. 257).

pořadí	slovní tvar	frekvence
1	*v*	14
2	*a*	12
3	*by*	11
4	*alkoholu*	9
5	*kouření*	7
6	*procent*	6
7	*na*	6
8	*za*	4
9	*se*	4
(…)	(…)	(…)
245	*zdravotnických*	1

Tabulka 2.5: Ranková frekvenční distribuce slovních tvarů v článku *Kouření v restauracích by mohlo být zakázáno od ledna 2016* (text č. 257).

V obou tabulkách, Tab. 2.4 a 2.5, $h = 6$ a nad h-bodem se nacházejí dvě tematická slova, v obou případech s totožným pořadím $r' = 4$ a $r' = 5$. Pokud by nebyl započítán vliv frekvence, byla by tematická koncentrace obou textů identická. Jak ale vidíme, výraz ‚hasiči' má vyšší frekvenci, tudíž i jeho podíl na tematické koncentraci je vyšší, než je tomu u výrazu ‚alkoholu'. V případě započítání frekvence je určení míry vlivu jednotlivých slov na tematickou koncentraci textu bezpochyby adekvátnější, srov. Tab. 2.6.

Vzhledem k tomu, že s narůstající délkou textu se zvyšuje i hodnota h-bodu, je třeba tematickou váhu jednotlivých slov normalizovat. Pokud bychom to neučinili, byla by jejich váha (a v důsledku toho i tematická koncentrace celých textů) závislá zejména na délce textu. Pro ilustraci této závislosti porovnejme texty o různé délce; konkrétně báseň V. Holana *Svítání* (text č. 87), která má $N = 134$ slov, $h = 5,57$, přičemž jediným

autosémantikem nad h-bodem je ‚chvíle' ($r' = 4$, $f = 9$), článek *V Beskydech blesk zapálil chatu, vítr lámal stromy* (text č. 267) ($N = 515$, $h = 6$) s autosémantiky ‚hasiči' ($r' = 4$, $f = 10$) a ‚voda' ($r' = 5$, $f = 7$), a povídku K. Čapka *Pád rodu Votických* (text č. 797; $N = 2443$, $h = 17$), v níž jsou nad h-bodem autosémantika ‚pan' ($r' = 6$, $f = 35$), ‚dr.' ($r' = 8$, $f = 30$), ‚Mejzlík' ($r' = 9$, $f = 29$), ‚archivář' ($r' = 11$, $f = 25$), ‚pane' ($r' = 4$, $f = 9$).

slovo	$h - r'$	$(h - r')f(r')$
hasiči	2	20
voda	1	7
alkoholu	2	18
kouření	1	7

Tabulka 2.6: Vzdálenost tematických slov nad h-bodem z Tab. 2.4 a 2.5 a hodnoty této vzdálenosti po vynásobení frekvence. Důležité nejsou v tomto případě hodnoty samotné (viz níže), ale rozdíly hodnot mezi slovy se stejným pořadím a rozdílnou frekvencí: srov. ‚hasiči' vs. ‚alkoholu'.

Na základě vzorce (2.4) dostáváme tematické váhy jednotlivých slovních tvarů, srov. Tab 2.7.

text	slovní tvar	N	h	r'	$f(r')$	$(h–r')f(r')$
Svítání	*chvíle*	134	5,57	4	9	14,13
V Beskydech...	*hasiči*	515	6	4	10	20
V Beskydech...	*voda*	515	6	5	7	7
Pád rodu...	*pan*	2443	17	6	35	385
Pád rodu...	*dr.*	2443	17	8	30	270
Pád rodu...	*Mejzlík*	2443	17	9	29	232
Pád rodu...	*archivář*	2443	17	11	25	150
Pád rodu...	*pane*	2443	17	4	9	117

Tabulka 2.7: Tematická váha jednotlivých slov textů rozdílné délky.

Závislost tematické váhy slova, tak jak byla doposud určena, na délce textu je evidentní. Navíc, pokud budeme definovat tematickou koncentraci celého textu jako součet tematických vah jednotlivých tematických slov (viz níže), pak je závislost na délce textu ještě očividnější, srov. součty hodnot posledního sloupce Tab. 2.7:
Svítaní = 14,13;
V Beskydech... = 27;

Pád rodu... = 1154.
Je tedy evidentní, že je nutné tematickou váhu nějak normalizovat. Popescu et al. (2009a) navrhují každou hodnotu vypočítanou na základě vzorce (2.4) vydělit sumou rozdílu vzdáleností $(h - r)$ u všech slov nad h-bodem a nejvyšší frekvencí slova v textu $f(1)$. Sumu všech vzdáleností vypočítáme

$$\sum_{r=1}^{h}(h-r) = h^2 - \sum_{r=1}^{h} r = h^2 - \frac{h(h+1)}{2} = \frac{h(h-1)}{2}. \tag{2.5}$$

Vydělíme-li tematickou váhu slova, tj. $(h–r')f(r')$, touto sumou vynásobenou nejvyšší frekvencí slova v textu $f(1)$, můžeme definovat (stanovit) index tematické váhy slova TV

$$TV_{slovo} = 2\frac{(h-r')f(r')}{h(h-1)f(1)}. \tag{2.6}$$

V případě textů z Tab. 2.7 dostáváme po normalizaci pro jednotlivé slovní tvary tematické váhy uvedené v Tab. 2.8 (poslední sloupec).

text	slovní tvar	N	h	r'	$f(r')$	$f(1)$	TV
Svítání	*chvíle*	134	5,57	4	9	11	0,10096
V Beskydech...	*hasiči*	515	6	4	10	18	0,07407
V Beskydech...	*voda*	515	6	5	7	18	0,02593
Pád rodu...	*pan*	2443	17	6	35	73	0,03878
Pád rodu...	*dr.*	2443	17	8	30	73	0,02720
Pád rodu...	*Mejzlík*	2443	17	9	29	73	0,02337
Pád rodu...	*archivář*	2443	17	11	25	73	0,01511
Pád rodu...	*pane*	2443	17	4	9	73	0,00478

Tabulka 2.8: Tematická váha jednotlivých slovních tvarů v textech různé délky po normalizaci podle vzorce (2.6).

Tematická koncentrace celého textu je pak dána součtem hodnot tematických vah jednotlivých tematických slov, tj.

$$TK_{text} = \sum TV_{slovo} = \sum_{j=1}^{T} 2\frac{(h-r'_{(j)})f(r'_{(j)})}{h(h-1)f(1)}, \tag{2.7}$$

kde T je počet tematických slov nad h-bodem a $r'_{(j)}$ je pořadí j-tého tematického slova nad h-bodem, $j = 1, 2, ..., T$.

Ilustrujme si výpočet takto normalizované tematické váhy jednotlivých slovních tvarů a následně i celkové tematické koncentrace na příkladu textu *V Beskydech blesk*

zapálil chatu, vítr lámal stromy (text č. 267), viz Tab. 2.4.

$$\begin{aligned}\mathrm{TV}_{V\ Beskydech...} &= \mathrm{TV}_{hasiči} + \mathrm{TV}_{voda} = 2\frac{(6-4)10}{6(6-1)18} + 2\frac{(6-5)7}{6(6-1)18} = 0,07407 + 0,02593\\ &= 0,1.\end{aligned}$$

Zde je třeba upozornit na to, že se nejedná o jediný možný způsob normalizace. Zejména násobení podílu všech vzdáleností hodnotou nejvyšší frekvence slova $f(1)$ by mohlo být nahrazeno například násobením sumou všech frekvencí nad h-bodem, případně nejvyšší frekvencí autosémantika nad h-bodem. Je ale otázkou, zda by tyto modifikace znamenaly vylepšení analýzy. Pokud tato normalizace splňuje svůj účel, tj. eliminuje vliv délky textu na hodnotu tematické koncentrace (viz kap. 5), lze ji hodnotit jako účelnou.

Tematická koncentrace textů z Tab. 2.7 a 2.8, stanovená na základě vzorce (2.7), je pak následující:
$\mathrm{TK}_{Svítání} = 0,10096$;
$\mathrm{TK}_{V\ Beskydech...} = 0,1$;
$\mathrm{TK}_{Pád\ rodu...} = 0,10924$.
Všechny tři texty, byť se výrazně lišící svou délkou, jsou přibližně stejně tematicky koncentrované.

Na první pohled by se mohlo zdát překvapivé, že text, který má více autosémantik nad h-bodem, v našem případě *Pád rodu Votických* (text č. 797), nemá vyšší tematickou koncentraci než texty s výrazně menším počtem tematických slov. Vyšší počet autosémantik nad h-bodem v delších textech je dán tím, že s narůstající délkou textu roste hodnota h-bodu: s ní se zvětšuje celkový počet slov nad tímto bodem, tudíž se zvyšuje hodnota dělitele, prostřednictvím něhož se normalizuje tematická váha tematického slova, viz vzorec (2.6). Větší počet autosémantik nad h-bodem tedy automaticky neznamená větší tematickou koncentraci textu. Rozhodující roli hraje jejich postavení ve frekvenční struktuře textu, tj. jejich pořadí a frekvence. Pokud by například v povídce *Pád rodu Votických* byl výraz ‚pan' nejfrekventovanějším slovním tvarem, tj. $r' = 1$ a $f = 73$, jeho tematická váha by byla více než třikrát větší, než je v reálném textu, srov.

$$\mathrm{TV}_{pan\ (\text{hypoteticky})} = 2\frac{(17-1)73}{17(17-1)73} = 0,11765.$$

Mimochodem, text s teoreticky nejvyšší hodnotou tematické koncentrace, $\mathrm{TK} = 1$, je takový text, v němž se nad h-bodem nacházejí výhradně slova autosémantická. Taková situace, vzhledem k vlivu gramatiky, je představitelná především u extrémně krátkých textů, které mají velmi nízkou hodnotu h-bodu. Navíc se musí jednat o text, ve kterém se minimálně opakují slova, což je třeba případ Skácelovy básně *Příliš čistý sníh* (text č. 205), srov. Tab. 2.9 a Obr. 2.3:

Příliš čistý sníh

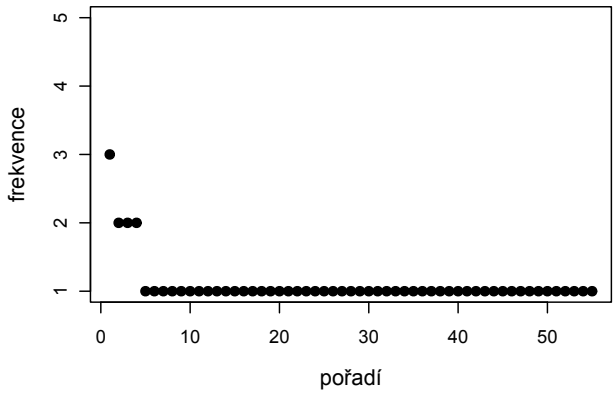
Vždycky, když padne první sníh,
mráz zamkne tůně na tři zámky,
zahodí klíče do studánky
sekerou třikrát rubané,
bývá mi smutno jako nikdy.
Jako by vítr z duše svál
poslední lístek prudce bílý

a všechno čisté polím dal.
A zaplakal bych plný studu.
Čistota jasná na polích.
To tiché nebe... Jednou budu...
Umřeme všichni pro ten sníh.

pořadí	slovní tvar	frekvence
1	*sníh*	3
2	*na*	2
3	*jako*	2
4	*a*	2
5	*čisté*	1
6	*polím*	1
(...)	(...)	(...)
55	*zaplakal*	1

Tabulka 2.9: Ranková frekvenční distribuce slovních tvarů v básni J. Skácela *Příliš čistý sníh* (text č. 205).

Je však třeba zdůraznit, že texty s hodnotou TK = 1 jsou výjimkou: jsou výsledkem souhry několika okolností, především délky textu a specifika žánru, které v tomto případě umožňuje minimální opakování slov. Minimální opakování slov v textu se projevuje vysokými hodnotami poměru počtu různých slov (typů) k celkovému počtu všech slov (tokenů) v textu, jde o tzv. type-token poměr (TTR). V případě básně *Příliš čistý sníh* dosahuje vysoké TTR = 0,92; pro srovnání, ještě kratší báseň *Odvaha k tomu* (text č. 200), viz Tab. 2.1 a Obr. 2.1 (N = 52, TK = 0), má TTR = 0,84. Ve zkoumaném vzorku 1168 textů se objevil jediný text s hodnotou TK = 1 (a to jak v případě analýzy založené na slovních tvarech, tak lemmatech). Jen pro srovnání: průměrná hodnota tematické koncentrace v daném vzorku je TK = 0,0297 (směrodatná odchyl-



Obrázek 2.3: Grafické znázornění rankové frekvenční distribuce slovních tvarů v básni J. Skácela *Příliš čistý sníh* (viz Tab. 2.9).

ka sd = 0,0791) pro slovní tvary, TK = 0,0415 (směrodatná odchylka sd = 0,0886) pro lemmatizované texty.

V poznámce pod čarou 4 (s. 15) bylo upozorněno, že je třeba při výpočtu TK pracovat s hodnotou průměrného pořadí slova, nikoliv hodnotou absolutní. Jde o to, že slova se stejnou frekvencí jsou v rankové frekvenční distribuci řazena buď náhodně, nebo podle nějaké konvence, například podle abecedy, podle pořadí výskytu slova v textu apod. Pokud bychom s průměrným pořadím nepracovali, mohli bychom u jednoho a téhož textu dospět k různým hodnotám TK pouze v důsledku náhodného seřazení slov se stejnou frekvencí, případně v důsledku konvence, která nemá v tomto případě žádný rozumně interpretovatelný význam (např. abecední řazení). Pro ilustraci sledujme rankovou frekvenční distribuci slovních tvarů z novinového článku *Soud zamítl Berdychovu žádost o podmínečné propuštění* (text č. 263), viz Tab. 2.10.

Na základě vzorce (2.1) nejdříve stanovíme hodnotu h-bodu, $h = 7$. Výrazy v 6.–8. pořadí mají stejnou frekvenci $f = 7$ a jsou uspořádány podle abecedy, stejně tak výrazy ve 4.–5. pořadí s frekvencí $f = 8$. Při tomto uspořádání se nad h-bodem vyskytuje jediné tematické slovo *Berdych* ($r = 4$). Pokud by však ranková frekvenční distribuce byla u slov se stejnou frekvencí uspořádána podle jiné konvence (např. v inverzním abecedním pořadí), objevily by se nad h-bodem dvě tematická slova, tj. *Berdych* ($r = 5$) a *trestu* ($r = 6$). Při původním uspořádání by byla tematická váha tematického slova *Berdych* $TV_{Berdych_1} = 0,08163$, při druhém uspořádání by byla o třetinu menší $TV_{Berdych_2} = 0,05442$, přičemž rozdíl obou hodnot závisí na faktoru (tj. způsobu

pořadí	slovní tvar	frekvence
1	*že*	14
2	*a*	12
3	*se*	11
4	*Berdych*	8
5	*za*	8
6	*z*	7
7	*na*	7
8	*trestu*	7
9	*je*	6
10	*v*	6
(...)	(...)	(...)
373	*neuvěřil*	1

Tabulka 2.10: Ranková frekvenční distribuce slovních tvarů v článku *Soud zamítl Berdychovu žádost o podmínečné propuštění* (text č. 263).

abecedního řazení), který je vzhledem k tematickým charakteristikám textu naprosto irelevantní. Je tedy evidentní, že je nutné pracovat s průměrným pořadím slov:

$$\bar{r}_{\text{slovo}} = \frac{\sum r_{f_i}}{n_{f_i}}, \tag{2.8}$$

kde r_{f_i} je pořadí slova o frekvenci f_i a n_{f_i} je počet slov s frekvencí f_i. Na základě tohoto vzorce dostáváme pro tematická slova z Tab. 2.10 hodnoty

$$\bar{r}_{Berdych} = \frac{r_{\text{Berdych}} + r_{za}}{2} = \frac{4+5}{2} = 4,5$$

a

$$\bar{r}_{trestu} = \frac{r_z + r_{na} + r_{trestu}}{3} = \frac{6+7+8}{3} = 7.$$

V tomto případě se nad h-bodem nachází jediné tematické slovo, jehož tematická váha je $TV_{Berdych} = 0,06803$. Tato hodnota odpovídá také tematické koncentraci celého textu.

V souvislosti s použitím průměrných hodnot pořadí vyvstává také otázka týkající se určování h-bodu: nebylo by smysluplnější určovat h-bod vzhledem k průměrnému pořadí? Na první pohled se zdá, že by takový přístup byl „racionálnější“ a teoreticky obhajitelnější než dosavadní metoda. Při bližší analýze povahy rankové frekvenční distribuce se však ukazuje, že mohou nastat případy, které by v případě použití průměrných hodnot vedly k tomu, že by v rankové frekvenční distribuci nebylo vůbec možné h-bod určit. V Tab. 2.11 je ranková frekvenční distribuce Holanovy básně *Ale čas* (text č. 73). V případě použití absolutní hodnoty pořadí je h-bod roven 2. Pokud

bychom však použili průměrné hodnoty, není možné h-bod stanovit, protože pro výrazy s $f = 2$ je $\bar{r} = 3,5$, tj. hodnota frekvence je nižší než hodnota průměrného pořadí, a nedojde tedy k „protnutí" hodnot frekvence a pořadí. Jinými slovy, v případě vzorce (2.2) není splněna podmínka, že pokud i a j jsou pořadí slov a $f(i)$ a $f(j)$ jsou jejich frekvence, pak i je takové číslo, pro které $i < f(i)$. Jestliže takové číslo neexistuje, není možné h-bod určit.

pořadí	průměrné pořadí	slovní tvar	frekvence
1	3,5	*v*	2
2	3,5	*tak*	2
3	3,5	*to*	2
4	3,5	*je*	2
5	3,5	*ale*	2
6	3,5	*čas*	2
7	27,5	*celé*	1
8	27,5	*chrámy*	1
9	27,5	*mi*	1
10	27,5	*řekl*	1
(...)	(...)	(...)	(...)
48	27,5	*že*	1

Tabulka 2.11: Ranková frekvenční distribuce slovních tvarů v básni V. Holana *Ale čas* (text č. 73).

2.2 Statistické testování rozdílů tematické koncentrace textu

Metoda analýzy tematické koncentrace byla od počátku koncipována tak, aby umožnila nejen tuto vlastnost kvantifikovat, ale především aby bylo možné jejím prostřednictvím statisticky testovat rozdíly mezi jednotlivými texty (srov. předpoklad (d) v kapitole 1). Pro aplikaci testu je třeba znát nejen samotné hodnoty TK, ale i její varianci. Popescu a Altmann (2011) odvodili vzorec pro výpočet variance,

$$Var(TK) = \left(\frac{2}{h(h-1)f(1)}\right)^2 \cdot \left(\sum_{j=1}^{T} f(r'_{(j)})\right) \cdot m_{2r'}, \tag{2.9}$$

kde $m_{2r'}$ je rozptyl (druhý centrální moment) tematických slov nad h-bodem, tj.

$$m_{2r'} = \frac{\sum_{j=1}^{T}(r'_{(j)} - m_{1r'})^2 f(r'_{(j)})}{\sum_{j=1}^{T} f(r'_{(j)})}, \tag{2.10}$$

kde $m_{1r'}$ je první počáteční moment, tj.

$$m_{1r'} = \frac{\sum_{j=1}^{T} r'_{(j)} \cdot f(r'_{(j)})}{\sum_{j=1}^{T} f(r'_{(j)})}. \tag{2.11}$$

Rozdíly hodnot TK jednotlivých textů je možné testovat prostřednictvím asymptotického u-testu[5],

$$u = \frac{TK_1 - TK_2}{\sqrt{Var(TK_1) + Var(TK_2)}}. \tag{2.12}$$

Pokud chceme porovnávat skupiny textů, použijeme průměrné hodnoty TK a do jmenovatele vzorce (2.12) dosadíme namísto variance hodnotu podílu průměru rozptylů TK v jednotlivé skupině s^2 a počtu textů n v této skupině, tj.

$$u = \frac{\overline{TK}_1 - \overline{TK}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}. \tag{2.13}$$

Celý postup testování budu ilustrovat na porovnání textů *V Beskydech blesk zapálil chatu, vítr lámal stromy* (Tab. 2.4, text č. 267) a *MF výrazně zlepšilo pro letošek odhad růstu ekonomiky na 2,7 %* (text č. 258). Tematická koncentrace prvního textu je $TK_{\textit{V Beskydech...}} = 0,1$ (viz výpočet na s. 18), přičemž tematickými slovy jsou ‚hasiči' ($r' = 4$, $f(r') = 10$) a ‚voda' ($r' = 5$, $f(r') = 7$). Pro výpočet variance potřebujeme znát také hodnotu h-bodu ($h = 6$) a nejvyšší frekvenci slovního tvaru v textu $f(1) = 18$. V případě druhého textu platí $TK_{\textit{MF výrazně...}} = 0,278571$, tematickými slovy jsou ‚procenta' ($r' = 2$, $f(r') = 12$) a ‚ekonomiky' ($r' = 4,5$, $f(r') = 7$), $h = 6$, $f(1) = 14$. Na základě vzorce (2.11) získáme hodnoty prvního počátečního momentu,

$$m_{1r'\textit{V Beskydech...}} = \frac{4 \cdot 10 + 5 \cdot 7}{10 + 7} = 4,4118,$$

$$m_{1r'\textit{MF výrazně...}} = \frac{2 \cdot 12 + 4,5 \cdot 7}{12 + 7} = 2,9211.$$

Hodnoty druhého centrálního momentu jsou podle vzorce (2.10)

$$m_{2r'\textit{V Beskydech...}} = \frac{(4 - 4,4118)^2 10 + (5 - 4,4118)^2 7}{10 + 7} = 0,24221,$$

$$m_{2r'\textit{MF výrazně...}} = \frac{(2 - 2,9211)^2 12 + (4,5 - 2,9211)^2 7}{12 + 7} = 1,45429.$$

[5] Ve statistice je označován zřejmě častěji jako z-test. V této knize se držím konvence, která převažuje v kvantitativní lingvistice.

Variance tematických koncentrací obou textů podle vzorce (2.9) odpovídají následujícím hodnotám:

$$\mathrm{Var}(\mathrm{TK}_{V\ Beskydech\ldots}) = \left(\frac{2}{6\cdot(6-1)\cdot 18}\right)^2 \cdot (10+7)\cdot 0,24221 = 0,000056,$$

$$\mathrm{Var}(\mathrm{TK}_{MF\ výrazně\ldots}) = \left(\frac{2}{6\cdot(6-1)\cdot 14}\right)^2 \cdot (12+7)\cdot 1,45429 = 0,000627.$$

Nyní je již možné pomocí asymptotického u-testu (2.13) testovat rozdíly mezi oběma texty,

$$u = \frac{\mathrm{TK}_{V\ Beskydech\ldots} - \mathrm{TK}_{MF\ výrazně\ldots}}{\sqrt{\mathrm{Var}(\mathrm{TK}_{V\ Beskydech\ldots}) + \mathrm{Var}(\mathrm{TK}_{MF\ výrazně\ldots})}} = \frac{0,1-0,27857}{\sqrt{0,000056+0,000627}} = -6,83.$$

Pokud zvolíme hladinu významnosti $\alpha = 0,05$, pak je rozdíl signifikantní, jestliže $|u| > 1,96$. Mezi analyzovanou dvojicí textů je tedy signifikantní rozdíl tematických koncentrací.

12
Závěr

Metoda měření tematické koncentrace textu, jak je představena v této knize, má podle mého názoru následující výhody: 1) je srozumitelně lingvisticky interpretovatelná, 2) kromě možnosti kvantifikace celkové zaměřenosti autora na dané téma či témata umožňuje detekovat výrazy (ať už ve formě slovních tvarů či lemmat), které reprezentují hlavní téma textu, 3) je v určitém intervalu (viz kap. 5) nezávislá na délce textu, 4) umožňuje statisticky testovat rozdíly mezi jednotlivými texty i skupinami textů, 5) umožňuje analyzovat dynamický vývoj textu (viz kap. 6), přičemž rozdíly tohoto vývoje mezi texty lze také statisticky testovat. Díky těmto vlastnostem je možné ji velmi dobře využít pro stylometrické studie nejrůznějšího charakteru. Nepochybně existuje celá řada dalších vlastností textu a způsobů jejich měření, které vykazují podobné vlastnosti. Jejich adekvátní aplikaci by však měla předcházet důkladná znalost toho, co a jak se měří.

V této knize jsem sledoval dva základní cíle: jednak prozkoumat vlastnosti tematické koncentrace a jejího měření, jednak se pokusit ukázat základní rysy toho, jak by měla vypadat prezentace de facto jakékoliv vlastnosti textu a jejího měření, jež je založena na kvantifikaci. Samozřejmě že zde prezentovaný způsob není jediný možný a bezpochyby by se dal (a doufám i bude) vylepšovat. V každém případě jsem ale přesvědčen o tom, že před aplikací každé metody tohoto typu je nutné důkladně prozkoumat, jak se chová a) vzhledem k délce textu a b) vzhledem k volbě jazykových jednotek. Bez těchto znalostí je třeba být k výsledkům každého textologického měření velmi ostražitý.

Pokud chceme lépe porozumět obecným vlastnostem textů, žánrů, stylů atp., je třeba sledovat vzájemné vztahy mezi jejich jednotlivými charakteristikami, pokusit se tyto vztahy modelovat a hledat teoretická zdůvodnění těchto vztahů. U tematické koncentrace jsem se pokusil tento směr výzkumu ukázat vzhledem k měření tzv. slovního bohatství (kap. 7) a částečně i analýze klíčových slov (kap. 8). V obou případech se ukázalo, jak je tento úkol nesnadný, a to zejména proto, že nemáme dostatečné znalosti o fungování daných metod. Osobně se domnívám, že právě modelování vztahů mezi různými vlastnostmi textu představuje jednu z výzev současné textologii.

V závěrečné části knihy (kap. 9 a 10) prezentuji několik možných způsobů, jak aplikovat tematickou koncentraci v kvantitativně založené textologii. Přestože se jedná o dílčí studie, jejichž hlavním cílem je poukázat na možnosti využití metody, výsledky naznačují, že tematická koncentrace je vlastností, která se dá dobře interpretovat s ohledem na znalosti současné stylistiky a textologie.

Vlastnosti každého textu odrážejí obrovskou komplexitu verbálního chování lidí, které není jednoduché zachytit, popsat a vysvětlit. Zaměřenost se na téma je jen jedním z dílčích aspektů tohoto chování, byť asi ne zcela zanedbatelným. Pokusil jsem se zde tento aspekt prozkoumat a popsat tak, aby se stal buď dobře použitelným pro další výzkum, nebo alespoň srozumitelně kritizovatelným. Ať už bude používán, nebo se stane inspirací pro kritické přehodnocení toho, jak analyzovat tematické charakteristiky textu, splnila tato práce z mého pohledu svůj účel.

Summary

The purpose of this book is to present a systematic analysis of a method to measure a thematic text property, termed thematic concentration, and to introduce ways of applying this method in textology. The method is based on frequency characteristics of a text. Select properties of rank frequency distribution of words are used to detect thematic words, i.e. words representing central topics of the text. Moreover, the method allows to quantify the thematic weight of these words and, consequently, to quantify a degree of the thematic concentration of the whole text. Differences between the thematic concentrations of particular texts (or groups of texts) can be statistically tested.

In order to overcome the limitations of the original method, as well as to reflect different goals of these textological studies, this book introduces various modifications, such as the secondary thematic concentration and the proportionally thematic concentration.

In any quantitative textological research, the results are strongly influenced by the choice of the language unit which is used for the measurement. However, the impact of this choice has not been taken into consideration in a majority of studies of this kind. To avoid this shortcoming, the relationships between this choice and the particular methods of analysis of the thematic concentration are investigated. Specifically, word forms, lemmas, and coreferential units are applied.

The length of the text is another factor which can fundamentally influence the quantitative text analysis. As for thematic concentration, the interval in which the length of the text has no impact on the thematic concentration is derived empirically. Moreover, this interval corresponds to the theoretical assumptions presented in this book.

The thematic concentration is not an isolated text property and it is obvious that it should be related to other text properties. However, because of the absence of a text theory based on which it would be possible to predict these relationships, an exploration in this research area has been up to now rather heuristic. This book studies relationships between the thematic concentration and the vocabulary richness, as well as between the thematic concentration and the keyword analysis.

The final part of the book is devoted to the application of this method in textology for analysis of the associated structure of a text and for classification of texts. As for the former, the method allows detection of statistically significant associations among thematic words in a text. Regarding the latter, particular registers such

as fiction, scientific texts, journalistic texts, etc. differ significantly with regard to the thematic concentration. The method can also be used for the analysis of authorship.

Literatura

S. Adolphs. *Introducing Electronic Text Analysis: A Practical Guide for Language and Literary Studies*. Routledge, London – New York, 2006.

G. Altmann a V. Burdinski. Towards a law of word repetitions in text blocks. *Glottometrika*, 4: 146–167, 1982.

E. Bejček, E. Hajičová, J. Hajič, P. Jínová, V. Kettnerová, V. Kolářová, M. Mikulová, J. Mírovský, A. Nedoluzhko, J. Panevová, L. Poláková, M. Ševčíková, J. Štěpánek a Š. Zikánová. *Prague Dependency Treebank 3.0. Data/software*. Univerzita Karlova v Praze, MFF, ÚFAL, Prague, 2013.

Ch. Bernet. Faits lexicaux. Richesse du vocabulaire. In P. Thoiron et al., red., *Etudes sur la richesse et la structure lexicale*, s. 1–11. Champion, Paris, 1988.

D. Biber a S. Conrad. *Register, Genre, and Style*. Cambridge University Press, Cambridge, 2009.

D. Biber, S. Johansson, G. Leech, S. Conrad a E. Finegan. *Grammar of Spoken and Written English*. Longman, Harlow, 1999.

M. Bondi a M. Scott. *Keyness in Texts*. Benjamins, Amsterdam, 2010.

M. B. Brown a A. B. Forsythe. Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346):364–367, 1974.

G. Caldarelli. *Scale-Free Networks: Complex Webs in Nature and Technology*. Oxford University Press, 2007.

M. A. Covington a J. D. McFall. Cutting the Gordian Knot: The Moving Average Type-Token Ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100, 2010.

B. Cvek. Relativismus ve světle nejnovější filosofie přírodních věd. *Filosofický časopis*, 59:269–276, 2011a.

B. Cvek. Filosofie přírodních věd na začátku 21. století. 1. Jak se pohled na vědu měnil za posledních asi dvě stě let? *Vesmír*, 90:724—725, 2011b.

B. Cvek. Filosofie přírodních věd na začátku 21. století. 2. Dějinnost vědy. *Vesmír*, 91:48–49, 2012a.

B. Cvek. Filosofie přírodních věd na začátku 21. století. 3. Otázky o povaze lidského poznání. *Vesmír*, 91:113–114, 2012b.

V. Cvrček a D. Kováříková. Možnosti a meze korpusové lingvistiky. *Naše řeč*, 94:113–133, 2011.

V. Cvrček a O. Richterová, red. 'pojmy:chi2'. In *Příručka ČNK, 12.9.2013*, 2013a. URL http://wiki.korpus.cz/doku.php?id=pojmy:chi2&rev=1378999273.

V. Cvrček a O. Richterová, red. 'pojmy:txtype_group'. In *Příručka ČNK, 12.9.2013*, 2013b. URL http://wiki.korpus.cz/doku.php/pojmy:txtype_group?rev=1379083398&vecdo=cite.

V. Cvrček a O. Richterová, red. 'pojmy:txtype'. In *Příručka ČNK, 12.9.2013*, 2013c. URL http://wiki.korpus.cz/doku.php?id=pojmy:txtype&rev=1379083369.

V. Cvrček a O. Richterová, red. 'pojmy:asociacni_miry'. In *Příručka ČNK, 21.1.2015*, 2015a. URL http://wiki.korpus.cz/doku.php/pojmy:asociacni_miry?redirect=1#log_likelihood.

V. Cvrček a O. Richterová, red. 'manualy:kwords'. In *Příručka ČNK, 21.1.2015*, 2015b. URL http://wiki.korpus.cz/doku.php?id=manualy:kwords&rev=1421859814.

V. Cvrček a O. Richterová, red. 'cnk:syn2005'. In *Příručka ČNK, 21.1.2015*, 2015c. URL http://wiki.korpus.cz/doku.php?id=cnk:syn2005&rev=1422001415.

V. Cvrček a O. Richterová, red. 'cnk:syn2010'. In *Příručka ČNK, 21.1.2015*, 2015d. URL http://wiki.korpus.cz/doku.php?id=cnk:syn2010&rev=1422000944.

V. Cvrček a P. Vondřička. *KWords*. FF UK, Praha, 2013. URL http://kwords.korpus.cz.

R. Čech. Language and ideology: Quantitative thematic analysis of New Year speeches given by Czechoslovak and Czech presidents (1949-2011). *Quality & Quantity*, 48(2):899–910, 2014a.

R. Čech. Jen popis s čísly? Perspektivy korpusové lingvistiky. *Naše řeč*, 97:171–184, 2014b.

R. Čech. Text length and the lambda frequency structure of the text. In G. K. Mikros a J. Mačutek, red., *Sequences in language and text*, s. 71–87. Mouton de Gruyter, Berlin – Boston, 2015.

R. Čech, J. Davidová Glogarová a J. David. Kvantitativně lingvistické metody a jejich využití v historické sémantice. In J. David, R. Čech, L. Radková, J. Davidová Glogarová a H. Šústková, red., *Slovo a text v historickém kontextu – perspektivy historickosémantické analýzy jazyka*, s. 32–84. Host, Brno, 2013a.

R. Čech, I. I. Popescu a G. Altmann. Methods of Analysis of a Thematic Concentration of the Text. *Czech and Slovak Linguistic Review*, s. 4–21, 2013b.

R. Čech, E. Kelih a J. Mačutek. Impact of Semantics on Case Diversification. In *Contributed talk, QUALICO 2014, Olomouc, Czech Republic, May 29 - June 1, 2014*, 2014a.

R. Čech, I. I. Popescu a G. Altmann. *Metody kvantitativní analýzy (nejen) básnických textů*. Univerzita Palackého v Olomouci, Olomouc, 2014b.

R. Čech, R. Garabik a G. Altmann. Testing the Thematic Concentration of Text. *Journal of Quantitative Linguistics*, 2015. (accepted).

M. Čechová, M. Krčmová a E. Minářová. *Současná stylistika*. Nakladatelství Lidové noviny, Praha, 2008.

J. David, R. Čech, L. Radková, J. Davidová Glogarová a H. Šústková. *Slovo a text v historickém kontextu – perspektivy historickosémantické analýzy jazyka*. Host, Brno, 2013.

J. Davidová Glogarová a R. Čech. Tematická koncentrace textu – některé aspekty autorského stylu Ladislava Jehličky. *Naše řeč*, 96:234–245, 2013.

J. Davidová Glogarová, J. David a R. Čech. Analýza tematické koncentrace textu – komparace publicistiky Ladislava Jehličky a Karla Čapka. *Slovo a slovesnost*, 74:41–54, 2013.

J. Demel. *Grafy a jejich aplikace*. Academia, Praha, 2002.

K. Ejiri a A. E. Smith. Proposal of a New 'Constraint Measure' for Text. In R. Köhler a B. B. Rieger, red., *Contributions to Quantitative Linguistics*, s. 195–211. Kluwer, Dordrecht, 1993.

Y. Esterková. *Lingvistická analýza novoročních projevů prezidenta Václava Havla*. Rigorózní práce, Ostravská univerzita v Ostravě, 2013. URL https://theses.cz/id/w4jgu8/.

R. Ferrer i Cancho a R.V. Solé. Two Regimes in the Frequency of Words and the Origin of Complex Lexicons: Zipf's Law Revisited. *Journal of Quantitative Linguistics*, 8:165–173, 2001.

P. K. Feyerabend. *Rozprava proti metodě*. Aurora, Praha, 2001.

A. Granas a J. Dugundji. *Fixed Point Theory*. Springer Science & Business Media, 2003.

S. Gries. *Statistics for Linguistics with R: A Practical Introduction*. Mouton de Gruyter, Berlin, 2009.

R. Grotjahn. *Linguistische und statistische Methoden in Metrik und Textwissenschaft*. Brockmeyer, Bochum, 1979.

R. Grotjahn. The Theory of Runs as an Instrument for Research in Quantitative Linguistics. *Glottometrika*, 2:11–43, 1980.

P. Guiraud. *Les catactères statistiques du vocabulaire*. Presses Universitaires de France, Paris, 1954.

P. Guiraud. *Problèmes et mèthodes de la statistique linguistique*. Reidel, Dordrecht, 1959.

J. Hajič. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*, svazek 1. Charles University Press, Prague, 2004.

B. Havránek. Úkoly spisovného jazyka a jeho kultura. In B. Havránek a M. Weingart, red., *Spisovná čeština a jazyková kultura*, s. 32–84. Melantrich, Praha, 1932.

B. Havránek. K funkčnímu rozvrstvení spisovného jazyka. *Časopis pro moderní filologii*, 28:409–416, 1942.

B. Havránek, J. Bělič, M. Helcl, A. Jedlička, V. Křístek a F. Trávníček, red. *Slovník spisovného jazyka českého*. Academia, Praha, 1989.

G. Herdan. *Type-Token Mathematics*. Moulton, The Hague, 1960.

G. Herdan. *The Advanced Theory of Language as Choice and Chance*. Springer, New York, 1966.

C. E. Hess, K. M. Sefton a R. G. Landry. Sample Size and Type-Token Ratios for Oral Language of Preschool Children. *Journal of Speech and Hearing Research*, 29:129–134, 1986.

C. E. Hess, K. M. Sefton a R. G. Landry. The Reliability of Type-Token Ratios for the Oral Language of School Age Children. *Journal of Speech and Hearing Research*, 32:536–540, 1989.

J. E. Hirsch. An Index to Quantify an Individual's Research Output. *Proceedings of the National Academy of Sciences of the USA*, 102(46):16569–16572, 2005.

Z. Hladká. Slovo. In P. Karlík, M. Nekula a J. Pleskalová, red., *Encyklopedický slovník češtiny*, s. 424. Nakladatelství Lidové noviny, Praha, 2002.

T. Honorè. Some Simple Measures of Richness of Vocabulary. *ALLC Bulletin*, 7:172–177, 1979.

L. Hřebíček. Text as a Construct of Aggregations. In R. Köhler a B. B. Rieger, red., *Contributions to Quantitative Linguistics*, s. 33–39. Kluwer, Dordrecht, 1993.

L. Hřebíček. *Lectures on Text Theory*. Oriental Institute, Prague, 1997.

L. Hřebíček. *Variation in Sequences*. Oriental Institute, Prague, 2000.

L. Hřebíček. *Vyprávění o lingvistických experimentech s textem*. Academia, Praha, 2002.

J. Chloupek, M. Čechová, M. Krčmová a E. Minářová. *Stylistika češtiny*. SPN, Praha, 1990.

J. Chromý. Korpus a reprezentativnost. *Naše řeč*, 97:185–193, 2014.

M. Jelínek. Styl publicistický. In P. Karlík, M. Nekula a J. Pleskalová, red., *Encyklopedický slovník češtiny*, s. 458–460. Nakladatelství Lidové noviny, Praha, 2002a.

M. Jelínek. Styl prozaický. In P. Karlík, M. Nekula a J. Pleskalová, red., *Encyklopedický slovník češtiny*, s. 458. Nakladatelství Lidové noviny, Praha, 2002b.

T. Jelínek. Nové značkování v Českém národním korpusu. *Naše řeč*, 91:13–20, 2008.

E. Kelih, A. Rovenchak a S. Buk. Analysing h-point in Lemmatised and Non-Lemmatised Texts. In G. Altmann, R. Čech, J. Mačutek a L. Uhlířová, red., *Empirical Approaches to Text and Language Analysis*, s. 81–93. RAM-Verlag, Ludenscheid, 2014.

W. Kirk a B. Sims, red. *Handbook of Metric Fixed Point Theory*. Springer Science & Business Media, 2001.

R. Köhler. *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Brockmeyer, Bochum, 1986.

R. Köhler. Synergetic Linguistics. In R. Köhler, G. Altmann a R. G. Piotrowski, red., *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*, s. 760–774. Mouton de Gruyter, Berlin – New York, 2005.

R. Köhler a G. Altmann. Quantitative Linguistics. In P. C. Hogan, red., *The Cambridge Encyclopedia od the Language Sciences*, s. 695–697. Cambridge University Press, New York, 2011.

R. Köhler a G. Altmann. *Kvantitativní lingvistika. Vybrané problémy* 2. Univerzita Palackého v Olomouci, Olomouc, 2014.

R. Köhler, G. Altmann a R. G. Piotrowski, red. *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*. Mouton de Gruyter, Berlin – New York, 2005.

J. Králík. The Representativeness of Czech Corpora. *International Journal of Corpus Linguistics*, 10:357–366, 2005.

J. Králík. Srovnávání nesrovnatelného. *Korpus – Gramatika – Axiologie*, 4:48–52, 2013.

K. Krippendorff. *Content Analysis. An Introduction to Its Methodology*, svazek 3. SAGE Publications, Inc., Los Angeles –- London –- New Delhi –- Singapore –- Washington DC, 2013.

M. Křen. *Odraz jazykových změn v synchronních korpusech*. Nakladatelství Lidové noviny, Praha, 2013.

M. Křen, T. Bartoň, V. Cvrček, M. Hnátková, T. Jelínek, J. Kocek, R. Novotná, V. Petkevič, P. Procházka, V. Schmiedtová a H. Skoumalová. *SYN2010: žánrově vyvážený korpus psané češtiny*. Ústav Českého národního korpusu FF UK, Praha, 2010. URL http://www.korpus.cz.

M. Kubát. Moving Window Type-Token Ratio and Text Length. In G. Altamnn, R. Čech, J. Mačutek a L. Uhlířová, red., *Empirical Approaches to Text and Language Analysis*, s. 105–113. RAM-Verlag, Lüdenscheid, 2014.

M. Kubát a J. Milička. Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics*, 20(4):339–349, 2013.

M. Kubát, V. Matlach a R. Čech. *QUITA. Quantitative Index Text Analyzer*. RAM-Verlag, Lüdenscheid, 2014.

T. S. Kunh. *Struktura vědeckých revolucí*. Oikoymenh, Praha, 1997.

A. Lee, R. Prasad, A. Joshi, N. Dinesh a B. Weber. Complexity of Dependencies in Discourse: Are Dependencies in Discourse More Complex than in Syntax? In J. Hajič a J. Nivre, red., *Proceedings of the 5th Workshop on Treebanks and Linguistic Theories (TLT 2006)*, s. 79–90. Prague, 2006.

G. Leech. New Resources, or Just Better Old Ones? The Holy Grail of Representativeness. In M. Hundt, N. Nesselhauf a C. Biewer, red., *Corpus Linguistics and the Web*, s. 133–149. Rodopi, Amsterdam – New York, 2007.

B. Mandelbrot. An Information Theory of the Statistical Structure of Language. In W. Jackson, red., *Communication Theory*, s. 486–502. Butterworth, London, 1953.

G. Martynenko. Measuring Lexical Richness and Its Harmony. In P. Grzybek, E. Kelih a J. Mačutek, red., *Text and Language. Structures • Functions • Interrelations. Quantitative Perspectives*, s. 125–132. Praesens, Wien, 2010.

V. Matlach. *Kvantitativně lingvistický software*. Diplomová práce, UP Olomouc, 2014. URL http://theses.cz/id/fz87uj.

N. Menard. *Mesure de la richesse lexicale*. Slatkine, Paris, 1983.

G. K. Mikros a K. Perifanos. Authorship Identification in Large Email Collections: Experiments Using Features that Belong to Different Linguistic Levels. In *Proceedings of PAN 2011 Lab, Uncovering Plagiarism, Authorship, and Social Software Misuse held in conjunction with the CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation, 19-22 September 2011, Amsterdam*, 2011.

G. K. Mikros a K. Perifanos. Authorship Attribution in Greek Tweets Using Multilevel Author's N-Gram Profiles. In E. Hovy, V. Markman, C. H. Martell a D. Uthus, red., *Papers from the 2013 AAAI Spring Symposium "Analyzing Microtext", 25-27 March 2013, Stanford, California*, s. 17–23, AAAI Press, Palo Alto, California, 2013.

M. Mikulová, A. Bémová, J. Hajič, E. Hajičová, J. Havelka, V. Kolářová, L. Kučová, M. Lopatková, P. Pajas, J. Panevová, M. Razímová, P. Sgall, J. Štěpánek, Z. Urešová, K. Veselá a Z. Žabokrtský. *Anotace na tektogramatické rovině pražského závislostního korpusu. Anotátorská příručka*. UFAL MFF UK, Praha, 2006. URL http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/t-layer/html/index.html.

G. A. Miller. Some Effects of Intermittent Silence. *The American Journal of Psychology*, s. 311–314, 1957.

G. A. Miller a N. Chomsky. Finitary Models of Language Users. In R. D. Luce, R. Bush a E. Galanter, red., *Handbook of mathematical psychology*, svazek 2, s. 419–491. Wiley, New York, 1963.

G. A. Miller, E. B. Newman a E. A. Friedman. Length-Frequency Statistics for Written English. *Information and Control*, 1(4):370–389, 1958.

J. Mírovský, L. Mladová a Š. Zikánová. Connective-Based Measuring of the Inter Annotator Agreement in the Annotation of Discourse in PDT. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, China*, s. 775–781, 2010.

A. Nedoluzhko. *Rozšířená textová koreference a asociační anaphora. Koncepce anotace českých dat v Pražském závislostním korpusu*. Ústav formálni a aplikované lingvistiky, Praha, 2011.

A. Nedoluzhko a J. Mírovský. *Annotating Extended Textual Coreference and Bridging Relations in the Prague Dependency Treebank. Annotation manual.* Technical report 44, UFAL MFF UK, Prague, 2011.

A. Nedoluzhko, J. Mírovský a M. Novák. A Coreferentially Annotated Corpus and Anaphora Resolution for Czech. In *Computational Linguistics and Intellectual Technologies. ABBYY, Moscow, Russia*, s. 467–475, 2013.

M. Newman. *Networks: an Introduction*. Oxford University Press, Oxford –- New York, 2011.

E. Panas. The Generalized Torquist: Specification and Estimation of a New Vocabulary Text-Size Function. *Journal of Quantitative Linguistics*, 8:233–252, 2001.

J. Panevová. Koreference. In P. Karlík, M. Nekula a J. Pleskalová, red., *Encyklopedický slovník češtiny*, s. 233–234. Nakladatelství Lidové noviny, Praha, 2002.

J. Panevová a kolektiv autorů. *Mluvnice současné češtiny 2. Syntax češtiny na základě anotovaného korpusu*. Karolinum, Praha, 2014.

V. Petkevič. Reliable Morphological Disambiguation of Czech: Rule-Based Approach is Necessary. In M. Šimková, red., *Insight into the Slovak and Czech Corpus Linguistics*, s. 26–44. Veda, Bratislava, 2006.

R. G. Piotrowskij. *Text, Computer, Mensch*. Brockmeyer, Bochum, 1984.

M. Popel a Z. Žabokrtský. TectoMT: Modular NLP Framework. In *Proceedings of IceTAL, 7th International Conference on Natural Language Processing, Reykjavík, Iceland, August 17, 2010*, s. 293–304, 2010. URL http://ufal.mff.cuni.cz/~popel/papers/2010_icetal.pdf.

I. I. Popescu. Text Ranking by the Weight of Highly Frequent Words. In P. Grzybek a R. Köhler, red., *Exact Methods in the Study of Language and Text*, s. 555–565. Mouton de Gruyter, Berlin –- New York, 2007.

I. I. Popescu a G. Altmann. Thematic Concentration in Texts. In E. Kelih, V. Levickij a Y. Matskulyak, red., *Issues in Quantitative Lingustics*, svazek 2, s. 110–116. RAM-Verlag, Lüdenscheid, 2011.

I. I. Popescu, G. Altmann, P. Grzybek, B. D. Jayaram, R. Köhler, V. Krupa, J. Mačutek, R. Pustet, L. Uhlířová a M. N. Vidya. *Word Frequency Studies*. Mouton de Gruyter, Berlin – New York, 2009a.

I. I. Popescu, J. Mačutek a G. Altmann. *Aspects of Word Frequencies*. RAM-Verlag, Lüdenscheid, 2009b.

I. I. Popescu, G. Altmann a R. Köhler. Zipf´s Law – Another View. *Quality and Quantity*, 44: 713–731, 2010.

I. I. Popescu, R. Čech a G. Altmann. *The Lambda-Structure of Texts*. RAM-Verlag, Lüdenscheid, 2011.

I. I. Popescu, R. Čech a G. Altmann. Some Geometric Properties of Slovak Poetry. *Journal of Quantitative Linguistics*, 19(2):121–131, 2012.

W. V. O. Quine. *Hledání pravdy*. Herrmann & synové, Praha, 1991.

A. Rapoport. Zipf's Law Re-visited. In H. Guiter a M. V. Arapov, red., *Studies on Zipf's Law*, s. 1–28. Brockmeyer, Bochum, 2011.

D. A. Ratkowsky a L. Hantrais. Tables for Comparing the Richness and Structure of Vocabulary in Texts of Different Length. *Computers and Humanities*, 9:69–75, 1975.

R. Rorty. Zkoumání jako rekontextualizace. In H. Guiter a M.V. Arapov, red., *Studies on Zipf's Law*, s. 147–171. Filosofia, Praha, 1998.

R. Rorty. *Filosofie a zrcadlo přírody*. Academia, Praha, 2012.

H. Sanada. Thematic Concentration in Japanese Prose. In I. Obradovic, E. Kelih a R. Köhler, red., *Methods and Applications of Quantitative Linguistics. Selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO), Belgrade, Serbia, April 26-29, 2012*, s. 130–140. University of Belgrade, Belgrade, 2013.

M. Scott. *WordSmith Tools version 6*. Lexical Analysis Software, Liverpool, 2011.

M. Scott a Ch. Tribble. *Textual Patterns. Key words and Corpus Analysis in Language Education*. John Benjamins, Amsterdam – Philadelphia, 2006.

H. Simon. On a Class of Skew Distribution Functions. *Biometrika*, 42:435–440, 1955.

J. Sinclair. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford, 1991.

D. Spoustová, J. Hajič, J. Votrubec, P. Krbec a P. Květoň. The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing ACL 2007, Praha*, s. 67–74, 2007.

U. Strauss, F. Fan a G. Altmann. *Kvantitativní lingvistika. Vybrané problémy* 1. Univerzita Palackého v Olomouci, Olomouc, 2014.

M. Stubbs. *Text and Corpus Analysis*. Wiley, Oxford, 1996.

M. Tešitelová. On the So-called Vocabulary Richness. *Prague Studies in Mathematical Linguistics*, 3:103–120, 1972.

M. Těšitelová. *Quantitative Linguistics*. Academia / John Benjamins, Praha / Amsterdam – Philadelphia, 1992.

M. Těšitelová, red. *Kvantitativní charakteristiky současné české publicistiky*. ÚJČ ČSAV, Praha, 1982.

M. Těšitelová, red. *Kvantitativní charakteristiky současné odborné češtiny*. ÚJČ ČSAV, Praha, 1983a.

M. Těšitelová, red. *Kvantitativní charakteristiky gramatických jevů v současné administrativě*. ÚJČ ČSAV, Praha, 1983b.

M. Těšitelová, red. *Psaná a mluvená odborná čeština z kvantitativního hlediska*. ÚJČ ČSAV, Praha, 1983c.

M. Těšitelová, red. *Současná česká administrativa z hlediska kvantitativního*. ÚJČ ČSAV, Praha, 1985.

J. Tuldava. Stylistics, Author Identification. In R. Köhler, G. Altmann a R. G. Piotrowski, red., *Studies on Zipf's Law*, s. 368–387. Mouton de Gruyter, Berlin-New York, 2005.

A. Tuzzi, I. I. Popescu a G. Altmann. *Quantitative Analysis of Italian Texts*. RAM-Verlag, Lüdenscheid, 2010.

F. J. Tweedie a R. H. Baayen. How Variable May a Constant Be? Measure of Lexical Richness in Perspective. *Computers and the Humanities*, 32:323–352, 1998.

L. Uhlířová. Length vs. Order: On Word Length and Clause Length from the Perspective of Word Order. In G. Altmann, J. Mikk, P. Saukkonen a G. Wimmer, red., *Linguistic structures. To honor J. Tuldava*, s. 266–275. Zwets, Lisse, 1997.

B. C. van Fraassen. *The Empirical Stance*. Yale University Press, 2002.

K. Veselovská a R. Čech. Opinion Target Identification Using Thematic Concentration of the Text. In *Contributed talk, QUALICO 2014, Olomouc, Czech Republic, May 29 - June 1, 2014*, 2014.

M. Weitzman. How Useful is the Logarithmic Type/Token Ratio? *Journal of Linguistics*, 7:237–243, 1971.

A. Wilson. Vocabulary Richness and Thematic Concentration in Internet Fetish Fantasies and Literary Short Stories. *Glottotheory*, 2(2):97–107, 2009.

G. Wimmer, G. Altmann, L. Hřebíček, S. Ondrejovič a S. Wimmerová. *Úvod do analýzy textov*. VEDA, Bratislava, 2003.

G. Wimmer. The Type-Token Relation. In R. Köhler, G. Altmann a R. G. Piotrowski, red., *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*, s. 361–368. Mouton de Gruyter, Berlin – New York, 2005.

L. Wittgenstein. *The statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge, 1944.

L. Wittgenstein. *Filosofická zkoumání*. Filosofický ústav AV ČR, Bratislava, 1993.

A. Ziegler a G. Altmann. *Denotative Textanalyse*. Praesens, Wien, 2002.

Š. Zikanová, L. Mladová, J. Mírovský a P. Jínová. Typical Cases of Annotators' Disagreement in Discourse Annotations in Prague Dependency Treebank. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta, 2002–2006*, 2010.

G. K. Zipf. *The Psycho-Biology of Language. An Introduction to Dynamic Philology*, svazek 2. Houghton-Mifflin / MIT Press, Boston / Cambridge, 1935.

G. K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, 1949.

Rejstřík