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 Institute of Formal and Applied Linguistics
as the 14th publication in the series
Studies in Computational and Theoretical Linguistics.

ISBN TODO-XXX-YYYY-ZZZZ-A

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.

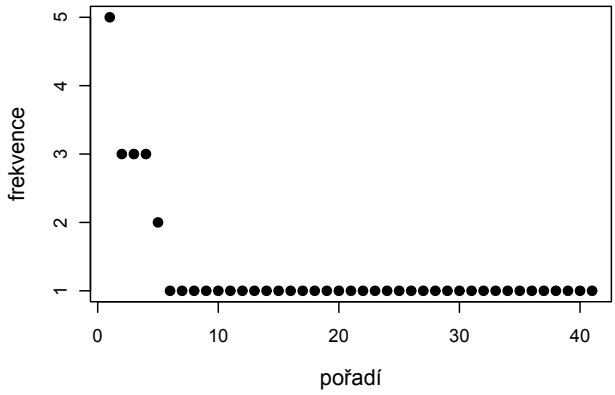
Tabulka 2.1: Ranková frekvenční distribuce slovních tvarů v básni J. Skácela *Odvaha k tomu* (text č. 200).

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

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:

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

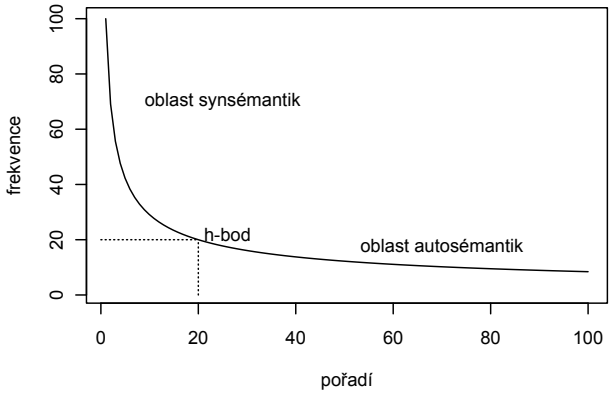
kde i a j jsou pořadí slov a $f(i)$ a $f(j)$ jsou jejich frekvence, přičemž $i < j$, kde i je největší takové číslo, pro které $i < f(i)$, a j je nejmenší takové číslo, pro které $j > f(j)$. V případě básně *Odvaha k tomu* je hodnota h rovna 3, protože třetí slovní tvar v pořadí má frekvenci $f = 3$ (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

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

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

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

h-bod byl použit 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.

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

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

Jak je vidět z Tab. 2.3, $h = 3$, přičemž nejfrekventovanějším slovním tvarem je výraz 'smuténka', který má frekvenci $f = 4$, 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

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

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.

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

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

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 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

[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.

se mohou výrazně lišit. Proto Popescu et al. (2009a) navrhují vzdálenost vynásobit právě frekvencí daného tematického slova, čímž činí analýzu přesnější:

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

kde $\mathrm{f}(\mathrm{r}')$ 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).

Tabulka 2.5: Ranková frekvenční distribuce slovních tvarů v 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

V obou tabulkách, Tab. 2.4 a 2.5, $\mathrm{h} = 6$ a nad h-bodem se nacházejí dvě tematická slova, v obou případech s totožným pořadím $\mathrm{r}' = 4$ a $\mathrm{r}' = 5$. 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 $\mathrm{N} = 134$ slov, $\mathrm{h} = 5,57$, přičemž jediným autosémantikem nad h-bodem je 'chvíle' ($\mathrm{r}' = 4$, $\mathrm{f} = 9$), článek *V Beskydech blesk zapálil chatu, vítr lámal stromy* (text č. 267) ($\mathrm{N} = 515$, $\mathrm{h} = 6$) s autosémantiky 'hasiči' ($\mathrm{r}' = 4$, $\mathrm{f} = 10$) a 'voda' ($\mathrm{r}' = 5$, $\mathrm{f} = 7$), a povídku K. Čapka *Pád rodu Votických* (text č. 797; $\mathrm{N} = 2443$, $\mathrm{h} = 17$), v níž jsou nad h-bodem autosémantika 'pan' ($\mathrm{r}' = 6$, $\mathrm{f} = 35$), 'dr.' ($\mathrm{r}' = 8$, $\mathrm{f} = 30$), 'Mejzlík' ($\mathrm{r}' = 9$, $\mathrm{f} = 29$), 'archivář' ($\mathrm{r}' = 11$, $\mathrm{f} = 25$), 'pane' ($\mathrm{r}' = 4$,

$f = 9$).

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'.

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

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

Tabulka 2.7: Tematická váha jednotlivých slov textů rozdílné délky.

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

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).

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

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

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.

$$TV_{V\ Beskydech...} = TV_{hasiči} + TV_{voda} = 2\frac{(6-4)10}{6(6-1)18} + 2\frac{(6-5)7}{6(6-1)18} = 0,07407 + 0,02593$$
$$= 0,1.$$

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:
$TK_{Svítání} = 0,10096$;
$TK_{V\ Beskydech...} = 0,1$;
$TK_{Pád\ rodu...} = 0,10924$.
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.

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

Mimochodem, text s teoreticky nejvyšší hodnotou tematické koncentrace, $TK = 1$, 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

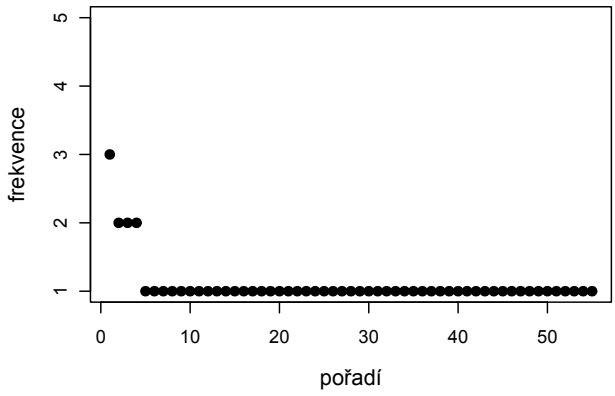
a všechno čisté polím dal.
A zaplakal bych plný studu.
Čistota jasná na polích.
To tiché nebe... Jednou budu...
Umřeme všichni pro ten sníh.

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

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

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 odchylka 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



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

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 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

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

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

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

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.

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 prostřed-

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

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

nictvím 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],

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

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

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)

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

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

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

$$Var(TK_{\textit{V Beskydech...}}) = \left(\frac{2}{6 \cdot (6-1) \cdot 18}\right)^2 \cdot (10 + 7) \cdot 0,24221 = 0,000056,$$

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

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

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

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.

3

Jiné způsoby měření tematické koncentrace textu

Metoda měření tematické koncentrace, tak jak je popsána v kap. 2, samozřejmě není ani jediným ani „nejlepším“ způsobem, jak zachytit tematické charakteristiky textu. Teoreticky existuje zřejmě nekonečně mnoho způsobů, jak tematické vlastnosti textu modelovat. Opíraje se o historicko-pragmatické pojetí filozofie vědy, reprezentované například Feyrabendem (2001), Kuhnem (1997), Rortym (2012), částečně i van Fraassenem (2002), v českém prostředí například Cvekem (2011a,b, 2012a,b), předpokládám, že hodnota *žádné* vědecké metody nemůže být stanovena poukazem na to, že modeluje nějaké *skutečné* vlastnosti sledovaného systému, tj. jeho podstatu. Totéž platí o pojmech (srov. Wittgenstein 1993; Quine 1991, Rorty 1998 aj). V rámci tohoto pohledu na povahu vědeckého bádání nejsou pojmy ničím více než námi vytvořenými nástroji, které nám umožňují „manipulovat“ s realitou, přičemž „manipulací“ je myšleno celé spektrum činností: od konverzace přes průmyslovou výrobu až po vědecký experiment. Hodnota pojmů, metod, které s danými pojmy pracují, ba i celých pojmových systémů, pak není dána jejich *pravdivostí*, tj. schopností odrážet nějaké skutečné vlastnosti „objektivní reality“, ale jejich *užitečností*. Jak však intersubjektivně stanovit, co je užitečné a co ne? Zdá se, že jediným spolehlivým kritériem není racionální diskuze, nýbrž praxe, tj. experiment. Můžeme zřejmě donekonečna diskutovat, které pojmové systémy „lépe“ vyjadřují podstatu například syntaktických vztahů (např. složkové, dependenční), aniž bychom vůbec kdy našli nějaké spolehlivé kritérium. V případě experimentálního přístupu však máme jasný korektiv. Jistě lze například zajímavě debatovat o vztahu tematické koncentrace a délky textu (případně absenci tohoto vztahu) a argumentovat ve prospěch toho či onoho postoje celou řadou „racionálních“ argumentů. Pokud však provedeme experiment (za jasně definovaných podmínek), který ukáže, že mezi sledovanými vlastnostmi je (či není) vztah, nemůžeme říct, že s výsledky nesouhlasíme, protože naše racionální argumenty „svědčí“ o opaku. Jinými slovy, praxe se oddiskutovat nedá.

Takto přistupuji i ke zde prezentované analýze tematické koncentrace. Na jejím počátku samozřejmě stála určitá racionální úvaha týkající se vlastností frekvenční struktury textu. Provedené výzkumy (tj. praxe) však záhy ukázaly, že některé její limity zbytečně omezují možnosti její aplikace při analýze textů. Proto byly navrženy její modifikace (Čech et al. 2015), které rozšiřují pole možností analýzy tematických charakteristik textu.

V prvé řadě jde o to, že výše navržený způsob měření (kap. 2) kategoriálně rozděluje analyzované texty do dvou skupin:

1) tematicky koncentrované, tj. v případě, že se nad h-bodem objeví nějaké autosémantikum, pak platí $\mathsf{TK} > 0$;
2) tematicky nekoncentrované (zjednodušeně bychom je mohli označit za tematicky „neutrální"), tj. v případě, že se nad h-bodem žádné autosémantikum neobjeví, pak platí $\mathsf{TK} = 0$.

Z teoretického hlediska to samozřejmě nemusí být problém – texty s hodnotou $\mathsf{TK} = 0$ prostě nemají vyhraněné téma, které se projevuje ve frekvenčních charakteristikách, proto není nutné výše navržený způsob měření TK (kap. 2) odmítat jako nevhodnou metodu. Vše záleží na badatelském cíli, k němuž metodu používáme.

Dále je třeba vzít na vědomí, že a) h-bod reprezentuje jen přibližnou hranici mezi autosémantiky a synsémantiky (tudíž text, u nějž se v rankové frekvenční distribuci vyskytne autosémantikum bezprostředně za h-bodem, je tematicky vyhraněnější než text, u nějž se v rankové frekvenční distribuci autosémantikum vyskytne mnohem dále od něj), b) o textech s hodnotou $\mathsf{TK} = 0$ se nedozvídáme nic jiného, než že jsou tematicky „neutrální", což může být v určitých případech z praktického hlediska dost neuspokojivé.

Další problém spočívá v otázce aplikace statistických testů. U textů s $\mathsf{TK} > 0$ se nad h-bodem často (zejména u textů kratších) vyskytuje pouze jediné autosémantikum. V tom případě hodnota $\mathrm{Var}(\mathsf{TK}) = 0$ (viz vzorec (2.9)), což znamená, že nelze statisticky testovat rozdíly mezi dvojicí textů, z nichž každé má nad h-bodem pouze jediné autosémantikum – srov. vzorec (2.11), u nějž by byla v takovém případě ve jmenovateli nula. Testovat dvojici textů, z nichž jeden $\mathrm{Var}(\mathsf{TK}) = 0$ a druhý $\mathrm{Var}(\mathsf{TK}) > 0$ (tj. v textu se objevují alespoň dvě autosémantika s rozdílnou frekvencí), sice lze, ale z hlediska statistiky se nejedná o příliš vhodný postup.

Z těchto důvodů byly navrženy (Čech et al. 2015) variantní metody analýzy tematické koncentrace: *sekundární tematická koncentrace textu* (kap. 3.1) a *proporcionální tematická koncentrace textu* (kap. 3.2). Tyto metody se snaží eliminovat některé z výše uvedených nedostatků TK.

V následujících částech bude prezentována aplikace těchto metod při analýze 15 textů – jedná se o novinové sloupky K. Čapka z cyklu *Jak se co dělá* (texty č. 974–988) – přičemž cílem je jednak sledovat výhody/nevýhody jednotlivých metod, jednak porovnat výsledky měření. Na rozdíl od kap. 2, kde byly představeny principy metody měření tematické koncentrace, zde budou použity lemmatizované texty, protože lemma se jeví být vhodnější jednotkou pro analýzu tohoto druhu (srov. kap. 4). Dále, ve shodě s naprostou většinou doposud publikovaných studií budou za tematická slova (to platí jak pro slovní tvary, tak lemmata) považována nikoliv všechna autosémantika nad h-bodem, ale pouze substantiva a jejich predikáty prvního řádu, tj. adjektiva a verba (s výjimkou sloves 'být', 'mít', 'moci', 'muset', 'smět').

Hodnoty TK a $\mathrm{Var}(\mathsf{TK})$ 15 výše uvedených Čapkových textů jsou uvedeny v Tab. 3.1. Jak je vidět, z těchto 15 textů je možné prostřednictvím statistických testů adek-

vátně porovnat pouze sedm: dva texty vykazují $TK = 0$, šest textů má sice $TK > 0$, ale nad h-bodem se vyskytuje pouze jedno autosémantikum, tudíž $Var(TK) = 0$.

Tabulka 3.1: Hodnoty TK a $Var(TK)$ v 15 lemmatizovaných textech K. Čapka (texty č. 974–988).

text	počet tematických lemmat	TK	$Var(TK)$
Jak se dělají noviny	1	0,019006	0
Z čeho se skládají noviny	1	0,083916	0
O redakci	2	0,011031	0,00000077
Jak vzniká číslo ranních novin	0	0	0
Další činitelé	2	0,038499	0,00000450
Jak se dělá film	1	0,246154	0
Krátký, ale nutný výklad o lidech	2	0,409091	0,00247655
Honba za námětem	2	0,166667	0,00019693
Čtyři filmové náměty	3	0,033605	0,00000407
Od námětu k scenáriu	2	0,020219	0
Stavíme	1	0,008627	0
Točíme	3	0,047309	0,00000068
Jak se tedy dělá film	1	0,022409	0
V dílnách a laboratořích	0	0	0
Premiéra	2	0,383754	0,00250182

3.1 Sekundární tematická koncentrace textu

Metoda analýzy sekundární tematické koncentrace textu (dále STK) se od původní metody (viz kap. 2) liší pouze v tom, že hranice mezi synsémantiky a autosémantiky není určena h-bodem, ale jeho dvojnásobkem, tj. $2h$. Vychází se přitom z toho, že h-bod reprezentuje pouze přibližnou hranici mezi těmito dvěma skupinami slov, tudíž z teoretického hlediska je takové posunutí h-bodu obhajitelné. Samozřejmě by bylo možné namítnout, proč se h násobí hodnotou 2, ale nikoliv třeba 1,5, 1,8 či 2,1 atd. Zde je třeba otevřeně přiznat, že se jedná o pragmatické rozhodnutí, jehož cílem je eliminovat počet textů s hodnotou, v nichž se nad pevným bodem nevyskytuje žádné, případně pouze jedno tematické slovo. STK se vypočítá podle vzorce

$$STK = \sum_{j=1}^{T} \frac{(2h - r'_{(j)})f(r'_{(j)})}{h(2h-1)f(1)}, \tag{3.1}$$

(T je počet tematických slov, jejichž pořadí je menší než $2h$), tzn. jedná se o analogický postup jako v případě vzorce (2.7), s tím rozdílem, že jde o sumu všech tematických

slov nad 2h. Její variance je pak

$$Var(STK) = \frac{\left[\sum_{j=1}^{T} f(r'_{(j)})\right] m_{2r'}}{[h(2h-1)f(1)]^2}, \tag{3.2}$$

kde $m_{2r'}$ je rozptyl (druhý centrální moment) tematických slov nad 2h-bodem, srov. vzorec (2.10).

Aplikací tohoto postupu na výše uvedených 15 textů K. Čapka získáváme výsledky prezentované v Tab. 3.2. Na první pohled je patrné, že všechny texty mají hodnotu $STK > 0$ a že kromě jednoho textu je možné u všech textů vypočítat varianci, a tudíž i testovat rozdíly mezi nimi, viz vzorec (2.12). STK v tomto ohledu přináší nepochybnou výhodu oproti TK.

Tabulka 3.2: Hodnoty STK a $Var(STK)$ v 15 lemmatizovaných textech K. Čapka (texty č. 974–988).

text	počet tematických lemmat	STK	$Var(STK)$
Jak se dělají noviny	5	0,037465	0,00000061
Z čeho se skládají noviny	2	0,097756	0.00002414
O redakci	12	0,020599	0,00000006
Jak vzniká číslo ranních novin	2	0,009983	0,00000006
Další činitelé	6	0,071850	0,00000117
Jak se dělá film	2	0,153846	0,00001142
Krátký, ale nutný výklad o lidech	4	0,295756	0,00003627
Honba za námětem	4	0,139589	0,00000426
Čtyři filmové náměty	11	0,035260	0,00000026
Od námětu k scenáriu	4	0,044635	0,00000010
Stavíme	1	0,019210	0
Točíme	9	0,047579	0,00000027
Jak se tedy dělá film	5	0,106658	0,00000218
V dílnách a laboratořích	5	0,016667	0,00000061
Premiéra	2	0,258697	0.00000725

3.2 Proporcionální tematická koncentrace textu

Další alternativou původního způsobu měření je proporcionální tematická koncentrace (dále PTK), která se vypočítá podle vzorce

$$PTK = \frac{1}{N_h} \sum_{r'<h} f(r'), \tag{3.3}$$

kde N_h je frekvence všech slov nad h-bodem a $f(r')$ je frekvence tematického slova nad h-bodem. PTK tedy vyjadřuje proporci frekvencí tematických slov nad h-bodem vzhledem k sumě frekvencí všech slov nad tímto bodem. Jak je ze vzorce patrné, prostřednictvím této metody se neřeší problém nepřítomnosti tematických slov nad h-bodem. To znamená, že pokud se v původním výběru (Tab. 3.1) vyskytovaly dva texty s $TK = 0$, tak při použití proporcionální tematické koncentrace budou mít ty samé texty opět hodnotu PTK rovnou nule, viz Tab. 3.3. Na rozdíl od TK je však možné (protože jde o proporci) vypočítat variance i pro texty, v nichž se nad h-bodem vyskytuje pouze jediné slovo,

$$Var(PTK) = \frac{PTK(1 - PTK)}{N_h}. \tag{3.4}$$

Při aplikaci PTK na výše uvedených 15 textů K. Čapka získáváme výsledky prezentované v Tab. 3.3. Z tabulky je jasné, že kromě dvou tematicky „neutrálních" textů (tj. $PTK = 0$) je možné statisticky testovat rozdíly mezi všemi ostatními texty.

Tabulka 3.3: Hodnoty PTK a $Var(PTK)$ v 15 lemmatizovaných textech K. Čapka (texty č. 974–988).

text	počet tematických lemmat	PTK	$Var(PTK)$
Jak se dělají noviny	1	0,070270	0,00353150
Z čeho se skládají noviny	1	0,155172	0,00147296
O redakci	2	0,054890	0,00004028
Jak vzniká číslo ranních novin	0	0	0
Další činitelé	2	0,136150	0,00040980
Jak se dělá film	1	0,235294	0,00233676
Krátký, ale nutný výklad o lidech	2	0,428571	0,00376766
Honba za námětem	2	0,233871	0,00100098
Čtyři filmové náměty	3	0,111702	0,00010094
Od námětu k scenáriu	2	0,102564	0,00034217
Stavíme	1	0,069620	0,00032225
Točíme	3	0,130795	0,00014707
Jak se tedy dělá film	1	0,105263	0,00083347
V dílnách a laboratořích	0	0	0
Premiéra	2	0,400000	0,00413793

Na první pohled se zdá logické, že nejlepším způsobem měření tematických charakteristik by mohla být kombinace STK a PTK, tj. metoda *sekundární proporcionální tematické koncentrace textu* (SPTK), protože by vzrostla pravděpodobnost, že budou eliminovány

jak nulové hodnoty samotné SPTK, tak i variance $Var(SPTK)$. Konkrétně

$$SPTK = \frac{1}{N_{2h}} \sum_{r'<2h} f(r'), \tag{3.5}$$

a

$$Var(SPTK) = \frac{SPTK(1 - SPTK)}{N_{2h}}. \tag{3.6}$$

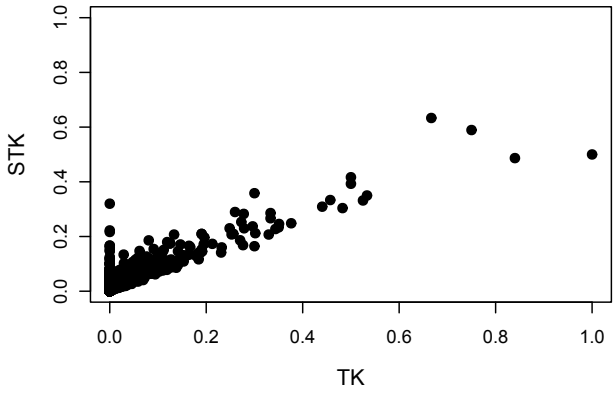
Vycházejíce z tohoto předpokladu, Čech et al. (2015) aplikovali i tuto metodu. Korelační analýza výsledků měření jednotlivých indexů (TK, STK, PTK, SPTK) však ukázala, že SPTK zřejmě měří jiné charakteristiky textu než TK, STK a PTK. Proto se dále touto metodou v této knize nebudu zabývat.

3.3 Porovnání odlišných způsobů měření tematické koncentrace

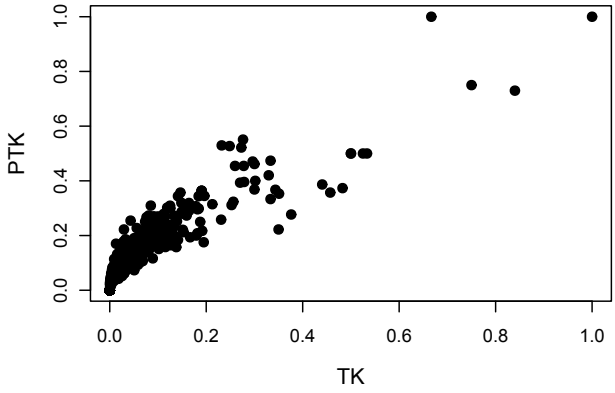
Různé testy, jak je ve statistice dobře známo, mohou vést k odlišným výsledkům. To samé platí i při aplikaci různých metod, jimiž se měří stejná vlastnost pozorovaného systému. Jedním ze způsobů, jak zjistit, zda různé metody měří to samé, byť s určitými odchylkami, je aplikace korelační analýzy. Vzhledem k tomu, že metody měření sekundární tematické koncentrace a proporcionální tematické koncentrace byly použity a srovnávány pouze jednou, a to na souboru pouhých 20 textů (srov. Čech et al. 2015), zaměřím se na důkladnější porovnání všech tří indexů. Konkrétně, budou porovnány hodnoty TK, STK a PTK u 1168 textů (jedná se o všechny texty uvedené v Příloze), a to jak nelemmatizovaných, tak lemmatizovaných, a vzájemný vztah mezi jednotlivými způsoby měření bude vyhodnocen pomocí Kendallova koeficientu korelace.

Tabulka 3.4: Korelační koeficienty mezi hodnotami jednotlivých indexů u 1168 textů (nelemmatizovaných i lemmatizovaných). Ve všech případech jsou výsledky signifikantní ($p < 0,001$), tj. mezi hodnotami jednotlivých indexů je monotónní závislost. Vzhledem k povaze dat (data nevykazují normální rozdělení) byl použit neparametrický Kendallův test.

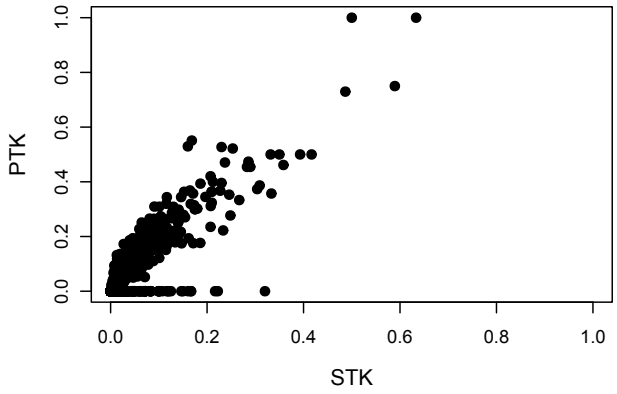
	τ
$TK_{nelem.}$–$STK_{nelem.}$	0,668
$TK_{nelem.}$–$PTK_{nelem.}$	0,938
$STK_{nelem.}$–$PTK_{nelem.}$	0,666
$TK_{lem.}$–$STK_{lem.}$	0,684
$TK_{lem.}$–$PTK_{lem.}$	0,889
$STK_{lem.}$–$PTK_{lem.}$	0,670



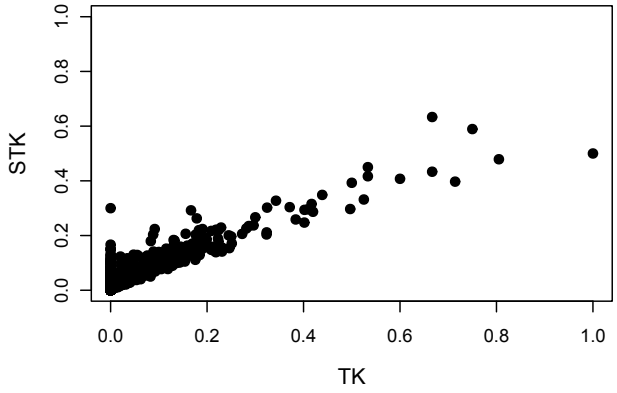
Obrázek 3.1: Vztah mezi TK a STK u 1168 nelemmatizovaných textů.



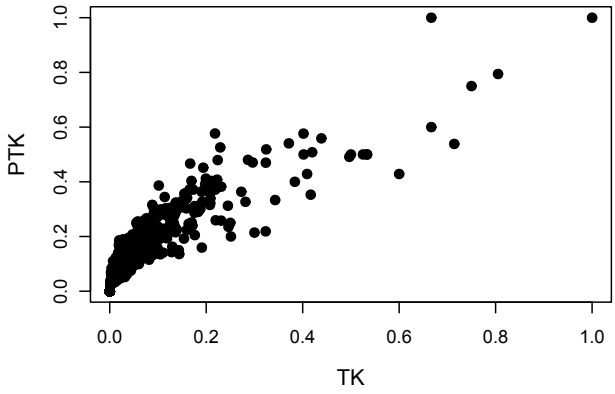
Obrázek 3.2: Vztah mezi TK a PTK u 1168 nelemmatizovaných textů.



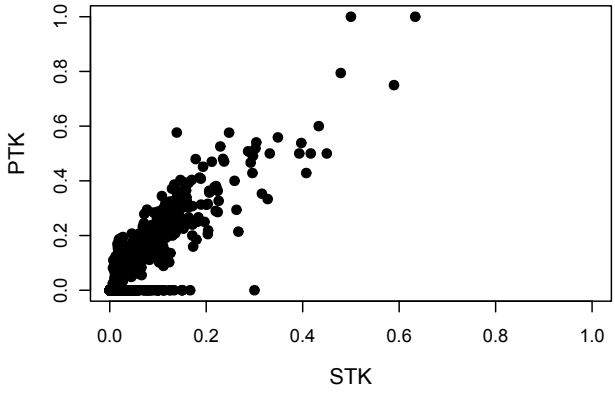
Obrázek 3.3: Vztah mezi STK a PTK u 1168 nelemmatizovaných textů.



Obrázek 3.4: Vztah mezi TK a STK u 1168 lemmatizovaných textů.



Obrázek 3.5: Vztah mezi TK a PTK u 1168 lemmatizovaných textů.



Obrázek 3.6: Vztah mezi STK a PTK u 1168 lemmatizovaných textů.

Jak je patrné z Tab. 3.4 a Obr. 3.1–6, u všech dvojic se projevuje vysoká míra korelace, což znamená, že všechny zvolené indexy měří stejnou vlastnost textu (v tomto případě tematickou zaměřenost autora na dané téma či témata), byť každá trochu jiným způsobem.

4

Tematická koncentrace a jazykové jednotky

Je nepochybné, že text, ať mluvený či psaný, je empiricky pozorovatelný objekt, který je možné dělit na určité části, tj. jednotky. V principu existuje nekonečně mnoho možností, jak text segmentovat. Samozřejmě jen několik málo je takových, které jsou smysluplné, přičemž kritériem smysluplnosti je zřejmě užitečnost. V „běžné" řeči například bez problému používáme pojem 'slovo', aniž bychom jej nějak *přesně* definovali – z kontextu je zpravidla jasné, o co nám jde. V případě odborných analýz jsme nuceni takovou definici podat, protože na ní obvykle záleží adekvátní interpretace výsledků bádání. Ale i v tomto případě se jedná o definice pracovní (např. vymezíme slovo jako akustickou, grafickou či významovou jednotku), jež nemají ambici na „ontologický" status, tj. nepředpokládá se, že takto vymezený pojem nějakým způsobem koresponduje s podstatou slova jako na člověku nezávislém fenoménu. Všechny snahy, v jejichž rámci se různí lingvisté o podobná vymezení pokusili, skončily nezdarem, srov. heslo „slovo" v Encyklopedickém slovníku češtiny: „Intuitivně vymezená základní jaz[*yková*] jednotka, vzhledem k formální a funkční různorodosti obtížně definovatelná. (Slovo *s*[*lovo*] pochází z běžného jazyka a ponechává si vágnost jeho pojmů i při pokusech o terminologizaci.) Tradičně uváděné charakteristiky nejsou obligatorní pro všechna *s*[*lova*] a zároveň zčásti odpovídají i jiným jaz[*ykovým*] jednotkám, lze jim proto přiznat pouze relativní platnost" (Hladká 2002, s. 424). Ve světle myšlenek tzv. postanalytické filosofie jazyka, jak je reprezentována např. Wittgensteinem (1993), Quienem 1991 či Rortym (1998, 2012) aj., není takový výsledek překvapením. Stručně řečeno, neexistuje žádné kritérium, na jehož základě by mohlo být rozhodnuto, zda daný pojem koresponduje/nekoresponduje s realitou. Z toho plyne jeden důležitý závěr: neexistují žádná na člověku nezávislá „data". Za data v tomto smyslu nelze považovat ani „amorfní" zvuk či soubor skvrn na papíře. V okamžiku, kdy mluvíme o 'zvuku' či 'skvrnách', totiž aplikujeme na tuto „amorfní" realitu (jež bezpochyby existuje, jak ukazuje praxe – např. silný zvuk vede k prasknutí ušních bubínků a hluchotě) námi vytvořené pojmy. A stejné je to v případě všech ostatních pojmů, jako jsou 'slovo', 'hláska', 'foném' atp. Jinými slovy, všechna data jsou našimi konceptuálními konstrukcemi, které nám umožňují manipulovat (v nejširším slova smyslu) s realitou. Data tak nejsou ničím jiným než klasifikací, jež nemá pravdivostní hodnotu.

V tomto duchu přistupuji i k volbě jazykových jednotek při analýze tematické koncentrace textu. Volba jazykové jednotky je volbou pragmatickou, a jak bude ukázáno v následujících řádcích, ani pragmatická kritéria nedovolují stanovit „nejlepší" jednotku (ve smyslu nejužitečnější) pro analýzu tohoto typu. S jistou nadsázkou si dovolím

tvrdit: „všechno je špatně“. V případě jakékoliv volby se totiž objevují problémy, které jsou v současné době neřešitelné. To ovšem neznamená, že je nutné celou analýzu tematické koncentrace zavrhnout. Naopak, znalost limitů metody znamená, že dané metodě lépe rozumíme a můžeme se například vyhnout neadekvátním interpretacím.

4.1 Slovní tvar a lemma

Vyjděme z předpokladu, že text je možné segmentovat na slova, přičemž slova mohou být reprezentována jednak jednotlivými tvary (např. ’pes’, ’psa’, ’psovi’), jednak základními podobami lexému, tzv. lemmaty (lemma ’pes’ reprezentuje množinu všech tvarů tohoto slova). Tento předpoklad je zcela v souladu s lingvistickou praxí, srov. mluvnice a slovníky. Dále, jak jednotlivé slovní tvary, tak lemmata mohou být nositeli tematických charakteristik textu, přičemž jejich frekvence nějakým způsobem reflektuje zaměření autora na hlavní téma či témata (srov. *Úvod*). Technicky (i z hlediska automatické analýzy textu) je samozřejmě nejjednodušší pracovat se slovními tvary. Ale i v jejich případě musí být jasně určeno, jakým způsobem je slovní tvar definován. Například můžeme slovní tvar vymezit jako každou sekvenci písmen, jejíž hranicí je mezera, jako jednotku gramatickou atp. V prvním případě dostáváme u výrazu ’budu zpívat’ dvě různé jednotky, ve druhém jednu. V této knize jsou slovní tvary definovány jako grafické jednotky, jejichž hranicí je mezera.

Na první pohled se volba slovních tvarů, jako jednotek reprezentujících téma(ta) textu, zejména u jazyků s bohatou flexí jistě nejeví jako nejvhodnější. Na druhou stranu, distribuce slovních tvarů je v rámci jednotlivých lemmat pravidelná (srov. Čech et al. 2014b), což znamená, že pro porovnávání tematických koncentrací jednotlivých textů není ani tato volba a priori volbou špatnou. Celkově je však eliminace vlivu flexe při analýze tematických charakteristik textu žádoucí, tudíž se jako vhodnější jednotky jeví například lemmata – dá se totiž předpokládat, že množina všech tvarů vyjadřuje dané téma komplexněji než jednotlivé instance daného lexému. Tento fakt by se měl mimo jiné projevit

a) v celkově menším počtu textů s nulovou hodnotou tematické koncentrace,
b) ve vyšší hodnotě TK, STK a PTK lemmatizovaných textů.

Ani v případě volby lemmatu však nestojíme před jednoznačnou volbou – za lemmata lze například považovat:

1) množinu všech tvarů bez ohledu na jakýkoliv jiný fakt; v tom případě ale výraz[1] ’stát’ patří pod jedno lemma ’STÁT’, ať se už reprezentuje substantivum či verbum;
2) množinu všech tvarů, přičemž se bere v úvahu slovní druh; v tom případě bude existovat jednak substantivní lemma ’STÁT’, jednak slovesné lemma ’STÁT’,

[1] Lemmatizovány jsou zde jednotlivé výrazy, které jsou definovány jako sekvence písmen oddělených mezerou. Analytický tvar slovesa ’narodil jsem se’ je tedy po lemmatizaci reprezentován třemi různými lemmaty: ’NARODIT’, ’BÝT’ a ’SE’. Tento přístup je ve shodě s lemmatizací používanou v Českém národním korpusu, více viz Hajič (2004), Petkevič (2006), Spoustová et al. (2007), Jelínek (2008).

3) množinu všech tvarů reprezentujících stejný význam daného lexému; v tom případě budou existovat např. lemmata 'HLAVA' (1) – ve významu část těla, 'HLAVA' (2) – ve významu osoby mající hlavní vliv v dané komunitě ('hlava rodiny'), 'HLAVA' (3) – ve významu části motoru atd. (mimochodem, *Slovník spisovného jazyka českého* (Havránek et al. 1989) uvádí 14 různých významů u hesla 'hlava').

V podobném třídění lze samozřejmě pokračovat i dále (je otázkou, jak např. naložit s víceslovnými výrazy, vlastními jmény atp.). Která volba je nejlepší, je však velmi obtížné určit. S narůstající granularitou narůstá jednak chybovost při automatické analýze, jednak klesá míra mezianotátorské shody, pokud se data zpracovávají manuálně. Je také obtížné určit, jak daleko vlastně při snaze zachytit jemné významové rozdíly zajít.

V následujících řádcích se zaměřím na porovnání tematických charakteristik nelemmatizovaných a lemmatizovaných textů a ověřím výše uvedené předpoklady týkající se eliminace nulových hodnot a očekávaných vyšších hodnot tematické koncentrace u lemmatizovaných textů. V této analýze bude použit nejjednodušší způsob lemmatizace (viz bod 1) výše). Ten se na první pohled jistě nejeví jako nejvhodnější, na druhou stranu výsledky tohoto způsobu lemmatizace při kvantitativních textologických analýzách nejsou signifikantně rozdílné například v porovnání se specializovaným (a mnohem sofistikovanějším) lemmatizátorem *Treex* (Popel a Žabokrtský 2010), jak uvádí Matlach (2014, s 47).

V Tab. 4.1 a 4.2 jsou porovnány počty textů s hodnotou $\mathrm{TK} = 0$ a $\mathrm{TK} > 0$ (resp. $\mathrm{STK} = 0$ a $\mathrm{STK} > 0$)[2] u 1168 nelemmatizovaných a lemmatizovaných českých textů. Z obou tabulek je evidentní, že počet textů s hodnotou daného indexu vyšší než nula je větší u lemmatizovaných textů – v případě TK jde o nárůst 210 textů, v případě STK o 185 textů vzhledem k nelemmatizovaným textům. Prostřednictvím chí-kvadrát testu byla testována významnost rozdílu a v obou případech se jedná o signifikantní rozdíl. V tomto ohledu lze tedy konstatovat, že použití lemmatizace vede k signifikantně významnému nárůstu počtu textů s hodnotou nenulové tematické koncentrace (při všech způsobech měření).

Tabulka 4.1: Porovnání počtu textů s $\mathrm{TK} = 0$ a $\mathrm{TK} > 0$ u nelemmatizovaných a lemmatizovaných textů. Mezi oběma skupinami je signifikantní rozdíl ($\chi^2 = 75,53$, $\mathrm{df} = 1$, p-hodnota $< 0,001$).

	$\mathrm{TK} = 0$	$\mathrm{TK} > 0$
nelemmatizované	696	472
lemmatizované	486	682

[2] V případě PTK dostáváme stejný počet textů s $\mathrm{PTK} = 0$ jako v případě TK, proto ho zde neanalyzuji zvlášť – co v tomto případě platí pro TK, platí i pro PTK.

Tabulka 4.2: Porovnání počtu textů s STK $= 0$ a STK > 0 u nelemmatizovaných a lemmatizovaných textů. Mezi oběma skupinami je signifikantní rozdíl ($\chi^2 = 86,96$, df $= 1$, p-hodnota $< 0,001$).

	STK $= 0$	STK > 0
nelemmatizované	343	825
lemmatizované	158	1010

Lemmatizace by měla vést nejen k určité eliminaci textů s nulovou tematickou koncentrací, ale i k nárůstu hodnot jednotlivých indexů (ve srovnání s nelemmatizovanými texty).[3] Proto se zaměřuji i na sledování průměrných hodnot jednotlivých indexů u obou typů textů a testuji rozdíly mezi těmito hodnotami. Výsledky jsou uvedeny v Tab. 4.3.

Tabulka 4.3: Průměrné hodnoty TK, STK a PTK u nelemmatizovaných a lemmatizovaných textů.

	nelemmatizované	lemmatizované
TK	0,02971	0,04147
s^2(TK)	0,00627	0,00786
STK	0,03555	0,05003
s^2(STK)	0,00388	0,00452
PTK	0,06149	0,09149
s^2(PTK)	0,01167	0,01452

Z Tab. 4.3 je patrné, že lemmatizované texty vykazují vyšší průměrné hodnoty u všech sledovaných indexů. Prostřednictvím u-testu, viz vzorec (2.12), je možné statisticky testovat rozdíly mezi průměrnými hodnotami nelemmatizovaných a lemmatizovaných textů:

$$|u| = \frac{|\overline{\text{TK}}_{\text{nelemmat.}} - \overline{\text{TK}}_{\text{lemmat.}}|}{\sqrt{\frac{s^2_{\text{nelemmat.}}}{n_{\text{nelemmat.}}} + \frac{s^2_{\text{lemmat.}}}{n_{\text{lemmat.}}}}} = \frac{|0,02971 - 0,04147|}{\sqrt{\frac{0,006269}{1168} + \frac{0,007863}{1168}}} = 3,38,$$

$$|u| = \frac{|\overline{\text{STK}}_{\text{nelemmat.}} - \overline{\text{STK}}_{\text{lemmat.}}|}{\sqrt{\frac{s^2_{\text{nelemmat.}}}{n_{\text{nelemmat.}}} + \frac{s^2_{\text{lemmat.}}}{n_{\text{lemmat.}}}}} = \frac{|0,03555 - 0,05003|}{\sqrt{\frac{0,003882}{1168} + \frac{0,004523}{1168}}} = 5,4,$$

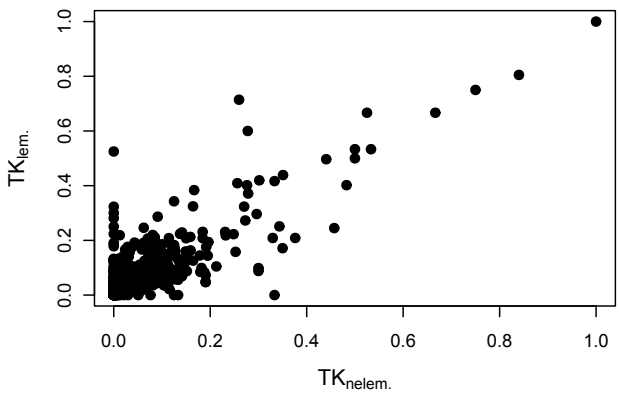
$$|u| = \frac{|\overline{\text{PTK}}_{\text{nelemmat.}} - \overline{\text{PTK}}_{\text{lemmat.}}|}{\sqrt{\frac{s^2_{\text{nelemmat.}}}{n_{\text{nelemmat.}}} + \frac{s^2_{\text{lemmat.}}}{n_{\text{lemmat.}}}}} = \frac{|0,06149 - 0,09149|}{\sqrt{\frac{0,011667}{1168} + \frac{0,014518}{1168}}} = 6,34,$$

[3] Určitou roli zde hraje také vliv lemmatizace na hodnotu h-bodu, srov. Kelih et al. (2014).

Ve všech případech je mezi průměrnými hodnotami TK, STK a PTK nelemmatizovaných a lemmatizovaných textů signifikantní rozdíl ($|u| > 1,96$; hladina významnosti $\alpha = 0,05$). To znamená, že lemmatizace (jako prostředek eliminace vlivu flexe) transformuje text do takové podoby, že umožňuje zachytit statisticky významně rozdílnější tematické charakteristiky textu (alespoň v průměru, podrobněji viz níže). Vysoký korelační koeficient mezi hodnotami TK, STK a PTK nelemmatizovaných a lemmatizovaných textů (viz Tab. 4.4 a Obr. 4.1–3) zase ukazuje na souvislost mezi hodnotami sledovaných indexů.

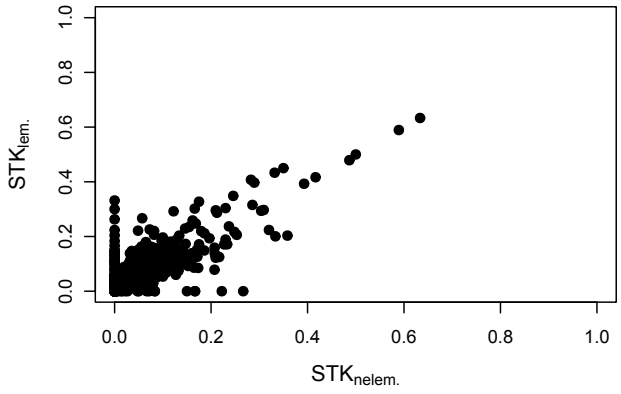
Tabulka 4.4: Korelační koeficienty mezi hodnotami TK, STK a PTK u nelemmatizovaných a lemmatizovaných textů. Ve všech případech jsou výsledky signifikantní (p-hodnota $< 0,001$), tj. mezi hodnotami jednotlivých indexů u nelemmatizovaných a lemmatizovaných textů je monotónní závislost. Vzhledem k povaze dat (data nevykazují normální rozdělení) byl použit neparametrický Kendallův test.

	τ
$TK_{nelem.}$–$TK_{lem.}$	0,648
$STK_{nelem.}$–$STK_{lem.}$.	0,621
$PTK_{nelem.}$–$PTK_{lem.}$	0,644

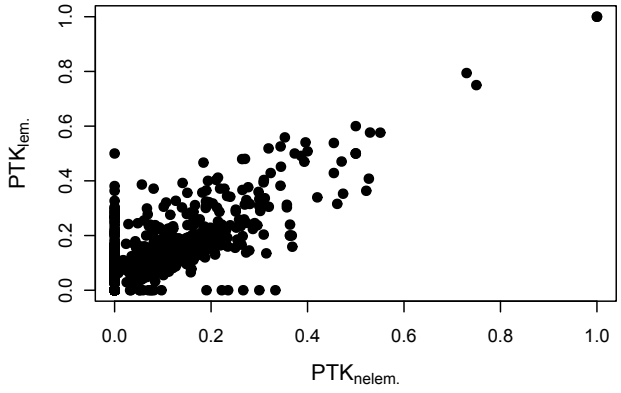


Obrázek 4.1: Vztah mezi hodnotami TK u 1024 nelemmatizovaných a lemmatizovaných textů.

Výsledky ukazují na jednoznačnou celkovou tendenci, tj. lemmatizace znamená vyšší

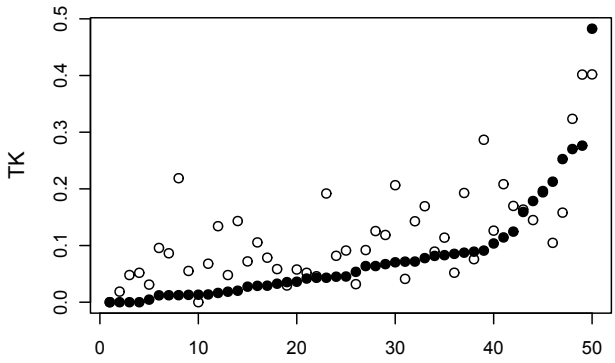


Obrázek 4.2: Vztah mezi hodnotami STK u 1024 nelemmatizovaných a lemmatizovaných textů.



Obrázek 4.3: Vztah mezi hodnotami PTK u 1024 nelemmatizovaných a lemmatizovaných textů.

hodnoty tematické koncentrace. U jednotlivých textů však může nastat situace, kdy má lemmatizovaný text nižší hodnotu. Pro ilustraci sledujme nejdříve hodnoty TK u nelemmatizovaných a lemmatizovaných 50 odborných textů (texty č. 323–373). I když



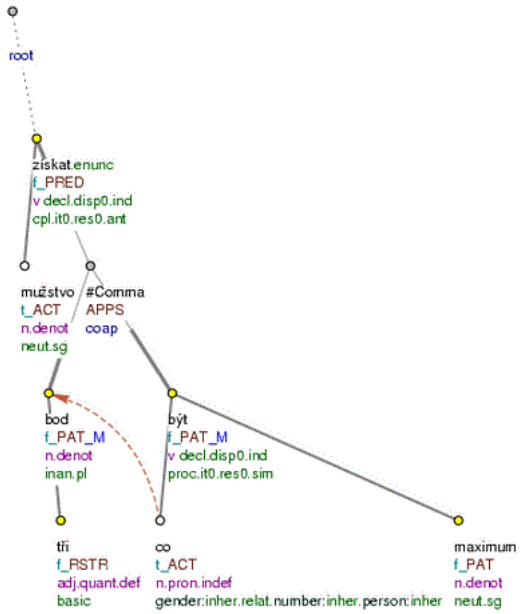
Obrázek 4.4: Hodoty TK u 50 nelemmatizovaných a lemmatizovaných odborných textů (texty č. 323–372). Světlejší body reprezentují lemmatizované texty.

jsou výsledky prezentované na Obr. 4.4 v souladu s celkovou tendencí, v daném souboru se vyskytuje osm textů, v nichž je hodnota TK u nelemmatizovaných textů vyšší než u textů lemmatizovaných. Jedná se o texty, kde se nad ħ-bodem vyskytuje velmi frekventované slovo (či slova) bez flexe, zpravidla jde o oborné termíny (např. 'Treg' v textu V. Hořejšího *Jak vyrobit regulační T-lymfocyty*, text č. 332) nebo o slova s velkou tvarovou homonymií (např. slovo 'záření' v textu F. Vožeha *Elektrosmog – co o něm dosud víme a nevíme?*, text č. 370).

4.2 Koreferenční jednotka

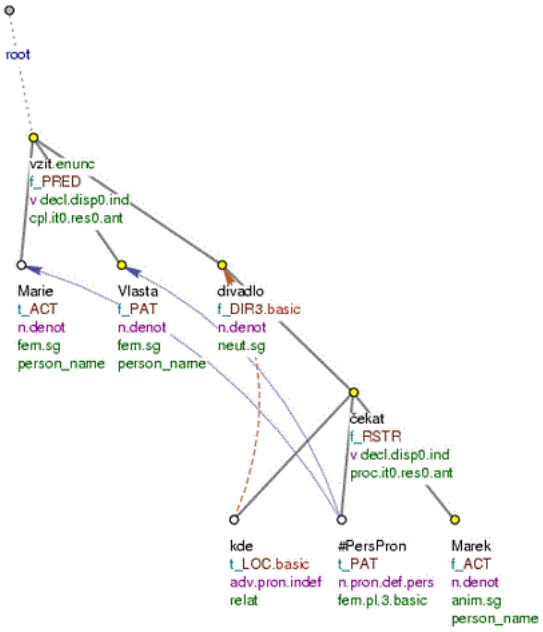
Termín koreference obecně označuje referenci dvou nebo více jednotek ke stejnému denotátu. Koreference se zpravidla dělí na gramatickou, při níž se použití koreferenčních jednotek řídí gramatickými pravidly (srov. Panevová et al. 2014, kap. 5; Nedoluzhko 2011), a textovou, jež se vyjadřuje anaforickými a kataforickými jazykovými prostředky, přičemž hranice mezi oběma typy není ostrá (Panevová 2002). Určení koreferenčních členů je mnohdy velmi problematické (to se týká zejména textové koreference), protože významnou roli hraje také situační kontext. Tento fakt významně omezuje možnosti použití koreferenčních jednotek při analýze tematické koncentrace, protože automatické nástroje se vyznačují relativně vysokou chybovostí a u manuální anotace se zase nepochybně projeví v nejednotnosti anotací u různých anotátorů (srov. Lee et al. 2006; Mírovský et al. 2010; Zikánová et al. 2010; Nedoluzhko et al. 2013).

Stejně jako v případě lemmatizovaných textů, i zde se zdá rozumné předpokládat, že texty, v nichž jsou jednotky společným referentem sloučeny pod jednu koreferenční jednotku, budou vykazovat vyšší hodnoty tematické koncentrace nejen než texty nelemmatizované, ale i lemmatizované. Tento předpoklad bude ověřen na analýze deseti textů z *Pražského závislostního korpusu PDT* 3.0 (Bejček et al. 2013), přičemž pro stanovení koreferenčních jednotek bude použita anotace tohoto korpusu (Nedoluzhko a Mírovský 2011). Konkrétně bude použita gramatická i textová koreference, při určování koreferenčních jednotek budou přitom započítávány i povrchově nevyjádřené jednotky zaznamenané na tzv. tektogramatické (hloubkové) struktuře věty. Pro ilustraci zde uvádím příklad použité anotace pro zachycení jak gramatické, tak textové koreference, viz Obr. 4.5 a 4.6.



Obrázek 4.5: Příklad gramatické koreference v tektogramatickém stromu vyjadřujícím hloubkovou strukturu věty 'Mužstvo získalo tři body, což je maximum'. Jednotlivé uzly reprezentují tzv. tektogramatická lemmata. Výraz 'což' odkazuje k doplnění 'tři body', koreferenční vztah vede od uzlu pro výraz 'což' k uzlu pro slovo 'bod'. V tektogramatickém stromě je zachycena apozice mezi doplněním 'tři body' a klauzí 'což je maximum' (převzato z Mikulová et al. (2006), kap. 8.2, Obr. 8.13).

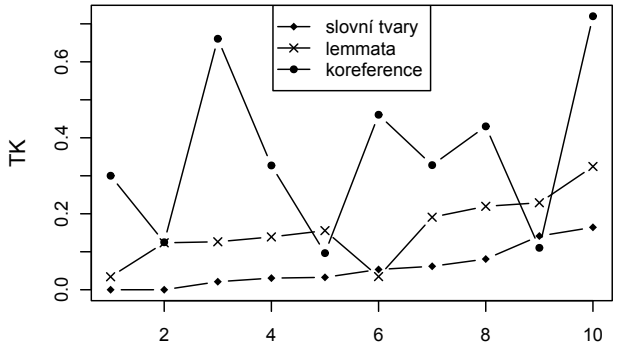
Pro analýzu tematické koncentrace koreferenčních jednotek bylo použito pět textů z hospodářského týdeníku *Českomoravský profit* (texty č. 1159–1163) a pět textů z *Lidových novin* (texty č. 1164–1168). Z každého textu byla vytvořena ranková frekvenční distribuce koreferenčních jednotek (tuto jednotku tvoří všechny koreferující jednotky tektogramatické roviny) a tektogramatických lemmat (tj. všechny výrazy, které nekoreferovaly, byly přiřazeny k patřičnému lemmatu) a byl určen h-bod. Všechny koreferenční jednotky, které obsahovaly alespoň jedno substantivum, adjektivum či verbum a jejichž pořadí bylo menší než hodnota h-bodu, byly označeny jako tematické koreferenční jednotky a u každého textu byly vypočítány indexy TK, STK a PTK (viz kap. 2.1, 3.1, 3.2). V Tab. 4.5 je pro ilustraci uvedena ranková frekvenční distribuce koreferenčně anotovaných jednotek a tektogramatických lemmat textu *Rusko zve zahraniční investory* (text č. 1167).



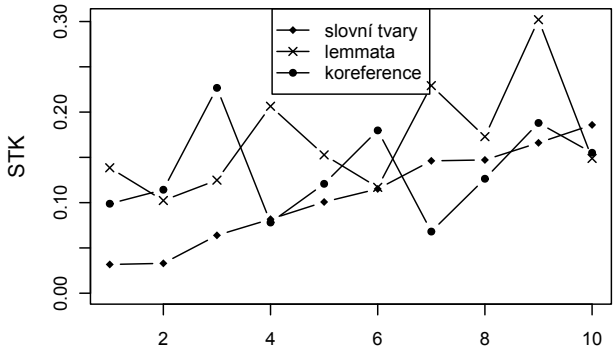
Obrázek 4.6: Příklad textové koreference v tektogramatickém stromu vyjadřujícím hloubkovou strukturu věty 'Marie vzala Vlastu do divadla, kde na ně čekal Marek'. Jednotlivé uzly reprezentují tzv. tektogramatická lemmata. Koreferovaným členem osobního zájmena 'na ně' (reprezentovaného v tektogramatickém stromě uzlem s tektogramatickým lemmatem # PersPron) jsou dva uzly ('Marie', 'Vlasta'), ke kterým je nutno odkázat jednotlivě (převzato z Mikulová et al. (2006), kap. 8.3.1.1, Obr. 8.93).

V Tab. 4.6–4.8 jsou pro porovnání prezentovány hodnoty TK, STK a PTK u nelemmatizovaných, lemmatizovaných a koreferenčně anotovaných textů. V případě TK a PTK je v souladu s výše uvedeným předpokladem nejvyšší hodnota obou indexů u koreferenčně anotovaných textů, v případě STK se daný předpoklad vůbec nepotvrdil. Pohled na grafické vyjádření vztahu mezi všemi indexy (Obr. 4.7–4.9) však naznačuje nízkou korelaci mezi koreferenčně anotovanými texty, na jedné straně, a texty nelemmatizovanými i lemmatizovanými, na straně druhé. To může být způsobeno tím, že transformace textu do podoby koreferenčních jednotek a tektogramatických lemmat představuje dost zásadní proměnu původního textu. Takto transformovaný text už možná není vhodný pro analýzu tematické koncentrace metodami popsanými v kap. 2 a 3.

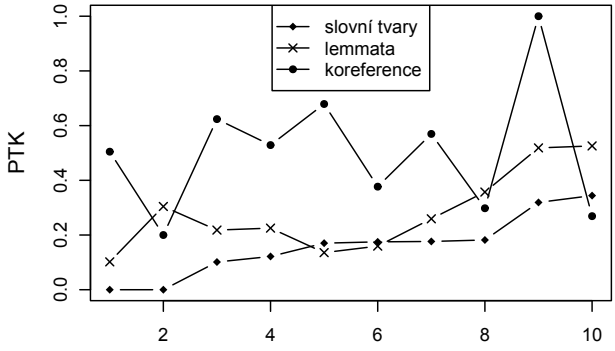
V prvé řadě je ale třeba podrobněji prozkoumat vztah mezi zjištěnými výsledky. Analogicky k postupu v kap. 4.1 budou proto nejdříve prostřednictvím u-testu, viz vzorec (2.13), statisticky testovány rozdíly mezi průměrnými hodnotami všech indexů. Takto bude ověřen původní předpoklad, tj. že koreferenční anotace by měla vykazovat významně vyšší hodnoty u všech měřených indexů. Následně bude vypočítán korelační koeficient – pokud se prokáže statisticky významná korelace, je možné předpokládat, že jsou měřeny stejné vlastnosti textu.



Obrázek 4.7: Hodoty TK u deseti nelemmatizovaných, lemmatizovaných a koreferenčně anotovaných publicistických textů. Viz Tab. 4.6. Texty jsou uspořádány ve vzestupném pořadí vzhledem k hodnotám TK u textů měřených na základě slovních tvarů.



Obrázek 4.8: Hodoty STK u deseti nelemmatizovaných, lemmatizovaných a koreferenčně anotovaných publicistických textů. Viz Tab. 4.7. Texty jsou uspořádány ve vzestupném pořadí vzhledem k hodnotám STK u textů měřených na základě slovních tvarů.



Obrázek 4.9: Hodoty PTK u deseti nelemmatizovaných, lemmatizovaných a koreferenčně anotovaných publicistických textů. Viz Tab. 4.8. Texty jsou uspořádány ve vzestupném pořadí vzhledem k hodnotám PTK u textů měřených na základě slovních tvarů.

Tabulka 4.5: Ranková frekvenční distribuce deseti nejfrekventovanějších koreferenčních jednotek (KJ) a tektogramatických lemmat (TL) v článku *Rusko zve zahraniční investory* (text č. 1167), $h = 8$. Hvězdičkou jsou označeny tematické koreferenční jednotky a tematická lemmata. V seznamu výrazů patřících do dané jednotky se objevují speciální výrazy tektogramatické roviny, jako jsou *#PersPron* (aktuální elipsa obligatorního aktantu), *#Gen* (nepřítomný všeobecný aktant).

r	jednotka	f	jednotlivé výrazy KJ nebo TL
1	KJ∗	18	*#PersPron si firem #PersPron podnikatelům si #PersPron jim investory jejich jich investorům investorů investorů podnikatelů investorů investory investory*
2	KJ∗	18	*Rusku velmoci Rusku Rusko Rusku ruský země Rusku Ruska federaci Rusku Rusku Ruská Rusku Rusku ruském ruské Ruska*
3	TL∗	18	*zahraniční zahraniční zahraniční zahraničních zahraničního zahraničních zahraničních zahraniční zahraničních zahraniční zahraničních zahraničních zahraničním zahraniční zahraničním zahraničních zahraniční zahraničního*
4	TL	16	*není Je je nejsou je jsou Není bylo je je být bude bude jsou jsou je*
5	TL	12	*#Gen #Gen #Gen #Gen #Gen #Gen #Gen #Gen #Gen #Gen #Gen*
6	TL	10	*a a a a a a a a a a*
7	KJ∗	9	*Černomyrdinovy Premiérův Černomyrdin on Černomyrdin jeho premiérovy Černomyrdinově premiérovi*
8	KJ	8	*#PersPron Opatření jež opatření která #PersPron #PersPron #PersPron*
9	TL	8	*lety let roce roku let let let let*
10	TL	6	*i i i i i i*

V Tab. 4.9–4.11 jsou prezentovány průměrné hodnoty všech indexů a výsledky statistických testů, pomocí nichž byly tyto hodnoty porovnávány. Ve všech třech případech mají lemmatizované texty signifikantně vyšší průměrné hodnoty ($|u| > 1,96$; hladina významnosti $\alpha = 0,05$), což je v souladu se zjištěními z předešlé kapitoly. V případě koreferenční anotace jsou průměrné hodnoty vyšší u TK a PTK, přičemž rozdíly jsou signifikantní. U STK vykazují nejvyšší hodnotu texty lemmatizované, průměrná hodnota tohoto indexu se u koreferenčně anotovaných textů signifikantně neliší jak vzhledem k textům lemmatizovaným, tak i nelemmatizovaným. Neočekávaně nízká hodnota STK u koreferenčně anotovaných textů je mimo jiné způsobena tím, že se v rankové frekvenční distribuci těchto textů vyskytují specifické uzly, například

Tabulka 4.6: Hodnoty TK v deseti publicistických textech (texty č. 1164–1168) měřené prostřednictvím slovních tvarů, lemmat a prostřednictvím koreferenční anotace. V šedých buňkách jsou nejvyšší hodnoty TK u každého textu, vzájemně porovnávány jsou jednotlivé způsoby měření, tj. údaje v jednotlivých řádcích. U každého textu je uvedeno identifikační číslo dokumentu (id) v *Pražském závislostním korpusu* PDT 3.0.

text	zdroj	id v PDT 3.0	slovní tvary	lemmata	korefer. anotace
Celní unie v ohrožení	ČMP	cmpr9410_001	0,053333	0,034632	0,460663
Stát – podnikatelé – nezaměstnanost	ČMP	cmpr9410_031	0,032634	0,155610	0,096429
Na život a na smrt – nejlépe po americku	ČMP	cmpr9413_029	0	0,034161	0,300132
Voda a teplo = peníze	ČMP	cmpr9413_049	0,030612	0,139037	0,327189
Podnikání v éteru	ČMP	cmpr9415_047	0,164310	0,324444	0,720000
Poklidné kompetence	LN	ln94200_105	0,061765	0,190927	0,328042
Je-li vypovídání smluv legální, je nutné novelizovat zákony	LN	ln94200_83	0	0,123589	0,125000
Podnikatelská banka nabírá dech	LN	ln94202_137	0,080808	0,219608	0,430070
Rusko zve zahraniční investory	LN	ln94202_20	0,021164	0,126254	0,660714
Jak statistický úřad počítá míru inflace	LN	ln94202_3	0,141429	0,229032	0,110318

tzv. zástupná lemmata, a ranková frekvenční distribuce těchto jednotek neodpovídá základnímu předpokladu analýzy tematické koncentrace, tj. že h-bod (případně 2h) v takové distribuci představuje hranici mezi autosémantiky a synsémantiky, viz Obr. 2.2. Výsledky korelační analýzy (Tab. 4.12) navíc ukazují, že koreferenčně anotované texty vykazují nízkou korelaci hodnot všech tří indexů vzhledem k dalším dvěma způsobům měření, srov. zejména vysoké p-hodnoty. Tato zjištění mě vedou k závěru, že koreferenční anotace v PDT 3.0 (spojená s použitím tektogramatických lemmat), tak jak zde byla použita, není vhodná pro analýzu tematické koncentrace. V tomto ohledu je třeba si zejména uvědomit, že tato anotace byla primárně vytvořena ke zcela jinému účelu. Pokud by se měla použít k měření tematické koncentrace, musela by být modifikována.

Tabulka 4.7: Hodnoty STK v deseti publicistických textech (texty č. 1164–1168) měřené prostřednictvím slovních tvarů, lemmat a prostřednictvím koreferenční anotace. V šedých buňkách jsou nejvyšší hodnoty STK u každého textu, vzájemně porovnávány jsou jednotlivé způsoby měření, tj. údaje v jednotlivých řádcích. U každého textu je uvedeno identifikační číslo dokumentu (id) v *Pražském závislostním korpusu* PDT 3.0.

text	zdroj	id v PDT 3.0	slovní tvary	lemmata	korefer. anotace
Celní unie v ohrožení	ČMP	cmpr9410_001	0,115152	0,116883	0,179766
Stát – podnikatelé –- nezaměstnanost	ČMP	cmpr9410_031	0,081731	0,206435	0,078056
Na život a na smrt – nejlépe po americku	ČMP	cmpr9413_029	0,032967	0,102355	0,114259
Voda a teplo = peníze	ČMP	cmpr9413_049	0,100733	0,152741	0,120791
Podnikání v éteru	ČMP	cmpr9415_047	0,166061	0,302020	0,188000
Poklidné kompetence	LN	ln94200_105	0,147154	0,172894	0,126374
Je-li vypovídání smluv legální, je nutné novelizovat zákony	LN	ln94200_83	0,031746	0,138503	0,098810
Podnikatelská banka nabírá dech	LN	ln94202_137	0,185859	0,148841	0,155048
Rusko zve zahraniční investory	LN	ln94202_20	0,063899	0,124843	0,226736
Jak statistický úřad počítá míru inflace	LN	ln94202_3	0,146167	0,229228	0,068067

4.3 Poznámka k agregátu/hrebu

Jednou z dalších možností segmentace textu je jeho rozdělení na tzv. agregáty (Hřebíček 1993, 1997, 2002) či hreby (Ziegler a Altmann 2002). Obecně se každý jednotlivý agregát či hreb skládá z množiny jednotek, které sdílejí nějakou společnou vlastnost. Agregátem či hrebem může být například množina vět, které obsahují aspoň jedno totožné lemma, množina slovních tvarů obsahujících stejný morfém, množina slabik obsahujících stejnou hlásku atp. Z hlediska analýzy tematické koncentrace se jeví jako nejadekvátnější přístup definovat agregát či hreb jako množinu slovních tvarů, které celé, či jejich část (morf) odkazují ke stejné entitě – Ziegler a Altmann (2002) tento přístup nazývají denotativní analýzou (podobně Wimmer et al. 2003, s. 297nn). Na první pohled je tento postup pro zkoumání tematických charakteristik textu velmi rozumný

Tabulka 4.8: Hodnoty PTK v deseti publicistických textech (texty č. 1164–1168) měřené prostřednictvím slovních tvarů, lemmat prostřednictvím koreferenční anotace. V šedých buňkách jsou nejvyšší hodnoty PTK u každého textu, vzájemně porovnávány jsou jednotlivé způsoby měření, tj. údaje v jednotlivých řádcích. U každého textu je uvedeno identifikační číslo dokumentu (id) v *Pražském závislostním korpusu* PDT 3.0.

text	zdroj	id v PDT 3.0	slovní tvary	lemmata	korefer. anotace
Celní unie v ohrožení	ČMP	cmpr9410_001	0,170213	0,135593	0,679012
Stát — podnikatelé — nezaměstnanost	ČMP	cmpr9410_031	0,181818	0,357143	0,297872
Na život a na smrt – nejlépe po americku	ČMP	cmpr9413_029	0	0,101852	0,504672
Voda a teplo = peníze	ČMP	cmpr9413_049	0,121622	0,225225	0,528846
Podnikání v éteru	ČMP	cmpr9415_047	0,319149	0,518519	1
Poklidné kompetence	LN	ln94200_105	0,175000	0,159420	0,376812
Je-li vypovídání smluv legální, je nutné novelizovat zákony	LN	ln94200_83	0	0,304348	0,200000
Podnikatelská banka nabírá dech	LN	ln94202_137	0,176471	0,259259	0,569620
Rusko zve zahraniční investory	LN	ln94202_20	0,101266	0,218045	0,623762
Jak statistický úřad počítá míru inflace	LN	ln94202_3	0,344086	0,525714	0,268966

Tabulka 4.9: Průměrné hodnoty TK u deseti publicistických textů (texty č. 1164–1168) a výsledky u-testu, viz vzorec (2.13). Tučně označené hodnoty vyjadřují signifikantní rozdíl ($|u| > 1,96$; hladina významnosti $\alpha = 0,05$).

	slovní formy	lemmata	koref. anotace
TK	0,058606	0,157729	0,355856
$s^2(TK)/n$	0,000314368	0,000786013	0,004751309
slovní formy	x		
lemmata	**2,99**	x	
koref. anotace	**4,18**	**2,66**	x

Tabulka 4.10: Průměrné hodnoty STK u deseti publicistických textů (texty č. 1164–1168) a výsledky u-testu, viz vzorec (2.13). Tučně označené hodnoty vyjadřují signifikantní rozdíl ($|u| > 1,96$; hladina významnosti $\alpha = 0,05$).

	slovní formy	lemmata	koref. anotace
STK	0,107147	0,169474	0,135591
$s^2(STK)/n$	0,000295546	0,000371605	0,00025962
slovní formy	x		
lemmata	**2,41**	x	
koref. anotace	1,21	1,35	x

Tabulka 4.11: Průměrné hodnoty PTK u deseti publicistických textů (texty č. 1164–1168) a výsledky u-testu, viz vzorec (2.13). Tučně označené hodnoty vyjadřují signifikantní rozdíl ($|u| > 1,96$; hladina významnosti $\alpha = 0,05$).

	slovní formy	lemmata	koref. anotace
PTK	0,158963	0,280512	0,504956
$s^2(PTK)/n$	0,001293383	0,002202294	0,005580281
slovní formy	x		
lemmata	**2,06**	x	
koref. anotace	**4,17**	**2,54**	x

Tabulka 4.12: Korelační koeficienty mezi hodnotami TK, STK a PTK u nelemmatizovaných, lemmatizovaných a koreferenčně anotovaných textů. Vzhledem k povaze dat (data nevykazují normální rozdělení) byl použit neparametrický Kendallův test. Tučně jsou označeny statisticky významné korelace (hladina významnosti $\alpha = 0,05$).

	τ	p-hodnota
$TK_{slov.\ formy} - TK_{lem.}$	**0,809**	0,0012
$TK_{slov.\ formy} - TK_{koref.}$	0,225	0,3692
$TK_{lem.} - TK_{koref.}$	0,111	0,7275
$STK_{slov.\ formy} - STK_{lem.}$	0,378	0,1557
$STK_{slov.\ formy} - STK_{koref.}$	0,244	0,3807
$STK_{lem.} - STK_{koref.}$	-0,200	0,4843
$PTK_{slov.\ formy} - STK_{lem.}$	**0,584**	0,0196
$PTK_{slov.\ formy} - STK_{koref.}$	0,044	0,8575
$PTK_{lem.} - STK_{koref.}$	-0,200	0,4843

– text je transformován pouze do tematických jednotek, přičemž tyto jednotky jsou vyhodnoceny postupem uvedeným v kap. 2 a 3 (srov. Čech et al. 2013b). Bohužel je však aplikovatelný jen na velmi krátké texty, protože vyžaduje ne příliš jednoduchou manuální anotaci. S rostoucí délkou textu je totiž velmi obtížné zaznamenat správně všechny jednotky patřící ke stejnému agregátu či hrebu (zejména na úrovni morfů). Dále předpokládám, že míra mezianotátorské shody (která u tohoto typu analýzy, pokud vím, nebyla doposud měřena) bude u delších textů nízká. Druhý důvod, který mě vede k tomu, že agregáty či hreby jako jednotky pro měření tematické koncentrace v této knize nepoužívám, je ten, že ranková frekvenční distribuce agregátů či jednotek neodpovídá základnímu předpokladu analýzy tematické koncentrace, tj. že h-bod (případně 2h) v takové distribuci představuje hranici mezi autosémantiky a synsémantiky, viz Obr. 2.2.

5
Tematická koncentrace a délka textu

Délka textu je v kvantitativní textologii faktorem, který výrazně ovlivňuje interpretaci kvantitativně založených způsobů měření a se kterým není lehké se adekvátně vyrovnat. V textologii obecně jde zejména o analýzu toho, čím se jednotlivé texty, případně skupiny textů – například žánry, texty určitého funkčního stylu –, liší (Těšitelová 1982, 1983a, 1983b, 1983c, 1985, 1992, s. 160nn). Může jít o distribuci slovních druhů, délku vět, frekvenční charakteristiky vybraných jednotek, slovní bohatství atp. Problém je ale v tom, že s rostoucí délkou textu zpravidla narůstá variabilita dat. Pozorované rozdíly mezi texty či skupinami textů pak nejsou výsledkem vlivu předpokládaných mechanismů (např. důsledkem žánru, autorství), ale délky textu. Jen málokdy má badatel možnost porovnávat stejně dlouhé texty, proto musí přistoupit k nějakému způsobu eliminace vlivu délky textu.

Nejjednodušším způsobem, jak daný problém vyřešit, je analyzovat pouze části textu, například počátečních 100 či 1000 slov – srov. nastavení *WordSmith Tools* pro analýzu poměru počtu výskytů různých slovních tvarů (types) a počtu všech slov v textu (tokens) (Scott 2011). Jak jsme uvedli v Čech et al. (2014a, s. 30), „[t]ento přístup je ovšem v mnohých ohledech neadekvátní. Zejména proto, že nebere v potaz tzv. homogenitu textu: představme si například detektivní povídku, ve které autor úmyslně použije velké množství nových slov v závěrečné části (např. v souvislosti s odhalením okolností zločinu); použít pro stanovení slovního bohatství počátečních sto či tisíc slov je v takovém případě jistě zavádějící." Jinou strategii při snaze o omezení vlivu délky volí Covington a McFall (2010), Kubát a Milička (2013) a Kubát (2014), kteří nejdou cestou transformace jednotlivých indexů (např. prostřednictvím logaritmizování), ale navrhují použít standardní, na délce textu evidentně závislé metody (jako je type-token poměr), novým způsobem. Konkrétně, sledovaný index aplikují na zvolenou velikost textu, tzv. okno (např. o délce 100 slov). Toto okno „posunují" vždy o jeden token od začátku textu do konce a počítají průměry zjištěných hodnot (více viz kap. 7.2). Dosavadní výsledky prezentované ve výše uvedených studiích ukazují, že se jedná o úspěšnou metodu eliminace vlivu délky.

Pokud uvažujeme o tematické koncentraci, je jistě na místě otázka po důvodech případné závislosti tematické koncentrace na délce textu. U extrémně krátkých textů, jako jsou například básně (srov. kap. 2), má autor nepochybně velkou možnost kontrolovat frekvenční charakteristiky použitého slovníku. Zde se dá tedy očekávat velká variabilita. Navíc, vzhledem k celkově nízkým frekvencím jednotlivých slov mají na hodnotu tematické koncentrace vliv i nepatrné změny frekvence slov – například

z Tab. 2.3 je patrné, že pokud by se slovní tvar 'smuténka' vyskytl jedenkrát méně, byla by hodnota $TK = 0$.

Ani u delších textů však není jednoduché odhadnout trend: s rostoucí délkou textu totiž sice roste frekvence synsémantik (vlivem gramatiky) i hodnota tzv. normalizační konstanty (tj. dělitel ve vzorci (2.5)), což by mělo vést k nižší hodnotě tematické koncentrace u delších textů. Na druhou stranu ale s délkou textu roste také hodnota h-bodu, tudíž narůstá pravděpodobnost, že se nad h-bodem objeví více autosémantik. Zdá se tedy, že zde fungují protikladné síly, které by měly vliv délky textu eliminovat, a že rozhodující roli zde bude hrát schopnost mluvčího či pisatele kontrolovat (v širokém slova smyslu) frekvenční charakteristiky slov v textu. Je ovšem otázka, u jak dlouhých textů je autor takové kontroly schopen – srov. pokus takovou hranici empiricky odvodit vzhledem k tzv. lambda frekvenční struktuře textu (Čech 2015). U extrémně dlouhých textů, jako jsou například mnohadílné romány psané několik let, pak vyvstává otázka, zda má vůbec smysl o tematické koncentraci (stejně jako jiné kvantitativní vlastnosti) uvažovat jako o vlastnosti textu, kterou má smysl použít pro analýzu rozdílů mezi texty.

V dalších řádcích bude zkoumán vztah délky textu a tematické koncentrace následujícím způsobem: 1) bude analyzováno všech 1168 textů dohromady; 2) bude analyzován kumulativní vývoj tematické koncentrace v jednotlivých textech. Ve všech případech budou použity lemmatizované texty.

5.1 Celková délka textu versus TK, STK a PTK

Na obrázcích 5.1–6 jsou graficky znázorněny vztahy mezi hodnotami TK, STK, PTK a délkou textu N u 1168 českých textů. Pro přehlednost je u každého grafu i varianta s logaritmizovaným měřítkem na ose x a y. Délka textu je počítána v počtu tokenů. V množině analyzovaných textů jsou texty o délce ležící v intervalu $N \in \langle 37; 96809 \rangle$ slov/tokenů.

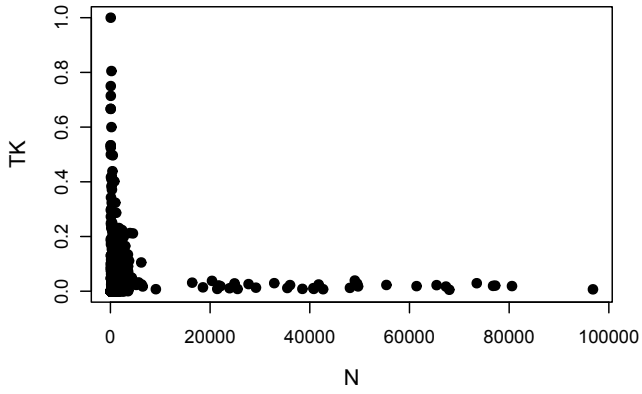
V případě TK a STK je vidět, že texty s $N < 200$ mají buď nulovou hodnotu daného indexu, nebo je tendence k vyšším hodnotám obou indexů – což je v souladu s tvrzeními uvedenými výše. U delších textů se projevuje velký rozptyl bez zjevné tendence.

U PTK se objevuje vztah mezi délkou textu N a minimální hodnotou tohoto indexu. To je dáno tím, že PTK vyjadřuje proporci frekvence tematických slov nad h-bodem vzhledem k frekvenci všech slov nad h-bodem, srov. vzorec (3.3): u velmi krátkých textů, kde se nad h-bodem vyskytuje málo slov, dochází k tomu, že je proporce frekvencí autosémantik velká i díky tomu, že hodnota h-bodu je malá a celkové množství slov nad h-bodem je menší než u textů delších – graficky se to projevuje jasnou linií minimálních hodnot u krátkých textů s $PTK > 0$. V tomto ohledu se pak PTK nejeví být vhodným nástrojem pro porovnávání textů s délkou $N < 2000$ slov, což je empiricky odvozená hranice, kde se přestává uvedená tendence projevovat (viz Obr. 5.12).

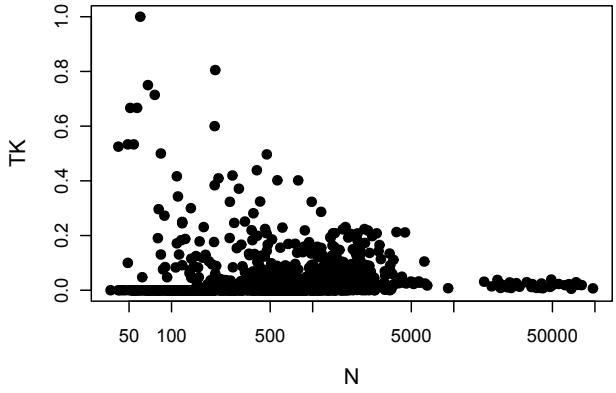
Bohužel se tím použití PTK evidentně omezuje a při aplikaci tohoto indexu je třeba brát v potaz případný vliv délky textu.

U všech třech sledovaných indexů je zjevné, že texty s délkou $N > 20000$ představují více méně homogenní skupinu – jedná se o romány, u kterých jsou hodnoty v relativně úzkém intervalu ($TK \in \langle 0,00531; 0,039\rangle$; $STK \in \langle 0,00663; 0,02960\rangle$; $PTK \in \langle 0,06277; 0,19473\rangle$). Jak jsem uvedl výše, je otázkou, zda má u tak dlouhých textů vůbec smysl tematickou koncentraci měřit, protože nelze předpokládat, že by autor byl schopen s jejich frekvenčními charakteristikami nějak „rozumně" manipulovat. Spíše by se dalo očekávat, že se zde bude projevovat na autorovi nezávislý mechanismus, srov.: „The longer the text, the more the writer loses his subconscious control over some proportions and keeps only the conscious control over contents, grammar, his aim, etc. But as soon as parts of control disappear, the text develops its own dynamics and begins to abide by some laws which are not known to the writer but work steadily in the background. The process is analogous to that in physics: if we walk, we consider our activity as something normal; but if we stumble, i.e. lose the control, gravitation manifests its presence and we fall. That means, gravitation does not work ad hoc in order to worry us maliciously, but it is always present, even if we do not realize it consciously. In writing, laws are present, too, and they work at a level which is only partially accessible. One can overcome their working, but one cannot eliminate them. On the other hand, if the writer slowly loses his control of frequency structuring, a new order begins to arise by selforganization or by some not perceivable background mechanism" (Popescu et al. 2012, 126–127). Každopádně se u této skupiny textů, jejichž délka $N > 20000$, v žádném ze sledovaných indexů neprojevuje výrazná závislost na délce.

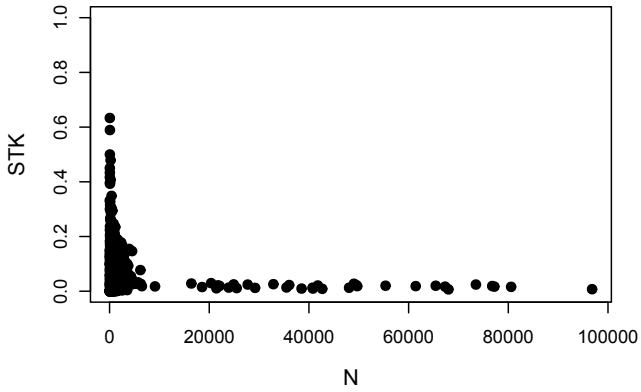
Ve studii zaměřené na vztah délky textu a lambda frekvenční struktury textu (Čech 2015) jsem se pokusil empiricky odvodit interval, v jehož rámci byl index lambda na délce textu nezávislý – pro 615 analyzovaných českých textů byl tento interval v rozsahu $N \in \langle 200; 6500\rangle$ slov/tokenů. Na základě této analýzy jsem se proto zaměřil na texty, jejichž délka leží v intervalu $N \in \langle 200; 6500\rangle$ slov/tokenů – výsledky jsou prezentovány na Obr. 5.7–12. V případě TK a STK se projevuje nezávislost na délce textů v tomto intervalu – přímka vyjadřující lineární vztah mezi oběma indexy a délkou je téměř vodorovná, nízký koeficient determinace svědčí o velkém rozptylu. V případě PTK se projevuje trend popsaný výše, navíc s narůstající délkou roste hodnota PTK. Z hlediska analýzy všech textů je tedy možné konstatovat, že nejvhodnějšími způsoby měření tematické koncentrace se jeví být indexy TK a STK, a to zejména pro texty, jejichž délka leží v intervalu cca 200–6500 slov/tokenů. U PTK se do cca $N = 2000$ projevuje jasná závislost na délce, což výrazně omezuje jeho použití. Protože značná část textů, které zde používám, má délku $N < 2000$, nebudu dále tento index aplikovat.



Obrázek 5.1: Vztah mezi hodnotou TK a délkou textu N u 1168 lemmatizovaných textů.



Obrázek 5.2: Vztah mezi hodnotou TK a délkou textu N u 1168 lemmatizovaných textů. Osa x je pro větší přehlednost výsledků logaritmizována.

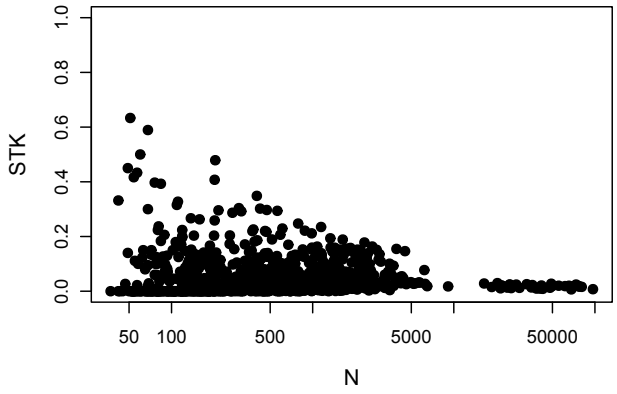


Obrázek 5.3: Vztah mezi hodnotou STK a délkou textu N u 1168 lemmatizovaných textů.

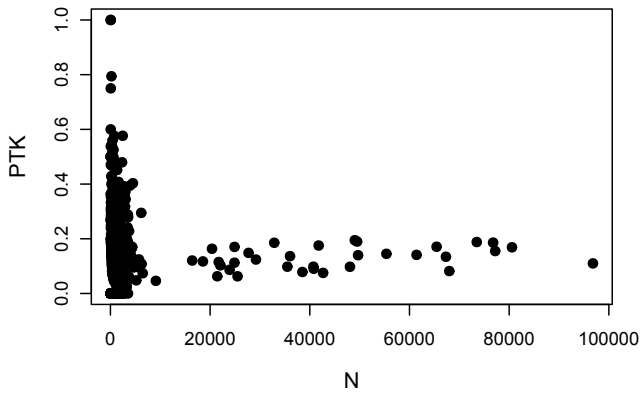
5.2 Kumulativní délka textu versus TK a STK

Analýza vztahu celkové délky textu a tematické koncentrace, jak byla prezentována v předchozí kapitole, sice neodhalila závislost mezi touto vlastností a indexy TK a STK (u indexu PTK je situace trochu komplikovanější, viz výše), ale při zkoumání vlivu délky textu na jakoukoliv jeho vlastnost musí být každý badatel velmi obezřetný. V prvé řadě jde o to, že při sledovaní této závislosti byla analyzována různorodá množina textů – jedná se o mix různých žánrů, autorů, stylů atd. Přestože se na první pohled může zdát výsledný obrázek jasný (srov. Obr. 5.1–5.12), bližší analýza může ukázat pravý opak, tj. závislost TK i STK na délce. V tomto ohledu je ilustrativní „osud“ tzv. lambda struktury textu, která se zdála být na základě analýzy více než tisíce textů pocházejících z různých jazyků evidentně na textu nezávislá (srov. Popescu et al. 2011). Podrobnější zkoumání (Čech 2015) však později odhalilo, že typologické rozdíly mezi jazyky zapříčinily nesprávnou interpretaci dat a že mezi daným indexem a délkou textů vztah existuje (byť svým způsobem neočekávaný – hodnota lambdy s délkou nejdříve stoupá, pak následuje interval relativní nezávislosti na délce a následně dochází k poklesu lambdy). Proto je třeba i v případě indexů tematické koncentrace možný vliv délky textu dále prozkoumat.

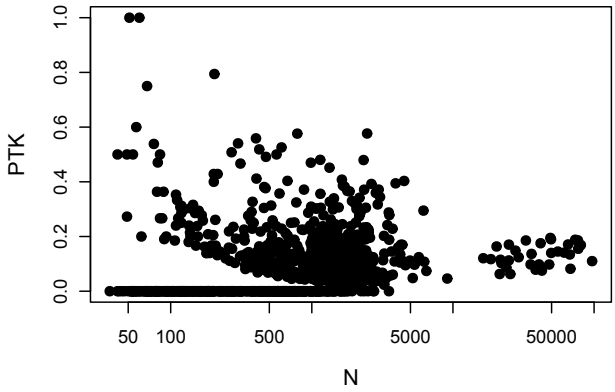
Zřejmě nejspolehlivější metodou, jak ověřovat (ne)závislost jakéhokoliv indexu a délky textu, je sledování vztahu mezi hodnotami daného indexu a kumulativní délkou textu. Výhodou tohoto přístupu je zejména to, že můžeme analyzovat homogenní textový útvar (jeden autor, jeden žánr, jeden styl, jedno téma). Dále tento přístup do-



Obrázek 5.4: Vztah mezi hodnotou STK a délkou textu N u 1168 lemmatizovaných textů. Osa x je pro větší přehlednost výsledků logaritmizována.



Obrázek 5.5: Vztah mezi hodnotou PTK a délkou textu N u 1168 lemmatizovaných textů.



Obrázek 5.6: Vztah mezi hodnotou PTK a délkou textu N u 1168 lemmatizovaných textů. Osa x je pro větší přehlednost výsledků logaritmizována.

voluje podrobně a přesně sledovat, jak se daný index postupně vyvíjí s rostoucí délkou textu.

V následující části bude analyzován vztah indexů TK a STK vzhledem k narůstající délce textu.[1] Při této analýze je třeba nejdříve definovat jednotky, na něž bude text segmentován. Vzhledem k tomu, že jde o analýzu tematických charakteristik, rozhodl jsem se dále pracovat s následujícími jednotkami:

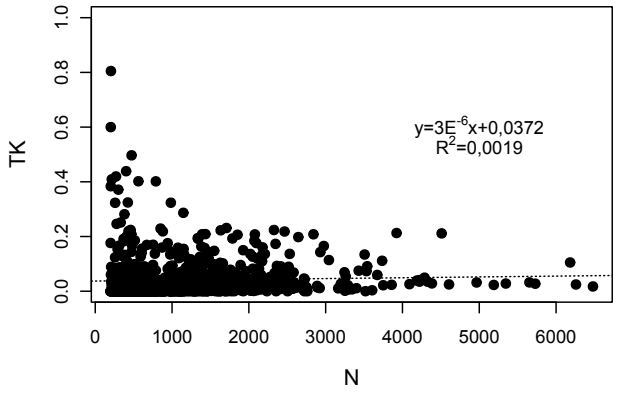
a) odstavec (v případě odborných textů, povídek a novinových zpráv),
b) strofa (v případě poezie),
c) kapitola (v případě románů).

Analyzované texty byly nejdříve segmentovány na tyto jednotky a následně byla měřena hodnota sledovaného indexu v kumulativně sloučených jednotkách daného textu. Konkrétně, například text B. Vachaly *Včely a med ve starém Egyptě* (text č. 366) byl segmentován do 19 odstavců.[2] Hodnota TK (resp. STK) byla nejprve změřena v prvním odstavci, následně byl přidán další odstavec a opět byla změřena hodnota daných indexů – takto se postupovalo až do konce textu, srov. Tab. 5.1. Na Obr. 5.13–14 jsou graficky znázorněny výsledky takto provedené analýzy u čtyř odborných textů[3]

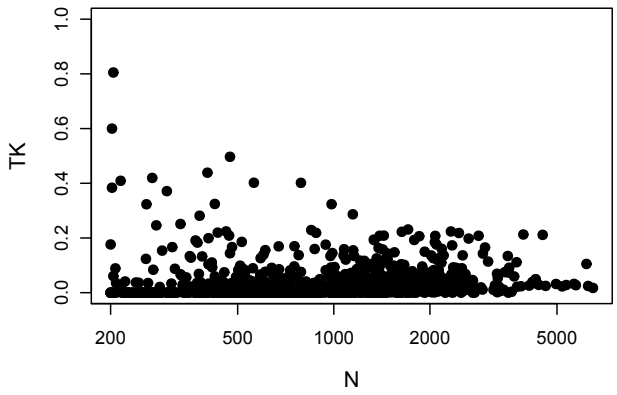
[1] Index PTK v této analýze nebudu používat, protože při kumulativní analýze se pracuje s relativně krátkými textovými jednotkami (viz níže); výsledky prezentované v předchozí kapitole ukázaly, že právě u krátkých textů je použití PTK problematické.

[2] Názvy kapitol či částí textu byly přiřazeny při segmentaci k odstavci, který po názvu bezprostředně následuje.

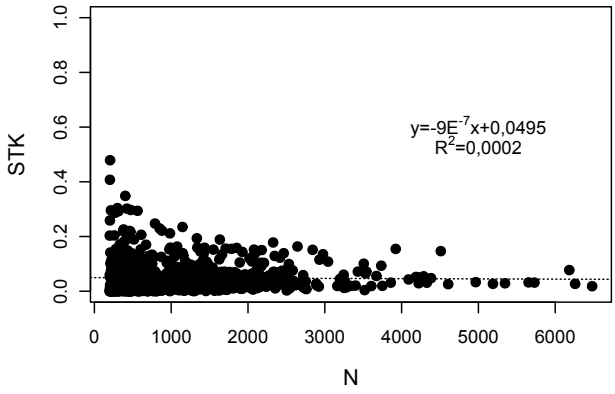
[3] Záměrně byly vybrány delší texty s co největším počtem odstavců.



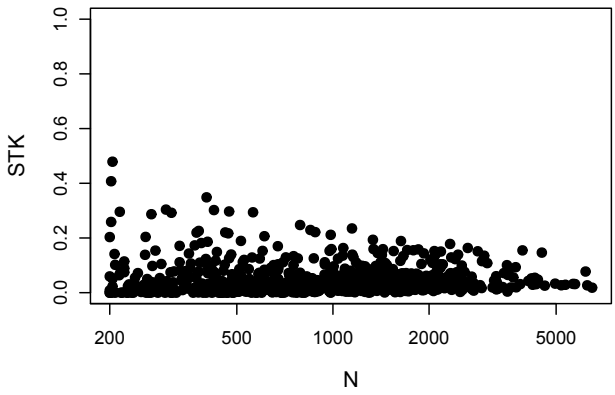
Obrázek 5.7: Vztah mezi hodnotou TK a délkou textu N u 887 lemmatizovaných textů, jejichž délka leží v intervalu $N \in \langle 200; 6500 \rangle$ slov.



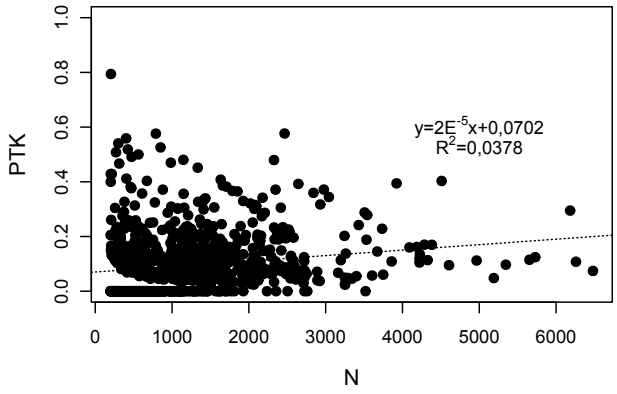
Obrázek 5.8: Vztah mezi hodnotou TK a délkou textu N u 887 lemmatizovaných textů, jejichž délka leží v intervalu $N \in \langle 200; 6500 \rangle$ slov. Osa x je pro větší přehlednost výsledků logaritmizována.



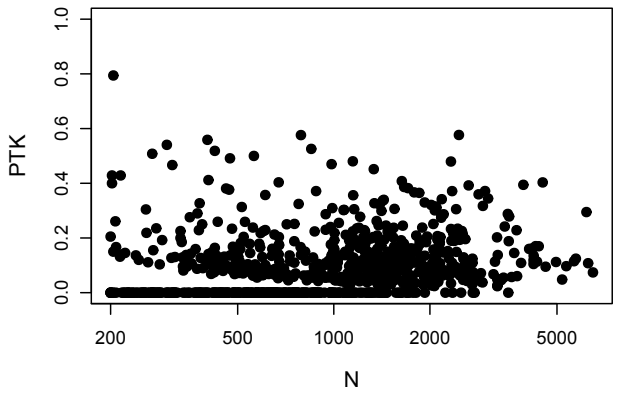
Obrázek 5.9: Vztah mezi hodnotou STK a délkou textu N u 887 lemmatizovaných textů, jejichž délka leží v intervalu $N \in \langle 200; 6500 \rangle$ slov.



Obrázek 5.10: Vztah mezi hodnotou STK a délkou textu N u 887 lemmatizovaných textů, jejichž délka leží v intervalu $N \in \langle 200; 6500 \rangle$ slov. Osa x je pro větší přehlednost výsledků logaritmizována.



Obrázek 5.11: Vztah mezi hodnotou PTK a délkou textu N u 887 lemmatizovaných textů, jejichž délka leží v intervalu $N \in \langle 200; 6500 \rangle$ slov.



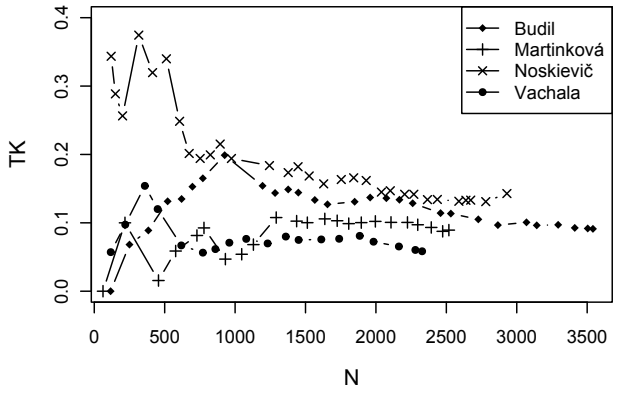
Obrázek 5.12: Vztah mezi hodnotou PTK a délkou textu N u 887 lemmatizovaných textů, jejichž délka leží v intervalu $N \in \langle 200; 6500 \rangle$ slov. Osa x je pro větší přehlednost výsledků logaritmizována.

Tabulka 5.1: Hodnoty TK a STK v kumulativně slučovaných odstavcích textu B. Vachaly *Včely a med ve starém Egyptě* (text č. 366).

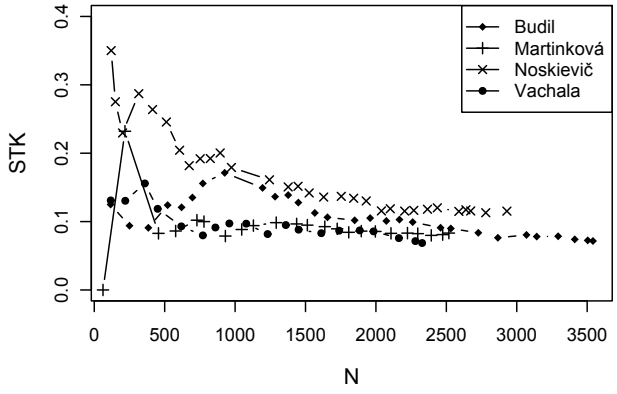
odstavec	N	TK	STK
1.	117	0,057143	0,130952
1.–2.	220	0,097403	0,130408
1.–3.	359	0,154062	0,155785
1.–4.	450	0,120130	0,118631
1.–5.	618	0,066964	0,093006
1.–6.	771	0,056397	0,079818
1.–7.	860	0,061587	0,091353
1.–8.	959	0,070947	0,097253
1.–9.	1079	0,076533	0,096950
1.–10.	1231	0,069853	0,081751
1.–11.	1360	0,079903	0,094921
1.–12.	1451	0,075000	0,088164
1.–13.	1612	0,075721	0,083003
1.–14.	1739	0,076667	0,086687
1.–15.	1885	0,080952	0,087101
1.–16.	1983	0,072392	0,085673
1.–17.	2163	0,065406	0,075828
1.–18.	2280	0,060337	0,071456
1.–19.	2328	0,058409	0,068631

Moderní totalitarismus a síla politické imaginace I. Budila (text č. 323), *Toxický účinek metanolu na lidský organismus* M. Martínkové (text č. 344), *Efektivní energetika* P. Noskieviče (text č. 346) a *Včely a med ve starém Egyptě* B. Vachaly (text č. 366). Již na první pohled je patrné, že nelze postulovat nějaký jednoznačný vztah mezi narůstající délkou textu a hodnotou TK či STK: v případě Budilova článku sledujeme postupný nárůst hodnot TK, které po dosažení vrcholu v osmém odstavci začínají klesat; u Noskieviče je evidentní celková klesající tendence; u Vachaly dochází v úvodu k nárůstu, který je následován poklesem, jenž je zhruba od třetiny textu vystřídán relativní stabilizací hodnot TK; u Martinkové lze sledovat několik oscilací, po nichž následuje stabilizace. Samozřejmě, že tendence k postupné stabilizaci hodnot je způsobena kumulativním měřením. Zejména oscilace však poukazují na absenci jednoznačného vztahu mezi délkou textu a danými indexy.

Abych se vyhnul neadekvátním generalizacím založeným na analýzách pouze jednoho žánru, použil jsem stejný postup i pro povídkové texty, novinové zprávy, poezii a romány. V případě povídek byly použity první čtyři texty z Čapkových *Povídek z druhé kapsy* (texty č. 809–812). Výsledky, prezentované graficky na Obr. 5.15–16, předně ukazují, že u povídek se projevují velmi malé intervaly hodnot TK a STK, ve kterých



Obrázek 5.13: Vztah mezi kumulativně měřenou hodnotou TK a délkou textu N u čtyř odborných textů (texty č. 323, 344, 346, 366). Každý text byl segmentován na odstavce. Jednotlivé body v grafu reprezentují kumulativně sloučené odstavce daného textu. Křivky vyjadřují průběh změn hodnot TK v kumulativně sloučených odstavcích v jednotlivých textech.



Obrázek 5.14: Analogie Obr. 5.13 pro STK.

se oscilace pohybují, ve srovnání s odbornými texty – srov. osy y na Obr. 5.13–14 a 5.15–16. Důležité je ale to, že i v těchto malých intervalech nelze zaznamenat žádnou jednoznačnou tendenci – to svědčí o tom, že vývoj hodnot sledovaných indexů nezávisí na délce textu, ale je dán jinými faktory.

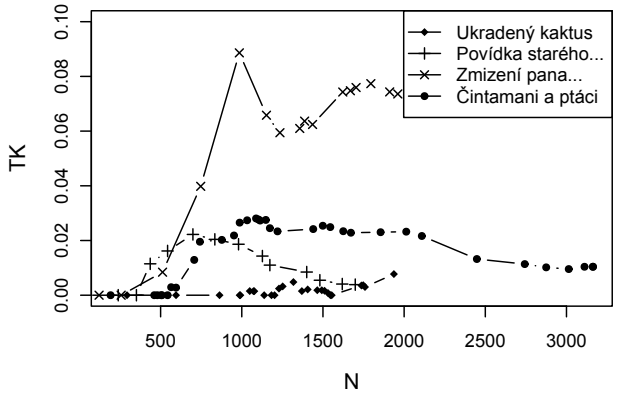
U novinových zpráv (texty č. 253, 261, 263, 269) se projevuje velký rozptyl a ani zde není jednoznačná tendence, srov. Obr 5.17–18.

Rozmanitá situace je i u Erbenových básní (texty č. 25, 29, 30, 31): na jedné straně na počátku silně tematicky koncentrovaný *Vodník* prudce směřující k nízkým hodnotám TK i STK, na straně druhé minimálně oscilující *Poklad a Záhořovo lože* či vzestupně-klesající průběh u *Vrby*.

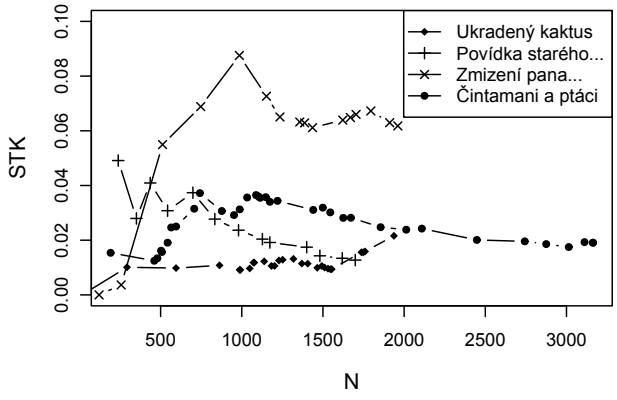
V případě románů (texty č. 552, 553, 554, 558) a jejich segmentace na kapitoly se stejně jako u povídek projevuje velmi malý interval sledovaných hodnot, navíc je evidentní, že s narůstající délkou textu už dochází k ustálení hodnot sledovaných indexů, což u takto dlouhých textů není překvapivé, viz kap. 5.1.

5.3 Poznámka ke vztahu délky textu a tematické koncentrace

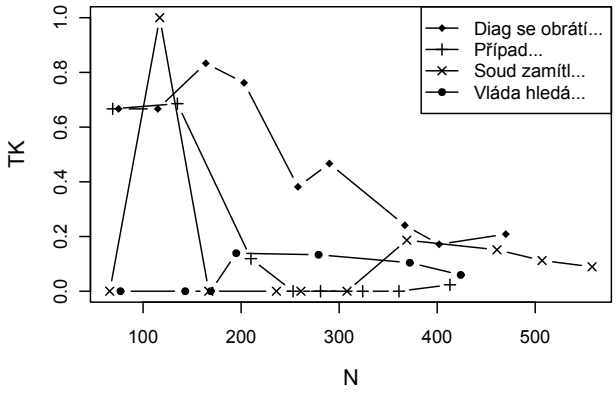
Výsledky prezentované v této kapitole ukazují na nezávislost indexů TK a STK na délce textu. Jak jsem uvedl výše, doposud navržené indexy byly, pokud je mi známo, na délce textu vždy závislé, a to i v případě, že byla tato závislost eliminována prostřednictvím nějaké transformace. TK a STK jsou indexy, které jsou na délce textu nezávislé, přičemž tato nezávislost není výsledkem lingvisticky nezdůvodněné transformace, jako je například použití dekadického logaritmu či druhé odmocniny. Je samozřejmé, že tyto výsledky je třeba ověřit i na dalších jazycích.



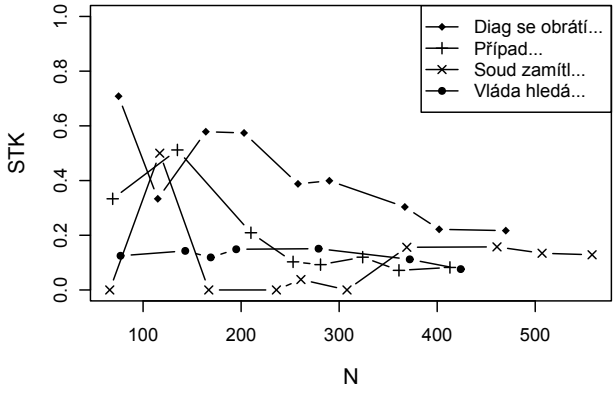
Obrázek 5.15: Vztah mezi kumulativně měřenou hodnotou TK a délkou textu N u prvních čtyř povídek K. Čapka ze sbírky *Povídky z druhé kapsy* (texty č. 809–812). Každý text byl segmentován na odstavce. Jednotlivé body v grafu reprezentují kumulativně sloučené odstavce daného textu. Křivky vyjadřují průběh změn hodnot TK v kumulativně sloučených odstavcích v jednotlivých textech.



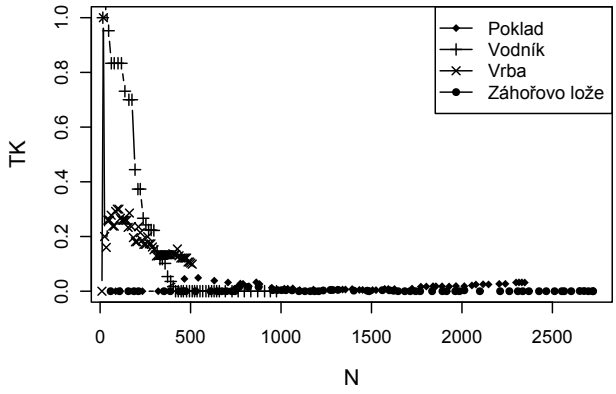
Obrázek 5.16: Analogie Obr. 5.15 pro STK.



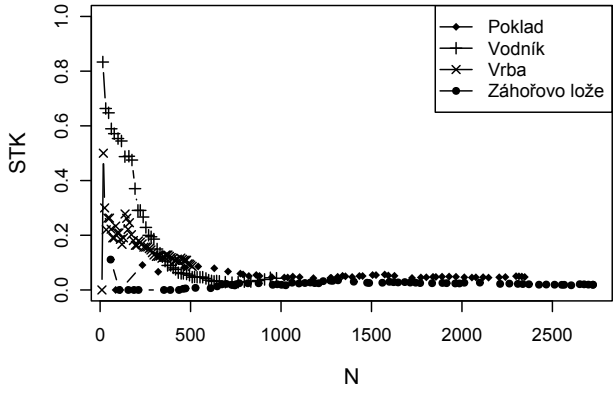
Obrázek 5.17: Vztah mezi kumulativně měřenou hodnotou TK a délkou textu N čtyř novinových zpráv (texty č. 253, 261, 263, 269). Každý text byl segmentován na odstavce. Jednotlivé body v grafu reprezentují kumulativně sloučené odstavce daného textu. Křivky vyjadřují průběh změn hodnot TK v kumulativně sloučených odstavcích v jednotlivých textech.



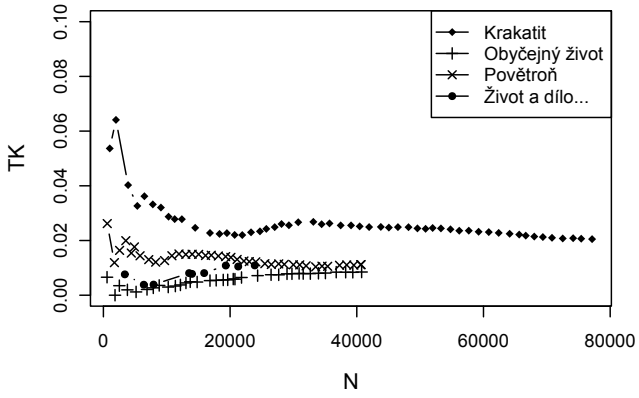
Obrázek 5.18: Analogie Obr. 5.17 pro STK.



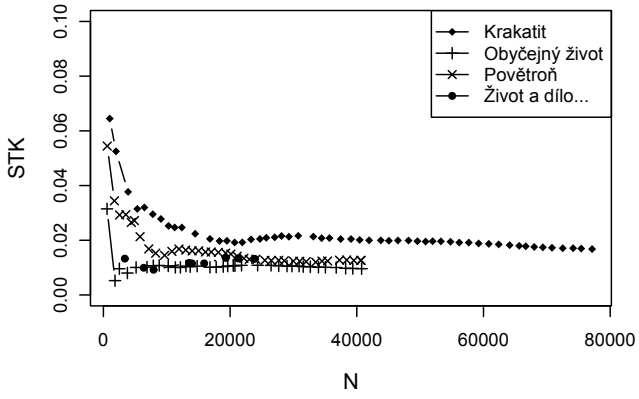
Obrázek 5.19: Vztah mezi kumulativně měřenou hodnotou TK a délkou textu N čtyř Erbenových básní (texty č. 25, 29, 30, 31). Každý text byl segmentován na strofy. Jednotlivé body v grafu reprezentují kumulativně sloučené strofy daného textu. Křivky vyjadřují průběh změn hodnot TK v kumulativně sloučených strofách v jednotlivých textech.



Obrázek 5.20: Analogie Obr. 5.19 pro STK.



Obrázek 5.21: Vztah mezi kumulativně měřenou hodnotou TK a délkou textu N čtyř textů K. Čapka (texty č. 552, 553, 554, 558). Každý text byl segmentován na kapitoly. Jednotlivé body v grafu reprezentují kumulativně sloučené kapitoly daného textu. Křivky vyjadřují průběh změn hodnot TK v kumulativně sloučených kapitolách v jednotlivých textech.



Obrázek 5.22: Analogie Obr. 5.21 pro STK.

6

Vývoj tematické koncentrace v textu

Text je tvořen sekvencemi jednotek (např. fonémů, slov, vět atd.), která vykazuje mnohé netriviální vlastnosti (srov. Altmann a Burdinski 1982; Grothajn 1979, 1980; Hřebíček 2000; Piotrowskij 1984; Uhlířová 1997; Wimmer et al. 2003). Dosavadní analýzy ukázaly, že sekvenční charakteristiky textu (např. sekvenční distribuce délek, frekvencí, iterace) jsou zřejmě výsledkem obecných mechanismů řídících verbální chování uživatelů jazyka (Köhler 2005). Je ale třeba zdůraznit, že výzkum v této oblasti je stále na počátku: jednak chybí obecná teorie textu, ke které by se daly zjištěné vlastnosti vztáhnout, jednak je třeba vzít na vědomí, že bylo stále provedeno příliš málo analýz, navíc na malém počtu jazyků. Problematická je i lingvistická interpretace mnohých způsobů měření, jako jsou Hurstův indikátor, Ljapunovův koeficient, Minkowského klobása či Fourierova analýza (srov. patřičné kapitoly Köhler a Altmann 2014; Strauss et al. 2014). Sekvenční analýza textu tak představuje jednu z neprobádaných oblastí, která může přinést nové pohledy na fungování nejen textu, ale i jazyka.

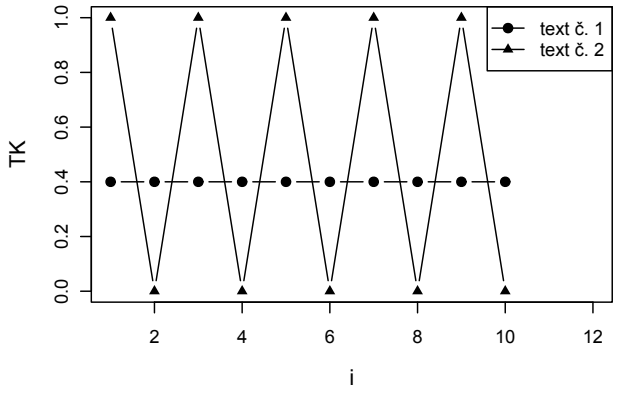
Z hlediska tematické koncentrace textu zkoumání sekvenčních charakteristik nabízí možnost sledovat, jak autor v průběhu textu s touto vlastností „nakládá“. Existují přitom dvě krajní možnosti:

1) hodnota TK (případně STK a PTK) je v průběhu textu konstantní;
2) hodnota TK (případně STK a PTK) v průběhu textu neustále osciluje mezi krajními hodnotami, tj. v intervalu $\langle 0; 1 \rangle$.

Oba případy jsou graficky znázorněny na Obr. 6.1, kde jsou zobrazeny hodnoty TK dvou hypotetických textů, které jsou rozděleny na deset úseků (je možné použít textové úseky se stejným počtem slov, odstavce, kapitoly atd.). Jednotlivé body grafu reprezentují hypotetické hodnoty TK v každém z těchto úseků, tak jak jdou v textu za sebou. Výsledkem je sekvence hodnot vyjadřující lineární vývoj TK v daném textu. V případě textu č. 1 se jedná o naprosto rovnoměrný vývoj, v případě textu č. 2 o vývoj extrémně nerovnoměrný.

6.1 Způsob měření vývoje tematické koncentrace v textu

Vývoj tematické koncentrace v textu je možné celkem jednoduše kvantifikovat na základě vlastnosti grafu, do kterého jsou zaneseny naměřené hodnoty jednotlivých textových úseků (Obr 6.1). Tato kvantifikace umožňuje jednotlivé texty porovnávat, včetně statistického testování naměřených rozdílů. Konkrétně, spojíme-li jednotlivé body grafu čarou, jak je to znázorněno na Obr. 6.1, získáme úsečku (text č. 1), nebo křivku (text č. 2). Čím větší je rozdíl TK (STK či PTK) u dvojice bodů spojených čarou, tím je



Obrázek 6.1: Vývoj hodnot TK u dvou hypotetických textů rozčleněných na deset úseků i. V případě textu č. 1 se jedná o naprosto rovnoměrný vývoj, v případě textu č. 2 o vývoj extrémně nerovnoměrný.

tato čára (spojnice) delší. Pro stanovení výpočtu celkové délky spojnice budu postupovat analogicky podle způsobu prezentovaného Hřebíčkem (2002, s. 153nn).

Pokud definujeme i jako hodnotu vyjadřující pořadí daného textového úseku v textu, pak vzdálenost pořadí mezi po sobě jdoucími úseky je rovna jedné: $(i+1)-i=1$, srov. Obr 6.2. Délku spojnice d_i mezi po sobě jdoucími body vyjadřujícími hodnoty TK pak lze jednoduše vypočítat podle Pythagorovy věty, protože tato spojnice je přeponou pravoúhlého trojúhelníku (Obr. 6.2):

$$d_i = [(TK_i - TK_{i+1})^2 + 1]^{1/2}. \tag{6.1}$$

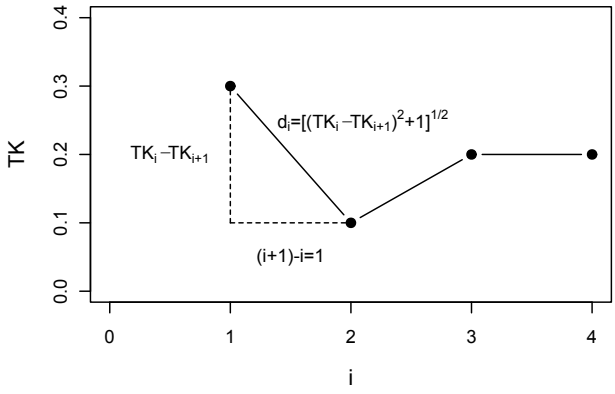
V případě prvních dvou hodnot na Obr. 6.2 ($TK_1 = 0,3$; $TK_2 = 0,1$) tak dostáváme

$$d_i = [(0,3 - 0,1)^2 + 1]^{1/2} = 1,019804.$$

Součtem jednotlivých vzdáleností d_i získáváme celkovou délku spojnice L

$$L = \sum_{i=1}^{N} d_i, \tag{6.2}$$

kde N je počet spojnic mezi jednotlivými body. Jelikož jsou jednotlivé texty různě dlouhé, budou mít nestejný počet úseků. Aby bylo možné porovnávat texty nestejné



Obrázek 6.2: Způsob výpočtu délky spojnice mezi dvěma body reprezentujícími vývoj TK.

délky, je nutné délku spojnice L vztáhnout k počtu úseků n, konkrétně vypočítat průměrnou délku spojnic mezi dvojicemi bodů v daném textu. Míru (ne)rovnoměrnosti vývoje tematické koncentrace[1] r lze pak definovat jako

$$r = \frac{L}{n}. \tag{6.3}$$

Pro texty s naprosto rovnoměrným vývojem (text č. 1 na Obr. 6.1) platí pro každou spojnici $d_i = [0^2 + 1]^{1/2} = 1$, tj. u těchto textů vždy $r = 1$. Pro texty s extrémně nerovnoměrným vývojem pak platí pro každou spojnici $d_i = [(1 - 0)^2 + 1]^{1/2} = 1,414214$, tj. u těchto textů vždy $r = 1,414214$. Hodnoty r všech textů musí tedy ležet v intervalu $\langle 1; 1,414214 \rangle$.

Výpočet L a r u reálných textů budu ilustrovat na příkladu vývoje TK u dvou článků, které vykazují opticky velmi odlišný vývoj, viz Obr. 6.3; jde o odborný text K. Kučery *K vokalizaci neslabičných předložek v současné češtině* (text č. 298) a esej V. Havla *Šest poznámek o kultuře* (text č. 451). Oba texty byly nejdříve segmentovány na úseky o délce 300 slov/tokenů. Zde je třeba zmínit důvody volby velikosti segmentu.

Na první pohled se jistě jako „přirozený" segment jeví odstavec, který většinou reprezentuje určitý tematicky uzavřený celek. Z hlediska analýzy tematické koncentrace se však jedná o příliš krátkou jednotku. Jak bylo ukázáno v kap. 5.1, tato metoda je vhodná pro texty o délce minimálně 250 slov/tokenů. Pokud analyzujeme texty kratší, je měření velice citlivé i na minimální frekvenční rozdíly. U všech textů

[1] To samozřejmě platí pro všechny indexy: TK, STK, PTK.

je délka naprosté většiny odstavců menší než 200 slov/tokenů, tudíž sledovat vývoj tematické koncentrace prostřednictvím odstavců není vhodné. Proto jsem se rozhodl texty mechanicky rozdělit na úseky obsahující 300 slov/tokenů, přičemž poslední úsek textu s délkou menší než 300 slov/tokenů jsem nebral v úvahu. Samozřejmě se nejedná o ideální řešení. Je třeba si ale uvědomit, že u naprosté většiny textů o délce $N \in \langle 200; 6500 \rangle$ slov/tokenů, což je interval, v jehož rámci jsou hodnoty TK i STK na délce textů nezávislé, zpravidla nelze formálně vymezit tematické úseky delší než odstavec a kratší než celý text. V některých případech jsou sice texty členěny na části (to se týká některých odborných článků), ale v případě esejů, povídek atd. se nic takového většinou neobjevuje. Z tohoto důvodu se jeví mechanické rozdělení textu jako přijatelné.

Tabulka 6.1: Hodnoty TK v po sobě jdoucích úsecích o délce 300 slov/tokenů v textech K. Kučery *K vokalizaci neslabičných předložek v současné češtině* (text č. 298) a V. Havla *Šest poznámek o kultuře* (text č. 451).

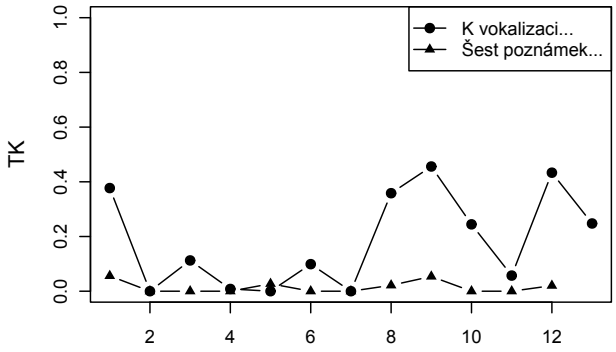
K vokalizaci... (text č. 298)			*Šest poznámek...* (text č. 451)		
úsek	TK	d_i	úsek	TK	d_i
1	0,377083		1	0,056034	
2	0	1,068734	2	0	1,001569
3	0,112500	1,006308	3	0	1
4	0,008202	1,005424	4	0	1
5	0	1,000034	5	0,027149	1,000368
6	0,098901	1,004879	6	0	1,000368
7	0	1,004879	7	0	1
8	0,358333	1,062263	8	0,021587	1,000233
9	0,455944	1,004753	9	0,053846	1,000520
10	0,244444	1,022121	10	0	1,001449
11	0,057143	1,017390	11	0	1
12	0,433333	1,068419	12	0,020343	1,000207
13	0,247619	1,017099			

V každém z úseků výše analyzovaných textů byla nejdříve změřena TK (Tab. 6.1 a Obr. 6.3). Na základě vzorce (6.1) byla následně vypočítána vzdálenost mezi jednotlivými úseky. Konkrétně, délka spojnice mezi prvním a druhým úsekem textu *K vokalizaci...* (text č. 298) je

$$d_1 = [(0,377083 - 0)^2 + 1]^{1/2} = 1,068734,$$

v textu *Šest poznámek...* (text č. 451)

$$d_1 = [(0,056034 - 0)^2 + 1]^{1/2} = 1,001569.$$



Obrázek 6.3: Hodnoty TK v po sobě jdoucích úsecích o délce 300 slov/tokenů v textech K. Kučery *K vokalizaci neslabičných předložek v současné češtině* (text č. 298) a V. Havla *Šest poznámek o kultuře* (text č. 451), srov. Tab. 6.1.

Analogicky pak byly vypočítány vzdálenosti mezi všemi dalšími úseky. Na základě vzorce (6.2) dostáváme délky křivek

$$L_{\textit{K vokalizaci...}} = \sum_{i=1}^{N} d_i = 12,282302$$

a

$$L_{\textit{Šest poznámek...}} = \sum_{i=1}^{N} d_i = 11,004714.$$

Průměrná míra (ne)rovnoměrnosti vývoje tematické koncentrace podle vzorce (6.3) je u obou textů následující:

$$r_{\textit{K vokalizaci...}} \frac{12,282302}{12} = 1,023525,$$

$$r_{\textit{Šest poznámek...}} = \frac{11,004714}{11} = 1,000429.$$

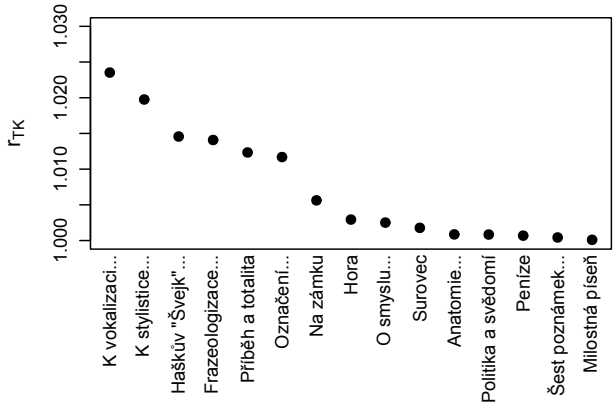
Na první pohled se zdá, že jde o rozdíl nepatrný. Je třeba si ale uvědomit, že hodnoty r se pohybují v intervalu $\langle 1; 1,414214 \rangle$. Adekvátní vyhodnocení rozdílů vývoje tematické koncentrace je sice možné až prostřednictvím statistických testů, srov. kap.

6.3, ale již pouhým vynesením hodnot r do grafu lze získat alespoň předběžný první vhled do rozsahu těchto rozdílů (Obr. 6.4 a 6.5).

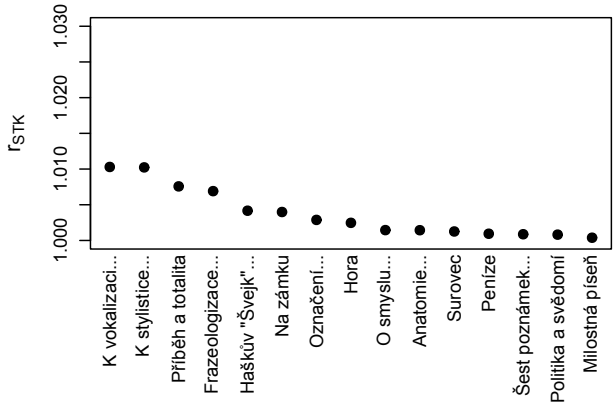
Postup popsaný v této kapitole jsem dále aplikoval na tři skupiny textů různých žánrů. Konkrétně na pět odborných textů z časopisu *Naše řeč* (texty č. 283, 286, 298, 305, 313), pět esejů V. Havla (texty č. 415, 432, 438, 444, 451) a pět povídek K. Čapka (texty č. 774, 775, 836, 837, 838). Každý text byl segmentován na úseky o velikosti 300 slov (tokenů), přičemž byly záměrně vybrány texty o délce $N > 3000$ slov/tokenů, abych získal u každého textu minimálně deset úseků. Poslední úsek textu, který obsahoval méně než 300 slov/tokenů, nebyl při analýze použit.

Tabulka 6.2: Průměrná míra (ne)rovnoměrnosti vývoje TK a STK.

	r_{TK}	r_{STK}
odborné		
Daneš, F. *Haškův „Švejk" a Vachkovo „Bidýlko" – dva milníky ve vývoji jazyka české prózy* (text č. 283)	1,014567	1,004167
Horálek, K. *K stylistice zvukových prostředků jazyka* (text č. 286)	1,019754	1,010232
Kučera, K. *K vokalizaci neslabičných předložek v současné češtině* (text č. 298)	1,023525	1,010293
Němec, I. *Frazeologizace slovesa dělati a jeho synonym* (text č. 305)	1,014082	1,006908
Štěpán, P. *Označení černé a bílé barvy v zeměpisných jménech v Čechách* (text č. 313)	1,011690	1,002880
eseje		
Havel V. *Anatomie jedné zdrženlivosti* (text č. 415)	1,000860	1,001436
Havel V. *O smyslu Charty 77* (text č. 432)	1,002515	1,001450
Havel V. *Politika a svědomí* (text č. 438)	1,000840	1,000816
Havel V. *Příběh a totalita* (text č. 444)	1,012336	1,007573
Havel V. *Šest poznámek o kultuře* (text č. 451)	1,000429	1,000871
povídky		
Čapek K. *Hora* (text č. 774)	1,002939	1,002468
Čapek K. *Milostná píseň* (text č. 775)	1,000103	1,000385
Čapek K. *Na zámku* (text č. 836)	1,005629	1,003985
Čapek K. *Peníze* (text č. 837)	1,000688	1,000943
Čapek K. *Surovec* (text č. 838)	1,001781	1,001258



Obrázek 6.4: Průměrná míra (ne)rovnoměrnosti vývoje TK.



Obrázek 6.5: Průměrná míra (ne)rovnoměrnosti vývoje STK.

Analýza výše vedených textů zejména ukázala, že hodnoty r u TK i STK se pohybují velmi blízko dolní hranice intervalu $\langle 1; 1,414214 \rangle$, tudíž u všech textů převažu-

je tendence rovnoměrného vývoje. Co se týká jednotlivých žánrů, nejvyšší hodnoty vykazují texty odborné, což je patrné zejména u TK. U ostatních textů (s výjimkou Havlova eseje *Příběh a totalita*) je hodnota r těsně nad hranicí naprosto rovnoměrného vývoje. Pro adekvátnější interpretaci zjištěných výsledků je nutné použít statistický test, viz následující kapitola.

6.2 Testování rozdílů vývoje tematické koncentrace v textu

Vývoj tematické koncentrace v textu lze vyjádřit posloupností délek křivky d_i, které se vyskytují mezi jednotlivými hodnotami TK (či STK), jak je uvedeno v Tab. 6.1. Rozdíly tohoto vývoje mezi dvěma texty lze testovat prostřednictvím Wilcoxonova-Mannova-Whitneyova statistického testu. Nulová hypotéza předpokládá, že vývoj sledované proměnné je v obou textech totožný. Pro texty uvedení v Tab. 6.1 dostáváme při použití Wilcoxonova-Mannova-Whitneyova statistického testu p-hodnotu = 0,0003, což vyjadřuje signifikantní rozdíl mezi těmito texty (na hladině významnosti $\alpha = 0,05$).

Výsledky prezentované na Obr. 6.4 a 6.5 naznačují, že vývoj tematické koncentrace se může lišit v závislosti na žánru, zejména odborné texty se zdají alespoň opticky vykazovat vyšší hodnoty r v případě TK. Aplikací Wilcoxonova-Mannova-Whitneyova testu, jehož prostřednictvím byl porovnáván vývoj TK a STK u každé dvojice textů z Tab. 6.2, můžeme u sledované skupiny textů zjistit, zda má žánr signifikantní vliv na vývoj TK a STK, nebo zda jsou rozdíly nevýznamné, tudíž je můžeme například přičíst vlivu náhody (na zvolené hladině významnosti). Níže prezentované výsledky a interpretace nemají samozřejmě obecnou platnost – mým cílem je zde především prezentovat možnosti této metody. Výsledky testů jsou uvedeny v Tab. 6.3 a 6.4 a na Obr. 6.6 a 6.7.

Z obou tabulek i grafů je na první pohled patrné, že signifikantní rozdíly se projevují častěji u TK než u STK. Tento výsledek není asi příliš překvapivý. Výhodou STK je sice to, že umožňuje porovnávat a testovat větší skupiny textů (viz kap. 3), na druhou stranu tato metoda do jisté míry nivelizuje rozdíly mezi texty výrazně tematicky zaměřenými a texty, u nichž se autor na dané téma či témata nezaměřuje tak intenzivně. Detailnější pohled na Tab. 6.3 a Obr. 6.6 ukazuje, že v analyzovaném vzorku textů se jednoznačně vyčleňuje skupina odborných textů, které se nejvíce odlišují od ostatních dvou žánrů. Mezi odbornými texty navzájem jsou navíc všechny rozdíly nesignifikantní. Podobně se chovají i povídky, u nichž jsou až na dvě výjimky opět rozdíly nesignifikantní; zde je však větší počet textů s nesignifikantními rozdíly vzhledem k esejům. Ve skupině esejů mají zvláštní postavení dva texty: v prvé řadě *Příběh a totalita* (text č. 444), který ve vývoji TK kromě dvou textů vykazuje nesignifikantní rozdíly se všemi sledovanými texty, a esej *Anatomie jedné zdrženlivosti* (text č. 415), který má naopak největší počet signifikantních rozdílů. Vysoký počet nesignifikantních rozdílů mezi eseji a povídkami svědčí o příbuznosti obou žánrových skupin.

Grafické znázornění (Obr. 6.6 a 6.7) dále umožňuje kvantifikovat jak míru homogenity/heterogenity v rámci jednoho žánru, tak míru podobnosti/rozdílnosti mezi

Tabulka 6.3: Výsledky Wilcoxonova-Mannova-Whitneyova testu, jehož prostřednictvím byl porovnáván vývoj TK u každé dvojice textů z Tab. 6.2. V tabulce jsou uvedeny p-hodnoty; pokud je $p < 0,05$, jde o signifikantní rozdíl (hladina významnosti $\alpha = 0,05$). Hodnoty se signifikantními rozdíly jsou v šedých buňkách.

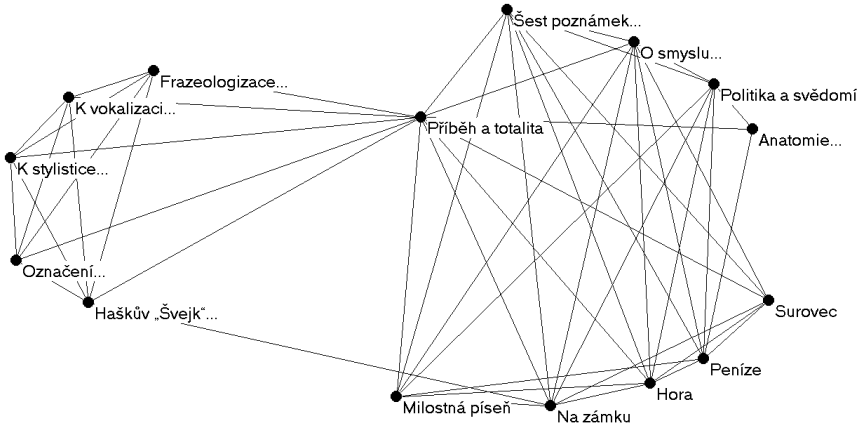
TK	*Označení...*	*K stylistice...*	*Haškův „Švejk“ ...*	*Frazeologizace...*	*K vokalizaci...*	*Anatomie...*	*Příběh a totalita*	*Politika a svědomí*
Označení...	x							
K stylistice...	0,2199	x						
Haškův „Švejk“...	0,8271	0,4118	x					
Frazeologizace...	0,8995	0,2471	0,8917	x				
K vokalizaci...	0,1572	0,7236	0,3140	0,2854	x			
Anatomie...	0,0000	0,0000	0,0000	0,0000	0,0000	x		
Příběh a totalita	0,2341	0,0759	0,3512	0,3261	0,0235	0,0006	x	
Politika a svědomí	0,0001	0,0000	0,0014	0,0004	0,0000	0,1322	0,0303	x
O smyslu...	0,0028	0,0004	0,0152	0,0241	0,0019	0,0158	0,1932	0,3258
Šest poznámek...	0,0005	0,0002	0,0120	0,0033	0,0003	0,0226	0,1577	0,5484
Na zámku	0,0073	0,0009	0,0644	0,0280	0,0024	0,0006	0,5201	0,0629
Hora	0,0050	0,0016	0,0346	0,0117	0,0006	0,0035	0,3141	0,1404
Peníze	0,0001	0,0000	0,0016	0,0012	0,0002	0,1710	0,0370	0,8293
Surovec	0,0094	0,0010	0,0398	0,0239	0,0018	0,0004	0,5868	0,0334
Milostná píseň	0,0000	0,0000	0,0023	0,0014	0,0001	0,0137	0,0920	0,6811

TK	*O smyslu...*	*Šest poznámek...*	*Na zámku*	*Hora*	*Peníze*	*Surovec*	*Milostná píseň*
O smyslu...	x						
Šest poznámek...	0,7786	x					
Na zámku	0,3491	0,5051	x				
Hora	0,5643	0,3860	0,9444	x			
Peníze	0,2388	0,1388	0,0258	0,0580	x		
Surovec	0,3612	0,2931	0,9330	0,6726	0,0180	x	
Milostná píseň	0,6357	0,1939	0,0311	0,1354	0,4613	0,0344	x

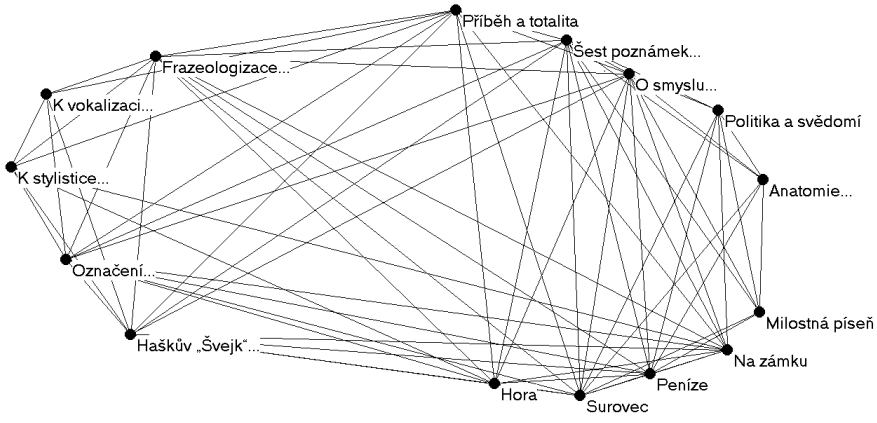
Tabulka 6.4: Výsledky Wilcoxonova-Mannova-Whitneyova testu, jehož prostřednictvím byl porovnáván vývoj STK u každé dvojice textů z Tab. 6.2. V tabulce jsou uvedeny p-hodnoty; pokud je $p < 0,05$, jde o signifikantní rozdíl (hladina významnosti $\alpha = 0,05$). Hodnoty se signifikantními rozdíly jsou v šedých buňkách.

STK	*Označení...*	*K stylistice...*	*Haškův „Švejk"...*	*Frazeologizace...*	*K vokalizaci...*	*Anatomie...*	*Příběh a totalita*	*Politika a svědomí*
Označení...	x							
K stylistice...	0,2388	x						
Haškův „Švejk"...	0,7203	0,2226	x					
Frazeologizace...	0,8201	0,4059	0,6114	x				
K vokalizaci...	0,0849	0,8517	0,2254	0,2189	x			
Anatomie...	0,0041	0,0006	0,0088	0,0141	0,0004	x		
Příběh a totalita	0,6532	0,3624	0,5701	0,6458	0,3889	0,0024	x	
Politika a svědomí	0,0091	0,0011	0,0094	0,0373	0,0005	0,6171	0,0079	x
O smyslu...	0,0925	0,0048	0,1207	0,2226	0,0031	0,2038	0,0562	0,4168
Šest poznámek...	0,1072	0,0129	0,0629	0,1507	0,0028	0,3381	0,0517	0,5770
Na zámku	0,4683	0,0715	0,4570	0,4985	0,0208	0,0232	0,2997	0,0529
Hora	0,5383	0,0788	0,5030	0,5499	0,0354	0,0067	0,3106	0,0182
Peníze	0,0620	0,0078	0,1048	0,1448	0,0011	0,1569	0,0771	0,3185
Surovec	0,0609	0,0072	0,2035	0,1366	0,0045	0,1047	0,1578	0,1670
Milostná píseň	0,0107	0,0012	0,0191	0,0345	0,0001	0,4133	0,0369	0,6188

STK	*O smyslu...*	*Šest poznámek...*	*Na zámku*	*Hora*	*Peníze*	*Surovec*	*Milostná píseň*
O smyslu...	x						
Šest poznámek...	0,6800	x					
Na zámku	0,4023	0,2689	x				
Hora	0,2009	0,1725	0,9451	x			
Peníze	0,7901	0,8089	0,2794	0,1970	x		
Surovec	0,4753	0,6490	0,4323	0,2954	0,7930	x	
Milostná píseň	0,5890	0,5691	0,0349	0,0213	0,5826	0,2642	x



Obrázek 6.6: Grafické vyjádření výsledků uvedených v Tab. 6.3. Dvojice textů, u nichž se projevil nesignifikantní rozdíl ve vývoji TK, jsou spojeny čarou. Vlevo v grafu jsou seskupeny odborné texty, vpravo nahoře eseje a vpravo dole povídky.



Obrázek 6.7: Grafické vyjádření výsledků uvedených v Tab. 6.4. Dvojice textů, u nichž se projevil nesignifikantní rozdíl ve vývoji STK, jsou spojeny čarou. Vlevo v grafu jsou seskupeny odborné texty, vpravo nahoře eseje a vpravo dole povídky.

jednotlivými žánry navzájem. Konkrétně, pokud by byly všechny rozdíly vývoje TK či STK mezi texty nesignifikantní, graf by obsahoval maximální možný počet hran e:

$$e_{max} = \frac{n(n-1)}{2}, \tag{6.4}$$

kde n je počet uzlů v grafu (zde jde o počet textů). Poměr pozorovaného počtu hran e a teoretického maximálního počtu hran e_{max} udává tzv. hustotu grafu (Newman 2011):

$$\rho = \frac{\sum_i e_i}{e_{max}}. \tag{6.5}$$

V případě maximálně propojeného grafu $\rho = 1$, v případě, že není propojen žádný uzel, $\rho = 0$. Aplikujeme-li toto měření na jednotlivé žánry, je možné stanovit míru homogenity/heterogenity dané skupiny textů. Každý soubor se skládá z pěti textů, proto podle vzorce (6.4) $e_{max} = 10$. Pro tři výše uvedené žánry pak platí:

$$\rho_{TK(\text{odborné})} = \frac{10}{10} = 1;$$

$$\rho_{TK(\text{eseje})} = \frac{7}{10} = 0,7;$$

$$\rho_{TK(\text{povídky})} = \frac{8}{10} = 0,8;$$

$$\rho_{STK(\text{odborné})} = \frac{10}{10} = 1;$$

$$\rho_{STK(\text{eseje})} = \frac{8}{10} = 0,8;$$

$$\rho_{STK(\text{povídky})} = \frac{8}{10} = 0,8.$$

U odborných textů je hustota grafu $\rho = 1$, jedná se tedy o žánr s maximální mírou homogenity.

Měření hustoty grafu lze s mírnou modifikací použít i pro porovnávání dvojic žánrů. V tomto případě porovnáváme deset textů, přičemž jde o porovnání dvou skupin, tj. vztahy v rámci skupiny musejí být ignorovány (jde o tzv. bipartitní graf). Proto v případě, že by každý text jedné skupiny byl propojen s každým textem druhé skupiny, platí

$$e_{max} = n_i \cdot n_j, \tag{6.6}$$

kde n_i a n_j je počet prvků v jednotlivých skupinách. Každý zde sledovaný soubor se skládá z pěti textů, proto podle vzorce (6.6) $e_{max} = 25$. Následně:

$$\rho_{TK(\text{odborné-eseje})} = \frac{5}{25} = 0,2;$$

$$\rho_{TK(\text{odborné-povídky})} = \frac{1}{25} = 0,04;$$

$$\rho_{TK(\text{eseje-povídky})} = \frac{19}{25} = 0,76;$$

$$\rho_{STK(\text{odborné-eseje})} = \frac{11}{25} = 0,44;$$

$$\rho_{STK(\text{odborné-povídky})} = \frac{14}{25} = 0,56;$$

$$\rho_{STK(\text{eseje-povídky})} = \frac{21}{25} = 0,84.$$

Tyto výsledky nabízejí detailnější pohled na vztahy mezi žánry. Zejména v případě TK je možné sledovat, do jaké míry jsou si žánry vzhledem k vývoji TK v textu podobné/rozdílné: eseje s povídkami vykazují nejvyšší míru podobnosti, zatímco mezi odbornými texty a povídkami je největší rozdíl. Odborné texty a povídky tak reprezentují žánrové protiklady. Opět zde ale připomínám, že tyto závěry jsou jen ilustrativní a slouží pro popis metody – analyzované skupiny obsahují pro obecnější závěry příliš málo textů. Na druhou stranu, malý počet textů umožňuje přehledně ilustrovat, jak metoda funguje, což je i cílem této kapitoly.

II TEMATICKÁ KONCETRACE A JINÉ VLASTNOSTI TEXTU

7

Tematická koncentrace a slovní bohatství textu

Slovní bohatství textu je na první pohled velmi jednoduše a dobře vymezitelná vlastnost textu: text, v němž se každé slovo vyskytne pouze jednou, lze považovat za lexikálně nejbohatší, na druhou stranu text, v němž by se stále opakovalo jediné slovo, lze označit za lexikálně nejchudší. Uvažujeme-li pak o vztahu slovního bohatství a tematické koncentrace (měřené zde popsanou metodou), dá se očekávat, že texty s vyšším slovním bohatstvím by měly být méně tematicky koncentrované a vice versa. V tomto ohledu se pak nejeví nic snazšího než v každém textu spočítat míru opakování slov (tzv. type-token poměr pro slovní tvary a lemma-token poměr pro lemmata) a porovnat ji s jednotlivými indexy tematické koncentrace. Výpočet type-token poměru je

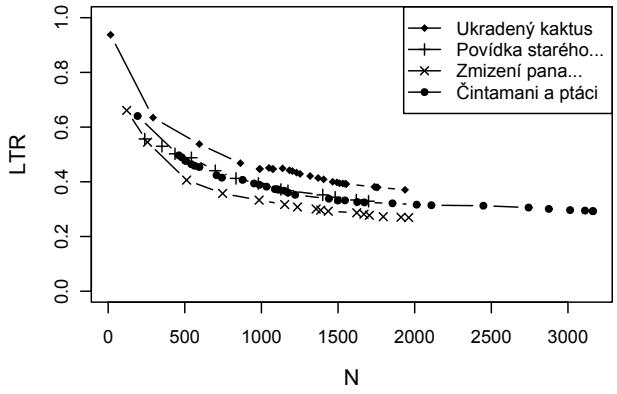
$$\mathrm{TTR} = \frac{V}{N}, \tag{7.1}$$

kde V je počet různých slovních tvarů a N je celkový počet slov/tokenů, analogicky lemma-token poměr

$$\mathrm{LTR} = \frac{L}{N}, \tag{7.2}$$

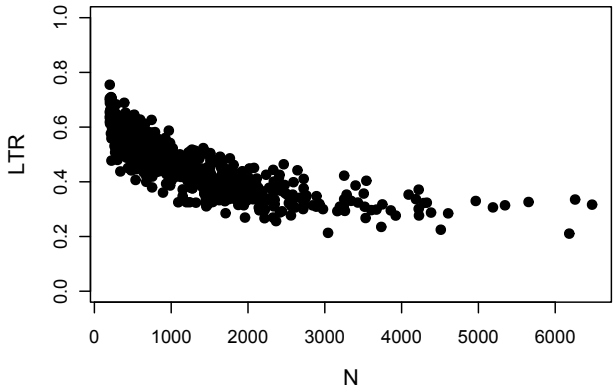
kde L je počet různých lemmat. Potíž je v tom, že míra opakování slov je závislá na délce textu: čím je text delší, tím se slovní tvary i lemmata častěji opakují, tudíž klesá hodnota TTR i LTR. Pro ilustraci sledujme vývoj LTR vzhledem k narůstající délce textu v prvních čtyřech povídkách z Čapkových *Povídek z druhé kapsy* (texty č. 809–812). Postup je analogický k výpočtu TK a kumulativní délky textu (kap. 5.2): všechny povídky byly segmentovány na odstavce; hodnota LTR byla nejprve změřena v prvním odstavci, následně byl přidán další odstavec a opět byla změřena LTR – takto bylo postupováno až do konce textu. Výsledky měření kumulativního vývoje LTR jsou prezentovány na Obr. 7.1. Závislost na délce textu je evidentní, zejména porovnáme-li výsledky s Obr. 5.15 a 5.16. Stejný výsledek dostáváme v případě, že porovnáváme skupiny textů: na Obr. 7.2 jsou znázorněny hodnoty LTR u 887 textů, jejichž délka leží v intervalu $N \in \langle 200; 6500 \rangle$, což je interval, který se ukázal jako vhodný pro použití indexů TK a STK (srov. Obr. 5.7–10).

Vliv délky textu na měření slovního bohatství je fundamentálním problémem této oblasti analýzy textu, a není proto divu, že existuje množství studií zabývajících se tímto fenoménem. S jistou mírou zjednodušení lze říct, že většina těchto studií



Obrázek 7.1: Vztah mezi kumulativně měřenou hodnotou LTR a délkou textu N u prvních čtyř povídek K. Čapka ze sbírky *Povídky z druhé kapsy*: *Ukradený kaktus* (text č. 809), *Povídka starého kriminálníka* (text č. 810), *Zmizení pana Hirsche* (text č. 811), *Čintamani a ptáci* (text č. 812). Každý text byl segmentován na odstavce. Jednotlivé body v grafu reprezentují kumulativně sloučené odstavce daného textu. Křivky vyjadřují průběh změn hodnot LTR (v kumulativně sloučených odstavcích v jednotlivých textech).

1) se snaží vliv délky textu určitým způsobem eliminovat, nejčastěji prostřednictvím nějaké transformace (např. Bernet 1988; Ejiri a Smith 1993; Guiraud 1954, 1959; Herdan 1960, 1966; Hess et al. 1986, 1989; Honoré 1979; Martynenko 2010; Menard 1983; Panas 2001; Popescu et al. 2009a; Popescu et al. 2011, 2012; Ratkowsky a Hantrais 1975; Těšitelová 1972; Tuldava, 1995; Tuzzi et al. 2010; Tweedie a Baayen 1998; Weitzman 1971; Yule 1944; poslední JQL), nebo
2) navrhuje čistě technický postup, jak eliminovat vliv délky textu pro původní TTR (Köhler, Galle 1993; Scott 2013; Covington, McFall 2010), nebo
3) dokonce revokuje pohled na TTR v tom smyslu, že odmítá vnímat tento index jako projev slovního bohatství a chápe jej jako model tzv. informačního vývoje textu (information flow), srov. Wimmer (2005), Popescu et al. (2009). Výsledky dosavadních analýz ukazují, že žádná doposud známá transformace původního TTR zcela nezbavuje tento index závislosti na délce textu. Některé návrhy jsou však relativně úspěšné, zvláště pokud se aplikují v intervalech pohybujících se v řádu stovek a tisíců slov. Úspěšně se daří eliminovat vliv délky také prostřednictvím metody „posouvání" TTR při jeho měření (Covington a McFall 2010; Kubát a Milička 2013; Kubát 2014).



Obrázek 7.2: Vztah mezi hodnotou LTR a délkou textu N u 704 lemmatizovaných textů, jejichž délka leží v intervalu $N \in \langle 200; 6500 \rangle$ slov.

Pro ověření předpokládaného vztahu mezi tematickou koncentrací a slovním bohatstvím – tj. předpokladem, že texty s vyšším slovním bohatstvím by měly být tematicky méně koncentrované než texty s menším slovním bohatstvím – jsem zvolil dva následující postupy:

1) z množiny navržených indexů slovního bohatství, které se snaží eliminovat vliv délky textu prostřednictvím nějaké transformace, použiju ten, jenž vykazuje nejmenší závislost na délce textu.[1] Konkrétně půjde o McIntoshovu transformaci indexu opakování slov RR_{Mc} (McIntosh 1967). Budou analyzovány pouze texty, jejichž délka se pohybuje v intervalu, kde je vliv délky na RR_{Mc} prakticky nulový. Hodnoty RR_{Mc} budou následně porovnány s TK a STK prostřednictvím Kendallova korelačního koeficientu (kap. 7.1).
2) Aplikuji postup navržený Covingtonem a McFallem (2010) a porovnám jimi navržený index průměrného průběžného type-token poměru (moving average type-token ratio) s TK a STK prostřednictvím Kendallova korelačního koeficientu (kap. 7.2). Protože pro analýzu tematické koncentrace je vhodnější používat lemmata, budou zpracovány lemmatizované texty (i v případě RR_{Mc}). Půjde tak de facto o průměrný průběžný lemma-token poměr (moving average lemma-token ratio: MALTR).

[1] Míru závislosti jsem ověřoval empiricky na vzorku 1168 textů (viz Příloha).

7.1 Tematická koncentrace textu a index opakováni slov

Index opakování RR (z anglického 'repeat rate') obecně vyjadřuje míru diverzity, homogenity či bohatství daného souboru a je velmi dobře znám například v biologii. Čím více stejných prvků daný soubor obsahuje, tím je homogennější (tedy chudší, méně diverzifikovaný) a naopak. V textologii je index opakování slov definován (Popescu et al. 2009) jako

$$RR = \sum_{r=1}^{V} p_r^2, \tag{7.3}$$

kde V je velikost slovníku, tj. počet různých slovních tvarů či lemmat v textu, a p_r pravděpodobnost výskytu slova r. Pravděpodobnost lze určit na základě relativní frekvence slovního tvaru či lemmatu v textu, tj.

$$p_r = \frac{f_r}{N}, \tag{7.4}$$

kde f_r jsou absolutní frekvence a N počet slov/tokenů v textu. Následně je možné vzorec (7.3) psát jako

$$RR = \frac{1}{N^2} \sum_{r=1}^{V} f_r^2. \tag{7.5}$$

Pro ilustraci předpokládejme, že máme text o délce $N = 1000$ slov/tokenů. V případě maximálně koncentrovaného slovníku by se celý takový text skládal pouze z jediného slova, které by se tisíckrát zopakovalo, tj.

$$RR = \frac{1000^2}{1000^2} = 1,$$

což je hodnota maximálně „chudého" slovníku (a maximálně silné tematické koncentrace textu). Na druhou stranu, pokud by se každé slovo v textu vyskytlo pouze jednou, tj. index opakování slov by byl nejmenší a slovník nejbohatší (tematická koncentrace textu zase nejnižší), pak pro RR v textu o délce $N = 1000$ slov platí

$$RR = \frac{1^2 + 1^2 + 1^2 + ... + 1^2}{1000^2} = \frac{1000}{1000^2} = 0,001.$$

Minimální hodnota RR je závislá na velikosti slovníku, tj.

$$RR_{min} = \frac{1}{N^2} \sum_{r=1}^{V} \left(\frac{N}{V}\right)^2 = \frac{1}{V}; \tag{7.6}$$

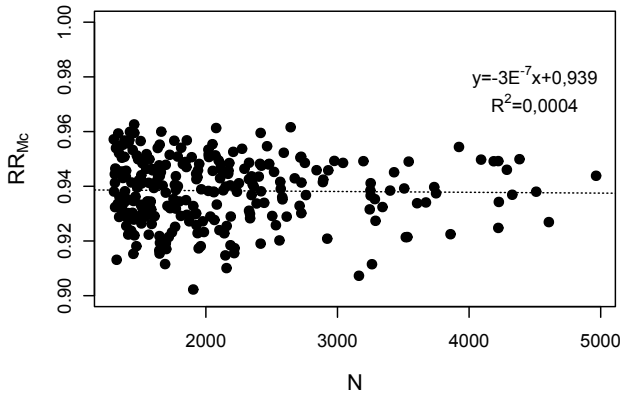
což znamená, že hodnota RR leží v intervalu $\langle 1/V; 1 \rangle$.

McIntosh (1967) navrhl úpravu tohoto indexu do podoby

$$RR_{Mc} = \frac{1 - \sqrt{RR}}{1 - 1/\sqrt{V}}, \tag{7.7}$$

což mimo jiné vedlo k výrazné eliminaci vlivu délky textu, byť ne úplné.

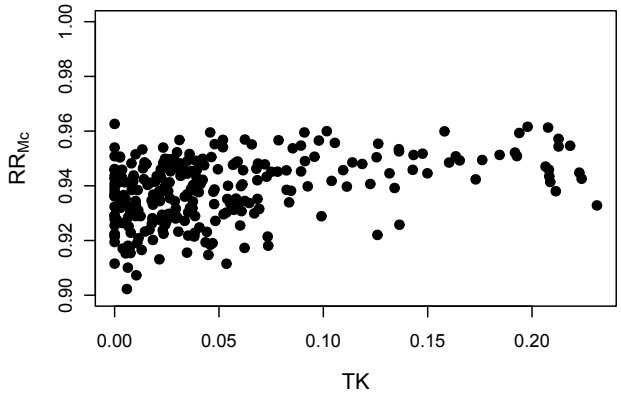
Aby bylo možné testovat hypotézu o vztahu tematické koncentrace a slovního bohatství textu, vyjádřeného prostřednictvím indexu opakování lemmat, je třeba nejdříve empiricky odvodit interval N, vyjadřující délku textu, v jehož rámci je RR_{Mc} na N nezávislý. Přitom je třeba respektovat, že použití indexů TK a STK je smysluplné pro texty o délce cca 200–6500 slov (srov. kap. 5.1). Z analyzovaného vzorku (viz Příloha 1) se jako vhodné jeví texty o délce $N \in \langle 1300; 5000 \rangle$ slov/tokenů, srov. Obr. 7.3.



Obrázek 7.3: Vztah mezi hodnotou RR_{Mc} a délkou textu N u 266 textů, jejichž délka leží v intervalu $N \in \langle 1300; 5000 \rangle$ slov/tokenů. Jedná se o interval, v němž je prakticky nulová závislost tohoto indexu na délce textu – přímka vyjadřující lineární závislost RR_{Mc} na N je téměř vodorovná, velmi nízká hodnota determinační koeficientu R^2 svědčí o velkém rozptylu a praktické nezávislosti obou veličin. Kendallův korelační koeficient má hodnotu $\tau = 0,026$; p-hodnota = 0,527, jde tedy o velmi nízkou korelaci, navíc výrazně nesignifikantní.

Pro ověření hypotézy byla aplikována korelační analýza a výsledky vyhodnoceny prostřednictvím neparametrického Kendallova testu, viz Obr. 7.4 a 7.5.

V případě obou indexů (TK i STK) se projevuje signifikantní vztah s RR_{Mc}, hypotéza tedy nebyla vyvrácena a je možné konstatovat, že texty s vyšší hodnotou tematické koncentrace (TK i STK) vykazují menší slovní bohatství než texty s tematickou

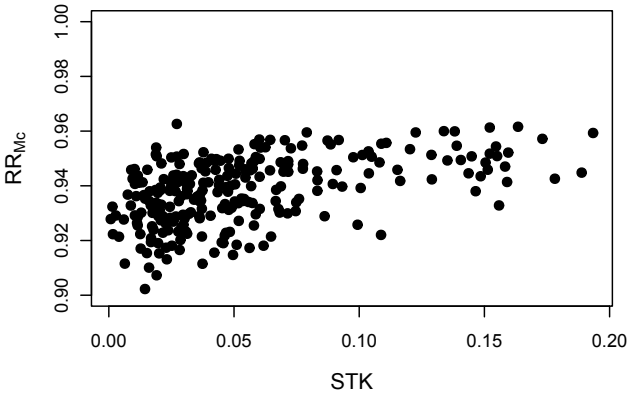


Obrázek 7.4: Vztah mezi hodnotou RR_{Mc} a TK u 266 lemmatizovaných textů, jejichž délka leží v intervalu $N \in \langle 1300; 5000 \rangle$ slov/tokenů. Výsledky jsou signifikantní ($\tau = 0,259$, p-hodnota $< 0,0001$), tj. mezi RR_{Mc} a TK je monotónní závislost.

koncentrací nižší. Je třeba zdůraznit, že vztah postulovaný hypotézou není v žádném případě nutný – indexy tematické koncentrace jsou odvozeny z vlastnosti frekvenční struktury textu, v níž má rozhodující roli h-bod. Teoreticky může nastat situace, kdy i text s velmi chudým slovníkem bude mít nulové hodnoty TK a STK, protože se nad h-bodem nevyskytne žádné autosémantikum. V tomto ohledu není překvapivé, že se silnější korelace projevuje v případě STK, u níž se vyskytuje větší počet tematických lemmat.

7.2 Tematická koncentrace textu a průměrný průběžný lemma-token poměr (MALTR)

Jednou z možností, jak eliminovat vliv délky textu nejen na TTR či LTR, ale v podstatě na každý index, je rozdělit texty na menší části o stejné délce, v každé z těchto částí změřit hodnotu sledovaného indexu a nakonec z jednotlivých dílčích hodnot vypočítat aritmetický průměr. Takto lze jednoduše analyzovat celý text, přičemž jeho celková délka nehraje roli. Tento přístup se označuje jako standardizovaný type token poměr a je například použit v softwaru *WordSmith Tools* (Scott 2013). Problém je ovšem v tom, že hranice mezi jednotlivými měřenými částmi jsou „umělé“ (10, 100, 200, 500… slov). Pro překonání tohoto problému Covington a McFall (2010) navrhli tzv. průměrný průběžný type-token poměr (MATTR), v němž je sice také arbitrárně zvolena určitá část textu (označována jako okno) pro měření daného indexu, ale ta



Obrázek 7.5: Vztah mezi hodnotou RR_{Mc} a STK u 266 lemmatizovaných textů, jejichž délka leží v intervalu $N \in \langle 1300; 5000 \rangle$ slov/tokenů. Výsledky jsou signifikantní ($\tau = 0,433$, p-hodnota $\sim$ 0,001), tj. mezi RR_{Mc} a STK je monotónní závislost.

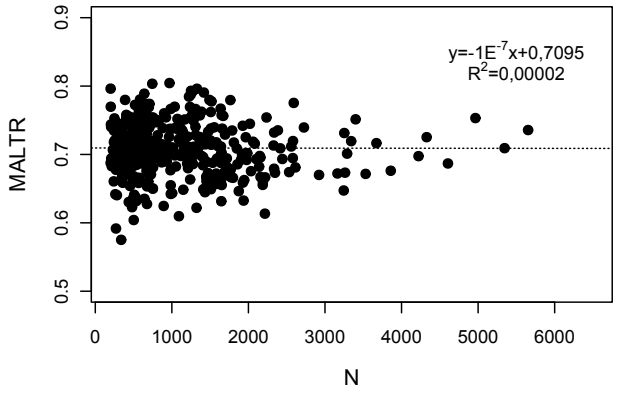
se následně posouvá vždy jen o jeden token. Takto je respektován text jakožto homogenní celek. Analogicky k původnímu MATTR lze definovat průběžný lemma-token poměr

$$MALTR(L) = \frac{\sum_{i=1}^{N-W} L_i}{W(N - W + 1)}, \tag{7.8}$$

kde W je arbitrárně zvolená velikost okna v tokenech (samozřejmě musí platit, že $W < N$), N je délka textu v tokenech a L_i je počet lemmat v jednotlivém okně.

Pro ověření hypotézy jsem použil lemmatizované texty K. Čapka různých žánrů. Konkrétně šlo o kapitoly románu (texty č. 589–642; 745–770), povídky (texty č. 771–841), cestopisy (texty č. 842–973), novinové sloupky (texty č. 989–1065) a dopisy (texty č. 1066 –1158). Z tohoto vzorku byly odebrány texty, jejichž délka neleží v intervalu $N \in \langle 200; 6500 \rangle$ slov/tokenů, ve výsledku tak bylo zpracováno 439 textů. Výpočet MALTR byl proveden pomocí softwaru MaWaTaTaRaD (Milička 2013). Jak je vidět na Obr. 7.6, mezi MALTR a délkou textu se neprojevuje žádný vztah, což potvrzují hodnoty Kendallova korelačního koeficientu ($\tau = -0,034$, p-hodnota = 0,277).

Na Obr. 7.7 a 7.8 je graficky znázorněn vztah mezi MALTR a TK, respektive STK. V obou případech se opticky jeví, že mezi MALTR a oběma indexy není žádný jednoznačný vztah. Korelační koeficienty jsou v obou případech nízké $\tau_{MALTR-TK} = -0,038$ a $\tau_{MALTR-STK} = -0,012$, Kendallův test potvrzuje v obou případech nesig-

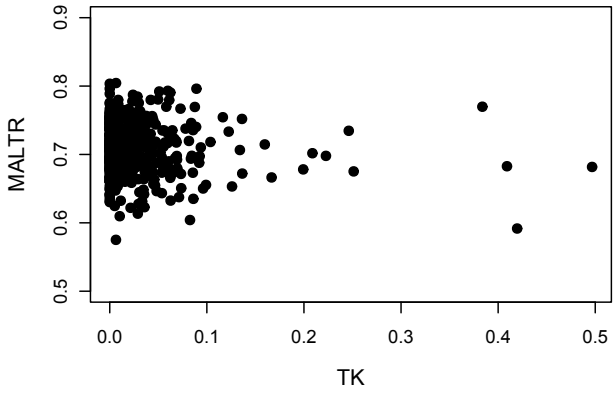


Obrázek 7.6: Vztah mezi hodnotou MALTR a délkou textu N u 439 textů, jejichž délka leží v intervalu $N \in \langle 200; 6500 \rangle$ slov/tokenů. Přímka vyjadřující lineární závislost MALTR na N je téměř vodorovná, velmi nízká hodnota determinační koeficientu R^2 svědčí o velkém rozptylu a praktické nezávislosti obou veličin. Kendallův korelační koeficient má hodnotu $\tau = -0,034$; p-hodnota = 0,277.

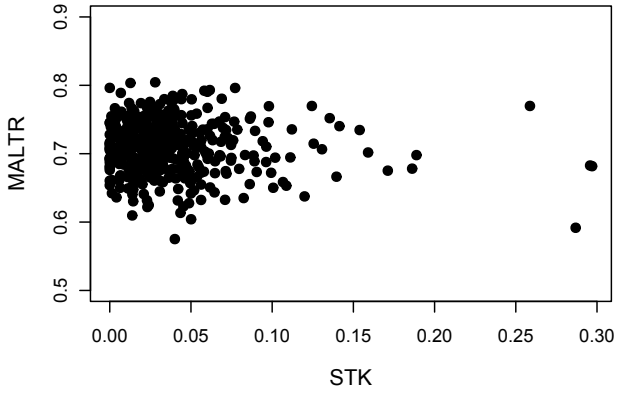
nifikantní korelaci. To znamená, že v případě měření slovního bohatství prostřednictvím MALTR je hypotéza zamítnuta.

Jedním z důvodů zamítnutí hypotézy může být nehomogennost zkoumaného vzorku. Proto znovu ověřím hypotézu na vzorcích homogennějších. V prvním případě rozdělím texty do skupin podle žánrů a budu sledovat vztah mezi MALTR a TK, respektive STK v rámci těchto skupin. Konkrétně půjde o kapitoly románu (texty č. 589 –642; 745–770), povídky (texty č. 771–841), cestopisy (texty č. 842–973), novinové sloupky (texty č. 989–1065) a dopisy (texty č. 1066–1158), přičemž z každého vzorku budou opět odebrány texty, jejichž délka neleží v intervalu $N \in \langle 200; 6500 \rangle$ slov/tokenů. Ve druhém případě budu sledovat korelaci u kapitol v rámci jedné knihy. Abych zajistil dostatečný počet porovnávaných hodnot, vybral jsem ty Čapkovy knihy, které obsahují alespoň 25 částí, knihy navíc musí být tematicky jednotné[2] (proto neberu v potaz například povídkové soubory). Konkrétně byly použity následující texty: *Krakatit* (texty č. 589–642), *Válka s mloky* (texty č. 745–770), *Anglické listy* (texty č. 842–876), *Cesta na sever* (texty č. 877–902), *Italské listy* (texty č. 903–927), *Obrázky z Holandska* (texty č. 928–944), *Výlet do Španěl* (texty č. 945–973), *Zahradníkův rok* (texty č. 1041–1066).

[2] Pojem tematická jednotnost samozřejmě není zcela jednoznačný – jako i v jiných případech jde o vlastnost, která má škálovitý charakter. *Krakatit* zřejmě reprezentuje homogennější soubor než cestopisné knihy či *Zahradníkův rok*, které na druhou stranu zase představují homogennější celky než povídkové soubory.



Obrázek 7.7: Vztah mezi hodnotou MALTR a TK u 439 lemmatizovaných textů. Mezi oběma indexy je velmi nízká korelace a je nesignifikantní, srov. hodnoty Kendallova korelačního koeficientu $\tau = -0,038$; p-hodnota = 0,25.



Obrázek 7.8: Vztah mezi hodnotou MALTR a STK u 439 lemmatizovaných textů. Mezi oběma indexy je velmi nízká korelace a je nesignifikantní, srov. hodnoty Kendallova korelačního koeficientu $\tau = -0,012$; p-hodnota = 0,738.

Tabulka 7.1: Korelační koeficienty mezi hodnotami MALTR a TK, resp. STK v rámci jednotlivých žánrů.

kapitoly románu (80 textů)	τ	p-hodnota
MALTR – TK	0,122	0,111
MALTR – STK	0,135	0,077
povídky (71 textů)	**τ**	**p-hodnota**
MALTR – TK	-0,019	0,819
MALTR – STK	-0,028	0,732
cestopisy (124 textů)	**τ**	**p-hodnota**
MALTR – TK	-0,169	0,009
MALTR – STK	-0,126	0,038
novinové sloupky (84 textů)	**τ**	**p-hodnota**
MALTR – TK	-0,089	0,253
MALTR – STK	-0,124	0,094
dopisy (80 textů)	**τ**	**p-hodnota**
MALTR – TK	-0,061	0,483
MALTR – STK	-0,037	0,632

Výsledky, viz Tab. 7.1, potvrzují převládající absenci vztahu mezi MALTR a TK i STK. Kromě vztahu MALTR – TK u cestopisů jsou navíc všechny další signifikantní vztahy na hranici zvolené hladiny významnosti (tedy v jakési „šedé zóně“). Na základě zde uvedených výsledků je tedy třeba hypotézu o vztahu MALTR a TK, respektive STK odmítnout.

7.3 Závěrečná poznámka ke vztahu tematické koncentrace a slovního bohatství textu

Výsledky analýzy vztahu tematické koncentrace a slovního bohatství nejsou zcela jednoznačné a otevírají celou řadu otázek. Slovní bohatství je sice pojmem, který je intuitivně dobře pochopitelný, velké množství různých indexů pro jeho měření je však projevem metodologické a teoretické nevyhraněnosti. Ve světle zde prezentovaných výsledků se možná zdá být oprávněná výše zmíněná reinterpretace type-token poměru (Wimmer 2005; Popescu et al. 2009), a to zřejmě i v případě průměrného prů-

Tabulka 7.2: Korelační koeficienty mezi hodnotami MALTR a TK, resp. STK u osmi textů K. Čapka.

Anglické listy	τ	p-hodnota
MALTR – TK	-0,060	0,651
MALTR – STK	0,006	0,973

Cesta na Sever	τ	p-hodnota
MALTR – TK	-0,159	0,287
MALTR – STK	-0,120	0,418

Výlet do Španěl	τ	p-hodnota
MALTR – TK	-0,003	0,984
MALTR – STK	0,057	0,666

Italské listy	τ	p-hodnota
MALTR – TK	-0,005	0,977
MALTR – STK	0,152	0,313

Krakatit	τ	p-hodnota
MALTR – TK	**0,218**	**0,020**
MALTR – STK	**0,191**	**0,042**

Válka s mloky	τ	p-hodnota
MALTR – TK	-0,095	0,512
MALTR – STK	0,022	0,896

Zahradníkův rok	τ	p-hodnota
MALTR – TK	0,140	0,384
MALTR – STK	-0,117	0,402

Obrázky z Holandska	τ	p-hodnota
MALTR – TK	**-0,451**	**0,023**
MALTR – STK	**-0,429**	**0,023**

běžného type-token poměru (MATTR), respektive lemma-token poměru (MALTR). Nevyvrácení hypotézy v případě McIntoshova indexu opakování slov naopak dovoluje setrvat na původním předpokladu, tj. že mezi oběma vlastnostmi textu existuje inverzní vztah. V každém případě se zde ukazuje, že pro řádné ověření hypotézy je třeba nejdříve důkladně teoreticky a metodologicky zpracovat problém slovního bohatství obecně. Slovní bohatství je třeba jednoznačně definovat, následně sledovat, jak se chová k základním vlastnostem textu, jako jsou délka, průběžný vývoj textu, vztah k jazykovým jednotkám atd. Pokud se to podaří, je možné se slovním bohatstvím pracovat jako s jevem, který lze dát do souvislosti s jinými vlastnostmi, což by umožnilo vytvořit teorii textu založenou na testovatelných hypotézách.

8

Tematická koncentrace a analýza klíčových slov

Tematická koncentrace vyjadřuje míru, s jakou je daný text zaměřen na dané téma či témata. Pomocí této metody však lze také určit, která slova reprezentují hlavní téma či témata textu a s jakou mírou – jde o tzv. tematická slova (kap. 2.1). V tomto ohledu lze tematickou koncentraci považovat za metodu blízkou tzv. analýze klíčových slov (Adolphs 2006; Bondi a Scott 2010; Scott a Tribble 2006; Sinclair 1991; Stubbs 1996). Cílem této kapitoly je 1) srovnat metodologické aspekty obou metod, 2) porovnat jejich výsledky.

8.1 Srovnání metod analýzy klíčových slov a tematické koncentrace

Kvantitativní analýza klíčových slov[1] je založena na porovnání frekvence jednotlivých slov v textu s frekvencemi v tzv. referenčním korpusu. Pokud se dané slovo vyskytne v textu signifikantně častěji než v referenčním korpusu, je považováno za slovo klíčové. Pro určení významnosti se používají různé statistické testy, např. chí-kvadrát (Cvrček a Richterová 2013a) či log-likelihood test (Cvrček a Richterová 2015a). Je zřejmé, že zásadní roli v této metodě má volba referenčního korpusu. V případě, že se jedná o analýzu předem jasně vymezené skupiny textů (např. literární dílo jednoho autora), v níž lze vymezit vlastnosti tzv. základního souboru (populace) a následně pak stanovit i reprezentativní vzorek textů, je volba referenčního korpusu celkem neproblematická. Zpravidla se ale předpokládá, že referenční korpus by měl reprezentovat *obecný úzus*. Jednotlivé texty jsou pak porovnávány právě vzhledem k tomuto úzu, srov. popis nástroje *KWords* (Cvrček a Vondřička 2013) určeného pro analýzu klíčových slov, vyvinutého v Českém národním korpusu: „Aplikace *KWords* slouží k analýze textů na základě jejich srovnání s obecným územ (referenčním korpusem). Jejím cílem je identifikovat tzv. klíčová slova (keywords), což jsou slovní tvary, která se ve zkoumaném textu objevují významně častěji než v referenčním korpusu, který má zrcadlit běžný jazykový úzus. Tato klíčová slova pak slouží jako základ pro textovou analýzu a interpretaci.“ (Cvrček a Richterová 2015b). Tento přístup je založen na předpokladu, že je v principu možné vytvořit tzv. obecný reprezentativní korpus textů, který by odrážel vlastnosti jazyka jako takového (případně jedné z jeho forem – psané nebo mluvené), srov. Králík a Šulc (2005), Leech (2007), Cvrček, Kováříková (2011), Křen (2013). Proti tomuto předpokladu však existují velmi vážné principiální

[1] K nekvantitativnímu pojetí viz pozn. 1 z kap. 2.

námitky (Králík 2013, Chromý 2014, Čech 2014b), které ve svém důsledku problematizují nikoliv samotnou metodu analýzy klíčových slov, ale právě tu interpretaci, jež předpokládá, že lze identifikovat slova vyskytující se významně častěji než v běžném úzu.

Výhodou analýzy klíčových slov je eliminace „neadekvátního" vlivu frekvence (bráno vzhledem k analýze výrazů reprezentujících hlavní témata textu). Pokud se totiž nějaké slovo vyskytuje s vysokou relativní frekvencí jak ve sledovaném textu, tak i referenčním korpusu, není prostřednictvím této metody určeno jako klíčové. Například při analýze tematické koncentrace novoročních prezidentských projevů (Čech 2014a) se jako tematické slovo s poměrně velkou tematickou vahou téměř ve všech projevech objevovalo lemma 'rok'. Je zřejmé, že vysoká frekvence tohoto lemmatu není dána tím, že by se jednotliví prezidenti zaměřovali na tento fenomén, ale je dána specifičností žánru – všichni prezidenti zpravidla přejí občanům úspěšný 'rok', bilancují, co se v předešlém 'roce' odehrálo, atp. V případě analýzy klíčových slov a za předpokladu, že referenční korpus tvoří všechny novoroční projevy, se lemma 'rok' mezi klíčovými slovy vůbec neobjeví. Dokonce i v případě, kdy je referenčním korpusem obecný korpus češtiny SYN2010 (Křen et al. 2010), slovní tvar 'rok' zdaleka nevykazuje nejvyšší hodnoty klíčovosti.[2] Na druhou stranu však analýza klíčových slov preferuje slova, která se sice v textu vyskytnou s malou frekvencí, ale jelikož se vůbec nevyskytují v referenčním korpusu, jsou označena jako klíčová, navíc s velmi vysokou hodnotou klíčovosti (viz níže); konkrétně, vezmeme-li *jakýkoliv* český text a vložíme-li do něj třikrát výraz, který se v referenčním korpusu nevyskytuje – například výraz 'oslových' pro referenční korpus SYN2010 –, bude pokaždé tento výraz označen jako klíčové slovo (za předpokladu, že je nastavena minimální frekvence klíčových slov f = 3, což je standardní nastavení aplikace *KWords*).

Při porovnání obou metod je třeba především brát v potaz, že každá z nich byla vytvořena primárně k jinému účelu: cílem analýzy klíčových slov je určení množiny slov reprezentujících hlavní témata textu, včetně výpočtu míry jejich klíčovosti, zatímco cílem analýzy tematické koncentrace je vyhodnocení vlastnosti textu jako celku, přičemž detekce tematických slov je jen jedním z dílčích kroků. Další rozdíl spočívá v možnosti statistického testování rozdílů mezi texty v případě tematické koncentrace, což u analýzy klíčových slov není možné, pokud je mi známo. Dále, nutnost použití referenčního korpusu u klíčových slov může představovat jak výhodu (viz eliminace „neadekvátního" vlivu frekvence), tak i nevýhodu, zejména v případě, že buď nemáme jasně vymezený okruh textů, které budou zkoumány, nebo akceptujeme kritický postoj vůči pojetí tzv. reprezentativních obecných korpusů. Nelze tedy říct, že jedna metoda je lepší než druhá – vše záleží na konkrétním badatelském cíli a, do jisté míry, i na zvolených teoreticko-metodologických východiscích.

[2] Vzhledem k tomu, že nástroj *KWords* analyzuje pouze slovní tvary, nikoliv lemmata, není možné tento závěr zcela jednoznačně vztáhnout k studii Čech (2014a), v níž jsou použita lemmata. Ale celková tendence je jasná i z porovnání slovních tvarů.

8.2 Srovnání výsledků analýzy klíčových slov a tematické koncentrace

Vyjděme z předpokladu, že cílem badatele je snaha zjistit, která slova reprezentují hlavní téma či témata v textu. Takový záměr může mít třeba i ryze komerční důvody: například jde o to určit, které výrazy reprezentují hlavní témata v uživatelských recenzích různých výrobků (Veselovská a Čech 2014). V následujících řádcích budu sledovat, k jakým výsledkům lze dojít prostřednictvím obou metod, a tyto výsledky porovnám. Pro analýzu klíčových slov použiji aplikaci Českého národního korpusu *KWords* (REF.), pro analýzu tematické koncentrace indexy TK a STK (kap. 2.1 a 3.1). Aplikace *KWords* neumožňuje zpracovat lemmata, proto budou jako jednotky použity slovní tvary, a to i při analýze tematické koncentrace, aby byly výsledky porovnatelné. Procedura vyhodnocení klíčovosti je následující: „Text vložený uživatelem se nejprve roztokenizuje způsobem, který je identický s tokenizací korpusových dat. V druhém kroku je spočtena frekvence všech slov v analyzovaném textu (s výjimkou těch, které uživatel z analýzy vyloučí prostřednictvím tzv. stop-listu, například předložky, spojky, čísla apod.). Následuje porovnání frekvencí v textu a v referenčním korpusu. Pro jednotky, u nichž byl zaznamenán statisticky signifikantní rozdíl (podle zvoleného statistického testu – chí-kvadrát či log-likelihood), je dále vypočítána hodnota DIN (difference index) vypovídající o relevanci daného rozdílu:

$$\mathrm{DIN} = 100\frac{\mathtt{RelFq(Ttxt)} - \mathtt{RelFq(RefC)}}{\mathtt{RelFq(Ttxt)} + \mathtt{RelFq(RefC)}}, \tag{8.1}$$

kde $\mathtt{RelFq(Ttxt)}$ je relativní frekvence jevu ve zkoumaném textu (target text) a $\mathtt{RelFq(RefC)}$ je relativní frekvence téhož jevu v referenčním korpusu.“ (Cvrček a Richterová 2015b).

Vzhledem k tomu, že ve statistice standardně používané hladiny významnosti ($\alpha = 0,05$ či $\alpha = 0,01$) jsou příliš „silné“, tj. vedou k tomu, že analýza vrací desítky klíčových slov, použiji nejnižší možnou hladinu významnosti, kterou lze v *KWords* nastavit, konktrétně $\alpha = 0,0001$. Dále je v aplikaci třeba zadat minimální frekvenci slova v textu, v nabídce se objevují možnosti ležící v intervalu $f_{min} \in \langle 2; 5\rangle$. Tento bod analýzy se mi jeví jako velmi problematický. Je totiž evidentní, že u delších textů, například v řádu tisíců či desetitisíců slov, má nastavení tohoto parametru výrazně jiný vliv na výsledky než u textu v řádu desítek či stovek slov. V důsledku toho pak počet klíčových slov bude významně ovlivňován celkovou délkou textu. Zde by se zdálo smysluplnější pracovat s relativní frekvencí slova, tu ale u aplikace *KWords* nelze nastavit. Pro minimální frekvenci slova volím standardní nastavení aplikace $f_{min} = 3$.

Za účelem porovnání obou metod budou vybrány texty, jejichž délka se pohybuje v intervalu $N \in \langle 200; 6000\rangle$ slov/tokenů, což je interval, v němž jsou TK i STK na délce textu nezávislé (srov. Obr. 5.10–12). Abych získal první představu o možném vlivu délky textů na analýzu klíčových slov (či o absenci tohoto vlivu), na úvod porovnám dva texty, jejichž délka se liší o jeden řád. Konkrétně novinový článek *V Beskydech blesk zapálil chatu, vítr lámal stromy* (text č. 267) ($N = 515$), srov. Tab 2.4, a esej V. Havla

O smyslu Charty 77 (text č. 432) ($N = 5189$). Výsledky, prezentované v Tab. 8.1 a 8.2, ukazují, že

1) v obou textech se slova, která byla prostřednictvím měření TK a STK označena jako tematická, objevují v seznamu klíčových slov;
2) analýza klíčových slov (na základě výše uvedeného nastavení) detekuje větší počet výrazů než tematická koncentrace;
3) při analýze klíčových slov hraje zvlášť významnou roli nastavení minimální frekvence; například u textu *V Beskydech blesk zapálil chatu, vítr lámal stromy* (text č. 267) by bylo při nastavení $f_{min} = 2$ určeno 36 klíčových slov, v případě $f_{min} = 4$ určena čtyři klíčová slova a při $f_{min} = 5$ určena tři klíčová slova; u textu *O smyslu Charty 77* (text č. 432) by bylo při nastavení $f_{min} = 2$ určeno 106 klíčových slov, v případě $f_{min} = 4$ určeno 50 klíčových slov a při $f_{min} = 5$ určeno 37 klíčových slov;
4) v případě delšího textu je určeno pětkrát více slov jako klíčových než u textu kratšího.

Tyto dílčí výsledky potvrzují vliv délky textu na počet klíčových slov. V kap. 5 jsem zmínil, že délka textu je faktorem, který problematizuje použití značné části indexů. Pokud se ukáže, že analýza klíčových slov je na délce textu závislá, je třeba při interpretaci výsledků vždy vzít tento fakt v úvahu.

Velmi dobrým způsobem, jak ověřit vztah délky textu a nějakého indexu, je sledovat jeho vývoj v daném textu. Analogicky k postupu uvedeném v kap. 5.2 jsem zkoumal vztah mezi kumulativní délkou textu a počtem klíčových slov u prvních čtyř povídek K. Čapka ze sbírky *Povídky z druhé kapsy* (texty č. 809–812) a také mezi kumulativní délkou textu a počtem tematických slov jak v rámci TK, tak i STK u stejných textů (Obr. 8.1–3). Výsledky poukazují na strmý lineární vztah mezi délkou textu a počtem klíčových slov – pro všechny analyzované texty má koeficient determinace, vyjadřující míru shody mezi daty a lineární funkcí, hodnoty $R^2 > 0,9$, což svědčí o vysoké míře shody. V případě TK a STK je patrný také určitý nárůst, ovšem zdaleka ne tak strmý (srov. hodnoty na ose y), a s výjimkou povídky *Zmizení pana Hirsche* nelze ani vztah mezi délkou a oběma indexy dobře modelovat prostřednictvím funkce s lingvisticky interpretovatelnými parametry.

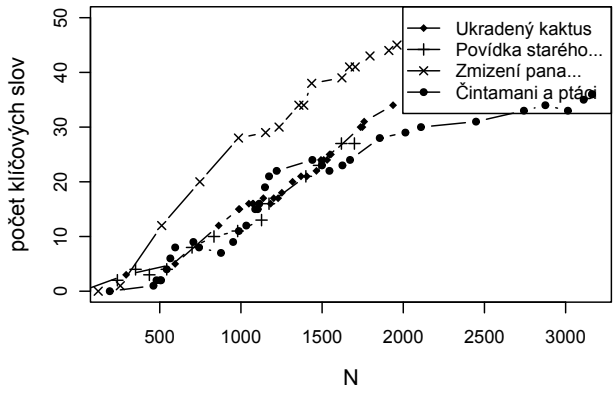
Pro ověření tendencí, které se objevily u výše uvedených textů, jsem analyzoval dalších 80 textů K. Čapka (texty č. 745–770; 877–902; 974–1001). Zde jsem se zaměřil na vztah celkové délky textu a počtu klíčových, respektive tematických slov. V případě klíčových slov opět získáváme lineární závislost s vysokým korelačním koeficientem ($\tau = 0,723$, p-hodnota $< 0,0001$). U tematických slov můžeme také sledovat jistý nárůst, u TK je ovšem nesignifikantní ($\tau = 0,143$, p-hodnota = 0,103), u STK signifikantní ($\tau = 0,305$, p-hodnota = 0,0002). Z těchto výsledků je patrné, že rostoucí počet klíčových slov u delších textů je způsoben vlastností měření, nikoliv intencí autora. Těžko lze totiž očekávat, že by se ve většině textů s délkou například cca 2000 slov (což je mimochodem přibližně stejný počet slov od začátku této kapitoly do tohoto místa) objevilo kolem 40 klíčových slov a že by zejména tento počet lineárně

Tabulka 8.1: Výsledky analýzy klíčových slov a tematické koncentrace v textu *V Beskydech blesk zapálil chatu, vítr lámal stromy* (text č. 267) (N = 515). DIN označuje míru rozdílu (viz vzorec (8.1)), f(text) frekvenci výrazu v textu, f(SYN2010) frekvenci výrazu v referenčním korpusu SYN2010. Pro minimální frekvenci slova je použito standardní nastavení aplikace $f_{min} = 3$. TV a STV jsou tematické váhy slova (viz vzorec 2.6) v rámci TK a STK. V šedých buňkách jsou označena klíčová slova, která se vyskytla jako tematická jak v rámci TK, tak STK, tučně jsou označena slova, která se vyskytla jako tematická pouze v rámci TK.

r	klíčová slova	DIN	f(text)	f(SYN2010)	r	tematická slova (TK)	TV	r	tematická slova (STK)	STV
1	*zaplavila*	99,9003	5	500	1	*hasiči*	0,07407	1	*hasiči*	0,067340
2	*sklepů*	99,8604	3	420	2	*voda*	0,02593	2	*voda*	0,041246
3	***hasiči***	99,6260	10	3755				3	*zaplavila*	0,016835
4	*okrese*	99,6090	3	1178				4	*vodu*	0,003367
5	*hasičů*	99,2505	3	2262						
6	*čtk*	99,2036	3	2404						
7	***voda***	98,4719	7	10803						
8	*vodu*	98,0432	4	7922						
9	*vítr*	97,9534	3	6217						
10	*stromy*	97,7398	3	6873						
11	*kraji*	96,3715	3	11111						
12	*mluvčí*	94,1398	3	18151						

Tabulka 8.2: Výsledky analýzy klíčových slov a tematické koncentrace v textu *O smyslu Charty 77* (text č. 432) (N = 5189). DIN označuje míru rozdílu (viz vzorec (8.1)), f(text) frekvenci výrazu v textu, f(SYN2010) frekvenci výrazu v referenčním korpusu SYN2010. Pro minimální frekvenci slova je použito standardní nastavení aplikace $f_{min} = 3$. TV a STV jsou tematické váhy slova (viz vzorec 2.6) v rámci TK a STK. V šedých buňkách jsou označena klíčová slova, která se vyskytla jako tematická jak v rámci TK, tak STK, tučně jsou označena slova, která se vyskytla jako tematická pouze v rámci TK.

r	klíčová slova	DIN	f(text)	f (SYN 2010)	r	tematická slova (TK)	TV	r	tematická slova (STK)	STV
1	*nedělitelnosti*	99,9610	3	12	1	*charta*	0,011189	1	*charta*	0,007741
2	*napřímení*	99,9537	4	19	2	*charty*	0,001277	2	*charty*	0,003055
3	***charta***	99,9367	42	273				3	*mravní*	0,002006
4	*chartou*	99,9188	6	50						
5	*pohnutka*	99,9059	3	29						
6	*spoluodpovědnost*	99,8929	3	33						
7	*bezmocí*	99,8669	3	41						
8	*chartu*	99,7373	6	162						
9	***charty***	99,6959	23	719						
10	*mravního*	99,6285	5	191						
11	*mravní*	99,4515	20	1129						
12	*říkajíc*	99,4462	3	171						
13	*eventuální*	99,0690	3	288						
14	*veskrze*	98,9500	3	325						
15	*totalitní*	98,6957	4	539						
(...)	(...)	(...)	(...)	(...)						
73	*jen*	39,0090	25	225164						

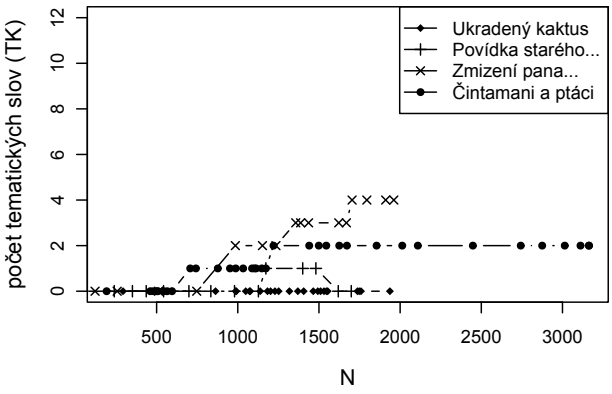


Obrázek 8.1: Vztah mezi počtem klíčových slov a kumulativní délkou textu N u prvních čtyř povídek K. Čapka ze sbírky *Povídky z druhé kapsy: Ukradený kaktus* (text č. 809), *Povídka starého kriminálníka* (text č. 810), *Zmizení pana Hirsche* (text č. 811), *Čintamani a ptáci* (text č. 812). Každý text byl segmentován na odstavce. Jednotlivé body v grafu reprezentují kumulativně sloučené odstavce daného textu. Křivky vyjadřují průběh změn počtu klíčových slov v kumulativně sloučených odstavcích v jednotlivých textech.

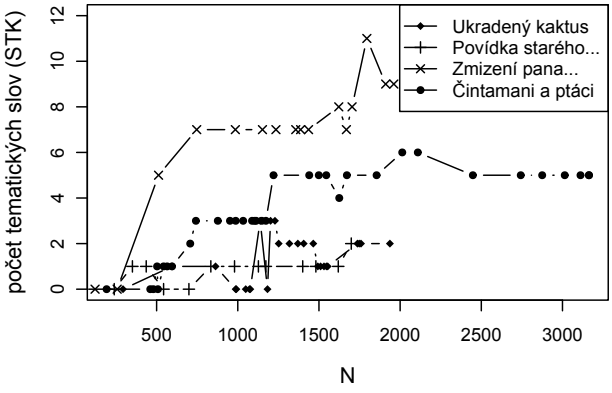
rostl v důsledku toho, že s narůstající délkou textu se autorovi nabízí možnost prezentovat více témat, respektive výrazů reprezentujících daná témata. U tematických slov je situace lingvisticky mnohem srozumitelnější. Celkový počet tematických slov u obou indexů se zdá být „rozumný" v tom smyslu, že se dá dobře předpokládat, že odráží intenci autora. To nepřímo potvrzují analýzy vztahu počtu tematických slov a kumulativní délky, kde je vidět nejen nárůst, ale i pokles, případně i ustálení počtu tematických slov na jedné hodnotě. Signifikantní nárůst v případě STK se projevuje v řádu jednotek slov, což interpretuji, na rozdíl od případu klíčových slov, jako důsledek toho, že se autorovi nabízí více prostoru pro rozšíření počtu dominujících témat. Navíc, nízká hodnota koeficientu determinace $R^2 = 0,2938$ svědčí o velkém rozptylu, zvláště porovnáme-li ji s hodnotou u klíčových slov $R^2 = 0,8217$.

8.3 Závěrečná poznámka ke vztahu tematické koncentrace a analýzy klíčových slov

Analýza klíčových slov stejně jako tematická koncentrace umožňuje v textech detekovat výrazy, které nějakým způsobem souvisejí s hlavním tématem či tématy textu. Obě jsou založeny na vyhodnocení frekvenčních charakteristik výrazů. Za hlavní ne-

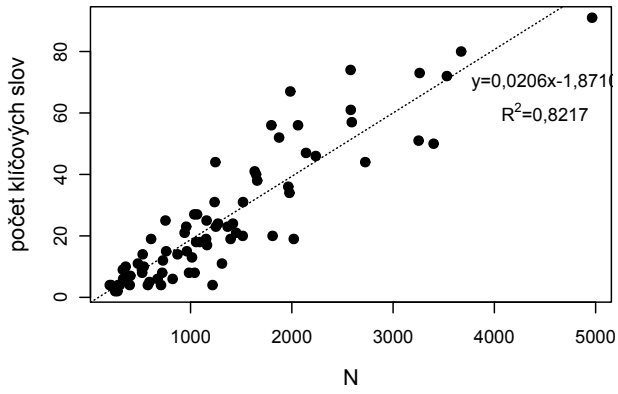


Obrázek 8.2: Analogie Obrázku 8.1 pro vztah mezi počtem tematických slov v rámci TK a kumulativní délkou textu N.

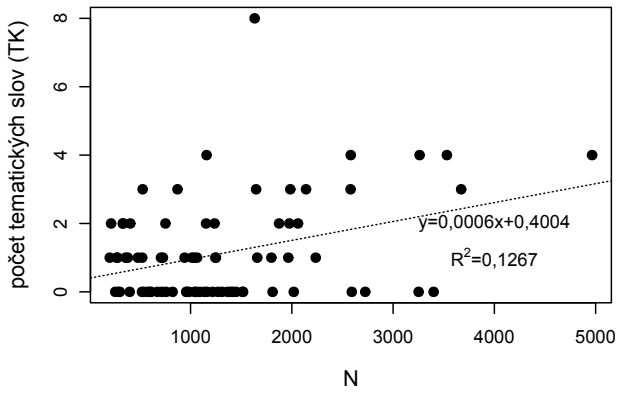


Obrázek 8.3: Analogie Obrázku 8.1 pro vztah mezi počtem tematických slov v rámci STK a kumulativní délkou textu N.

dostatek analýzy klíčových slov považuji její závislost na délce textu, i když s velkou pravděpodobností je tento nedostatek odstranitelný tím, že by se namísto absolutní

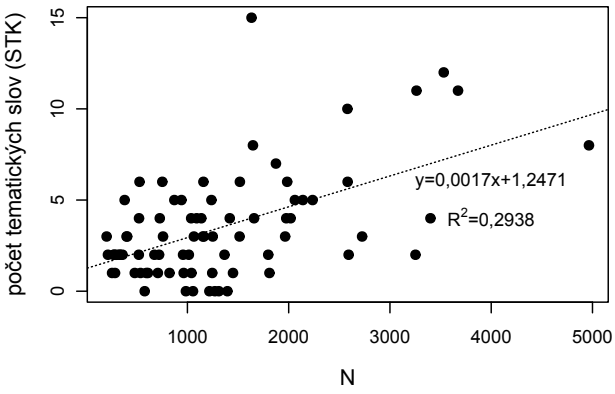


Obrázek 8.4: Vztah mezi počtem klíčových slov a délkou textu N u 80 textů K. Čapka (texty č. 745–770; 877–902; 974–1001).



Obrázek 8.5: Vztah mezi počtem tematických slov v rámci TK a délkou textu N u 80 textů K. Čapka (texty č. 745–770; 877–902; 974–1001).

minimální frekvence zadala hodnota minimální relativní frekvence. Určení tematic-



Obrázek 8.6: Vztah mezi počtem tematických slov v rámci STK a délkou textu N u 80 textů K. Čapka (texty č. 745–770; 877–902; 974–1001).

kých slov prostřednictvím TK a STK se jeví jako „robustnější“ metoda v tom smyslu, že nevyžaduje ad hoc rozhodnutí, jako je již výše zmíněné zadání minimální frekvence, či volbu referenčního korpusu.

III VYUŽITÍ TEMATICKÉ KONCETRACE V TEXTOLOGII

9

Asociativní tematická struktura textu

Analýza tematické koncentrace umožňuje v textu detekovat slova, která reprezentují hlavní téma, tzv. tematická slova (viz kap. 2.1 a 8). Jednotlivá tematická slova mohou, ale nemusí být ve vzájemném vztahu. Na jedné straně je možné si představit text, v němž se vyskytují dvě tematická slova, přičemž jedno dominuje v úvodní části textu, druhé slovo v části závěrečné a obě slova se nikdy společně nevyskytnou v jedné větě či odstavci; na straně druhé text, v němž se dvě tematická slova vyskytují vždy společně v rámci věty či odstavce. Pokud by byla asociativnost tematických slov definována na základě společného výskytu v rámci věty či odstavce, v prvním textu by obě tematická slova byla neasociována, kdežto v druhém by byla asociována maximálně. Měření asociativnosti tak umožňuje sledovat vztahy mezi tematickými slovy a prostřednictvím grafu vytvořit tematickou asociativní strukturu textu. Dále je možné stanovit míru asociativnosti mezi jednotlivými dvojicemi tematických slov a následně určit celkovou míru tematické asociativnosti textu.

9.1 Měření asociace tematických slov

Metoda měření asociace tematických slov je založena jednak na zpracování tematických charakteristik textu jako celku, konkrétně jde o určení tematických slov, jednak na následném vyhodnocení chování tematických slov v rámci zvolených segmentů daného textu. Za segmenty je možné zvolit věty, odstavce, strofy básně atp., v analýze prezentované níže je segment reprezentován větou. V souladu s poznatky z kap. 5 je pro tento způsob analýzy vhodné volit texty o délce $N \in \langle 200; 6500 \rangle$ slov/tokenů, což je délka, v jejímž rámci jsou hodnoty TK a STK na délce textu nezávislé.

Nejjednodušší způsob určení míry asociace je zřejmě následující. Pro dvojici slov A a B, u kterých je měřena asociativnost, se spočítá počet vět x, ve kterých se v textu společně vyskytují, a tento počet se vydělí počtem vět M, v nichž se vyskytuje z dané dvojice aspoň jedno slovo, tj.

$$a_{(A,B)} = \frac{x}{M}. \tag{9.1}$$

Tento postup měření je velmi jednoduchý a umožňuje získat určitý vhled do tematické asociační struktury textu. Jeho výhodou je, že míra asociativnosti a leží v intervalu $\langle 0; 1 \rangle$ a že umožňuje porovnávat texty různé délky. Asociativnost je zde však založena na prostém souvýskytu dvojice výrazů v rámci zvoleného segmentu, nebere se v úvahu pravděpodobnost toho, že se oba výrazy mohou vyskytnout společně. Míra pravděpodobnosti záleží totiž nejen na počtu vět, v nichž se daná slova vyskytují

společně a zvlášť, ale i na celkové frekvenci slov v textu. V tomto ohledu představuje výpočet pomocí vzorce (9.1) dost velké zjednodušení (či „zploštění) pohledu na asociace.

Wimmer et al. (2003) a Tuzzi et al. (2010) pracují s obecnou metodou měření asociativnosti, která je založena na vyhodnocení pravděpodobnosti souvýskytu sledovaných výrazů v daném segmentu a která umožňuje stanovit, zda je pozorovaný souvýskyt signifikantní, či nikoliv (na zvolené hladině významnosti). V principu jde o to detekovat jako asociované jen takové dvojice slov, které se vyskytnou ve větě společně *navzdory* kombinatorickým důsledkům plynoucím z frekvenčních charakteristik. To, že se tak děje právě navzdory očekávané pravděpodobnosti, lze interpretovat jako důsledek jiných faktorů, například sémantických charakteristik výrazů, intence autora atd. V následujících řádcích aplikuji jimi navrženou metodu pro analýzu asociace tematických lemmat.

Postup výpočtu asociativnosti budu ilustrovat na odborném článku *Nepředstavitelně krátké laserové impulsy* (text č. 365). Text obsahuje $N = 79$ vět a čtyři tematická lemmata – 'laserový', 'impuls', 'délka', 'výkon' –, která byla jako tematická označena prostřednictvím indexu TK. Pro každou dvojici tematických lemmat byl vypočítán počet vět, ve kterých se společně daná dvojice vyskytla. V Tab. 9.1 jsou uvedeny jednotlivé dvojice tematických lemmat, které jsou seřazeny podle míry asociativnosti na základě vzorce (9.1).

Tabulka 9.1: Jednotlivé dvojice tematických lemmat v *Nepředstavitelně krátké laserové impulsy* (text č. 365). x označuje počet vět, ve kterých se dvojice v textu vyskytla společně, m frekvenci prvního lemmatu z dané dvojice v textu, n frekvenci druhého lemmatu z dané dvojice, a míru asociace podle vzorce (9.1) a N celkový počet vět (zde $N = 79$).

dvojice tematických lemmat	x	m	n	M	a	mn/N
impuls – délka	11	32	18	30	0,367	7,3
laserový – impuls	13	36	32	40	0,325	14,6
laserový – výkon	7	36	18	38	0,184	8,2
impuls – výkon	6	32	18	36	0,167	7,3
laserový – délka	6	36	18	38	0,158	8,2
délka – výkon	4	18	18	29	0,138	4,1

Při aplikaci postupu, jenž bere v úvahu pravděpodobnost, je třeba pro každou dvojici tematických lemmat vypočítat, jaká je na základě frekvence jednotlivých lemmat pravděpodobnost jejich souvýskytu za daných podmínek. Dvě lemmata lze následně považovat za asociovaná, pokud je hodnota pozorovaného počtu souvýskytů dvou lemmat větší než hodnota očekávaná, tj. pokud

$$x > \frac{mn}{N}, \tag{9.2}$$

a pokud je pravděpodobnost P tohoto souvýskytu nižší než zvolená hladina významnosti α. Pravděpodobnost se vypočítá podle vzorce[1]

$$P(X \geq x) = \sum_{k=x}^{\min(m,n)} \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}. \tag{9.3}$$

V Tab. 9.1 je pouze jediná dvojice lemmat, která splňuje podmínku vyjádřenou ve vzorci (9.2), konkrétně 'impuls' – 'délka'. Na základě vzorce (9.3) pak pro tuto dvojici získáváme

$$\begin{aligned}
P(X \geq x) &= \sum_{k=11}^{18} \frac{\binom{32}{k}\binom{79-32}{18-k}}{\binom{79}{18}} \\
&= \frac{\left(\frac{32!}{11!(32-11)!}\right)\left(\frac{47!}{7!(47-7)!}\right) + \left(\frac{32!}{12!(32-12)!}\right)\left(\frac{47!}{6!(47-6)!}\right)}{\frac{79!}{18!(79-18)!}} + \\
&+ \frac{\left(\frac{32!}{13!(32-13)!}\right)\left(\frac{47!}{5!(47-5)!}\right) + \left(\frac{32!}{14!(32-14)!}\right)\left(\frac{47!}{4!(47-4)!}\right)}{\frac{79!}{18!(79-18)!}} + \\
&+ \frac{\left(\frac{32!}{15!(32-15)!}\right)\left(\frac{47!}{3!(47-3)!}\right) + \left(\frac{32!}{16!(32-16)!}\right)\left(\frac{47!}{2!(47-2)!}\right)}{\frac{79!}{18!(79-18)!}} + \\
&+ \frac{\left(\frac{32!}{17!(32-17)!}\right)\left(\frac{47!}{1!(47-1)!}\right) + \left(\frac{32!}{18!(32-18)!}\right)\left(\frac{47!}{0!(47-0)!}\right)}{\frac{79!}{18!(79-18)!}} + \\
&= 0,041,
\end{aligned}$$

což je hodnota nižší než hladina významnosti $\alpha = 0,05$. Proto jsou obě lemmata na této hladině významnosti označena jako asociovaná. V případě ostatních dvojic lemmat je jejich souvýskyt důsledkem relativně vysoké frekvence lemmat v daném počtu vět, nikoliv vlivem asociace, jak je definovaná výše.

Vzhledem k vlivu vysoké frekvence lemmat, jež jsou jako tematická označena na základě indexu TK, není až tak překvapivé, že se vzájemně neasociují. Pro poznání asociativní tematické struktury textu se tak jeví vhodnější pracovat s tematickými lemmaty, která jsou jako tematická označena na základě indexu STK. V případě textu *Nepředstavitelně krátké laserové impulsy* (text č. 365) se vyskytuje 11 takových tematických lemmat: 'laserový', 'impuls', 'délka', 'výkon', 'laser', 'vlnový', 'energie', 'krátký', 'médium', 'prostředí', 'rezonátor'. Na základě výše uvedeného postupu jsou jako asociované označeny dvojice lemmat v Tab. 9.2. Míra asociativnosti je nepřímo úměrná pravděpodobnosti $P(X > x)$.

[1] Pro detailní postup výpočtu viz Wimmer et al. (2003, s. 196n).

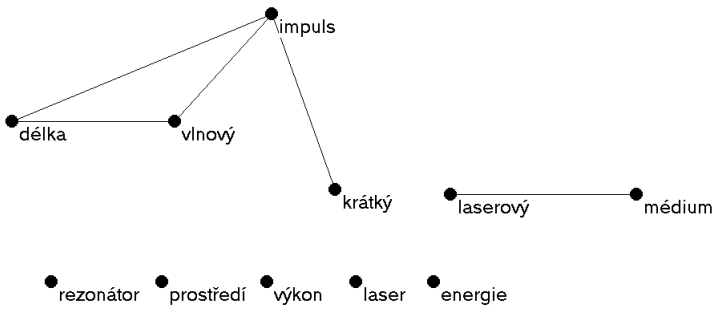
Tabulka 9.2: Jednotlivé dvojice tematických lemmat v *Nepředstavitelně krátké laserové impulsy* (text č. 365). x označuje počet vět, ve kterých se dvojice v textu vyskytla společně, m frekvenci prvního lemmatu z dané dvojice v textu, n frekvenci druhého lemmatu z dané dvojice, a míru asociace podle vzorce (9.1) a N celkový počet vět (zde N = 79).

dvojice tematických lemmat	x	m	n	mn/N	$P(X > x)$
délka – vlnový	11	18	12	2,7	< 0,001
impuls – krátký	8	32	10	4,1	0,010
laserový – médium	7	36	9	4,1	0,043
impuls – vlnový	8	32	12	4,9	0,047
impuls – délka	11	32	18	7,3	0,041

Na základě tohoto způsobu měření je následně možné vyhodnotit asociativní strukturu celého textu, viz následující kapitola.

9.2 Měření asociativní tematické struktury textu

Asociativní vztahy mezi lemmaty se dají jednoduše znázornit prostřednictvím grafu: každá dvě asociovaná lemmata se spojí čarou, viz Obr. 9.1. Grafické znázornění



Obrázek 9.1: Grafické vyjádření asociativní tematické struktury textu *Nepředstavitelně krátké laserové impulsy* (text č. 365) na základě údajů z Tab. 9.2.

je nejen prostředkem přehledného vyjádření těchto vztahů, ale na základě vlastností grafu (srov. Demel 2002; Caldareli 2007; Newman 2011) je dále možné kvantifikovat asociativní tematickou strukturu textu. Tato kvantifikace se dá použít pro porovnání asociativních tematických struktur různých textů. Velmi jednoduchý způsob předsta-

vuje měření tzv. hustoty grafu.[2] Pokud jsou všechny prvky grafu propojeny, jedná se o graf s maximální hustotou, není-li propojen žádný uzel, graf má hustotu minimální. V případě analýzy asociativní tematické struktury textu lze nazvat text plně tematicky asociovaným tehdy, jsou-li propojena všechna lemmata navzájem, a tematicky neasociovaným, nejsou-li spojena žádná lemmata. Vydělíme-li pozorovaný počet hran v grafu maximálním možným počtem hran, získáme hodnotu hustoty grafu, která vyjadřuje míru asociativní tematické struktury textu C, tj.

$$C = \frac{k}{\frac{r(r-1)}{2}} = \frac{2k}{r(r-1)}, \tag{9.4}$$

kde k je pozorovaný počet hran a r je počet prvků grafu, zde počet tematických lemmat. Hodnoty C leží v intervalu $\langle 0; 1\rangle$ Pro graf na Obr. 9.1 dostáváme

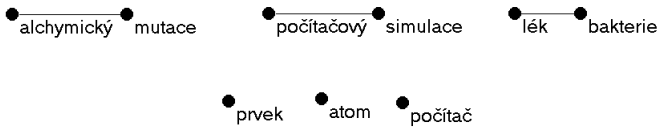
$$C_{\textit{Nepředstavitelně...}} = \frac{2 \cdot 5}{11(11-1)} = 0,091.$$

Tímto způsobem lze analyzovat a porovnávat jednotlivé texty. V Tab. 9.3 a na Obr. 9.2–9.5 jsou asociativní tematické struktury pěti odborných textů (texty č. 327, 340, 348, 360, 365). Všechny grafy jsou sestaveny tak, aby nejvýše byla lemmata s největším počtem asociací, následují slova s nižším počtem atd. Takto je možné jednoduše opticky vyhodnotit asociační strukturu daného textu. Ve všech případech jsou hodnoty C nízké, což znamená, že tematická lemmata v těchto textech nemají tendenci se asociovat.

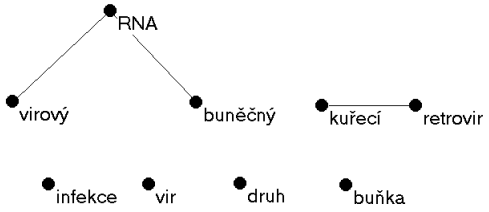
Tabulka 9.3: Míra asociativní tematické struktury u pěti odborných textů (texty č. 327, 340, 348, 360, 365). n vyjadřuje délku textu, k počet hran v grafu, r počet tematických lemmat (podle STK) a C hodnotu asociativní tematické struktury.

text	n	k	r	C
Počítačová alchymie (text č. 340)	1410	3	9	0,083
Proč jsme nevyhynuli... (text č. 360)	976	3	9	0,083
Smrt... (text č. 348).	1892	5	12	0,076
Nepředstavitelně... (text č. 365)	1340	5	11	0,091
Pulzující... (text č. 327)	1147	3	12	0,045

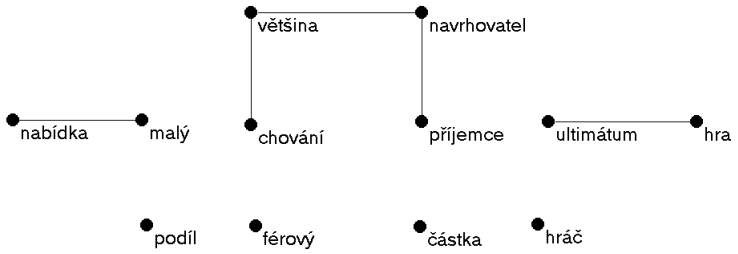
[2] Měření hustoty grafu bylo použito i v kap. 7.3.



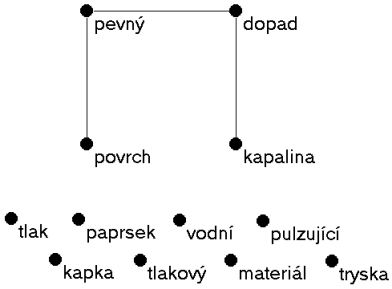
Obrázek 9.2: Grafické vyjádření asociativní tematické struktury textu *Počítačová alchymie* (text č. 340).



Obrázek 9.3: Grafické vyjádření asociativní tematické struktury textu *Proč jsme nevyhynuli na virové infekce?* (text č. 360).



Obrázek 9.4: Grafické vyjádření asociativní tematické struktury textu *Smrt (a částečné vzkříšení) Homo economicus* (text č. 348).



Obrázek 9.5: Grafické vyjádření asociativní tematické struktury textu *Pulzující vodní paprsek – technologie budoucnosti?* (text č. 327).

10

Tematická koncentrace a klasifikace textů

Je dobře známo, že texty vykazují kvantifikovatelné vlastnosti (např. délka slov či vět, distribuce slovních druhů či větných členů atd.), které lze dobře použít pro stylovou, žánrovou nebo v principu vlastně jakoukoliv jinou klasifikaci textů (srov. Těšitelová 1982, 1983a,b,c, 1985; Biber a Conrad 2009). Při textových analýzách tohoto typu se vychází z předpokladu, že pragmatické faktory (mezi něž lze řadit faktory stylotvorné, žánrové, autorské atp.) významným způsobem ovlivňují volbu jazykových prostředků, a to i vzhledem k četnosti použití těchto prostředků. Výhodou kvantitativních přístupů při typologii textů je to, že významně redukují vliv subjektivismu a že zejména umožňují rozdíly mezi texty či skupinami textů přesněji porovnávat. Jako čtenář můžu mít například zkušenost, že v odborných textech se zřejmě vyskytuje více substantiv a více delších vět než v textech publicistických. Jen na základě kvantifikace je však možné určit, zda má čtenářská zkušenost odpovídá realitě a do jaké míry se obě skupiny textů vzhledem ke sledovaným parametrům liší. Prostřednictvím aplikace vhodných statistických metod (Wimmer et al. 2003; Popescu et al. 2009a) lze navíc jít za hranice pouhého popisu. Konkrétně, použití statistických testů dovoluje hlubší interpretaci změřených rozdílů v tom smyslu, že je možné za daných podmínek určit, zda je sledovaný rozdíl významný, či ne. Takto lze eliminovat vliv náhody při interpretaci rozdílů, a dospět tak k hlubšímu poznání vztahů mezi texty (či skupinami textů) vzhledem ke sledovaným vlastnostem. Přepokládám, že metoda analýzy tematické koncentrace textu je vhodnou metodou pro klasifikaci textů, protože:

1) umožňuje rozdíly mezi texty nejen měřit, ale i statisticky testovat (kap. 2.2, 3.1 a 3.3);
2) jsou její indexy TK a STK v intervalu $\langle 200; 6500 \rangle$ slov/tokenů nezávislé na délce textu;
3) umožňuje použití jak slovních forem, tak i lemmat, přičemž výsledky založené na obou typech jednotek signifikantně korelují. To samozřejmě neznamená, že je možné navzájem porovnávat výsledky pocházející z nelemmatizovaných a lemmatizovaných textů. Oba typy jednotek lze ovšem použít s vědomím, že měří stejnou vlastnost textu (byť každá trochu jinak);
4) umožňuje texty porovnávat na základě asociace tematických slov, přičemž asociativnost je definována jako nenáhodný souvýskyt v rámci daného segmentu (např. věta, odstavec atd.), srov. kap. 9;
5) umožňuje texty porovnávat na základě vývoje tematické koncentrace v textu, srov. kap. 6.

Analýza tematické koncentrace je dále vhodným nástrojem pro klasifikaci také proto, že měří vlastnost textu, která se evidentně mění v závislosti na pragmatických faktorech: autor se v závislosti na kontextu rozhoduje, s jakou intenzitou se na určité téma či témata zaměří. Kontextem, který jeho rozhodování (a v důsledku jeho verbální chování) může zásadním způsobem ovlivnit, je například vliv žánru, cíl komunikace, osobní dispozice atd.

V následujících řádcích aplikuji měření tematické koncentrace na různé způsoby klasifikace textů. Nejprve se zaměřím na dva způsoby třídění vzhledem k obecným stylovým charakteristikám, následně na autorský styl a nakonec představím, jak je možné použít měření tematické koncentrace u textů delších, než je interval vhodný pro použití indexů TK a STK (tj. textů s délkou $N > 6500$ slov/tokenů).

10.1 Tematická koncentrace a textové skupiny

Česká lingvistická tradice se při obecné klasifikaci jazykových projevů opírá o tzv. teorii funkční stylů (Havránek 1932, 1942). Tato teorie, která je ovšem spíše určitým způsobem třídění textů než teorií ve smyslu souboru tvrzení, na jejichž základě je možné postulovat predikce, vychází z předpokladu, že výběr a uspořádání výrazových prostředků každého projevu odpovídá funkci tohoto projevu. Vymezení funkčních stylů se pak v jednotlivých případech liší vzhledem ke zvolené míře obecnosti. Na nejobecnější rovině se jazykové projevy zpravidla dělí na čtyři styly: prostěsdělovací (hovorový), odborný, publicistický a umělecký (Chloupek et al. 1990)[1]. Podobně jsou texty členěny i v zahraničních přístupech, za všechny viz použití tzv. registry u Biber et al. (1999).

Tento způsob klasifikace vede k třídění textů do intuitivně velmi dobře vymezitelných skupin. V praxi se samozřejmě setkáváme s texty, které stojí na pomezí těchto základní stylů, případně jsou do nich jen obtížně zařaditelné. Ale jako první „hrubý" vhled do typologie jazykových projevů se zdá být tento přístup užitečný. Je na něm kupříkladu založeno členění textů v Českém národním korpusu, kde se na nejobecnější rovině psané texty dělí na tři textové skupiny 1) beletrie (BEL), 2) odborné texty (ODB) a 3) publicistika (PUB); více viz 'Txtype_group (textová skupina)' (Cvrček a Richterová 2013b).

Při aplikaci měření tematické koncentrace za účelem klasifikace textů vycházím z třídění, které je použito v *Českém národním korpusu* (proto také používám termín 'textová skupina'). Předpokládám přitom, že jednotlivé textové skupiny by se měly signifikantně lišit vzhledem k tematické koncentraci: dá se zřejmě čekat, že autor odborného článku se bude věnovat určitému tématu s mnohem větší intenzitou než autor beletristického textu, podobně lze očekávat větší tematickou zaměřenost u pu-

[1] Mezi lingvisty nepanuje shoda ve vymezení tzv. primárních stylů, například Čechová et al. (2008) pracují s šesti primárními styly. Vymezení počtu těchto stylů je závislé na zvolené míře obecnosti kritérií, na jejichž základě jsou styly definovány.

blicistiky než u beletrie. Pro analýzu byly použity texty uvedené v Tab. 10.1, délka všech textů leží v intervalu $N \in \langle 200; 6500 \rangle$ slov/tokenů.

Tabulka 10.1: Textové skupiny a čísla textů, které byly přiřazeny do jednotlivých skupin.

textová skupina	počet textů	čísla textů
beletrie (BEL)	126	1, 2, 9, 12, 20, 21, 22, 24–32, 39, 44, 58, 114, 119, 125, 126, 128, 130, 131, 135, 140, 142–146, 161, 169, 172, 174, 175, 179-184, 186, 188, 189, 191, 236, 238, 240, 243, 244, 249, 251, 771–841
odborné texty (ODB)	100	273–372
publicistika (PUB)	148	253–272, 414–426, 428, 430–446, 448–462, 974–1066

Je třeba zdůraznit, že analyzovaný vzorek textů v žádném případě neodráží vlastnosti češtiny jako celku, jedná se jen o sondu. Není totiž mým cílem zde prezentovat důkladnou stylometrickou analýzu češtiny, ale představit možnosti, které nabízí metoda měření tematické koncentrace – to je možné i na výše uvedeném vzorku textů. Dále, pokud se uvažuje o případné korespondenci jakéhokoliv analyzovaného vzorku textů s vlastnostmi jazyka jako celku, je nutné se vyrovnat s problematikou tzv. reprezentativnosti. Přestože například mnozí korpusoví lingvisté přepokládají, že pracují s reprezentativními vzorky, srov. Cvrček, Kováříková (2011), dále charakteristiky korpusů SYN2005 (Cvrček a Richterová 2015c), SYN2010 (Cvrček a Richterová 2015d), existují vážné námitky poukazující na to, že z *principiálních* důvodů není možné vytvořit reprezentativní vzorek textů, odrážející vlastnosti daného jazyka *jako celku* (Chromý 2014; Čech 2014b). V každém případě lze níže uvedené výsledky interpretovat jen jako prvotní vhled do případného použití tematické koncentrace pro analýzu textových skupin.

Pro porovnání textových skupin byly použity indexy TK a STK. Nejjednodušší způsob porovnávání skupin textů spočívá ve srovnání průměrných hodnot daných indexů a následném vyhodnocení významnosti rozdílu prostřednictvím u-testu, viz vzorec (2.12), kap. 2.2. Z Tab. 10.2 a 10.3 je zřejmé, že nejvyšší hodnoty TK a STK vykazují texty odborné (což jistě není překvapivé), u nichž se projevuje signifikantní rozdíl v porovnání s oběma dalšími textovými skupinami. Podobně jsou v souladu s očekáváním také vyšší hodnoty TK a STK u textů publicistických než beletristických. Zde je již rozdíl menší, dokonce v případě TK je nesignifikantní. Údaje z Tab. 10.2 a 10.3 umožňují přehledně graficky vyjádřit nejen „vzdálenost" mezi texty s ohledem na TK a STK, ale také s ohledem na velikost rozdílu vyjádřeného statistickým u-testem. Konkrétně, pokud u každé textové skupiny vydělíme součet testových hodnot u druhou odmocninou počtu porovnání skupin navzájem k, získáme váženou hodnotu u_v, kte-

rá vyjadřuje velikost rozdílu mezi danou textovou skupinou a ostatními skupinami:

$$u_\nu = \frac{\sum |u_i|}{\sqrt{k}}. \tag{10.1}$$

Tabulka 10.2: Průměrné hodnoty TK jednotlivých textových skupin a hodnoty testovacího kritéria u. Pro zvolenou hladinu významnosti $\alpha = 0,05$ je rozdíl signifikantní, pokud $|u| > 1,96$.

	BEL	ODB	PUB
TK (průměr)	0,0382	0,1183	0,0535
BEL	x		
ODB	**7,03**	x	
PUB	1,37	**5,92**	x

Tabulka 10.3: Průměrné hodnoty STK jednotlivých textových skupin a hodnoty testovacího kritéria u. Pro zvolenou hladinu významnosti $\alpha = 0,05$ je rozdíl signifikantní, pokud $|u| > 1,96$.

	BEL	ODB	PUB
STK (průměr)	0,0434	0,1082	0,0640
BEL	x		
ODB	**8,92**	x	
PUB	**2,65**	**5,75**	x

V Tab. 10.2 a 10.3 jsou pro každou textovou skupinu dvě porovnání, tj. $k = 2$, proto například u TK pro beletristické texty dostáváme vážený rozdíl

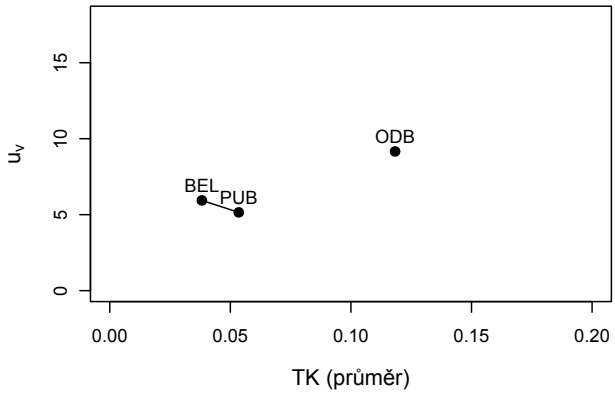
$$u_{\nu\ \mathrm{TK(BEL)}} = \frac{7,03 + 1,37}{\sqrt{2}} = 5,94.$$

Hodnota váženého rozdílu tak vyjadřuje celkovou míru podobnosti/rozdílnosti jednotlivých skupin textů vzhledem k ostatním skupinám, viz Tab. 10.4.

V případě TK se tak jasně ukazuje specifické postavení odborných textů. Rozdílné výsledky vztahu mezi beletrií a publicistikou u TK a STK jsou mimo jiné výsledkem toho, že u TK se objevuje mnohem více nulových hodnot než v případě STK, což do jisté míry nivelizuje rozdíly. Pro detailnější pohled na vztah mezi skupinami textů je tedy dobré volit skupiny s menším počtem nulových hodnot u daných indexů. V případě zde analyzovaného vzorku se STK jeví jako vhodnější nástroj pro klasifikaci textových skupin. Mimo jiné, nejnižší hodnota u_ν u publicistiky je dobře interpretovatelná

Tabulka 10.4: Vážené rozdíly u jednotlivých textových skupin.

TK	u_v	STK	u_v
PUB	5,15	PUB	5,94
BEL	5,94	BEL	8,18
ODB	9,16	ODB	10,37

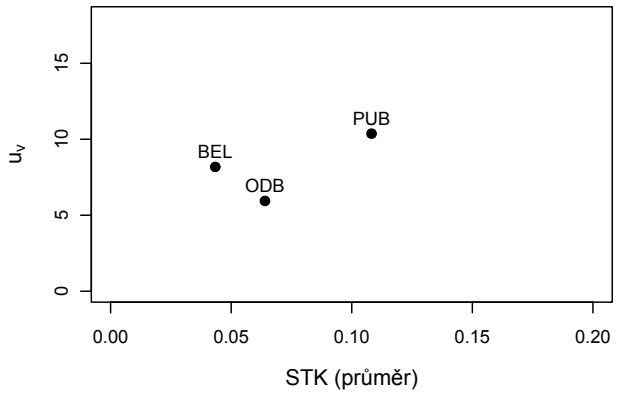


Obrázek 10.1: Grafické vyjádření průměrných hodnot TK a vážených rozdílů u_v u třech textových skupin. Čarou jsou spojeny skupiny, u nichž je nesignifikantní rozdíl TK (hladina významnosti $\alpha = 0,05$)

v souladu s poznatky tradiční stylistiky, která tvrdí, že publicistický styl je méně vyhraněný než například styl odborný nebo administrativní (Jelínek 2002a).

10.2 Tematická koncentrace a textové typy

Klasifikace textů uvedená v přechozí kapitole je jednou z klasifikací nejobecnějších. Texty zařazené do tak široce definovaných skupin však většinou představují nehomogenní vzorek. Proto je mnohdy užitečnější používat klasifikaci detailnější. Samozřejmě existuje celá řada způsobů, jak texty dále třídit. Zde budu opět vycházet z klasifikace, již používá *Český národní korpus*, konkrétně půjde o dělení textů na tzv. textové typy, které v *Českém národním korpusu* představuje základní třídění textů z hlediska registru (Cvrček a Richterová 2013c). Textový typ je z pohledu obecnosti na nižší úrovni než textová skupina a na vyšší s ohledem na žánry. V *Českém národním korpusu* se



Obrázek 10.2: Grafické vyjádření průměrných hodnot STK a vážených rozdílů u_v u třech textových skupin. Mezi sledovanými skupinami jsou signifikantní rozdíly STK (hladina významnosti $\alpha = 0,05$).

používá 14 textových typů, které jsou přiřazeny ke třem výše uvedeným textovým skupinám:

- beletrie
  - básně
  - písně
  - romány
  - soubory povídek, jednotlivá povídka
  - literatura faktu
  - dramatické texty, scénáře
  - jiné imaginativní texty
- odborná literatura
  - vědeckonaučná literatura
  - populárněnaučná literatura
  - učebnice
  - abecedně, systematicky a jinak uspořádaná díla
  - administrativa
- publicistika
  - publicistika (noviny a neodborné časopisy)
  - rozmanité

Na základě tohoto třídění jsem pro zde prezentovanou analýzu pracoval s šesti textovými typy (Tab. 10.5). Dále jsem zavedl textový typ řečnický, abych mohl do vzorku zařadit i prezidentské novoroční projevy. Pro analýzu byly použity texty, jejichž délka leží v intervalu $N \in \langle 200; 6500 \rangle$ slov/tokenů. V případě románů proto nejsou analyzovány romány jako celek (pro tak dlouhé texty jsou metody TK a STK nevhodné, viz kap. 5), nýbrž jsou zpracovány jednotlivé kapitoly zvlášť. Analogicky k postupu uve-

Tabulka 10.5: Textové skupiny a čísla textů, které byly přiřazeny do jednotlivých skupin.

textový typ	počet textů	čísla textů
poezie (VER)	55	1, 2, 9, 12, 20–22, 24–32, 39, 44, 58, 114, 119, 125, 126, 128, 130, 131, 135, 140, 142–146, 161, 169, 172, 174, 175, 179-184, 186, 188, 189, 191, 236, 238, 240, 243, 244, 249, 251
povídky (COL)	71	771–841
romány – jednotlivé kapitoly (NOV)	210	559–715, 717–757, 759–770
publicistika (PUB)	148	253–272, 414–426, 428, 430–446, 448–462, 974–1066
řečnické projevy (REC)	64	463–526
populárněnaučné (POP)	50	323–372
vědecké (SCI)	50	273–322

deném v předchozí kapitole byly zpracovány jednotlivé textové typy, tj. pro každý typ byla vypočítána průměrná hodnota TK a STK, rozdíly mezi jednotlivými typy byly testovány podle vzorce (2.12) a na základě vzorce (10.1) byly stanoveny vážené rozdíly. Výsledky jsou uvedeny v Tab. 10.6–7 a na Obr. 10.3–4.

Výsledky analýzy zejména potvrdily specifické postavení odborných textů, které mají v případě obou typů nejvyšší hodnoty TK a STK, přičemž rozdíly mezi oběma typy SCI a POP jsou v obou případech nesignifikantní (na hladině významnosti $\alpha = 0,05$). Očekávaná je jistě i blízkost povídek (COL) a kapitol románů (NOV), stejně jako tendence publicistiky (PUB) se vymezit jako samostatný typ, u nějž se projevují signifikantní rozdíly vzhledem k ostatním typům (tato tendence se projevuje především u STK). Na druhou stranu se jako překvapivé jeví postavení poezie (VER), které má nejnižší hodnotu vážených rozdílů u_v a v obou grafech nejvyšší počet hran, což

Tabulka 10.6: Průměrné hodnoty TK jednotlivých textových typů a hodnoty testovacího kritéria u. Pro zvolenou hladinu významnosti $\alpha = 0,05$ je rozdíl signifikantní, pokud $|u| > 1,96$.

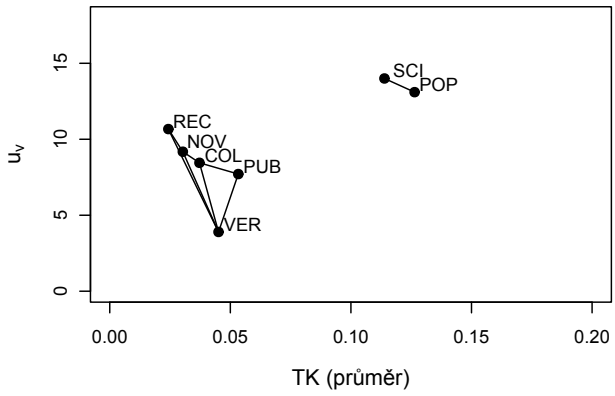
	VER	COL	NOV	PUB	REC	POP	SCI
TK (průměr)	0,0451	0,0372	0,0303	0,0533	0,0243	0,1264	0,1139
VER	x						
COL	0,34	x					
NOV	0,64	1,31	x				
PUB	0,34	1,87	**2,81**	x			
REC	0,90	**2,57**	1,43	**3,62**	x		
POP	**3,15**	**7,00**	**7,71**	**5,14**	**8,26**	x	
SCI	**2,80**	**7,61**	**8,61**	**5,10**	**9,36**	0,82	x

Tabulka 10.7: Průměrné hodnoty STK jednotlivých textových typů a hodnoty testovacího kritéria u. Pro zvolenou hladinu významnosti $\alpha = 0,05$ je rozdíl signifikantní, pokud $|u| > 1,96$.

	VER	COL	NOV	PUB	REC	POP	SCI
STK (průměr)	0,0510	0,0422	0,0367	0,0640	0,0360	0,1195	0,1003
VER	x						
COL	0,62	x					
NOV	1,01	1,34	x				
PUB	0,84	**3,30**	**4,25**	x			
REC	1,03	1,45	0,07	**4,35**	x		
POP	**4,30**	**9,28**	**10,10**	**5,79**	**10,21**	x	
SCI	**3,25**	**8,66**	**9,72**	**4,42**	**9,87**	1,96	x

Tabulka 10.8: Vážené rozdíly u jednotlivých typů textu. Texty jsou seřazeny podle rostoucí hodnoty u_v.

TK	u_v	STK	u_v
VER	3,90	VER	4,51
PUB	7,71	PUB	9,37
COL	8,45	COL	10,06
NOV	9,18	NOV	10,81
REC	10,67	REC	11,01
POP	13,10	SCI	15,46
SCI	14,00	POP	17,00



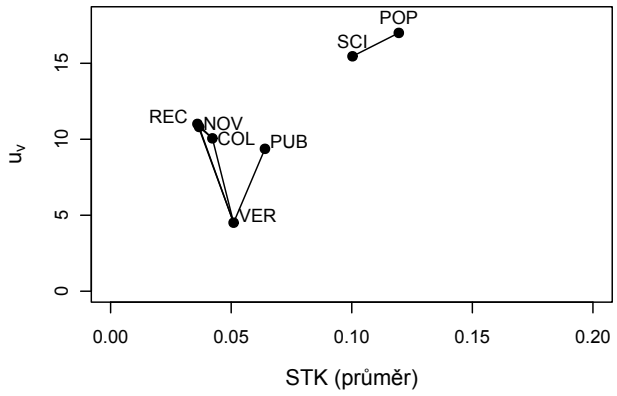
Obrázek 10.3: Grafické vyjádření průměrných hodnot TK a vážených rozdílů u_v u analyzovaných typů textů. Čarou jsou spojeny typy textů, u nichž je nesignifikantní rozdíl TK (hladina významnosti $\alpha = 0,05$).

vyjadřuje nejvyšší míru podobnosti s jinými typy – kromě SCI a POP má všude nesignifikantní rozdíly. Zdá se tedy, že navzdory velké variabilitě výrazových prostředků nehraje míra zaměření se na dané téma či témata v tomto typu textů významnější „diferenciační“ roli.

10.3 Tematická koncentrace a autorský styl

Každý jazykový projev nese specifické rysy, které odrážejí vliv jedinečnosti autora. Problematika analýzy autorství je široce rozpracována a výsledky těchto analýz jsou velmi přesvědčivé (srov. např. Mikros a Perifanos 2011, 2013). Jelikož je určování autorství založeno na sledování velkého množství charakteristik textu, není možné očekávat, že by tematická koncentrace mohla být samostatně pro určování autorství použita. To ostatně není ani její ambicí. Na druhou stranu ale nic nebrání tomu, aby byla tato metoda aplikována při sledování rozdílů mezi jednotlivými autory.

Pokud chceme sledovat rozdíly v autorském stylu, je nutné porovnávat texty, které sdílejí co nejvíce společných vlastností, tj. texty jednoho žánru: například zvlášť soukromé dopisy, odborné články, lyrickou poezii, deníkové záznamy či politické projevy. Je totiž třeba v co největší míře eliminovat vliv jiných faktorů, než je právě autorství. Metoda měření tematické koncentrace už byla v tomto směru úspěšně použita, a to při analýze publicistických textů K. Čapka, J. Durycha a L. Jehličky a novoročních prezidentských projevů československých a českých prezidentů (Davidová et al.



Obrázek 10.4: Grafické vyjádření průměrných hodnot STK a vážených rozdílů u_v u analyzovaných typů textu. Čarou jsou spojeny typy textů, u nichž je nesignifikantní rozdíl STK (hladina významnosti $\alpha = 0,05$).

2013; David et al. 2013; Čech 2014a). Protože v té době ještě nebyl zpracován koncept STK (ten až v Čech et al. 2015), byl ve výše uvedených studiích použit pouze index TK. Ten má však, jak je uvedeno v kap. 3, své limity – zejména jde o to, že mnohé texty mají nulovou hodnotu TK, což je sice z teoretického hlediska v pořádku (jedná se vzhledem k tematické koncentraci o neutrální texty), pro porovnávání autorů je to však spíše hendikep. To se ukázalo například při analýze prezidentských projevů, kde sedm ze 13 projevů V. Havla mělo $TK = 0$. Proto v následujících řádcích bude použit pro porovnání autorského stylu prezidentských projevů (texty č. 463–526) index STK.

Je třeba zdůraznit, že novoroční prezidentské projevy se pro porovnávání autorských stylů výborně hodí, protože splňují požadavek co největší eliminace jiných faktorů ovlivňujících vlastnosti jednotlivých textů, jedná se totiž o velmi specifický žánr. Na druhou stranu se může jevit jako problematická samotná otázka autorství projevů – v mnohých případech si autoři zřejmě nepsali projevy sami (srov. Esterková 2013). Autorství je ale možné definovat také jako projev politické odpovědnosti za text. A právě v tomto smyslu je k němu přistupováno i zde.

U každého projevu byla nejdříve změřena hodnota STK, poté byly u jednotlivých prezidentů výsledky zprůměrovány a následně tyto průměry porovnány prostřednictvím u-testu, viz vzorec (2.12).

Výsledky prezentované v Tab. 10.9–10 a Obr. 10.5 v prvé řadě poukazují na výraznou specifičnost projevů V. Havla. Mezi průměrnou hodnotou STK jeho projevů

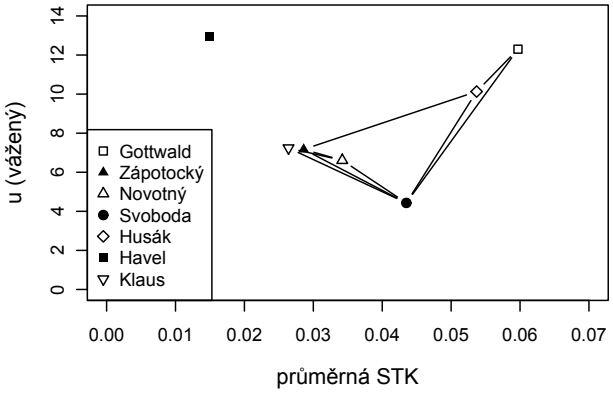
Tabulka 10.9: Průměrné hodnoty STK prezidentských novoročních projevů a hodnoty testovacího kritéria u. Pro zvolenou hladinu významnosti $\alpha = 0,05$ je rozdíl signifikantní, pokud $|u| > 1,96$.

	Gottwald	Zápotocký	Novotný	Svoboda	Husák	Havel	Klaus
STK (průměr)	0,0597	0,0286	0,0342	0,0435	0,0537	0,0150	0,0264
Gottwald	x						
Zápotocký	**6,18**	x					
Novotný	**4,79**	1,08	x				
Svoboda	1,78	1,67	1,03	x			
Husák	1,15	5,00	**3,67**	1,19	x		
Havel	**9,93**	**3,20**	**4,16**	**3,30**	**8,63**	x	
Klaus	**6,30**	0,44	1,45	1,88	**5,17**	**2,50**	x

a průměrnými hodnotami projevů všech ostatních prezidentů jsou vždy signifikantní rozdíly (na hladině významnosti $\alpha = 0,05$). Tato specifičnost se projevuje nízkou STK, což znamená, že Havel má výraznou tendenci se silně nezaměřovat na probíraná témata, je spíše tematicky „roztříštěný“. V opozici k jeho stylu stojí projevy K. Got-

Tabulka 10.10: Vážené rozdíly u_v u jednotlivých prezidentů. Prezidenti jsou seřazeni podle rostoucí hodnoty u_v.

STK	u_v
Svoboda	4,43
Novotný	6,61
Zápotocký	7,17
Klaus	7,24
Husák	10,13
Gottwald	12,30
Havel	12,95

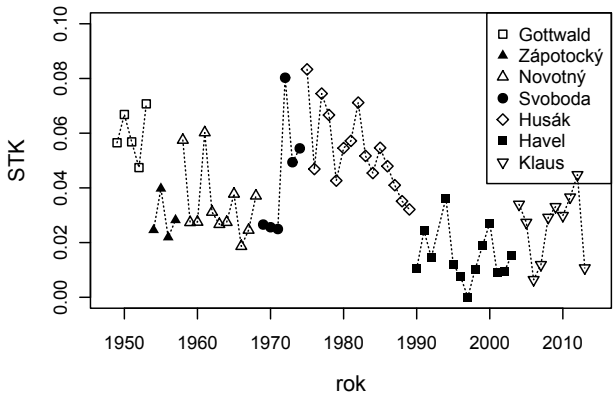


Obrázek 10.5: Grafické vyjádření průměrných hodnot STK a vážených rozdílů u_v u jednotlivých prezidentů. Čarou jsou spojeni prezidenti, u nichž je nesignifikantní rozdíl STK (hladina významnosti $\alpha = 0,05$).

twalda a G. Husáka, jež vykazují nejvyšší hodnoty STK. Ve studii Čech (2014a) byl postulován vztah mezi tematickou koncentrací a ideologií (totalitní vs. demokratická), přičemž tento vztah lze použít i pro interpretaci rozdílů autorského stylu Havla, jako jednoho z nejvýraznějších představitelů demokratických prezidentů na jedné straně, a Gottwalda a Husáka, jako představitelů dvou výrazných období komunistické totality – 50. let 20. století a tzv. normalizace 70. let – na straně druhé (k podrobnostem viz výše zmíněné studie). Jinými slovy, předpokládá se (a výsledky to prozatím potvr-

zují), že kromě samotného autorského stylu má na nízké hodnoty STK u Havlových projevů výrazný vliv i politický kontext (konkrétně ideologie), stejně jako na vysoké hodnoty STK u Gottwalda a Husáka. Velmi malé (a nesignifikantní) rozdíly STK mezi projevy Klause, Novotného a Zápotockého a malé rozdíly testového kritéria u mezi těmito projevy navzájem poukazují na malý vliv autorství. Jestli existuje něco jako „jádro" žánru (pro něj by platilo, že žánr je silnějším faktorem než individuální autorské rozdíly), tak je v případě prezidentských novoročních projevů reprezentováno právě těmito třemi prezidenty.

Specifičnost autorství je možné také sledovat vzhledem k tzv. stabilitě individuálního stylu. Pokud by měl autor maximálně stabilní styl, byly by hodnoty STK ve všech jeho textech stejné. V praxi však samozřejmě dochází k tomu, že se v textech tyto hodnoty liší i u jednotlivých autorů, srov. Obr. 10.6 (v případě maximální stability by u každého autora byla spojnicí mezi všemi hodnotami STK vodorovná úsečka).



Obrázek 10.6: Hodnoty STK v jednotlivých novoročních projevech československých a českých prezidentů.

Míru (ne)stability stylu lze definovat na základě měření rozptylu (variance): čím větší je variabilita naměřených hodnot od střední hodnoty (např. průměru či mediánu), tím je autorský styl nestabilnější. V tomto ohledu vykazují největší stabilitu projevy A. Zápotockého, nejmenší pak L. Svobody, viz Tab. 10.11.

V případě A. Zápotockého a L. Gottwalda však máme k dispozici jen čtyři, respektive pět projevů, což je dáno jejich relativně krátkým funkčním obdobím, takže malý rozptyl může být projevem docela malého počtu měření – u těchto prezidentů jsou

Tabulka 10.11: Rozptyl STK v projevech jednotlivých prezidentů. Čím větší je jeho hodnota, tím je autorský styl nestabilnější. Prezidenti jsou seřazeni ve vzestupném pořadí.

	$Var(STK)$
Zápotocký	0,000062
Gottwald	0,000085
Havel	0,000092
Klaus	0,000158
Novotný	0,000178
Husák	0,000217
Svoboda	0,000491

výsledky spíše jen ilustrativní. Podobně opatrně je třeba přistupovat i při interpretaci projevů L. Svobody (šest projevů).

Prostřednictvím statistických testů je dále možné testovat rozdíly jednotlivých rozptylů v daných výběrech, a sledovat tak, zda se jednotliví prezidenti mezi sebou liší v míře stability, jak je definována výše. Pro STK je zvolen Brownův-Forsythův test, který nepředpokládá normální rozdělení dat (Brown a Forsythe 1974). Podobně jako u průměrných hodnot STK, i zde jsou porovnávány jednotlivé dvojice prezidentů.

Výsledky v Tab. 10.12 ukazují na vysokou míru shody mezi stabilitou projevů všech prezidentů – pouze v případě projevů L. Svobody a V. Havla se rozptyly signifikantně liší (na hladině významnosti $\alpha = 0,05$). Většina prezidentů si tedy udržuje v tomto specifickém žánru nesignifikantní rozdíly variancí STK i navzdory tomu, že se mezi nimi navzájem projevují signifikantní rozdíly průměrných hodnot tohoto indexu (srov. Tab. 10.9). U L. Svobody je patrné, že po jeho prvních třech projevech nastává výrazná změna, a je spíše otázkou pro historiky, čím mohla být způsobena (např. politickou situací, změnou skutečného autora textu vzhledem k jeho nemoci atp.). Je však třeba upozornit, že signifikantnost rozdílu rozptylů je závislá i na rozsahu souborů – například rozdíl rozptylů mezi Zápotockým a Svobodou je větší než mezi Havlem a Svobodou (Tab. 10.11), ale mezi Zápotockým a Svobodou je nesignifikantní, kdežto mezi Havlem a Svobodou signifikantní (Tab. 10.12). Tento výsledek je mimo jiné způsoben tím, že u Zápotockého máme málo projevů, proto test „toleruje“ i tento větší rozdíl, tj. pokládá ho za výsledek vlivu náhody a vyhodnotí ho jako nesignifikantní.

10.4 Analýza tematické koncentrace u „dlouhých“ textů[2]

Na první pohled se metoda měření tematické koncentrace pro analýzu delších textů, jako je třeba román, nehodí. Především se dá jen těžko očekávat, že u textů čítajících

[2] Jako „dlouhý“ zde označujeme text o délce $N > 6500$ slov, což je empiricky odvozená horní hranice intervalu, na níž je TK a STK na délce textu nezávislá (srov. kap. 5).

Tabulka 10.12: p-hodnoty Brownova-Forsythova testu, jehož prostřednictvím byly testovány rozdíly rozptylu STK mezi projevy jednotlivých prezidentů. Na hladině významnosti $\alpha = 0,05$ byl zjištěn jediný signifikantní rozdíl: mezi projevy Havla a Svobody.

	Gottwald	Zápotocký	Novotný	Svoboda	Husák	Havel	Klaus
Gottwald	x						
Zápotocký	0,7412	x					
Novotný	0,6935	0,5679	x				
Svoboda	0,0986	0,0938	0,1634	x			
Husák	0,3038	0,2275	0,5627	0,1975	x		
Havel	0,9617	0,6962	0,5949	**0,0230**	0,1611	x	
Klaus	0,5673	0,4192	0,9521	0,1193	0,1611	0,5000	x

desítky nebo stovky tisíc slov/tokenů je autor schopen „manipulovat“ s touto vlastností textu. Analýza vztahu jednotlivých indexů tematické koncentrace a délky textu (kap. 5) tento předpoklad potvrdila: 1) texty obsahující desítky tisíc slov/tokenů a více vykazují velmi malý interval hodnot všech indexů v porovnání s texty kratšími; 2) v případě sledování vztahu mezi kumulativní délkou textu a sledovanými

indexy se ukazuje, že u dlouhých textů se hodnoty těchto indexů zhruba od délky 10000 slov/tokenů ustalují v rámci velmi malého intervalu (srov. Obr. 5.21–22). To znamená, že u dlouhých textů je tematická koncentrace (měřená jakýmkoliv výše popsaným indexem) jen důsledkem frekvenční struktury textu, na který již autorství, žánr, volba tématu atd. nemá vliv. Je to dobře pochopitelné, pokud si uvědomíme základní princip, na němž měření tematické koncentrace stojí, tj. na rozdělení rankové frekvenční distribuce na oblast synsémantik a autosémantik prostřednictvím h-bodu (kap. 2). S narůstající délkou textu totiž roste hodnota h-bodu a v případě dlouhých textů už tento bod nemůže vymezovat hranici mezi synsémantiky a autosémantiky, protože synsémantik je v jazyce omezené množství. To je třeba si uvědomit při aplikaci všech metod, které využívají h-bod pro vymezení hranice mezi autosémantiky a synsémantiky (Popescu et al 2009a,b).

Pokud ovšem přistoupíme k dlouhému textu jako k souboru tematických jednotek – například kapitol – je možné analogicky k analýze textových skupin a typů (kap. 10.1–2) vyhodnotit i tematickou koncentraci takového textu, tj. vypočítat průměrnou hodnotu tematických jednotek a stanovit takto tematickou koncentraci celého románu. Velkou výhodou tohoto postupu je, že na rozdíl od textových skupin a typů zde pracujeme s homogenním textovým materiálem: tematické jednotky, jako jsou kapitoly, jsou dílem jednoho autora, patří zpravidla do jednoho žánru, tematicky jsou si přinejmenším podobné či na sebe aspoň navazují, často vznikají v relativně krátkém časovém období atp.

Pro ilustraci bude porovnáno šest prozaických textů K. Čapka, konkrétně *Krakatit*, *Hordubal*, *Povětroň*, *Obyčejný život*, *Válka s mloky* a *První parta*. Každý text byl rozdělen na kapitoly a v každé kapitole byla změřena hodnota TK a STK. Kapitoly, jejichž délka se nevyskytovala v intervalu $N \in \langle 200; 6500 \rangle$ slov/tokenů, nebyly započítány – v sledovaném vzorku šlo pouze o dvě kapitoly: kap. 14 v románu *Válka s mloky* ($N = 16405$) a kapitola 39 v novele *Povětroň* 39 ($N = 69$). Následně byla vypočítána průměrná hodnota indexu pro daný román a tyto průměrné hodnoty porovnány prostřednictvím u-testu, viz vzorec 2.13. Zpracovány byly texty č. 559–715, 717–757, 759–770.

Výsledky, prezentované v Tab. 10.13–15 a na Obr. 10.7–8, poukazují na překvapivě velké rozdíly TK i STK mezi jednotlivými Čapkovými prozaickými texty. Vzhledem k tomu, že všechny analyzované texty je možné řadit k jednomu funkčnímu stylu – stylu prozaickému (Jelínek 2002b) – a že se navíc jedná o texty jednoho autora, dala by se očekávat mnohem větší míra podobnosti. Prostřednictvím měření hustoty grafu, srov. vzorec (7.5), je možné míru podobnosti sledovaného vzorku kvantifikovat:

$$\rho_{TK} = \frac{5}{15} = 0,33,$$

$$\rho_{STK} = \frac{3}{15} = 0,2.$$

Zde naměřené hodnoty lze předběžně interpretovat jako důsledek „volnosti“ prozaického stylu, který dává autorovi možnost realizovat jeho estetický záměr prostřed-

Tabulka 10.13: Průměrné hodnoty TK prozaických textů K. Čapka a hodnoty testovacího kritéria u. Pro zvolenou hladinu významnosti $\alpha = 0,05$ je rozdíl signifikantní, pokud $|u| > 1,96$.

	Krakatit	*Hordubal*	*Povětroň*	*Obyčejný život*	*Válka s mloky*	*První parta*
TK (průměr)	0,0369	0,0454	0,0093	0,0042	0,0582	0,0436
počet slov / tokenů	77198	32868	40722	40798	65441	49492
Krakatit	x					
Hordubal	1,01	x				
Povětroň	**5,85**	**4,78**	x			
Obyčejný život	**7,41**	**5,57**	**2,16**	x		
Válka s mloky	**2,20**	1,06	**5,72**	**6,47**	x	
První parta	1,04	0,21	**6,54**	**7,90**	1,43	x

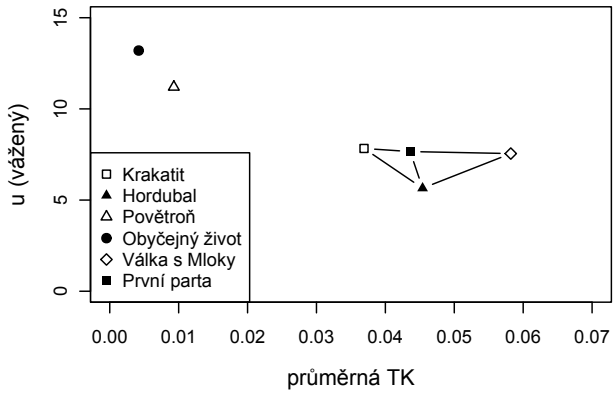
Tabulka 10.14: Průměrné hodnoty STK prozaických textů K. Čapka a hodnoty testovacího kritéria u. Pro zvolenou hladinu významnosti $\alpha = 0,05$ je rozdíl signifikantní, pokud $|u| > 1,96$.

	Krakatit	*Hordubal*	*Povětroň*	*Obyčejný život*	*Válka s mloky*	*První parta*
STK (průměr)	0,0426	0,0602	0,0220	0,0127	0,0632	0,0453
počet slov/tokenů	77198	32868	40722	40798	65441	49492
Krakatit	x					
Hordubal	**2,35**	x				
Povětroň	**4,95**	**5,36**	x			
Obyčejný život	**8,26**	**6,92**	**2,86**	x		
Válka s mloky	**2,46**	0,17	**5,35**	**6,82**	x	
První parta	0,53	1,95	**5,20**	**8,20**	**2,08**	x

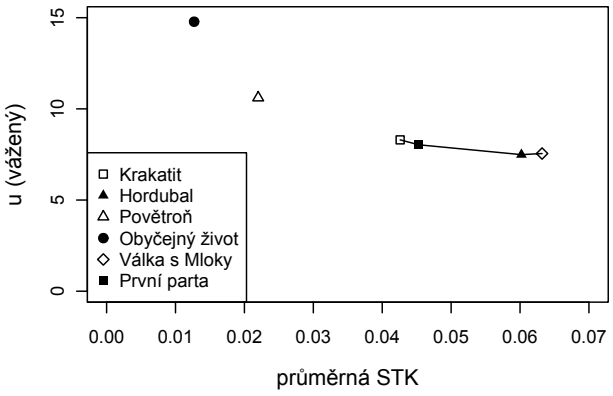
nictvím široké palety jazykových prostředků (Jelínek 2002b), a to zajisté i vzhledem k jejich frekvenčním charakteristikám, jež mají rozhodující vliv na hodnoty tematické koncentrace. O tom, zda se jedná o rys typický pro Čapka, nebo jde o fenomén univerzálnější povahy, mohou rozhodnout jen analýzy dalších autorů.

Tabulka 10.15: Vážené rozdíly u_ν u jednotlivých prozaických textů K. Čapka.

TK	u_ν	STK	u_ν
Hordubal	5,65	*Hordubal*	7,49
Válka s mloky	7,55	*Válka s mloky*	7,55
První parta	7,66	*První parta*	8,03
Krakatit	7,83	*Krakatit*	8,30
Povětroň	11,20	*Povětroň*	10,61
Obyčejný život	13,20	*Obyčejný život*	14,78



Obrázek 10.7: Grafické vyjádření průměrných hodnot TK a vážených rozdílů u_ν u prozaických textů K. Čapka. Čarou jsou spojeny texty, u nichž je nesignifikantní rozdíl TK (hladina významnosti $\alpha = 0,05$).



Obrázek 10.8: Grafické vyjádření průměrných hodnot STK a vážených rozdílů u_v u prozaických textů K. Čapka. Čarou jsou spojeny texty, u nichž je nesignifikantní rozdíl STK (hladina významnosti $\alpha = 0,05$).

11

QUITA – software (nejen) pro analýzu tematické koncentrace

Pro analýzu tematické koncentrace textů prostřednictvím indexů TK, STK a PTK je mimo jiné možné použít volně dostupný software *QUITA – Quantitative Index Text Analyzer* (Matlach et al. 2014). Pomocí tohoto softwaru lze nejen velmi jednoduše provádět samotné výpočty indexů TK, STK a PTK, ale také testovat rozdíly daných indexů jak mezi jednotlivými texty, tak skupinami textů. Pro testování rozdílů jsou v softwaru použity statistické testy uvedené v této knize. Stejně tak ve shodě se zde prezentovaným přístupem jsou za tematická slova považována substantiva, adjektiva a verba (kromě sloves *být*, *mít*, *moci*, *muset*, *smět*), srov. kap. 2. Software také umožňuje tvorbu grafů, tabulek, export dat atd.

Detailní informace k softwaru je možné najít u Kubáta et al. (2014), v češtině je volně dostupná diplomová práce Matlacha (2014).

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.

Seznam obrázků

Seznam tabulek

A

Příloha

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
1	Amín	Biebl, K.	248	0	0,030303	0	0	0,015152	0	Biebl, K.: S lodí jež dováží čaj a kávu. In Cesta na Jávu. Praha 2001.
2	Cesta k lesu	Biebl, K.	219	0	0	0	0	0	0	ibid.
3	Javanky	Biebl, K.	37	0	0	0	0	0	0	ibid.
4	Na hoře Merbabu	Biebl, K.	102	0,125	0,175	0,3	0	0,15	0	ibid.
5	Na oceánu	Biebl, K.	163	0	0	0	0	0,005208	0	ibid.
6	Na procházce	Biebl, K.	55	0	0	0	0	0	0	ibid.
7	Návštěva	Biebl, K.	82	0	0,083333	0	0	0	0	ibid.
8	Pacienti	Biebl, K.	191	0	0	0	0	0	0	ibid.
9	Protinožci	Biebl, K.	227	0	0	0	0	0	0	ibid.
10	Pustil jsem, pustil svou bláznivou volnoběžku	Biebl, K.	109	0,35	0,233333	0,222222	0,171429	0,171429	0,24	ibid.
11	S lodí jež dováží čaj a kávu	Biebl, K.	84	0	0	0	0,130612	0,183673	0,266667	ibid.
12	S očima k nebi	Biebl, K.	357	0,082051	0,076923	0,150943	0,128106	0,089376	0,142857	ibid.
13	Soudní referát	Biebl, K.	111	0,125	0,175	0,3	0,342857	0,327381	0,333333	ibid.
14	Toké	Biebl, K.	138	0,14652	0,17094	0,357143	0,115385	0,152137	0,30303	ibid.
15	Tulácká	Biebl, K.	94	0	0	0	0	0	0	ibid.
16	V Africe	Biebl, K.	97	0	0	0	0	0	0	ibid.
17	V noci	Biebl, K.	69	0	0	0	0	0,028846	0	ibid.
18	Yorck	Biebl, K.	113	0	0,066667	0	0,130612	0,163265	0,266667	ibid.
19	Začarovaná studánka	Biebl, K.	72	0	0	0	0	0	0	ibid.
20	Žně	Biebl, K.	284	0	0,002222	0	0	0,011852	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
21	Dceřina kletba	Erben, K. J.	222	0,111888	0,151049	0,27907	0,041026	0,114219	0,145455	Erben, K. J.: Kytice z básní, 1874. Česká elektronická knihovna – Poezie 19. a počátku 20. století, ÚČL AV ČR, 2005.
22	Holoubek	Erben, K. J.	316	0	0,041667	0	0	0,037879	0	ibid.
23	Kytice	Erben, K. J.	115	0	0	0	0	0	0	ibid.
24	Lilie	Erben, K. J.	494	0,016317	0,068376	0,118644	0,057516	0,065155	0,112245	ibid.
25	Poklad	Erben, K. J.	2347	0,022607	0,029467	0,091093	0,031589	0,047339	0,108955	ibid.
26	Polednice	Erben, K. J.	202	0,277778	0,282738	0,454545	0,6	0,407407	0,428571	ibid.
27	Štědrý den	Erben, K. J.	679	0	0,023923	0	0	0,036765	0	ibid.
28	Svatební košile	Erben, K. J.	1465	0	0,003158	0	0	0,01094	0	ibid.
29	Vodník	Erben, K. J.	975	0	0,017857	0	0	0,042718	0	ibid.
30	Vrba	Erben, K. J.	509	0,3	0,164502	0,368421	0,098765	0,086081	0,159091	ibid.
31	Záhořovo lože	Erben, K. J.	2724	0,002963	0,012061	0,03212	0	0,019114	0	ibid.
32	Zlatý kolovrat	Erben, K. J.	1444	0,004651	0,02537	0,05	0,013249	0,051784	0,088235	ibid.
33	A nastává mi, tuším, vážná jízda	Gellner, F.	144	0	0	0	0	0	0	Gellner, F.: Radosti života, 1903. Česká elektronická knihovna – Poezie 19. a počátku 20. století, ÚČL AV ČR, 2005.
34	Bláznění vjelo do párů	Gellner, F.	98	0	0	0	0	0	0	ibid.
35	Buď matka boží pomocná	Gellner, F.	106	0	0	0	0	0	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
36	Což, páni spisovatelé	Gellner, F.	119	0	0	0	0	0	0	ibid.
37	Drobky pod stůl hází nám osud	Gellner, F.	123	0	0,1	0	0	0,053571	0	ibid.
38	Já jsem k tobě nepřišel	Gellner, F.	179	0	0,160714	0	0,020833	0,09537	0,138889	ibid.
39	Když skutečnost mou nezlomila bytost	Gellner, F.	256	0	0	0	0	0,007645	0	ibid.
40	Konečně je to možná věc	Gellner, F.	54	0	0	0	0	0	0	ibid.
41	Na dnešek měl jsem pěkný sen	Gellner, F.	161	0	0	0	0	0	0	ibid.
42	Napsala mi psaní	Gellner, F.	140	0	0,006696	0	0	0,002778	0	ibid.
43	Nečekám nic od reforem	Gellner, F.	57	0,525	0,331731	0,5	0,666667	0,433333	0,6	ibid.
44	Nic vyčítat ti nechci, moje milá	Gellner, F.	200	0	0,013393	0	0	0,008392	0	ibid.
45	Noc byla	Gellner, F.	183	0	0	0	0	0,013661	0	ibid.
46	Ožeň se, bratře	Gellner, F.	89	0,272727	0,253247	0,521739	0,272727	0,206061	0,363636	ibid.
47	Pomalu v revolver se ztrácí víra	Gellner, F.	97	0	0,071429	0	0	0,035714	0	ibid.
48	Rád věděl bych, proč právě nyní vzkvétá	Gellner, F.	147	0	0	0	0	0,006061	0	ibid.
49	Sobota, myslím byla snad	Gellner, F.	80	0	0	0	0	0	0	ibid.
50	Už se mi k smrti protiví	Gellner, F.	71	0	0	0	0	0	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
51	Všichni mi lhali	Gellner, F.	135	0,190476	0,152381	0,363636	0,072727	0,092308	0,24	ibid.
52	Vzdušné mé vidiny, nádherná těla,	Gellner, F.	80	0	0,32	0	0,190476	0,22381	0,363636	ibid.
53	A kdesi...	Havel, V.	142	0	0	0	0	0,04	0	Havel, V.: Záchvěvy. In Spisy I. Básně, antikódy. Praha, 1999
54	Až...	Havel, V.	79	0	0	0	0	0	0	ibid.
55	Až jednou...	Havel, V.	58	0	0	0	0	0	0	ibid.
56	Básník...	Havel, V.	72	0	0	0	0	0,15	0	ibid.
57	Bylo...	Havel, V.	118	0	0	0	0	0	0	ibid.
58	Často...	Havel, V.	270	0,029412	0,071895	0,222222	0	0,039103	0	ibid.
59	Čeho...	Havel, V.	67	0	0	0	0	0	0	ibid.
60	Chcete...	Havel, V.	69	0	0,013333	0	0	0	0	ibid.
61	Dnes...	Havel, V.	67	0	0	0	0	0	0	ibid.
62	Filosofové...	Havel, V.	62	0,190476	0,209524	0,363636	0,047619	0,122449	0,2	ibid.
63	Hlubokým...	Havel, V.	138	0	0,06	0	0,045714	0,119048	0,2	ibid.
64	Je...	Havel, V.	58	0	0,066667	0	0	0	0	ibid.
65	Každý...	Havel, V.	43	0	0	0	0	0	0	ibid.
66	Letní...	Havel, V.	80	0	0	0	0	0	0	ibid.
67	Mrtvý...	Havel, V.	175	0	0,05	0	0	0,024571	0	ibid.
68	Nejsem...	Havel, V.	68	0	0	0	0	0,019048	0	ibid.
69	Pražské...	Havel, V.	103	0	0	0	0	0,045918	0	ibid.
70	Ptali...	Havel, V.	42	0	0	0	0	0	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
71	Seděl...	Havel, V.	78	0	0	0	0	0	0	ibid.
72	Zas...	Havel, V.	78	0	0	0	0	0,010084	0	ibid.
73	Ale čas	Holan, V.	54	0	0,083333	0	0	0	0	Holan, V.: Bolest. In Spisy Vladimíra Holana. Lamento. Praha, 2000.
74	Eva	Holan, V.	123	0	0,028571	0	0	0,020833	0	ibid.
75	Je	Holan, V.	111	0	0	0	0	0,001008	0	ibid.
76	Jednoho rána	Holan, V.	74	0	0	0	0	0	0	ibid.
77	Jeskyně slov	Holan, V.	84	0,5	0,392857	0,5	0,5	0,392857	0,5	ibid.
78	Když prší v neděli	Holan, V.	90	0	0,107143	0	0	0,098571	0	ibid.
79	Letovisko	Holan, V.	93	0	0	0	0	0	0	ibid.
80	Lovec zmijí	Holan, V.	74	0	0	0	0	0	0	ibid.
81	Oči muže	Holan, V.	123	0	0	0	0	0,027778	0	ibid.
82	Po letech u maminky	Holan, V.	93	0	0,1	0	0	0,057143	0	ibid.
83	Po smrti sestry Růženy	Holan, V.	56	0	0	0	0	0	0	ibid.
84	Podzim II	Holan, V.	69	0	0,15	0	0	0	0	ibid.
85	Při nespavosti	Holan, V.	57	0	0	0	0	0,022222	0	ibid.
86	Slunce o hromnicích	Holan, V.	79	0	0	0	0	0	0	ibid.
87	Svítání	Holan, V.	134	0,100962	0,107095	0,1875	0,062413	0,06822	0,18	ibid.
88	Ti druzí	Holan, V.	60	0,333333	0,266667	0,333333	0	0	0	ibid.
89	Trojúhelník	Holan, V.	65	0	0,125	0	0	0,08	0	ibid.
90	Umírající básník	Holan, V.	83	0	0,014118	0	0	0,004762	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
91	Verše	Holan, V.	90	0,179487	0,132184	0,2	0,084656	0,074074	0,190476	ibid.
92	Vzpomínka	Holan, V.	181	0	0,069444	0	0	0,049242	0	ibid.
93	Barbarský zpěv	Hrubín, F.	189	0	0,046296	0	0	0,017857	0	Hrubín, F.: Země po polednách. In Básně. Brno, 2010.
94	Básníkova smrt	Hrubín, F.	129	0	0	0	0	0	0	ibid.
95	Člověk	Hrubín, F.	105	0	0	0	0	0	0	ibid.
96	De profundis	Hrubín, F.	134	0	0	0	0	0	0	ibid.
97	Elegie	Hrubín, F.	94	0	0	0	0	0	0	ibid.
98	Chvění listu	Hrubín, F.	107	0	0,015625	0	0	0	0	ibid.
99	Kráčející v polích	Hrubín, F.	79	0	0,035294	0	0	0,056471	0	ibid.
100	Letní smrt	Hrubín, F.	122	0	0	0	0	0	0	ibid.
101	Městu na pahorcích	Hrubín, F.	164	0	0	0	0	0	0	ibid.
102	Na konec písní	Hrubín, F.	72	0	0	0	0	0	0	ibid.
103	Napsáno mrtvým	Hrubín, F.	95	0	0	0	0	0	0	ibid.
104	Nokturno smrti	Hrubín, F.	144	0	0	0	0	0	0	ibid.
105	Osud	Hrubín, F.	72	0	0	0	0	0	0	ibid.
106	Píseň pocestného	Hrubín, F.	160	0	0	0	0	0	0	ibid.
107	Tvář bez podoby	Hrubín, F.	169	0,23088	0,141414	0,258065	0,23088	0,141414	0,258065	ibid.
108	Víno chvály	Hrubín, F.	107	0	0	0	0	0	0	ibid.
109	Železné rouno	Hrubín, F.	131	0,07619	0,126984	0,266667	0	0,06	0	ibid.
110	Země po polednách	Hrubín, F.	92	0	0,048485	0	0	0,046154	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
111	Žízeň	Hrubín, F.	115	0	0,071429	0	0	0,071429	0	ibid.
112	Zpěv marnosti	Hrubín, F.	117	0	0	0	0	0	0	ibid.
113	Ač dávno tak...	Kohout, P.	122	0	0	0	0	0,006944	0	Kohout, P.: Čas lásky a boje. Praha, 1953.
114	Čas boje	Kohout, P.	725	0	0,00989	0	0	0,013344	0	ibid.
115	Čas lásky	Kohout, P.	149	0	0	0	0	0,0344	0	ibid.
116	Dělnickému studentovi	Kohout, P.	156	0	0	0	0	0	0	ibid.
117	Dnes večer zpíval kos	Kohout, P.	126	0	0	0	0	0,021429	0	ibid.
118	Francouzské proměny	Kohout, P.	165	0	0	0	0	0,017857	0	ibid.
119	Na západním jezeře v Chang-Čou	Kohout, P.	204	0,84	0,486667	0,72973	0,805	0,478889	0,794118	ibid.
120	Jízda	Kohout, P.	119	0	0,1	0	0,25	0,196429	0,25	ibid.
121	Korejská vteřina	Kohout, P.	119	0	0	0	0,091429	0,22381	0,285714	ibid.
122	Nad novinami	Kohout, P.	84	0	0,026667	0	0	0	0	ibid.
123	Nemá tvář supa	Kohout, P.	185	0	0	0	0	0	0	ibid.
124	Odsouzený	Kohout, P.	119	0	0,061224	0	0	0,022959	0	ibid.
125	On bude s námi v Bukurešti	Kohout, P.	287	0	0,003233	0	0	0,053419	0	ibid.
126	Panně Orleánské tureckého lidu	Kohout, P.	237	0	0	0	0	0	0	ibid.
127	Píseň pro J. Fučíka	Kohout, P.	149	0	0	0	0,125	0,133929	0,285714	ibid.
128	Slovo k soudruhu Stalinovi	Kohout, P.	313	0	0	0	0	0	0	ibid.
129	Trest	Kohout, P.	145	0	0	0	0	0,055556	0	ibid.
130	V noci, která vítá Nový rok	Kohout, P.	337	0,056566	0,086667	0,148936	0,017544	0,04974	0,09589	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
131	Vlak, který vyjel v únoru 1948	Kohout, P.	453	0	0,024747	0	0	0,005411	0	ibid.
132	Vrahům Rosenbergových	Kohout, P.	119	0	0	0	0	0	0	ibid.
133	Cukrová balada	Nezval, V.	76	0	0	0	0	0	0	Nezval, V.: Pantomima. Praha 2004.
134	Exotická láska	Nezval, V.	118	0	0	0	0	0	0	ibid.
135	Jarmareční písnička o nevěrné lásce	Nezval, V.	208	0	0	0	0	0	0	ibid.
136	Jarní	Nezval, V.	99	0	0	0	0	0	0	ibid.
137	Klára	Nezval, V.	65	0	0	0	0	0	0	ibid.
138	Maceška	Nezval, V.	144	0,3	0,358333	0,461538	0,088154	0,203306	0,315789	ibid.
139	Památce Mikuláše Lenina	Nezval, V.	185	0	0	0	0	0	0	ibid.
140	Podivuhodný kouzelník – zpěv 1	Nezval, V.	438	0	0	0	0	0,006244	0	ibid.
141	Podivuhodný kouzelník – zpěv 2	Nezval, V.	188	0	0,005952	0	0	0,079365	0	ibid.
142	Podivuhodný kouzelník – zpěv 3	Nezval, V.	302	0	0,015152	0	0	0,042011	0	ibid.
143	Podivuhodný kouzelník – zpěv 4	Nezval, V.	641	0	0,008868	0	0	0,011719	0	ibid.
144	Podivuhodný kouzelník – zpěv 5	Nezval, V.	422	0	0,014957	0	0,013605	0,031005	0,082474	ibid.
145	Podivuhodný kouzelník – zpěv 6	Nezval, V.	1212	0,029014	0,039468	0,063025	0,02579	0,044827	0,05694	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
146	Podivuhodný kouzelník – zpěv 7	Nezval, V.	446	0,006803	0,040278	0,076923	0	0,035167	0	ibid.
147	Poetika	Nezval, V.	104	0	0	0	0	0	0	ibid.
148	Rosina Lodolla	Nezval, V.	119	0,457143	0,333333	0,357143	0,244898	0,20068	0,3125	ibid.
149	Týden v barvách	Nezval, V.	154	0	0,081633	0	0,083333	0,089286	0,238095	ibid.
150	Vápeníci	Nezval, V.	72	0	0,146667	0	0	0,133333	0	ibid.
151	Vlak jenž projel parkem	Nezval, V.	78	0	0,216667	0	0	0,125	0	ibid.
152	Zátiší v peřinách	Nezval, V.	116	0	0,033333	0	0	0,053571	0	ibid.
153	Báseň léta	Orten, J.	118	0	0	0	0	0	0	Orten, J.: Ohnice. In Spisy IV. Knihy veršů. Praha 1995.
154	Báseň šera	Orten, J.	189	0,109091	0,141414	0,206897	0,036364	0,107438	0,139535	ibid.
155	Co jsem odpověděl kanárkovi	Orten, J.	61	0	0	0	0	0	0	ibid.
156	Co mi řekl kanárek	Orten, J.	65	0	0	0	0	0	0	ibid.
157	Cvičení o smrti	Orten, J.	87	0	0	0	0	0,1	0	ibid.
158	Cvičení před spaním	Orten, J.	79	0	0	0	0	0	0	ibid.
159	Cvičení v šeru	Orten, J.	79	0	0	0	0	0,1	0	ibid.
160	Dětská	Orten, J.	114	0	0	0	0	0	0	ibid.
161	Dvojitá tma	Orten, J.	251	0	0	0	0	0,002493	0	ibid.
162	Klenba	Orten, J.	93	0	0	0	0	0	0	ibid.
163	Po smrti	Orten, J.	104	0	0	0	0	0	0	ibid.
164	Podzim	Orten, J.	63	0	0	0	0	0,15	0	ibid.
165	První báseň	Orten, J.	178	0	0,004464	0	0	0	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
166	První noc	Orten, J.	105	0	0	0	0	0	0	ibid.
167	Předjaří	Orten, J.	68	0	0	0	0	0,3	0	ibid.
168	Stará knížka	Orten, J.	84	0	0	0	0	0	0	ibid.
169	Ta krajina	Orten, J.	318	0,181818	0,129477	0,208333	0,087468	0,077797	0,130435	ibid.
170	Touha	Orten, J.	164	0	0,05	0	0	0,019481	0	ibid.
171	Vločky	Orten, J.	124	0	0	0	0	0	0	ibid.
172	Život	Orten, J.	574	0	0,017281	0	0	0,023518	0	ibid.
173	Báseň nejpokornější	Seifert, J.	86	0	0	0	0	0	0	Seifert, J.: Město v slzách. In Dílo Jaroslava Seiferta, sv. 1. Praha, 2001.
174	Báseň plná odvahy a víry	Seifert, J.	318	0	0	0	0	0,006208	0	ibid.
175	Báseň úvodní	Seifert, J.	214	0	0,016807	0	0,018674	0,063725	0,131579	ibid.
176	Děti z předměstí	Seifert, J.	197	0	0	0	0	0,028571	0	ibid.
177	Dobrá zvěst	Seifert, J.	156	0	0,028571	0	0,047619	0,09949	0,2	ibid.
178	Hříšné město	Seifert, J.	76	0,25974	0,289562	0,454545	0,714286	0,397072	0,538462	ibid.
179	Chudý	Seifert, J.	203	0	0,00303	0	0	0	0	ibid.
180	Konec války	Seifert, J.	226	0	0	0	0	0,006818	0	ibid.
181	Město v slzách	Seifert, J.	204	0,059524	0,057292	0,16129	0,060606	0,051136	0,15	ibid.
182	Modlitba na chodníku	Seifert, J.	467	0	0,002313	0	0	0	0	ibid.
183	Monolog bezrukého vojáka	Seifert, J.	290	0,116667	0,112346	0,225806	0,153846	0,104895	0,192308	ibid.
184	Na vojenském hřbitově	Seifert, J.	255	0	0	0	0	0,009324	0	ibid.
185	Plátno v kinu	Seifert, J.	185	0	0	0	0	0	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
186	Prosinec 1920	Seifert, J.	309	0	0,0069	0	0	0,001061	0	ibid.
187	Řeč davu	Seifert, J.	151	0	0,029762	0	0,02442	0,059829	0,166667	ibid.
188	Revoluce	Seifert, J.	234	0	0	0	0	0	0	ibid.
189	Stvoření světa	Seifert, J.	350	0	0,007018	0	0	0,020468	0	ibid.
190	V Getsemanské zahradě	Seifert, J.	130	0	0,071429	0	0	0,10119	0	ibid.
191	V předměstské uličce	Seifert, J.	370	0	0,006349	0	0	0,018939	0	ibid.
192	Zpívaná modlitba	Seifert, J.	112	0	0	0	0	0,125	0	ibid.
193	Dcery písně	Skácel, J.	102	0	0	0	0	0	0	Skácel, J.: Smuténka. Praha, 2001.
194	Dopis	Skácel, J.	53	0	0	0	0	0	0	ibid.
195	Druhá báseň o Brně	Skácel, J.	54	0,5	0,416667	0,5	0,533333	0,416667	0,5	ibid.
196	Elegie	Skácel, J.	49	0	0	0	0,1	0,14	0,272727	ibid.
197	Jiná báseň o srdci	Skácel, J.	58	0	0,12	0	0	0,1	0	ibid.
198	Kdo nás napomene	Skácel, J.	68	0	0,1	0	0	0,095238	0	ibid.
199	Nahým a mokrým navrch	Skácel, J.	61	0	0	0	0	0	0	ibid.
200	Odvaha k tomu	Skácel, J.	52	0	0	0	0	0	0	ibid.
201	Ovečky	Skácel, J.	49	0,533333	0,35	0,5	0,533333	0,45	0,5	ibid.
202	Podzim s Moravany	Skácel, J.	81	0	0	0	0	0	0	ibid.
203	Pořád	Skácel, J.	79	0	0	0	0	0,059524	0	ibid.
204	Převozné pro Chárona	Skácel, J.	50	0	0	0	0	0	0	ibid.
205	Příliš čistý sníh	Skácel, J.	60	1	0,5	1	1	0,5	1	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
206	Přísloví	Skácel, J.	47	0	0	0	0	0,026667	0	ibid.
207	Smuténka	Skácel, J.	51	0,666667	0,633333	1	0,666667	0,633333	1	ibid.
208	Trnka	Skácel, J.	87	0	0	0	0,07619	0,126984	0,266667	ibid.
209	Trubači	Skácel, J.	53	0	0	0	0	0	0	ibid.
210	Třmeny	Skácel, J.	107	0	0,064815	0	0,083333	0,17963	0,185185	ibid.
211	Verše	Skácel, J.	55	0	0	0	0	0,111111	0	ibid.
212	Zlatá brána	Skácel, J.	52	0	0	0	0	0	0	ibid.
213	Dláždění	Wolker, J.	125	0	0,061517	0	0	0,086124	0	Wolker, J.: Host do domu. Brumovice, 2012.
214	Hoj	Wolker, J.	93	0	0,1	0	0	0,083333	0	ibid.
215	Host do domu	Wolker, J.	178	0	0	0	0	0	0	ibid.
216	Kamna	Wolker, J.	89	0	0	0	0	0	0	ibid.
217	Návrat	Wolker, J.	140	0	0	0	0	0	0	ibid.
218	Nemocná milá	Wolker, J.	100	0	0	0	0	0	0	ibid.
219	Noční déšť	Wolker, J.	154	0	0	0	0	0,111111	0	ibid.
220	Okno	Wolker, J.	117	0,125	0,175	0,3	0,182857	0,171429	0,307692	ibid.
221	Pokora	Wolker, J.	42	0	0	0	0,525	0,331731	0,5	ibid.
222	Poštovní schránka	Wolker, J.	80	0	0	0	0	0	0	ibid.
223	Poutníci	Wolker, J.	82	0	0	0	0	0,035714	0	ibid.
224	Rekruti	Wolker, J.	155	0	0	0	0	0	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
225	Smrt	Wolker, J.	88	0	0	0	0	0	0	ibid.
226	Svatodušní svátky	Wolker, J.	110	0	0	0	0	0,044643	0	ibid.
227	V parku před polednem	Wolker, J.	82	0	0,022222	0	0	0	0	ibid.
228	Vězeň	Wolker, J.	121	0	0	0	0	0	0	ibid.
229	Zamilovaný	Wolker, J.	94	0	0	0	0	0,071429	0	ibid.
230	Ze soboty na neděli	Wolker, J.	136	0	0	0	0	0,052469	0	ibid.
231	Žebráci	Wolker, J.	168	0	0	0	0	0,011111	0	ibid.
232	Žně	Wolker, J.	77	0	0,12	0	0	0,1	0	ibid.
233	Dětský pasionál	Zahradníček, J.	113	0	0,028571	0	0	0,042857	0	Zahradníček, J.: Pokušení smrti. In Dílo I. Praha 1991.
234	Domov	Zahradníček, J.	156	0	0	0	0	0,083333	0	ibid.
235	Duše	Zahradníček, J.	198	0	0	0	0	0,014815	0	ibid.
236	Jejich stín	Zahradníček, J.	1254	0,008573	0,020243	0,057416	0,020057	0,044707	0,048632	ibid.
237	Lítost	Zahradníček, J.	187	0	0	0	0	0,003209	0	ibid.
238	Má podoba	Zahradníček, J.	276	0	0,018182	0	0	0,022792	0	ibid.
239	Nářek těla	Zahradníček, J.	132	0	0	0	0	0,021429	0	ibid.
240	Nevinnost	Zahradníček, J.	227	0	0,006667	0	0	0	0	ibid.
241	Paměť	Zahradníček, J.	177	0	0	0	0	0,013889	0	ibid.
242	Podoba smrti	Zahradníček, J.	171	0	0	0	0	0,04784	0	ibid.
243	Pohrobci	Zahradníček, J.	217	0	0	0	0	0	0	ibid.
244	Pokušení smrti	Zahradníček, J.	200	0	0,035714	0	0	0,059524	0	ibid.
245	Proměny	Zahradníček, J.	148	0	0,009091	0	0	0,00974	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
246	Předjaří	Zahradníček, J.	163	0	0	0	0	0	0	ibid.
247	Tesknice	Zahradníček, J.	132	0	0,055556	0	0	0,047619	0	ibid.
248	Tvář zla	Zahradníček, J.	199	0	0,031746	0	0	0,111111	0	ibid.
249	Vděčnost mrtvým	Zahradníček, J.	228	0	0	0	0	0,02381	0	ibid.
250	Večer dětí	Zahradníček, J.	172	0	0	0	0	0	0	ibid.
251	Země stínů	Zahradníček, J.	222	0	0,071429	0	0	0,074747	0	ibid.
252	Ztracené okamžiky	Zahradníček, J.	184	0	0,044444	0	0,10582	0,135802	0,208333	ibid.
253	Diag se obrátí kvůli výroku arbitrů na zahraniční exekuční soudy	ČTK	470	0,37619	0,248352	0,276923	0,208705	0,216927	0,376623	České noviny. Zpravodajský server ČTK. http://www.ceskenoviny.cz/
254	Druhá arbitráž kvůli Blance potrvá dva měsíce, zaplatí ji IDS	ČTK	402	0,350922	0,245921	0,352941	0,438735	0,348561	0,558824	ibid.
255	EU dosáhla předběžné dohody o podobě nových sankcí	ČTK	387	0	0,116162	0	0,132609	0,181496	0,25	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
256	Islamisté v Iráku zbourali mešitu a svatyni ze 14. století	ČTK	198	0	0	0	0,043269	0,077586	0,166667	ibid.
257	Kouření v restauracích by mohlo být zakázáno od ledna 2016	ČTK	374	0,119048	0,179654	0,301887	0,183088	0,219679	0,289855	ibid.
258	MF výrazně zlepšilo pro letošek odhad růstu ekonomiky na 2,7 %	ČTK	300	0,278571	0,229978	0,395833	0,371292	0,303769	0,540541	ibid.
259	OSN: Sestřelení letadla nad Ukrajinou je možná válečný zločin	ČTK	259	0	0	0	0,323232	0,203896	0,21875	ibid.
260	Podle Obamy je nutné okamžité a bezpodmínečné příměří v Gaze	ČTK	312	0,058333	0,122222	0,184211	0,166667	0,292125	0,466667	ibid.
261	Případ neoprávněné restituentky podle TI odkryl další podvody	ČTK	413	0	0,044872	0	0,023443	0,082313	0,111111	ibid.
262	Sobotka: Speciální zákon pro NP Šumava není ideálním řešením	ČTK	380	0	0,072739	0	0,281385	0,225774	0,326923	ibid.
263	Soud zamítl Berdychovu žádost o podmínečné propuštění	ČTK	558	0,068027	0,1146	0,150943	0,089286	0,128646	0,238095	ibid.
264	Spor o autorství Klausovy amnestie možná skončí smírem	ČTK	473	0,194805	0,171329	0,175439	0,144048	0,125833	0,135802	ibid.
265	USA obvinily Rusko z porušení dohody o omezení jaderných zbraní	ČTK	354	0,011111	0,05303	0,115385	0,133644	0,143162	0,275862	ibid.
266	Útoky v Gaze zabily přes 100 Palestinců, zasáhly dům vůdce Hamasu	ČTK	460	0	0,081818	0	0,224041	0,220276	0,380952	ibid.
267	V Beskydech blesk zapálil chatu, vítr lámal stromy	ČTK	515	0,1	0,128788	0,269841	0,185714	0,189286	0,313253	ibid.
268	Vedení armády čekají od začátku srpna výrazné změny	ČTK	400	0,029412	0,10223	0,155556	0,020408	0,123214	0,102564	ibid.
269	Vláda hledá další peníze na platy, aby 3,5 % dostali všichni	ČTK	424	0,083333	0,118371	0,175439	0,059524	0,076236	0,163934	ibid.
270	Zeman v Týdnu: Sobotka bruselskou schůzkou o komisaři ponížil ČR	ČTK	383	0,047511	0,106627	0,159091	0,038095	0,10989	0,126984	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
271	Žalobce zrušil stíhání dvou úředníků z ministerstva průmyslu	ČTK	187	0	0	0	0	0,166667	0	ibid.
272	Železniční trať na Kolínsku byla po dvou dnech plně zprůjezdněna	ČTK	158	0	0	0	0,178571	0,262755	0,294118	ibid.
273	K otázce kodifikační pravomoci	Adam, R.	1504	0	0,020202	0	0,097855	0,087347	0,226115	Naše řeč. http://nase-rec.ujc.cas.cz/
274	Máte něco proti mámině sezení v tátovu křesle?	Adam, R.	1209	0,024409	0,055072	0,068627	0,1173	0,100143	0,27758	ibid.
275	Znovu a šířeji o formě kodifikace	Adam, R.	1844	0,040332	0,041183	0,108108	0,058952	0,05789	0,114173	ibid.
276	Morbus professionalis (K motivovanosti českých názvů nemocí)	Bozděchová, I.	2465	0,232074	0,160113	0,529563	0,218241	0,138941	0,576596	ibid.
277	Relační (desubstantivní) adjektiva v odborné lékařské terminologii	Bozděchová, I.	2644	0,079751	0,074548	0,140884	0,197881	0,163429	0,392193	ibid.
278	Za ještě tvrdší kodifikační diktát?	Cvrček, V.	1317	0,023511	0,064468	0,125	0,023262	0,060556	0,111748	ibid.
279	Příznakovost systémová a situačně-kontextová	Čechová, M.	3430	0,003482	0,024192	0,028614	0,075024	0,071642	0,241885	ibid.
280	Herbář z kodexu vodňanského a jeho ortografické zvláštnosti	Černá, A.	4091	0,012052	0,041244	0,054688	0,025798	0,042953	0,16004	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
281	Vyjadřování množství léčiv ve staročeských lékařských textech	Černá, A.	4287	0,04756	0,042262	0,135076	0,049479	0,054494	0,170128	ibid.
282	Haškův „Švejk“ a Vachkovo „Bidýlko“ – dva milníky ve vývoji jazyka české prózy	Daneš, F.	4222	0	0,011152	0	0,03501	0,041977	0,119026	ibid.
283	Účelové přívlastky s předložkami pro, na, k	Daneš, F.	2718	0,005158	0,026505	0,05814	0,046343	0,058713	0,124324	ibid.
284	Věříš tomu? Vůbec (ne)	Daneš, F.	2712	0,000434	0,006753	0,024963	0,004624	0,016773	0,02924	ibid.
285	Dva příspěvky k odvozování sloves	Dokulil, M.	3507	0,110572	0,076215	0,251163	0,134142	0,100582	0,287849	ibid.
286	K stylistice zvukových prostředků jazyka	Horálek, K.	4383	0,003388	0,01652	0,023756	0,028495	0,047812	0,169717	ibid.
287	Recepce textu, jeho analýza a interpretace	Hrbáček, J.	2841	0,184404	0,140972	0,298361	0,20808	0,151336	0,359564	ibid.
288	Onomaziologické funkce pojmenování barev ve Stifterově povídce Bergkristall	Jaklová, A.	1926	0,080769	0,104264	0,273196	0,14987	0,143675	0,329832	ibid.
289	Text a obraz v „billboardové“ reklamě	Jaklová, A.	2019	0,027462	0,052961	0,106796	0,136175	0,120275	0,320833	ibid.
290	Jak nesrovnávat překlady	Janiš, V.	1020	0	0,029778	0	0,060287	0,057986	0,099448	ibid.
291	Vy neznáte Cajzla? (K původu a významu přezdívky)	Kloferová, S.	2416	0,033757	0,068289	0,101737	0,090909	0,122559	0,305317	ibid.
292	K jazykovým a právním aspektům přechylování příjmení v češtině	Knappová, M.	2347	0,12589	0,094396	0,227444	0,172999	0,129008	0,371302	ibid.
293	K jazykové stránce jednoho rozhlasového pořadu	Kořenský, J.	1368	0	0,019799	0	0,065705	0,08858	0,208861	ibid.
294	Výskyt variantních tvarů podstatných jmen v češtině, I-II	Kořenský, J.	1626	0,041643	0,075724	0,109661	0,103968	0,11636	0,191579	ibid.
295	Pojmenování žen s formantem -ka ve staré češtině	Kouba, J.	1077	0,027222	0,055686	0,139706	0,159006	0,163175	0,302752	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
296	K pomnožným podstatným jménům v současné češtině	Kroupová, L.	2078	0,149967	0,12891	0,280255	0,207675	0,152137	0,315403	ibid.
297	Klasifikace sekundárních předložek z hlediska jejich tvoření	Kroupová, L.	1252	0,036657	0,049614	0,092937	0,104957	0,083021	0,224377	ibid.
298	K vokalizaci neslabičných předložek v současné češtině	Kučera, K.	3735	0,053931	0,057097	0,117584	0,11127	0,0932	0,2284	ibid.
299	K jazykové stránce kázání Sváti Karáska	Kvítková, N.	2411	0	0,008497	0	0,008045	0,035986	0,083994	ibid.
300	Jména vlastností ve větě	Macháčková, E.	2072	0,089744	0,074074	0,188953	0,184453	0,128837	0,300917	ibid.
301	Výrazy typu být předmětem jednání, stát se středem pozornosti	Macháčková, E.	1150	0,10168	0,071404	0,150754	0,154795	0,139047	0,356209	ibid.
302	Opakování a syntaktický paralelizmus v rozhovorech s muži a ženami v jednom z druhů institucionální komunikace	Müllerová, O.	4226	0,010768	0,025534	0,05145	0,037511	0,050287	0,135154	ibid.
303	K vyjadřování předmětu a příslovečných určení v češtině	Nebeská, I.	2178	0,127347	0,113273	0,239609	0,16026	0,150462	0,341751	ibid.
304	Několik poznámek k využívání formantu -ák ve slanzích a ve spisovném jazyce	Nekvapil, J.	1549	0,076923	0,11413	0,189474	0,147532	0,152142	0,306373	ibid.
305	Frazeologizace slovesa dělati a jeho synonym	Němec, I.	3923	0,158561	0,130136	0,308767	0,212714	0,154629	0,394354	ibid.
306	Zaniklá apelativa v současných pomístních jménech	Oliva, K.	2172	0,137931	0,085766	0,157647	0,109493	0,090944	0,284264	ibid.
307	Rozvití předmětová a příslovečná, doplňující a určující	Panevová, J.	2208	0,088765	0,075701	0,163482	0,136253	0,1037	0,286614	ibid.
308	K vývojovým tendencím českých místních jmen zakončených na -ice	Polívková, A.	943	0,191578	0,145536	0,217143	0,17562	0,111232	0,28692	ibid.
309	Poznámka k pojmu hyperkorektnost	Sgall, P.	1825	0,010217	0,033896	0,046997	0,068681	0,06957	0,159844	ibid.
310	Ke kategorii životnosti některých právních termínů a k jejich shodě	Šimandl, J.	1086	0	0,016565	0	0,067813	0,07488	0,14978	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
311	Dvojí překlad knihy D. H. Lawrence Milenec Lady Chatterleyové z perspektivy genderu	Širokovská, S.	4187	0,028517	0,031099	0,105068	0,038972	0,052229	0,162069	ibid.
312	K periferii současného souvětného systému se spojovacím výrazem co	Štěpán, J.	6183	0,073358	0,064455	0,25113	0,105075	0,077302	0,294753	ibid.
313	Označení černé a bílé barvy v zeměpisných jménech v Čechách	Štěpán, P.	4510	0,08508	0,090772	0,309483	0,211371	0,146503	0,403141	ibid.
314	Konkurence kladu a záporu v otázkách zjišťovacích	Štícha, F.	2365	0,05783	0,06717	0,111801	0,092522	0,068717	0,21682	ibid.
315	O jazyce soudních rozhodnutí	Štícha, F.	3197	0,006881	0,027057	0,065878	0,02689	0,04454	0,11329	ibid.
316	Slovosled v prózách Bohumila Hrabala	Štícha, F.	2422	0,042312	0,039739	0,152273	0,042892	0,052074	0,148936	ibid.
317	Přejaté odborné názvy a hledisko spisovnosti	Uher, F.	1661	0,011452	0,041919	0,056522	0,101662	0,133815	0,386423	ibid.
318	K poloze příklonek ve vedlejších větách spojkových	Uhlířová, L.	2169	0,073983	0,068037	0,162637	0,131698	0,103844	0,274062	ibid.
319	O frekvenci příslovečného určení v souvislém textu	Uhlířová, L.	2327	0,137432	0,110505	0,264463	0,223757	0,178138	0,479549	ibid.
320	Sloveso určité v aktuálním členění větném	Uhlířová, L.	2976	0,093561	0,087733	0,218698	0,165322	0,13523	0,37163	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
321	O českých zaříkávadlech: text prostoru a text jména	Velmezová, J.	2097	0,038746	0,061508	0,166667	0,176115	0,140575	0,31127	ibid.
322	Ještě k přízvukování prvotních jednoslabičných předložek	Zeman, J.	1708	0,184297	0,116585	0,343891	0,231092	0,155844	0,382155	ibid.
323	Moderní totalitarismus a síla politické imaginace	Budil, I.	3541	0,045372	0,040714	0,149618	0,0912	0,071616	0,278634	Vesmír. http://casopis.vesmir.cz/
324	Jak mořský fytoplankton ovlivňuje podnebí?	Burkartová, K.	698	0,053571	0,076562	0,107143	0,031853	0,111377	0,088889	ibid.
325	O původu válek	Duda, P.	712	0,012987	0,054356	0,170213	0,055165	0,129054	0,25	ibid.
326	Vodní klastry a nejrychlejší rychlovarná konvice na světě	Fárník, M.	1362	0,063838	0,066089	0,166667	0,125521	0,097633	0,261972	ibid.
327	Pulzující vodní paprsek – technologie budoucnosti?	Foldyna, J.	1147	0,091085	0,154286	0,27027	0,28661	0,234838	0,480263	ibid.
328	Křehké vztahy atomů helia	Friedrich, B.	1334	0,19617	0,196546	0,344828	0,193878	0,193407	0,451613	ibid.
329	Obezita a její léčba	Haluzík, M.	1760	0,02916	0,032951	0,064394	0,078541	0,070311	0,145553	ibid.
330	Kovy jako „skladiště“ vodíku	Havela, L.	2050	0,10365	0,086133	0,162791	0,126283	0,108837	0,251101	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
331	Vepsáno do půdy	Hladký, J.	1695	0,027425	0,043878	0,144262	0,071989	0,071602	0,126697	ibid.
332	Jak vyrobit regulační T-lymfocyty	Hořejší, V.	562	0,482609	0,303878	0,373333	0,401961	0,293925	0,5	ibid.
333	K čemu slouží vrstva hlenu na sliznicích?	Hořejší, V.	880	0,012422	0,048551	0,080808	0,218782	0,221369	0,371429	ibid.
334	Neutrofilní granulocyty bojují proti nádorům	Hořejší, V.	943	0,045013	0,067295	0,092437	0,081871	0,103463	0,229947	ibid.
335	Záhada aktivace γδT lymfocytů vyřešena	Hořejší, V.	1423	0,252564	0,207569	0,311404	0,158059	0,138153	0,336391	ibid.
336	O diceru	Imrichová, T.	1381	0,159263	0,130745	0,273092	0,163462	0,144911	0,246291	ibid.
337	Evoluce ve zkumavce	Kejnovský, E.	989	0,178548	0,129722	0,30719	0,144934	0,159388	0,309179	ibid.
338	Relikty světa RNA	Kejnovský, E.	789	0,276276	0,168345	0,550847	0,401652	0,247394	0,576159	ibid.
339	Jsou to opravdu vodíkové vazby, které stabilizují DNA?	Kolář, M.	984	0,270251	0,186248	0,393333	0,323554	0,211531	0,46988	ibid.
340	Počítačová alchymie	Kolář, M.	1410	0,028889	0,041504	0,06701	0,105523	0,110828	0,299035	ibid.
341	Když milenky s manželkami táhnou za jeden provaz	Lhotský, J.	1181	0,012289	0,04945	0,071006	0,08635	0,111034	0,23506	ibid.
342	Odvrácená tvář půdy	Ložek, V.	2114	0,067222	0,054858	0,154054	0,118645	0,077577	0,193483	ibid.
343	Trojdílný holocén ve světle poznatků z našich luhů a hájů	Ložek, V.	1186	0,03543	0,03193	0,087629	0,029461	0,030272	0,077922	ibid.
344	Toxický účinek metanolu na lidský organismus	Martínková, M.	2518	0,081704	0,070039	0,199584	0,089327	0,083183	0,22546	ibid.
345	Stopy v hlubokých mořích	Mikuláš, R.	1855	0	0,01005	0	0,051897	0,064448	0,176471	ibid.
346	Efektivní energetika	Noskievič, P.	2930	0,071726	0,065722	0,203419	0,142832	0,11527	0,31725	ibid.
347	Turmalín, minerál mnoha podob	Novák, M.	753	0,124554	0,086043	0,214286	0,169882	0,133528	0,251462	ibid.
348	Smrt (a částečné vzkříšení) Homo economicus	Nováková, J.	1892	0,012142	0,040333	0,084112	0,095771	0,104959	0,238	ibid.
349	Hormonální zombie	Petr, J.	941	0	0,042293	0	0,018701	0,066048	0,0553	ibid.
350	Chiméry přicházejí	Petr, J.	1852	0,070405	0,08342	0,190058	0,206492	0,158262	0,364937	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
351	Jak předávají otcové svá traumata potomkům?	Petr, J.	1456	0,043103	0,048156	0,08547	0,045788	0,079048	0,123288	ibid.
352	Skrytý půvab klasifikace	Pokorný, P.	1649	0,018538	0,028775	0,051587	0,047937	0,058182	0,141618	ibid.
353	Všechno je jinak	Pokorný, P.	1455	0	0,008454	0	0	0,027155	0	ibid.
354	Změny klimatu v Česku	Pretel, J.	2112	0,013669	0,038699	0,078947	0,067908	0,077492	0,206897	ibid.
355	Válku s nádorem mohou rozhodnout kolaboranti	Raudenská, M.	1392	0,114301	0,091342	0,195489	0,208242	0,148501	0,322165	ibid.
356	Oblaka – víc než polovina krásy světa	Soukupová, J.	1475	0,020281	0,045576	0,11245	0,143223	0,101256	0,149847	ibid.
357	Syndrom vyhoření a hormony	Stárka, L.	979	0,063796	0,048485	0,103093	0,091987	0,076478	0,176211	ibid.
358	Původ rakoviny	Storchová, Z.	1786	0,087392	0,089965	0,265018	0,192744	0,155063	0,366667	ibid.
359	Příčiny vzniku nádorového růstu	Svoboda, J.	1381	0,004559	0,055588	0,053942	0,031056	0,091816	0,12013	ibid.
360	Proč jsme nevyhynuli na virové infekce?	Svoboda, J.	976	0,016347	0,050292	0,073333	0,134135	0,15292	0,240175	ibid.
361	Porucha chování v REM spánku	Šonka, K.	3042	0,083017	0,078034	0,233154	0,113904	0,1081	0,344456	ibid.
362	Fluor v podzemní vodě na riftu v Etiopii	Šráček, O.	672	0,077806	0,118155	0,210526	0,169382	0,169924	0,403409	ibid.
363	Epigenetické procesy u eukaryotních buněk	Švorcová, J.	2481	0,036145	0,037945	0,080092	0,057704	0,056948	0,193303	ibid.
364	Mnohonožky na talíři?	Tuf, I. H.	789	0,089164	0,071981	0,116505	0,075893	0,072917	0,138211	ibid.
365	Nepředstavitelně krátké laserové impulsy	Turčičová, H.	1340	0,043162	0,078443	0,126829	0,191815	0,159527	0,327044	ibid.
366	Včely a med ve starém Egyptě	Vachala, B.	2328	0,032667	0,039538	0,186158	0,058409	0,068631	0,232975	ibid.
367	Aktualizace bonitovaných půdně ekologických jednotek	Vašků, Z.	1812	0,041759	0,045113	0,179487	0,051907	0,062533	0,208824	ibid.
368	V suterénu Vulkánovy dílny	Vítek, J.	1300	0,212679	0,173156	0,314488	0,104722	0,085066	0,134884	ibid.
369	Poliovirus stále řádí	Vondrejs, V.	704	0,013228	0,064198	0,081301	0	0,033445	0	ibid.
370	Elektrosmog – co o něm dosud víme a nevíme?	Vožeh, F.	2275	0,085341	0,072714	0,175627	0,052015	0,057672	0,15212	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
371	Fotosyntéza známá neznámá	Wilhelmová, N.	995	0	0,03652	0	0,047955	0,053378	0,148148	ibid.
372	Co dělat, když vyteče DNA ?	Zouhar, P.	673	0,071578	0,103128	0,195876	0,041152	0,103849	0,175439	ibid.
373	dopis 01	Hrubín, F.	179	0	0	0	0	0	0	Málková, I., Řehák, D.: Adresát František Hrubín. Dopisy F. Hrubína, J. Seiferta, J. Strnadla, E. Frynty. Brno, 2010.
374	dopis 02	Hrubín, F.	183	0	0	0	0	0	0	ibid.
375	dopis 03	Hrubín, F.	101	0	0	0	0	0	0	ibid.
376	dopis 04	Hrubín, F.	167	0	0	0	0	0	0	ibid.
377	dopis 05	Hrubín, F.	128	0	0	0	0	0	0	ibid.
378	dopis 06	Hrubín, F.	169	0	0	0	0	0	0	ibid.
379	dopis 07	Hrubín, F.	420	0	0	0	0	0,007143	0	ibid.
380	dopis 08	Hrubín, F.	448	0	0	0	0	0	0	ibid.
381	dopis 09	Hrubín, F.	124	0	0	0	0	0	0	ibid.
382	dopis 10	Hrubín, F.	445	0	0	0	0	0	0	ibid.
383	dopis 11	Hrubín, F.	418	0	0	0	0	0	0	ibid.
384	dopis 12	Hrubín, F.	153	0	0	0	0	0	0	ibid.
385	dopis 13	Hrubín, F.	420	0	0	0	0	0,011372	0	ibid.
386	dopis 14	Hrubín, F.	147	0	0	0	0	0	0	ibid.
387	dopis 15	Hrubín, F.	136	0	0	0	0	0	0	ibid.
388	dopis 16	Hrubín, F.	323	0	0	0	0	0	0	ibid.
389	dopis 17	Hrubín, F.	167	0	0	0	0	0,014815	0	ibid.
390	dopis 18	Hrubín, F.	160	0	0	0	0	0	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
391	dopis 19	Hrubín, F.	127	0	0	0	0	0	0	ibid.
392	dopis 20	Hrubín, F.	118	0	0	0	0	0	0	ibid.
393	Dopis 01	Havel, V.	758	0	0	0	0	0,001747	0	Havel, V.: Dopisy Olze. In Spisy V. Dopisy Olze. Praha, 1999.
394	Dopis 05	Havel, V.	1558	0	0	0	0,010055	0,016124	0,046838	ibid.
395	Dopis 07	Havel, V.	906	0	0	0	0	0,007985	0	ibid.
396	Dopis 09	Havel, V.	1295	0	0	0	0	0	0	ibid.
397	Dopis 10	Havel, V.	1171	0	0	0	0	0,0015	0	ibid.
398	Dopis 11	Havel, V.	1904	0,004813	0,013411	0,042373	0,002269	0,00879	0,029982	ibid.
399	Dopis 12	Havel, V.	638	0	0,006192	0	0	0,007656	0	ibid.
400	Dopis 13	Havel, V.	3519	0	0	0	0	0,004054	0	ibid.
401	Dopis 14	Havel, V.	2065	0	0	0	0,000941	0,005976	0,024096	ibid.
402	Dopis 15	Havel, V.	1204	0	0	0	0,002375	0,010321	0,04142	ibid.
403	Dopis 16	Havel, V.	1401	0	0	0	0,006521	0,011512	0,041379	ibid.
404	Dopis 17	Havel, V.	2504	0	0	0	0	0,002774	0	ibid.
405	Dopis 18	Havel, V.	575	0	0	0	0	0	0	ibid.
406	Dopis 19	Havel, V.	806	0	0	0	0,021667	0,027763	0,062802	ibid.
407	Dopis 20	Havel, V.	757	0	0	0	0	0,004924	0	ibid.
408	Dopis 21	Havel, V.	866	0	0	0	0	0,003768	0	ibid.
409	Dopis 27	Havel, V.	401	0	0	0	0	0	0	ibid.
410	Dopis 28	Havel, V.	800	0	0	0	0	0	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
411	Dopis 30	Havel, V.	505	0	0	0	0	0,003939	0	ibid.
412	Dopis 31	Havel, V.	1117	0	0	0	0,009327	0,021139	0,051118	ibid.
413	„Daj to sem!“	Havel, V.	748	0	0,001875	0	0	0,001316	0	Havel, V.: Články. http://www.vaclavhavel.cz/
414	Alfréd Radok. Nekrolog	Havel, V.	1275	0	0	0	0	0,015814	0	ibid.
415	Anatomie jedné zdrženlivosti	Havel, V.	9116	0	0,00374	0	0,007335	0,017538	0,046278	ibid.
416	Bělorusko, náš nový východní soused	Havel, V.	595	0	0,013333	0	0	0,051471	0	ibid.
417	Břemeno 21. srpna	Havel, V.	861	0	0	0	0,030824	0,052886	0,128834	ibid.
418	Byli jsme zbyteční?	Havel, V.	556	0	0	0	0	0	0	ibid.
419	Co je to Hrad? Prostě jímka!	Havel, V.	378	0,141414	0,096212	0,184211	0,068765	0,075321	0,228571	ibid.
420	Co si opravdu myslím o stíhačkách?	Havel, V.	319	0	0	0	0	0,011278	0	ibid.
421	Diktátoři porozumí jen síle	Havel, V.	502	0	0,026515	0	0	0,056818	0	ibid.
422	Dozrál čas	Havel, V.	861	0	0	0	0	0,01863	0	ibid.
423	Dvě poznámky o Chartě 77	Havel, V.	1603	0,007177	0,011107	0,054152	0,041174	0,032375	0,105386	ibid.
424	Globalizovaná odpovědnost	Havel, V.	216	0	0,071429	0	0,025	0,074444	0,138889	ibid.
425	Historická šance pro naši zemi – nepromarněme ji!	Havel, V.	1029	0,024762	0,033513	0,134146	0,027079	0,041821	0,110236	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
426	Jak volit mého nástupce? Přímo!	Havel, V.	813	0,005102	0,024702	0,066116	0,032526	0,057096	0,15	ibid.
427	K Falbrově lži	Havel, V.	150	0	0,05	0	0	0	0	ibid.
428	Litomyšlské znaky	Havel, V.	1634	0	0,005952	0	0	0,018909	0	ibid.
429	Moc bezmocných	Havel, V.	24882	0,008712	0,009413	0,069891	0,023703	0,023463	0,11254	Havel, V.: Eseje a jiné texty. http://www.vaclavhavel.cz/
430	Nepodléhejme rétorice diktátora Castra	Havel, V.	288	0	0	0	0	0,012698	0	ibid.
431	O jedné otázce	Havel, V.	1754	0,011834	0,021686	0,04244	0,058179	0,052942	0,116197	ibid.
432	O smyslu Charty 77	Havel, V.	5189	0,012465	0,012802	0,057118	0,022957	0,027037	0,048024	ibid.
433	Občan Krob	Havel, V.	487	0	0,003715	0	0	0,017287	0	Havel, V.: Články. http://www.vaclavhavel.cz/
434	Oblíbená hračka, nebo věc principu?	Havel, V.	1078	0,006944	0,028454	0,058824	0	0,031534	0	ibid.
435	Odpovědnost jako osud	Havel, V.	3605	0	0,002339	0	0,003417	0,018922	0,057258	ibid.
436	Originalita versus banalita	Havel, V.	1294	0	0	0	0	0,013475	0	ibid.
437	Podcenili jsme nebezpečí ničivého démona	Havel, V.	484	0	0	0	0	0	0	Havel, V.: Eseje a jiné texty. http://www.vaclavhavel.cz/
438	Politika a svědomí	Havel, V.	6480	0,003689	0,004556	0,023031	0,017356	0,018286	0,074039	ibid.
439	Poznámky ke hře Largo desolato	Havel, V.	1949	0	0,002058	0	0,031514	0,020973	0,052023	ibid.
440	Proces	Havel, V.	1772	0	0	0	0	0,011575	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
441	Proč jsem vrátil lustrační zákon	Havel, V.	904	0,006494	0,040341	0,068376	0,049374	0,039251	0,092784	ibid.
442	Proč radar přijmout	Havel, V.	601	0	0,006541	0	0	0,014011	0	ibid.
443	Průchod spravedlnosti	Havel, V.	415	0,1125	0,081439	0,176471	0,110491	0,097917	0,141026	ibid.
444	Příběh a totalita	Havel, V.	6259	0,000636	0,006652	0,018648	0,02454	0,026828	0,107776	Havel, V.: Eseje a jiné texty. http://www.vaclavhavel.cz/
445	Role českého prezidenta	Havel, V.	2052	0,018803	0,026348	0,053012	0,082312	0,064365	0,157556	Havel, V.: Články. http://www.vaclavhavel.cz/
446	Setkání s Gorbačovem	Havel, V.	879	0	0	0	0	0,002438	0	ibid.
447	Svazky StB jako staronová reality show	Havel, V.	188	0	0,05	0	0	0,022575	0	ibid.
448	Svět bez Zdeňka	Havel, V.	2021	0	0,005698	0	0,03588	0,029878	0,06288	ibid.
449	Světe, nezklamal jsi	Havel, V.	629	0	0	0	0	0,01299	0	ibid.
450	Svobodu uchopil Člověk v tísni jako závazek	Havel, V.	430	0	0,060606	0	0	0,069118	0	ibid.
451	Šest poznámek o kultuře	Havel, V.	3749	0,006964	0,010814	0,05042	0,021762	0,019549	0,060498	ibid.
452	Šifra socialismus	Havel, V.	752	0,006275	0,045882	0,083333	0,058333	0,066667	0,109756	ibid.
453	Testovací terén	Havel, V.	972	0	0,001094	0	0	0,006755	0	ibid.
454	Václav Havel a jeho čtyři vzpomínky na Jiřího Dienstbiera	Havel, V.	304	0	0,058889	0	0	0,020856	0	ibid.
455	Václav Havel píše o evropské ústavě	Havel, V.	240	0	0,005128	0	0,037296	0,054545	0,136364	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
456	Válka, nebo schůze?	Havel, V.	638	0	0,010684	0	0	0,011458	0	ibid.
457	Vyložené karty	Havel, V.	1037	0	0,005643	0	0	0,007297	0	ibid.
458	Z listopadu se můžeme poučit	Havel, V.	748	0	0	0	0	0,010765	0	ibid.
459	Zpívá celá rodina	Havel, V.	1026	0,00873	0,028571	0,074324	0,019903	0,049676	0,116183	ibid.
460	Ztráta paměti?	Havel, V.	1249	0	0,012763	0	0,069318	0,055193	0,151316	ibid.
461	Zvát či nezvat?	Havel, V.	800	0	0,008824	0	0	0,012266	0	ibid.
462	Život na vidrholci	Havel, V.	2402	0,00293	0,010144	0,034642	0,020185	0,029709	0,094595	ibid.
463	Prezidentský novoroční projev 1949	Gottwald, K.	1394	0,010256	0,020972	0,060606	0,061538	0,056497	0,134557	Od TGM k Zemanovi: Poslechněte si vánoční a novoroční projevy všech prezidentů. http://www.rozhlas.cz/
464	Prezidentský novoroční projev 1950	Gottwald, K.	2132	0,048592	0,045952	0,157667	0,082411	0,06683	0,173375	ibid.
465	Prezidentský novoroční projev 1951	Gottwald, K.	2150	0,012876	0,020717	0,049887	0,060544	0,056828	0,10789	ibid.
466	Prezidentský novoroční projev 1952	Gottwald, K.	1772	0,028563	0,027991	0,076923	0,047853	0,047414	0,148728	ibid.
467	Prezidentský novoroční projev 1953	Gottwald, K.	1645	0,006646	0,023569	0,04797	0,056709	0,070726	0,24505	ibid.
468	Prezidentský novoroční projev 1954	Zápotocký, A.	2569	0,002479	0,010711	0,030864	0,009006	0,024599	0,0625	ibid.
469	Prezidentský novoroční projev 1955	Zápotocký, A.	1566	0,010943	0,023833	0,052632	0,026278	0,039757	0,132832	ibid.
470	Prezidentský novoroční projev 1956	Zápotocký, A.	2892	0	0,003599	0	0,01402	0,021937	0,0401	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
471	Prezidentský novoroční projev 1957	Zápotocký, A.	2477	0	0,004874	0	0,015214	0,02824	0,087302	ibid.
472	Prezidentský novoroční projev 1958	Novotný, A.	1590	0,005804	0,021053	0,083969	0,025689	0,057471	0,123256	ibid.
473	Prezidentský novoroční projev 1959	Novotný, A.	2108	0	0,008418	0	0,003639	0,02736	0,083721	ibid.
474	Prezidentský novoroční projev 1960	Novotný, A.	2726	0	0,0013	0	0,008259	0,027495	0,057772	ibid.
475	Prezidentský novoroční projev 1961	Novotný, A.	1571	0	0,009881	0	0,042152	0,060215	0,112948	ibid.
476	Prezidentský novoroční projev 1962	Novotný, A.	2675	0	0,007428	0	0,022311	0,031147	0,074425	ibid.
477	Prezidentský novoroční projev 1963	Novotný, A.	1936	0	0,006154	0	0,008996	0,026679	0,035448	ibid.
478	Prezidentský novoroční projev 1964	Novotný, A.	2889	0	0,007504	0	0,019463	0,027425	0,071272	ibid.
479	Prezidentský novoroční projev 1965	Novotný, A.	2251	0,001762	0,020256	0,036496	0,017159	0,037759	0,078078	ibid.
480	Prezidentský novoroční projev 1966	Novotný, A.	3250	0	0,002587	0	0,008173	0,018645	0,049358	ibid.
481	Prezidentský novoroční projev 1967	Novotný, A.	2565	0	0,002611	0	0,008882	0,024576	0,079838	ibid.
482	Prezidentský novoroční projev 1968	Novotný, A.	2293	0,002673	0,013203	0,033482	0,013604	0,03708	0,062027	ibid.
483	Prezidentský novoroční projev 1969	Svoboda, L.	2059	0,002436	0,011438	0,037783	0,021346	0,026585	0,04918	ibid.
484	Prezidentský novoroční projev 1970	Svoboda, L.	2185	0	0,000636	0	0,015789	0,025608	0,062208	ibid.
485	Prezidentský novoroční projev 1971	Svoboda, L.	1551	0	0	0	0,002564	0,024956	0,068627	ibid.
486	Prezidentský novoroční projev 1972	Svoboda, L.	454	0,030435	0,034585	0,118644	0,074534	0,080268	0,131868	ibid.
487	Prezidentský novoroční projev 1973	Svoboda, L.	507	0	0,014545	0	0,045584	0,049383	0,115385	ibid.
488	Prezidentský novoroční projev 1974	Svoboda, L.	428	0	0,008264	0	0,017316	0,054446	0,112676	ibid.
489	Prezidentský novoroční projev 1975	Husák, G.	1510	0,016151	0,027976	0,097744	0,068376	0,083333	0,222222	ibid.
490	Prezidentský novoroční projev 1976	Husák, G.	1478	0	0,007328	0	0,023993	0,046887	0,125307	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
491	Prezidentský novoroční projev 1977	Husák, G.	1276	0	0,007334	0	0,054693	0,074471	0,150273	ibid.
492	Prezidentský novoroční projev 1978	Husák, G.	1581	0,015048	0,019646	0,052147	0,05564	0,066639	0,155462	ibid.
493	Prezidentský novoroční projev 1979	Husák, G.	1318	0	0,0091	0	0,039437	0,04265	0,123288	ibid.
494	Prezidentský novoroční projev 1980	Husák, G.	1370	0	0,004621	0	0,039249	0,054547	0,160819	ibid.
495	Prezidentský novoroční projev 1981	Husák, G.	1546	0,001316	0,021389	0,040741	0,03291	0,057181	0,168704	ibid.
496	Prezidentský novoroční projev 1982	Husák, G.	1150	0,010721	0,018404	0,065476	0,066834	0,071181	0,2	ibid.
497	Prezidentský novoroční projev 1983	Husák, G.	1125	0	0,017359	0	0,038066	0,051695	0,166667	ibid.
498	Prezidentský novoroční projev 1984	Husák, G.	1028	0,004344	0,016686	0,06338	0,026573	0,045421	0,125628	ibid.
499	Prezidentský novoroční projev 1985	Husák, G.	1362	0,002899	0,011419	0,044898	0,036727	0,054649	0,153846	ibid.
500	Prezidentský novoroční projev 1986	Husák, G.	1312	0	0,007709	0	0,029196	0,047935	0,14557	ibid.
501	Prezidentský novoroční projev 1987	Husák, G.	1485	0	0,01477	0	0,024145	0,040987	0,127717	ibid.
502	Prezidentský novoroční projev 1988	Husák, G.	771	0	0,010577	0	0	0,035131	0	ibid.
503	Prezidentský novoroční projev 1989	Husák, G.	854	0	0,007253	0	0,006	0,032121	0,055	ibid.
504	Prezidentský novoroční projev 1980	Havel, V.	2355	0	0	0	0	0,010419	0	ibid.
505	Prezidentský novoroční projev 1991	Havel, V.	2419	0	0,001849	0	0,024357	0,024369	0,073834	ibid.
506	Prezidentský novoroční projev 1992	Havel, V.	3284	0	0	0	0,005179	0,014495	0,046465	ibid.
507	Prezidentský novoroční projev 1994	Havel, V.	2752	0	0,008107	0	0,014308	0,036286	0,063804	ibid.
508	Prezidentský novoroční projev 1995	Havel, V.	3252	0	0,003362	0	0,001709	0,012042	0,023669	ibid.
509	Prezidentský novoroční projev 1996	Havel, V.	2760	0	0	0	0	0,007565	0	ibid.
510	Prezidentský novoroční projev 1997	Havel, V.	598	0	0	0	0	0	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
511	Prezidentský novoroční projev 1998	Havel, V.	1318	0	0,000259	0	0	0,010271	0	ibid.
512	Prezidentský novoroční projev 1999	Havel, V.	1725	0	0,003292	0	0,010027	0,018779	0,045977	ibid.
513	Prezidentský novoroční projev 2000	Havel, V.	2023	0	0,001753	0	0,002524	0,027138	0,033465	ibid.
514	Prezidentský novoroční projev 2001	Havel, V.	1595	0	0	0	0	0,008936	0	ibid.
515	Prezidentský novoroční projev 2002	Havel, V.	1928	0	0	0	0	0,00949	0	ibid.
516	Prezidentský novoroční projev 2003	Havel, V.	1940	0	0,000316	0	0,003564	0,015379	0,035225	ibid.
517	Prezidentský novoroční projev 2004	Klaus, V.	906	0	0,01076	0	0,035354	0,033977	0,080769	ibid.
518	Prezidentský novoroční projev 2005	Klaus, V.	971	0	0	0	0,014116	0,027354	0,055762	ibid.
519	Prezidentský novoroční projev 2006	Klaus, V.	841	0	0,003209	0	0	0,006418	0	ibid.
520	Prezidentský novoroční projev 2007	Klaus, V.	800	0	0	0	0	0,011905	0	ibid.
521	Prezidentský novoroční projev 2008	Klaus, V.	906	0	0,010483	0	0,022032	0,029197	0,064639	ibid.
522	Prezidentský novoroční projev 2009	Klaus, V.	866	0	0,001681	0	0,035761	0,033103	0,078512	ibid.
523	Prezidentský novoroční projev 2010	Klaus, V.	900	0	0,006762	0	0,038218	0,02985	0,076305	ibid.
524	Prezidentský novoroční projev 2011	Klaus, V.	884	0	0,005303	0	0,030051	0,036603	0,077273	ibid.
525	Prezidentský novoroční projev 2012	Klaus, V.	893	0	0,002902	0	0,040578	0,044785	0,076233	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
526	Prezidentský novoroční projev 2013	Klaus, V.	979	0	0	0	0	0,020706	0	ibid.
527	Cipísek	Čtvrtek, V.	24903	0,022621	0,015722	0,118577	0,028592	0,02467	0,170348	Čtvrtek, V.: Cipísek. Praha, 2008.
528	Manka	Čtvrtek, V.	27715	0,017458	0,013907	0,094918	0,026015	0,024271	0,148443	Čtvrtek, V.: Manka. Praha, 2004.
529	Rumcajs	Čtvrtek, V.	20377	0,031403	0,020738	0,136642	0,037817	0,029595	0,163563	Čtvrtek, V.: Rumcajs. Praha, 2002.
530	Osudy dobrého vojáka Švejka I	Hašek, J.	55372	0,011256	0,009195	0,073384	0,022805	0,020128	0,145032	Hašek, J.: Osudy dobrého vojáka Švejka za světové války I. Praha, 2011.
531	Osudy dobrého vojáka Švejka II	Hašek, J.	67314	0,007393	0,007631	0,065124	0,017019	0,016749	0,134211	Hašek, J.: Osudy dobrého vojáka Švejka za světové války II. Praha, 2011.
532	Osudy dobrého vojáka Švejka III	Hašek, J.	61422	0,009495	0,009073	0,078053	0,018568	0,018582	0,141125	Hašek, J.: Osudy dobrého vojáka Švejka za světové války III. Praha, 2011.
533	Osudy dobrého vojáka Švejka IV	Hašek, J.	22062	0,009502	0,010054	0,05086	0,018422	0,019376	0,102769	Hašek, J.: Osudy dobrého vojáka Švejka za světové války IV. Praha, 2011.
534	Obsluhoval jsem anglického krále	Hrabal, B.	67990	0,000964	0,001872	0,023634	0,00531	0,006632	0,082027	Hrabal, B.: Obsluhoval jsem anglického krále. Praha, 2007.
535	Ostře sledované vlaky	Hrabal, B.	18560	0,0053	0,006817	0,047752	0,014078	0,015277	0,117217	Hrabal, B.: Ostře sledované vlaky. Praha, 2000.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
536	Postřižiny	Hrabal, B.	29196	0,006396	0,006049	0,067223	0,01295	0,012213	0,12413	Hrabal, B.: Postřižiny. Praha, 2009.
537	Nesnesitelná lehkost bytí	Kundera, M.	73497	0,025841	0,017693	0,13423	0,0292	0,024347	0,188148	Kundera, M.: Nesnesitelná lehkost bytí. Brno, 2007.
538	Žert	Kundera, M.	96809	0,001545	0,00205	0,01766	0,006821	0,007403	0,109972	Kundera, M.: Žert. Brno, 2007.
539	Kalibův zločin	Rais, K. V.	49684	0,007849	0,007432	0,063729	0,017495	0,018428	0,13992	Rais, K. V.: Kalibův zločin Praha, 1976.
540	Západ	Rais, K. V.	76774	0,010648	0,010323	0,095076	0,019514	0,018925	0,186028	Rais, K. V.: Západ. Praha, 1960.
541	Báječná léta pod psa	Viewegh, M.	49033	0,044709	0,027098	0,172473	0,039	0,027376	0,194731	Viewegh, M.: Báječná léta pod psa. Praha, 1992.
542	Báječná léta s Klausem	Viewegh, M.	41779	0,022366	0,016866	0,123076	0,02459	0,020893	0,1753	Viewegh, M.: Báječná léta s Klausem. Praha, 2002.
543	Lekce tvůrčího psaní	Viewegh, M.	21723	0,026179	0,020586	0,107154	0,021006	0,021349	0,115055	Viewegh, M.: Lekce tvůrčího psaní. Praha, 2005.
544	Názory na vraždu	Viewegh, M.	25505	0,005173	0,007395	0,037769	0,00821	0,011318	0,062774	Viewegh, M.: Názory na vraždu. Praha, 1990.
545	Případ nevěrné Kláry	Viewegh, M.	38518	0,006181	0,006206	0,045455	0,008515	0,009647	0,078605	Viewegh, M.: Případ nevěrné Kláry. Praha, 2003.
546	Román pro ženy	Viewegh, M.	21438	0,000451	0,001766	0,00764	0,008399	0,010534	0,062848	Viewegh, M.: Román pro ženy. Praha, 2001.
547	Účastníci zájezdu	Viewegh, M.	80564	0,018233	0,013506	0,116973	0,018771	0,015936	0,168841	Viewegh, M.: Účastníci zájezdu. Praha, 1996.
548	Vybíjená	Viewegh, M.	48054	0,007937	0,006989	0,047008	0,012231	0,012648	0,097629	Viewegh, M.: Vybíjená. Praha, 2004.
549	Výchova dívek v Čechách	Viewegh, M.	42670	0,006465	0,005522	0,037669	0,007291	0,008669	0,075309	Viewegh, M.: Výchova dívek v Čechách. Praha, 2004.
550	Zapisovatelé otcovské lásky	Viewegh, M.	35494	0,005416	0,006717	0,037465	0,011706	0,013655	0,097921	Viewegh, M.: Zapisovatelé otcovské lásky. Praha, 1998.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
551	Hordubal	Čapek, K.	32868	0,016578	0,013888	0,11596	0,029424	0,025704	0,185492	Karel Čapek on-line. Společný projekt Městské knihovny v Praze, Ústavu Českého národního korpusu FF ÚK, Společnosti bratří Čapků a Památníku Karla Čapka. http://www.mlp.cz/cz/projekty/on-line-projekty/karel-capek/
552	Krakatit	Čapek, K.	77198	0,009592	0,007692	0,09398	0,020506	0,016775	0,155131	ibid.
553	Obyčejný život	Čapek, K.	40798	0,001367	0,002356	0,01391	0,008483	0,00961	0,090468	ibid.
554	Povětroň	Čapek, K.	40722	0,002325	0,004018	0,037326	0,011187	0,012656	0,097967	ibid.
555	První parta	Čapek, K.	49492	0,015344	0,012878	0,131928	0,027814	0,022399	0,189792	ibid.
556	Továrna na absolutno	Čapek, K.	36033	0,012706	0,010852	0,084339	0,022455	0,022544	0,136406	ibid.
557	Válka s mloky	Čapek, K.	65441	0,008877	0,008823	0,080546	0,022202	0,020021	0,170884	ibid.
558	Život a dílo skladatele Foltýna	Čapek, K.	23902	0,004178	0,005059	0,037491	0,010875	0,013186	0,086452	ibid.
559	Hordubal kap. 01	Čapek, K.	1314	0,025495	0,043409	0,106762	0,025225	0,046958	0,109785	ibid.
560	Hordubal kap. 02	Čapek, K.	1130	0,001876	0,008861	0,047619	0,006912	0,017322	0,042553	ibid.
561	Hordubal kap. 03	Čapek, K.	1437	0,036433	0,040065	0,147727	0,040131	0,04654	0,123348	ibid.
562	Hordubal kap. 04	Čapek, K.	1438	0,039515	0,050032	0,15411	0,054795	0,059777	0,19346	ibid.
563	Hordubal kap. 05	Čapek, K.	652	0,047619	0,088426	0,110092	0,077411	0,100857	0,212644	ibid.
564	Hordubal kap. 06	Čapek, K.	1020	0	0,006398	0	0,016637	0,031188	0,052239	ibid.
565	Hordubal kap. 07	Čapek, K.	1354	0,029438	0,041196	0,10443	0,030935	0,047057	0,092857	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
566	Hordubal kap. 08	Čapek, K.	1137	0,032091	0,057665	0,161417	0,072734	0,080174	0,202985	ibid.
567	Hordubal kap. 09	Čapek, K.	1014	0	0,009895	0	0	0,017934	0	ibid.
568	Hordubal kap. 10	Čapek, K.	1963	0	0,008449	0	0,016263	0,033071	0,090202	ibid.
569	Hordubal kap. 11	Čapek, K.	984	0,008015	0,028602	0,056995	0,029806	0,048116	0,136364	ibid.
570	Hordubal kap. 12	Čapek, K.	763	0	0,007039	0	0	0,020168	0	ibid.
571	Hordubal kap. 13	Čapek, K.	1252	0,046735	0,036553	0,134146	0,041634	0,052426	0,174242	ibid.
572	Hordubal kap. 14	Čapek, K.	200	0,02886	0,133838	0,178571	0,176309	0,203306	0,205128	ibid.
573	Hordubal kap. 15	Čapek, K.	653	0,049231	0,037959	0,107143	0,029203	0,03283	0,077419	ibid.
574	Hordubal kap. 16	Čapek, K.	428	0	0,08371	0	0,015803	0,065567	0,083333	ibid.
575	Hordubal kap. 17	Čapek, K.	1306	0,113636	0,080879	0,265455	0,100851	0,081008	0,197772	ibid.
576	Hordubal kap. 18	Čapek, K.	357	0,008696	0,038867	0,089552	0,02029	0,056983	0,107692	ibid.
577	Hordubal kap. 19	Čapek, K.	644	0,080357	0,065417	0,203883	0,066667	0,078431	0,157895	ibid.
578	Hordubal kap. 20	Čapek, K.	776	0,080409	0,088837	0,189394	0,137321	0,125313	0,324468	ibid.
579	Hordubal kap. 21	Čapek, K.	748	0	0,019298	0	0,00731	0,039302	0,05814	ibid.
580	Hordubal kap. 22	Čapek, K.	620	0	0,015942	0	0	0,021429	0	ibid.
581	Hordubal kap. 23	Čapek, K.	1082	0,02795	0,03897	0,116667	0,063447	0,067623	0,137072	ibid.
582	Hordubal kap. 24	Čapek, K.	1147	0,055887	0,059515	0,227451	0,060425	0,070391	0,224189	ibid.
583	Hordubal kap. 25	Čapek, K.	673	0,125	0,10084	0,309091	0,057639	0,083224	0,203488	ibid.
584	Hordubal kap. 26	Čapek, K.	653	0,059203	0,056926	0,157895	0,019048	0,044812	0,065574	ibid.
585	Hordubal kap. 27	Čapek, K.	1205	0,109782	0,077199	0,218978	0,075637	0,072309	0,163814	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
586	Hordubal kap. 28	Čapek, K.	610	0,139286	0,102778	0,252336	0,076423	0,083825	0,228916	ibid.
587	Hordubal kap. 29	Čapek, K.	545	0,039072	0,062132	0,107527	0,019841	0,045255	0,084034	ibid.
588	Hordubal kap. 30	Čapek, K.	5729	0,009666	0,016528	0,035333	0,02747	0,031493	0,124402	ibid.
589	Krakatit kap. 01	Čapek, K.	1004	0,064261	0,057465	0,150235	0,053665	0,064494	0,191824	ibid.
590	Krakatit kap. 02	Čapek, K.	982	0,034252	0,043833	0,131068	0,02265	0,046053	0,103571	ibid.
591	Krakatit kap. 03	Čapek, K.	1917	0,033154	0,023883	0,072072	0,025949	0,031999	0,057554	ibid.
592	Krakatit kap. 04	Čapek, K.	1456	0,010766	0,021739	0,056426	0,029273	0,030239	0,091549	ibid.
593	Krakatit kap. 05	Čapek, K.	1117	0,034674	0,034789	0,131687	0,024033	0,05217	0,140805	ibid.
594	Krakatit kap. 06	Čapek, K.	1331	0,050216	0,04116	0,127517	0,055314	0,051674	0,127072	ibid.
595	Krakatit kap. 07	Čapek, K.	1281	0,038596	0,024854	0,080569	0,029605	0,028981	0,074205	ibid.
596	Krakatit kap. 08	Čapek, K.	1207	0,016844	0,022577	0,060241	0,014104	0,025189	0,050704	ibid.
597	Krakatit kap. 09	Čapek, K.	969	0,007656	0,01832	0,064103	0,006433	0,027978	0,056995	ibid.
598	Krakatit kap. 10	Čapek, K.	1134	0,057285	0,046255	0,193798	0,059888	0,055532	0,177177	ibid.
599	Krakatit kap. 11	Čapek, K.	2103	0,01861	0,012237	0,057692	0,016704	0,017825	0,06891	ibid.
600	Krakatit kap. 12	Čapek, K.	2336	0,034216	0,020882	0,108856	0,030132	0,024514	0,09831	ibid.
601	Krakatit kap. 13	Čapek, K.	1506	0,007656	0,011848	0,056537	0,010167	0,019808	0,040964	ibid.
602	Krakatit kap. 14	Čapek, K.	1167	0,042683	0,045911	0,152439	0,021964	0,038237	0,093525	ibid.
603	Krakatit kap. 15	Čapek, K.	1224	0,032003	0,028438	0,131068	0,028505	0,038811	0,107143	ibid.
604	Krakatit kap. 16	Čapek, K.	1204	0,036188	0,039082	0,125	0,062925	0,061705	0,165493	ibid.
605	Krakatit kap. 17	Čapek, K.	1385	0,118788	0,099601	0,233677	0,084513	0,083281	0,191748	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
606	Krakatit kap. 18	Čapek, K.	1409	0,07161	0,074632	0,203804	0,053535	0,068223	0,171674	ibid.
607	Krakatit kap. 19	Čapek, K.	1005	0,098084	0,117151	0,230769	0,093781	0,096376	0,218391	ibid.
608	Krakatit kap. 20	Čapek, K.	1328	0,10444	0,089902	0,179245	0,089275	0,077191	0,20339	ibid.
609	Krakatit kap. 21	Čapek, K.	1073	0,153086	0,108699	0,212291	0,136191	0,135417	0,301961	ibid.
610	Krakatit kap. 22	Čapek, K.	1168	0,033944	0,026511	0,076142	0,017778	0,03883	0,06338	ibid.
611	Krakatit kap. 23	Čapek, K.	1474	0,086727	0,065362	0,209836	0,122495	0,089461	0,223005	ibid.
612	Krakatit kap. 24	Čapek, K.	2385	0,059722	0,044829	0,147638	0,064499	0,057329	0,189231	ibid.
613	Krakatit kap. 25	Čapek, K.	1309	0,013333	0,013684	0,057851	0,004879	0,013449	0,045732	ibid.
614	Krakatit kap. 26	Čapek, K.	1246	0,086436	0,06442	0,198473	0,116463	0,086875	0,228659	ibid.
615	Krakatit kap. 27	Čapek, K.	1764	0,04222	0,034054	0,088571	0,061234	0,05042	0,157113	ibid.
616	Krakatit kap. 28	Čapek, K.	1637	0,062879	0,045063	0,158845	0,072837	0,060212	0,179177	ibid.
617	Krakatit kap. 29	Čapek, K.	1365	0,016747	0,023265	0,053942	0,026465	0,034807	0,1	ibid.
618	Krakatit kap. 30	Čapek, K.	1491	0,024649	0,025281	0,105072	0,042174	0,044078	0,109694	ibid.
619	Krakatit kap. 31	Čapek, K.	1849	0,021368	0,029872	0,06006	0,021653	0,034035	0,076493	ibid.
620	Krakatit kap. 32	Čapek, K.	1269	0,045295	0,053619	0,163424	0,0279	0,053515	0,110749	ibid.
621	Krakatit kap. 33	Čapek, K.	1433	0	0,011785	0	0,030522	0,030819	0,079681	ibid.
622	Krakatit kap. 34	Čapek, K.	1671	0,035609	0,028594	0,09697	0,037616	0,044007	0,093607	ibid.
623	Krakatit kap. 35	Čapek, K.	1410	0	0	0	0	0,012549	0	ibid.
624	Krakatit kap. 36	Čapek, K.	1287	0,038788	0,035652	0,111554	0,026242	0,031298	0,090652	ibid.
625	Krakatit kap. 37	Čapek, K.	1117	0,101803	0,077846	0,218605	0,068603	0,068511	0,191176	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
626	Krakatit kap. 38	Čapek, K.	1308	0,041206	0,043281	0,128099	0,034783	0,050747	0,102102	ibid.
627	Krakatit kap. 39	Čapek, K.	1509	0,0329	0,021736	0,077869	0,021645	0,027056	0,0625	ibid.
628	Krakatit kap. 40	Čapek, K.	1386	0,047506	0,037456	0,12013	0,028833	0,036184	0,100257	ibid.
629	Krakatit kap. 41	Čapek, K.	1505	0,012614	0,014267	0,051205	0,022759	0,029637	0,086651	ibid.
630	Krakatit kap. 42	Čapek, K.	1554	0,019854	0,016117	0,060317	0,021662	0,02794	0,053165	ibid.
631	Krakatit kap. 43	Čapek, K.	1605	0	0,010925	0	0	0,020574	0	ibid.
632	Krakatit kap. 44	Čapek, K.	1566	0	0	0	0	0,001648	0	ibid.
633	Krakatit kap. 45	Čapek, K.	1729	0	0	0	0,002875	0,018411	0,029762	ibid.
634	Krakatit kap. 46	Čapek, K.	1510	0,015805	0,027028	0,064516	0,025679	0,030847	0,096203	ibid.
635	Krakatit kap. 47	Čapek, K.	1036	0	0	0	0,003809	0,012563	0,044369	ibid.
636	Krakatit kap. 48	Čapek, K.	1343	0	0,010965	0	0	0,028846	0	ibid.
637	Krakatit kap. 49	Čapek, K.	1361	0,020144	0,032262	0,140845	0,02931	0,037156	0,124031	ibid.
638	Krakatit kap. 50	Čapek, K.	1442	0,049365	0,039066	0,125413	0,035273	0,032748	0,094431	ibid.
639	Krakatit kap. 51	Čapek, K.	1696	0	0,007093	0	0	0,016829	0	ibid.
640	Krakatit kap. 52	Čapek, K.	1641	0,028571	0,025673	0,076389	0,02674	0,038571	0,091127	ibid.
641	Krakatit kap. 53	Čapek, K.	1276	0,00601	0,010406	0,043137	0,01016	0,026993	0,071839	ibid.
642	Krakatit kap. 54	Čapek, K.	1718	0,031333	0,025926	0,109694	0,038145	0,037177	0,122677	ibid.
643	Obyčejný život kap. 01	Čapek, K.	575	0,026786	0,03875	0,101124	0,006579	0,031476	0,055215	ibid.
644	Obyčejný život kap. 02	Čapek, K.	1260	0	0	0	0	0,002997	0	ibid.
645	Obyčejný život kap. 03	Čapek, K.	683	0	0,00963	0	0,01358	0,028686	0,061111	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
646	Obyčejný život kap. 04	Čapek, K.	1271	0	0,006318	0	0,01195	0,018167	0,038356	ibid.
647	Obyčejný život kap. 05	Čapek, K.	1372	0	0,007359	0	0,002233	0,01821	0,031401	ibid.
648	Obyčejný život kap. 06	Čapek, K.	1725	0	0,000124	0	0	0,004977	0	ibid.
649	Obyčejný život kap. 07	Čapek, K.	990	0	0,01438	0	0,009701	0,018565	0,075988	ibid.
650	Obyčejný život kap. 08	Čapek, K.	935	0	0,005343	0	0	0,018391	0	ibid.
651	Obyčejný život kap. 09	Čapek, K.	1424	0	0	0	0	0,011785	0	ibid.
652	Obyčejný život kap. 10	Čapek, K.	1124	0	0	0	0	0,004705	0	ibid.
653	Obyčejný život kap. 11	Čapek, K.	798	0	0	0	0	0,00915	0	ibid.
654	Obyčejný život kap. 12	Čapek, K.	875	0	0,00185	0	0	0,006653	0	ibid.
655	Obyčejný život kap. 13	Čapek, K.	674	0	0	0	0	0,008333	0	ibid.
656	Obyčejný život kap. 14	Čapek, K.	1093	0	0,003736	0	0,003226	0,012009	0,039286	ibid.
657	Obyčejný život kap. 15	Čapek, K.	2027	0	0,004943	0	0,001974	0,012214	0,026357	ibid.
658	Obyčejný život kap. 16	Čapek, K.	978	0	0	0	0	0,003444	0	ibid.
659	Obyčejný život kap. 17	Čapek, K.	1114	0,007215	0,012077	0,058366	0,007876	0,018994	0,042328	ibid.
660	Obyčejný život kap. 18	Čapek, K.	706	0	0,004983	0	0	0,014035	0	ibid.
661	Obyčejný život kap. 19	Čapek, K.	872	0	0,001339	0	0	0,002924	0	ibid.
662	Obyčejný život kap. 20	Čapek, K.	270	0	0	0	0	0,035888	0	ibid.
663	Obyčejný život kap. 21	Čapek, K.	1038	0	0	0	0	0,008403	0	ibid.
664	Obyčejný život kap. 22	Čapek, K.	2542	0	0,000787	0	0,007588	0,010223	0,046796	ibid.
665	Obyčejný život kap. 23	Čapek, K.	2123	0	0,001617	0	0,015287	0,011153	0,039506	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
666	Obyčejný život kap. 24	Čapek, K.	1205	0	0,012054	0	0	0,01326	0	ibid.
667	Obyčejný život kap. 25	Čapek, K.	1312	0	0,004444	0	0,025532	0,031395	0,103792	ibid.
668	Obyčejný život kap. 26	Čapek, K.	780	0	0	0	0	0,004094	0	ibid.
669	Obyčejný život kap. 27	Čapek, K.	1128	0	0,001396	0	0,015251	0,016805	0,043702	ibid.
670	Obyčejný život kap. 28	Čapek, K.	647	0	0	0	0	0,002674	0	ibid.
671	Obyčejný život kap. 29	Čapek, K.	1124	0	0	0	0	0,005107	0	ibid.
672	Obyčejný život kap. 30	Čapek, K.	1283	0	0,002517	0	0	0,012299	0	ibid.
673	Obyčejný život kap. 31	Čapek, K.	1099	0	0	0	0	0,00366	0	ibid.
674	Obyčejný život kap. 32	Čapek, K.	1706	0	0,005173	0	0,008779	0,01182	0,0553	ibid.
675	Obyčejný život kap. 33	Čapek, K.	1491	0	0,003432	0	0,008849	0,012005	0,032154	ibid.
676	Obyčejný život kap. 34	Čapek, K.	1126	0	0,00162	0	0,002162	0,009245	0,030691	ibid.
677	Obyčejný život kap. 35	Čapek, K.	1394	0	0,000272	0	0,004993	0,010936	0,034351	ibid.
678	Povětroň kap. 01	Čapek, K.	607	0	0,007123	0	0,02619	0,054444	0,083333	ibid.
679	Povětroň kap. 02	Čapek, K.	1132	0,032705	0,036453	0,111111	0,010317	0,032548	0,083815	ibid.
680	Povětroň kap. 03	Čapek, K.	818	0,05557	0,074514	0,159509	0,041055	0,055442	0,117117	ibid.
681	Povětroň kap. 04	Čapek, K.	983	0,011258	0,0313	0,052083	0	0,021733	0	ibid.
682	Povětroň kap. 05	Čapek, K.	866	0	0	0	0,013258	0,033473	0,094697	ibid.
683	Povětroň kap. 06	Čapek, K.	489	0	0,011842	0	0,006351	0,030514	0,068027	ibid.
684	Povětroň kap. 07	Čapek, K.	892	0	0,000478	0	0	0,006375	0	ibid.
685	Povětroň kap. 08	Čapek, K.	1373	0	0,000524	0	0	0,006646	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
686	Povětroň kap. 09	Čapek, K.	1093	0	0,000639	0	0	0,008487	0	ibid.
687	Povětroň kap. 10	Čapek, K.	1408	0	0,003865	0	0,013538	0,020614	0,071553	ibid.
688	Povětroň kap. 11	Čapek, K.	1153	0,059434	0,054195	0,147619	0,037983	0,050333	0,111111	ibid.
689	Povětroň kap. 12	Čapek, K.	1117	0,039133	0,042807	0,129353	0,025932	0,04212	0,097179	ibid.
690	Povětroň kap. 13	Čapek, K.	1073	0	0,011891	0	0,014205	0,032285	0,051903	ibid.
691	Povětroň kap. 14	Čapek, K.	1009	0	0,004848	0	0,005029	0,016906	0,045113	ibid.
692	Povětroň kap. 15	Čapek, K.	1024	0	0,001032	0	0,014354	0,024037	0,055556	ibid.
693	Povětroň kap. 16	Čapek, K.	1084	0,006015	0,011429	0,051887	0	0,015112	0	ibid.
694	Povětroň kap. 17	Čapek, K.	908	0	0,007636	0	0	0,00984	0	ibid.
695	Povětroň kap. 18	Čapek, K.	1119	0	0	0	0	0,006972	0	ibid.
696	Povětroň kap. 19	Čapek, K.	1149	0	0,00401	0	0	0,004109	0	ibid.
697	Povětroň kap. 20	Čapek, K.	829	0	0	0	0,022695	0,019149	0,063158	ibid.
698	Povětroň kap. 21	Čapek, K.	983	0	0	0	0	0,000278	0	ibid.
699	Povětroň kap. 22	Čapek, K.	1070	0	0,005815	0	0	0,006168	0	ibid.
700	Povětroň kap. 23	Čapek, K.	855	0	0,014312	0	0	0,022034	0	ibid.
701	Povětroň kap. 24	Čapek, K.	1023	0,002984	0,014438	0,060606	0,022364	0,039057	0,092527	ibid.
702	Povětroň kap. 25	Čapek, K.	1434	0	0,006231	0	0,00206	0,01703	0,036585	ibid.
703	Povětroň kap. 26	Čapek, K.	1020	0,014685	0,018069	0,0625	0,026449	0,031814	0,097561	ibid.
704	Povětroň kap. 27	Čapek, K.	1161	0	0,009936	0	0	0,009913	0	ibid.
705	Povětroň kap. 28	Čapek, K.	883	0	0	0	0	0,005821	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
706	Povětroň kap. 29	Čapek, K.	1435	0,014738	0,020104	0,083942	0,001377	0,01614	0,032432	ibid.
707	Povětroň kap. 30	Čapek, K.	1001	0	0	0	0	0,015338	0	ibid.
708	Povětroň kap. 31	Čapek, K.	1008	0	0,015774	0	0	0,013866	0	ibid.
709	Povětroň kap. 32	Čapek, K.	1512	0	0,006356	0	0,001122	0,017184	0,03112	ibid.
710	Povětroň kap. 33	Čapek, K.	998	0,090226	0,088492	0,231707	0,043956	0,076822	0,163866	ibid.
711	Povětroň kap. 34	Čapek, K.	890	0	0	0	0	0,005339	0	ibid.
712	Povětroň kap. 35	Čapek, K.	1876	0,028623	0,019821	0,06366	0,019808	0,023077	0,079767	ibid.
713	Povětroň kap. 36	Čapek, K.	1131	0	0,009831	0	0,001887	0,017724	0,039146	ibid.
714	Povětroň kap. 37	Čapek, K.	1123	0	0	0	0	0,005257	0	ibid.
715	Povětroň kap. 38	Čapek, K.	1086	0	0	0	0,004243	0,021036	0,038217	ibid.
716	Povětroň kap. 39	Čapek, K.	68	0	0,166667	0	0	0	0	ibid.
717	První parta kap. 01	Čapek, K.	1625	0	0	0	0	0,002503	0	ibid.
718	První parta kap. 02	Čapek, K.	1206	0,057781	0,04618	0,096899	0,059322	0,052813	0,099698	ibid.
719	První parta kap. 03	Čapek, K.	2400	0,018023	0,016176	0,05082	0,026834	0,027596	0,093855	ibid.
720	První parta kap. 04	Čapek, K.	1368	0,064789	0,047802	0,129909	0,073372	0,054084	0,122271	ibid.
721	První parta kap. 05	Čapek, K.	1202	0	0,011329	0	0,034889	0,031072	0,092697	ibid.
722	První parta kap. 06	Čapek, K.	1801	0,002514	0,018993	0,038147	0,007705	0,026527	0,035225	ibid.
723	První parta kap. 07	Čapek, K.	2484	0,016267	0,023953	0,083789	0,026156	0,034642	0,062958	ibid.
724	První parta kap. 08	Čapek, K.	1020	0,0487	0,040627	0,086538	0,048951	0,053988	0,116788	ibid.
725	První parta kap. 09	Čapek, K.	1189	0,02099	0,039117	0,097122	0,020097	0,03632	0,113757	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
726	První parta kap. 10	Čapek, K.	1734	0,052703	0,045433	0,153285	0,063344	0,050962	0,123616	ibid.
727	První parta kap. 11	Čapek, K.	2549	0,020268	0,027029	0,073846	0,031485	0,044354	0,129171	ibid.
728	První parta kap. 12	Čapek, K.	1496	0,040758	0,04882	0,118694	0,05344	0,073681	0,133462	ibid.
729	První parta kap. 13	Čapek, K.	1395	0,050769	0,045323	0,07309	0,027287	0,034017	0,086283	ibid.
730	První parta kap. 14	Čapek, K.	1722	0,009644	0,0323	0,076555	0,027191	0,034638	0,101818	ibid.
731	První parta kap. 15	Čapek, K.	2112	0,027153	0,04249	0,121331	0,059654	0,066457	0,195499	ibid.
732	První parta kap. 16	Čapek, K.	1090	0,136141	0,089608	0,19685	0,120868	0,087779	0,211538	ibid.
733	První parta kap. 17	Čapek, K.	1540	0,054311	0,04247	0,12973	0,06183	0,054089	0,146694	ibid.
734	První parta kap. 18	Čapek, K.	2547	0,036319	0,034541	0,136729	0,030763	0,035983	0,123984	ibid.
735	První parta kap. 19	Čapek, K.	1899	0,004451	0,028699	0,035714	0,023073	0,038225	0,099857	ibid.
736	První parta kap. 20	Čapek, K.	1513	0,013222	0,018932	0,088146	0,040781	0,038139	0,135198	ibid.
737	První parta kap. 21	Čapek, K.	2159	0,022519	0,018059	0,057199	0,0284	0,028097	0,081081	ibid.
738	První parta kap. 22	Čapek, K.	1721	0,05135	0,035132	0,087071	0,062618	0,051271	0,112186	ibid.
739	První parta kap. 23	Čapek, K.	639	0,069595	0,060976	0,136752	0,090909	0,090909	0,134615	ibid.
740	První parta kap. 24	Čapek, K.	2727	0,045233	0,026808	0,08567	0,043791	0,036502	0,075281	ibid.
741	První parta kap. 25	Čapek, K.	1003	0,083402	0,076381	0,234742	0,070501	0,076243	0,159696	ibid.
742	První parta kap. 26	Čapek, K.	2311	0,026859	0,031695	0,107595	0,016576	0,028733	0,066169	ibid.
743	První parta kap. 27	Čapek, K.	2292	0,033763	0,03338	0,094444	0,050068	0,047747	0,125714	ibid.
744	První parta kap. 28	Čapek, K.	2720	0,018201	0,021635	0,041892	0,022091	0,028238	0,10942	ibid.
745	Válka s mloky kap. 01	Čapek, K.	3672	0,059143	0,046915	0,145877	0,05945	0,055779	0,145174	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
746	Válka s mloky kap. 02	Čapek, K.	1874	0,04045	0,046846	0,104019	0,048261	0,051894	0,131498	ibid.
747	Válka s mloky kap. 03	Čapek, K.	3262	0,073393	0,056238	0,166667	0,0536	0,037409	0,137681	ibid.
748	Válka s mloky kap. 04	Čapek, K.	1985	0,019613	0,033755	0,11804	0,014547	0,027738	0,090237	ibid.
749	Válka s mloky kap. 05	Čapek, K.	1647	0,009127	0,044579	0,113801	0,035205	0,046288	0,154905	ibid.
750	Válka s mloky kap. 06	Čapek, K.	3531	0,095222	0,067497	0,186691	0,073342	0,064785	0,188244	ibid.
751	Válka s mloky kap. 07	Čapek, K.	2582	0,069579	0,061247	0,204819	0,068437	0,075952	0,195889	ibid.
752	Válka s mloky kap. 08	Čapek, K.	1418	0	0,012468	0	0,062424	0,060109	0,129568	ibid.
753	Válka s mloky kap. 09	Čapek, K.	1633	0,248492	0,229969	0,527108	0,222639	0,188845	0,407895	ibid.
754	Válka s mloky kap. 10	Čapek, K.	1158	0,148487	0,115482	0,319672	0,134011	0,13066	0,305136	ibid.
755	Válka s mloky kap. 11	Čapek, K.	1518	0	0,024211	0	0,02331	0,039021	0,054795	ibid.
756	Válka s mloky kap. 12	Čapek, K.	4965	0,019414	0,026643	0,108434	0,031839	0,033094	0,11217	ibid.
757	Válka s mloky kap. 13	Čapek, K.	1153	0,118457	0,078119	0,180328	0,05688	0,050695	0,128205	ibid.
758	Válka s mloky kap. 14	Čapek, K.	16405	0,016349	0,012056	0,076205	0,031269	0,028182	0,120521	ibid.
759	Válka s mloky kap. 15	Čapek, K.	941	0,053872	0,077671	0,103896	0,026715	0,047622	0,107692	ibid.
760	Válka s mloky kap. 16	Čapek, K.	1658	0,005243	0,023019	0,042904	0,078054	0,069748	0,17907	ibid.
761	Válka s mloky kap. 17	Čapek, K.	1250	0,017628	0,029726	0,06044	0,059829	0,061589	0,108949	ibid.
762	Válka s mloky kap. 18	Čapek, K.	1037	0,013393	0,039236	0,060811	0,058191	0,124434	0,255144	ibid.
763	Válka s mloky kap. 19	Čapek, K.	1271	0	0	0	0,050975	0,058232	0,127341	ibid.
764	Válka s mloky kap. 20	Čapek, K.	1450	0	0,005551	0	0,017706	0,036814	0,08	ibid.
765	Válka s mloky kap. 21	Čapek, K.	2019	0	0,00633	0	0,038202	0,037402	0,097889	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
766	Válka s mloky kap. 22	Čapek, K.	1245	0,033788	0,032099	0,078788	0,023925	0,04462	0,118182	ibid.
767	Válka s mloky kap. 23	Čapek, K.	1237	0,151521	0,131751	0,22093	0,087631	0,097987	0,203187	ibid.
768	Válka s mloky kap. 24	Čapek, K.	2592	0	0,009056	0	0,029893	0,037783	0,092532	ibid.
769	Válka s mloky kap. 25	Čapek, K.	2140	0,060308	0,050898	0,147287	0,038404	0,046788	0,17601	ibid.
770	Válka s mloky kap. 26	Čapek, K.	1798	0,006677	0,014008	0,043103	0,062344	0,053674	0,10991	ibid.
771	Boží muka: Šlépěj	Čapek, K.	1610	0,010751	0,019134	0,051829	0,029354	0,04062	0,120155	ibid.
772	Boží muka: Lída	Čapek, K.	1926	0	0,006777	0	0,009313	0,017465	0,040568	ibid.
773	Boží muka: Lída II	Čapek, K.	2614	0,013528	0,022544	0,043029	0,012402	0,021556	0,042627	ibid.
774	Boží muka: Hora	Čapek, K.	5346	0,035309	0,026581	0,115652	0,028037	0,02895	0,096945	ibid.
775	Boží muka: Milostná píseň	Čapek, K.	4222	0,031938	0,018997	0,084337	0,040634	0,028984	0,104818	ibid.
776	Boží muka: Elegie	Čapek, K.	2922	0,024904	0,021172	0,054286	0,011443	0,01688	0,037475	ibid.
777	Boží muka: Utkvění času	Čapek, K.	460	0	0	0	0	0,02388	0	ibid.
778	Boží muka: Historie beze slov	Čapek, K.	881	0,026239	0,030442	0,083333	0,026531	0,03507	0,060465	ibid.
779	Boží muka: Ztracená cesta	Čapek, K.	1322	0	0,002273	0	0,021433	0,023165	0,077093	ibid.
780	Boží muka: Nápis	Čapek, K.	1053	0,023041	0,03125	0,071749	0,02489	0,032841	0,057348	ibid.
781	Boží muka: Pokušení	Čapek, K.	985	0	0,017253	0	0,046612	0,056238	0,133333	ibid.
782	Boží muka: Odrazy	Čapek, K.	934	0,016667	0,017105	0,062176	0,01341	0,030034	0,101626	ibid.
783	Boží muka: Čekárna	Čapek, K.	1407	0,013774	0,030687	0,053381	0,053975	0,054019	0,128019	ibid.
784	Boží muka: Pomoc	Čapek, K.	1211	0,08319	0,064602	0,195312	0,069492	0,054888	0,185629	ibid.
785	Povídky z jedné kapsy: Případ Dr. Mejzlíka	Čapek, K.	1199	0,096653	0,08658	0,239726	0,062651	0,070987	0,23753	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
786	Povídky z jedné kapsy: Modrá chryzantéma	Čapek, K.	1742	0,010527	0,036069	0,084577	0,044188	0,048327	0,162029	ibid.
787	Povídky z jedné kapsy: Věštkyně	Čapek, K.	1434	0,329562	0,2072	0,420245	0,208715	0,159056	0,339545	ibid.
788	Povídky z jedné kapsy: Jasnovidec	Čapek, K.	1945	0,083601	0,068239	0,210526	0,062245	0,056196	0,184679	ibid.
789	Povídky z jedné kapsy: Tajemství písma	Čapek, K.	1928	0,031515	0,032818	0,096296	0,024219	0,029817	0,135135	ibid.
790	Povídky z jedné kapsy: Naprostý důkaz	Čapek, K.	1472	0	0,011723	0	0,006452	0,025188	0,033645	ibid.
791	Povídky z jedné kapsy: Experiment profesora Rousse	Čapek, K.	1645	0,025397	0,070981	0,051724	0,022142	0,056365	0,093366	ibid.
792	Povídky z jedné kapsy: Ztracený dopis	Čapek, K.	1856	0,062697	0,079607	0,153488	0,052195	0,058674	0,190713	ibid.
793	Povídky z jedné kapsy: Ukradený spis 139/VII, odd. C	Čapek, K.	2158	0,072209	0,071667	0,197719	0,060135	0,05795	0,203085	ibid.
794	Povídky z jedné kapsy: Muž, který se nelíbil	Čapek, K.	1457	0,164415	0,133438	0,290398	0,125912	0,108712	0,245439	ibid.
795	Povídky z jedné kapsy: Básník	Čapek, K.	1540	0,181953	0,132021	0,295858	0,099101	0,086181	0,237452	ibid.
796	Povídky z jedné kapsy: Případ pana Janíka	Čapek, K.	2532	0,130474	0,088294	0,223787	0,136459	0,099353	0,215464	ibid.
797	Povídky z jedné kapsy: Pád rodu Votických	Čapek, K.	2443	0,109237	0,109113	0,261364	0,083599	0,074847	0,23399	ibid.
798	Povídky z jedné kapsy: Rekord	Čapek, K.	2148	0,051096	0,069424	0,144981	0,044916	0,049585	0,184146	ibid.
799	Povídky z jedné kapsy: Případ Selvinův	Čapek, K.	2189	0,003759	0,026981	0,033557	0,010046	0,019761	0,084848	ibid.
800	Povídky z jedné kapsy: Šlépěje	Čapek, K.	2335	0,08653	0,076861	0,190217	0,066918	0,071392	0,189889	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
801	Povídky z jedné kapsy: Kupón	Čapek, K.	2559	0,013386	0,028616	0,071313	0,018351	0,028604	0,095187	ibid.
802	Povídky z jedné kapsy: Oplatkův konec	Čapek, K.	2347	0,0008	0,007708	0,031746	0,0185	0,036404	0,089655	ibid.
803	Povídky z jedné kapsy: Poslední soud	Čapek, K.	1261	0,072933	0,06735	0,159624	0,09211	0,096265	0,186391	ibid.
804	Povídky z jedné kapsy: Zločin v chalupě	Čapek, K.	1239	0,031937	0,040412	0,112281	0,033333	0,041887	0,122449	ibid.
805	Povídky z jedné kapsy: Zmizení herce Bendy	Čapek, K.	3245	0,098358	0,072169	0,20098	0,069249	0,060275	0,202277	ibid.
806	Povídky z jedné kapsy: Vražedný útok	Čapek, K.	1651	0,027199	0,044496	0,087558	0,043576	0,045079	0,1472	ibid.
807	Povídky z jedné kapsy: Propuštěný	Čapek, K.	1581	0,05867	0,035254	0,093168	0,057068	0,047889	0,147793	ibid.
808	Povídky z jedné kapsy: Zločin na poště	Čapek, K.	2215	0,04101	0,041794	0,09329	0,034687	0,042131	0,148352	ibid.
809	Povídky z druhé kapsy: Ukradený kaktus	Čapek, K.	1937	0	0,007618	0	0,007729	0,02159	0,060842	ibid.
810	Povídky z druhé kapsy: Povídka starého kriminálníka	Čapek, K.	1699	0	0,006185	0	0,003867	0,012721	0,0299	ibid.
811	Povídky z druhé kapsy: Zmizení pana Hirsche	Čapek, K.	1961	0,049167	0,069617	0,176238	0,073581	0,061796	0,205263	ibid.
812	Povídky z druhé kapsy: Čintamani a ptáci	Čapek, K.	3163	0,007653	0,017002	0,056416	0,0104	0,019082	0,067425	ibid.
813	Povídky z druhé kapsy: Příběh o kasaři a žháři	Čapek, K.	2157	0,005967	0,018621	0,034799	0,006351	0,016055	0,072127	ibid.
814	Povídky z druhé kapsy: Ukradená vražda	Čapek, K.	2218	0,004722	0,015605	0,033932	0,007173	0,022878	0,076623	ibid.
815	Povídky z druhé kapsy: Případ s dítětem	Čapek, K.	2416	0,070454	0,063936	0,22089	0,046937	0,045741	0,207188	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
816	Povídky z druhé kapsy: Grófinka	Čapek, K.	1648	0	0,007523	0	0,007592	0,0152	0,034991	ibid.
817	Povídky z druhé kapsy: Historie dirigenta Kaliny	Čapek, K.	1690	0	0,001658	0	0	0,006384	0	ibid.
818	Povídky z druhé kapsy: Smrt barona Gandary	Čapek, K.	1476	0,039393	0,05835	0,159236	0,018098	0,03629	0,078	ibid.
819	Povídky z druhé kapsy: Příběh sňatkového podvodníka	Čapek, K.	2187	0,046213	0,045234	0,143098	0,045876	0,051048	0,147305	ibid.
820	Povídky z druhé kapsy: Balada o Juraji Čupovi	Čapek, K.	1579	0	0,033356	0	0,036147	0,049033	0,117895	ibid.
821	Povídky z druhé kapsy: Povídka o ztracené noze	Čapek, K.	1588	0,005442	0,02067	0,04451	0,017857	0,022745	0,077358	ibid.
822	Povídky z druhé kapsy: Závrať	Čapek, K.	1585	0,083317	0,063677	0,189911	0,069055	0,067728	0,147793	ibid.
823	Povídky z druhé kapsy: Ušní zpověď	Čapek, K.	1276	0	0,015328	0	0,015583	0,031251	0,073991	ibid.
824	Povídky z druhé kapsy: O lyrickém zloději	Čapek, K.	1694	0	0,015256	0	0,01083	0,019683	0,065056	ibid.
825	Povídky z druhé kapsy: Soud pana Havleny	Čapek, K.	1772	0,030085	0,035547	0,062147	0,037287	0,045305	0,160584	ibid.
826	Povídky z druhé kapsy: Jehla	Čapek, K.	1432	0	0,005968	0	0,023065	0,023693	0,088795	ibid.
827	Povídky z druhé kapsy: Telegram	Čapek, K.	1534	0,094644	0,089495	0,1875	0,06079	0,074621	0,202381	ibid.
828	Povídky z druhé kapsy: Muž, který nemohl spát	Čapek, K.	1646	0	0,006033	0	0,012954	0,028278	0,085515	ibid.
829	Povídky z druhé kapsy: Sbírka známek	Čapek, K.	1904	0	0,001404	0	0,005879	0,014441	0,032951	ibid.
830	Povídky z druhé kapsy: Obyčejná vražda	Čapek, K.	1449	0	0,001319	0	0,005302	0,020109	0,06422	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
831	Povídky z druhé kapsy: Porotce	Čapek, K.	1766	0	0	0	0,006676	0,017277	0,061931	ibid.
832	Povídky z druhé kapsy: Poslední věci člověka	Čapek, K.	1149	0,019724	0,046746	0,049834	0,028887	0,043525	0,114865	ibid.
833	Trapné povídky: Otcové	Čapek, K.	1650	0	0,015878	0	0,009123	0,025791	0,045714	ibid.
834	Trapné povídky: Tři	Čapek, K.	2083	0,065208	0,042188	0,128834	0,047696	0,046402	0,129985	ibid.
835	Trapné povídky: Helena	Čapek, K.	3289	0,015899	0,013985	0,067143	0,021978	0,020781	0,048314	ibid.
836	Trapné povídky: Na zámku	Čapek, K.	5653	0,041012	0,033242	0,153967	0,032133	0,032127	0,114923	ibid.
837	Trapné povídky: Peníze	Čapek, K.	4606	0,025554	0,017321	0,071685	0,024781	0,025744	0,095009	ibid.
838	Trapné povídky: Surovec	Čapek, K.	4328	0,043948	0,029872	0,097292	0,034768	0,0311	0,113686	ibid.
839	Trapné povídky: Košile	Čapek, K.	2328	0,011478	0,012372	0,039698	0,011293	0,02027	0,033921	ibid.
840	Trapné povídky: Uražený	Čapek, K.	3859	0,01303	0,012292	0,060181	0,023437	0,031256	0,10879	ibid.
841	Trapné povídky: Tribunál	Čapek, K.	828	0,004973	0,042424	0,065041	0,010417	0,059492	0,060109	ibid.
842	Anglické listy 01	Čapek, K.	657	0	0	0	0	0,008537	0	ibid.
843	Anglické listy 02	Čapek, K.	483	0	0,030075	0	0,030537	0,06747	0,168067	ibid.
844	Anglické listy 03	Čapek, K.	730	0,039062	0,036765	0,106383	0,081544	0,063351	0,115183	ibid.
845	Anglické listy 04	Čapek, K.	720	0	0	0	0,003003	0,020768	0,045249	ibid.
846	Anglické listy 05	Čapek, K.	821	0	0	0	0	0,010136	0	ibid.
847	Anglické listy 06	Čapek, K.	482	0	0,006978	0	0,019875	0,040486	0,081081	ibid.
848	Anglické listy 07	Čapek, K.	549	0	0,013736	0	0,02188	0,029841	0,074074	ibid.
849	Anglické listy 08	Čapek, K.	494	0,010878	0,054843	0,08642	0,012613	0,034398	0,074468	ibid.
850	Anglické listy 09	Čapek, K.	669	0,021978	0,020408	0,079646	0,003361	0,030725	0,052941	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
851	Anglické listy 10	Čapek, K.	608	0	0	0	0	0,012963	0	ibid.
852	Anglické listy 11	Čapek, K.	623	0	0,004931	0	0	0,010802	0	ibid.
853	Anglické listy 12	Čapek, K.	599	0	0,009105	0	0,004689	0,032789	0,063492	ibid.
854	Anglické listy 13	Čapek, K.	658	0	0	0	0	0,013763	0	ibid.
855	Anglické listy 14	Čapek, K.	505	0	0,006063	0	0,019098	0,04647	0,085714	ibid.
856	Anglické listy 15	Čapek, K.	672	0	0,004396	0	0,042063	0,050926	0,14	ibid.
857	Anglické listy 16	Čapek, K.	621	0	0,009419	0	0	0,01439	0	ibid.
858	Anglické listy 17	Čapek, K.	627	0	0	0	0	0,014985	0	ibid.
859	Anglické listy 18	Čapek, K.	569	0	0,000723	0	0,026991	0,034712	0,096154	ibid.
860	Anglické listy 19	Čapek, K.	538	0,031786	0,033217	0,105263	0,021645	0,038337	0,104167	ibid.
861	Anglické listy 20	Čapek, K.	558	0	0,004579	0	0,005952	0,021875	0,0625	ibid.
862	Anglické listy 21	Čapek, K.	568	0	0,010382	0	0,010453	0,039668	0,07563	ibid.
863	Anglické listy 22	Čapek, K.	489	0,024607	0,037061	0,109756	0,025974	0,041958	0,092784	ibid.
864	Anglické listy 23	Čapek, K.	125	0	0,035714	0	0,1875	0,147321	0,3	ibid.
865	Anglické listy 24	Čapek, K.	68	0,75	0,589286	0,75	0,75	0,589286	0,75	ibid.
866	Anglické listy 25	Čapek, K.	109	0,333333	0,285714	0,473684	0,416667	0,315476	0,352941	ibid.
867	Anglické listy 26	Čapek, K.	45	0	0,166667	0	0	0	0	ibid.
868	Anglické listy 27	Čapek, K.	426	0	0	0	0	0,012121	0	ibid.
869	Anglické listy 28	Čapek, K.	573	0	0,005068	0	0	0,01929	0	ibid.
870	Anglické listy 29	Čapek, K.	619	0	0,00553	0	0	0,028571	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
871	Anglické listy 30	Čapek, K.	500	0	0	0	0	0,0301	0	ibid.
872	Anglické listy 31	Čapek, K.	667	0	0,003583	0	0,003069	0,016304	0,061224	ibid.
873	Anglické listy 32	Čapek, K.	406	0	0,018519	0	0,033333	0,071023	0,112676	ibid.
874	Anglické listy 33	Čapek, K.	347	0	0,005682	0	0,02381	0,037546	0,095745	ibid.
875	Anglické listy 34	Čapek, K.	893	0	0,009211	0	0,017273	0,045758	0,105485	ibid.
876	Anglické listy 35	Čapek, K.	2034	0,022032	0,025038	0,089686	0,038538	0,036107	0,107739	ibid.
877	Cesta na sever 01	Čapek, K.	594	0	0,005279	0	0,013273	0,021289	0,065089	ibid.
878	Cesta na sever 02	Čapek, K.	81	0,296296	0,237037	0,470588	0,296296	0,237037	0,470588	ibid.
879	Cesta na sever 03	Čapek, K.	1162	0	0,011607	0	0,015556	0,020443	0,050336	ibid.
880	Cesta na sever 04	Čapek, K.	823	0	0,00361	0	0,004545	0,020618	0,045226	ibid.
881	Cesta na sever 05	Čapek, K.	1137	0	0,005149	0	0,024845	0,031315	0,111498	ibid.
882	Cesta na sever 06	Čapek, K.	1217	0	0	0	0,011575	0,010997	0,049844	ibid.
883	Cesta na sever 07	Čapek, K.	1040	0	0,005414	0	0	0,010681	0	ibid.
884	Cesta na sever 08	Čapek, K.	520	0	0,014237	0	0,00559	0,032645	0,070312	ibid.
885	Cesta na sever 09	Čapek, K.	1310	0	0	0	0	0,001443	0	ibid.
886	Cesta na sever 10	Čapek, K.	1366	0	0,001612	0	0,018375	0,020949	0,052632	ibid.
887	Cesta na sever 11	Čapek, K.	285	0	0,013527	0	0,02108	0,044664	0,103448	ibid.
888	Cesta na sever 12	Čapek, K.	2236	0,003647	0,013848	0,033708	0	0,016306	0	ibid.
889	Cesta na sever 13	Čapek, K.	1976	0,023063	0,021355	0,078775	0,033597	0,025174	0,081882	ibid.
890	Cesta na sever 14	Čapek, K.	1395	0	0	0	0	0,011512	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
891	Cesta na sever 15	Čapek, K.	2725	0	0,001227	0	0,007498	0,01294	0,050938	ibid.
892	Cesta na sever 16	Čapek, K.	1055	0	0	0	0,008472	0,020624	0,041522	ibid.
893	Cesta na sever 17	Čapek, K.	1810	0	0,002897	0	0	0,010401	0	ibid.
894	Cesta na sever 18	Čapek, K.	1515	0	0,010281	0	0,007756	0,02437	0,070352	ibid.
895	Cesta na sever 19	Čapek, K.	1014	0,011268	0,013936	0,060914	0,008472	0,019551	0,045455	ibid.
896	Cesta na sever 20	Čapek, K.	578	0	0	0	0	0,021711	0	ibid.
897	Cesta na sever 21	Čapek, K.	256	0	0,022222	0	0	0,029091	0	ibid.
898	Cesta na sever 22	Čapek, K.	985	0	0	0	0	0,003169	0	ibid.
899	Cesta na sever 23	Čapek, K.	963	0	0,004565	0	0,062271	0,040275	0,116466	ibid.
900	Cesta na sever 24	Čapek, K.	674	0	0,007103	0	0,01176	0,037093	0,068182	ibid.
901	Cesta na sever 25	Čapek, K.	720	0	0,000916	0	0	0,01234	0	ibid.
902	Cesta na sever 26	Čapek, K.	537	0	0,004735	0	0	0,039886	0	ibid.
903	Italské listy 01	Čapek, K.	396	0	0	0	0	0,006279	0	ibid.
904	Italské listy 02	Čapek, K.	55	0	0	0	0	0	0	ibid.
905	Italské listy 03	Čapek, K.	661	0	0	0	0	0,010709	0	ibid.
906	Italské listy 04	Čapek, K.	483	0	0	0	0,016092	0,025601	0,081395	ibid.
907	Italské listy 05	Čapek, K.	652	0	0	0	0	0,007128	0	ibid.
908	Italské listy 06	Čapek, K.	625	0,02795	0,03655	0,1125	0,005435	0,038932	0,068182	ibid.
909	Italské listy 07	Čapek, K.	705	0	0	0	0	0,000672	0	ibid.
910	Italské listy 08	Čapek, K.	639	0	0,000595	0	0,019305	0,028604	0,077519	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
911	Italské listy 09	Čapek, K.	597	0,014118	0,06223	0,098765	0,041209	0,070795	0,130435	ibid.
912	Italské listy 10	Čapek, K.	656	0	0	0	0,015553	0,036156	0,076923	ibid.
913	Italské listy 11	Čapek, K.	567	0	0	0	0	0,013932	0	ibid.
914	Italské listy 12	Čapek, K.	657	0,004662	0,014957	0,074468	0	0,011905	0	ibid.
915	Italské listy 13	Čapek, K.	596	0	0,001654	0	0	0,024791	0	ibid.
916	Italské listy 14	Čapek, K.	661	0	0	0	0	0,005518	0	ibid.
917	Italské listy 15	Čapek, K.	744	0,004902	0,016968	0,07	0	0,012745	0	ibid.
918	Italské listy 16	Čapek, K.	699	0	0,01016	0	0,009097	0,02272	0,065217	ibid.
919	Italské listy 17	Čapek, K.	577	0	0	0	0	0,003816	0	ibid.
920	Italské listy 18	Čapek, K.	601	0	0	0	0,014652	0,023387	0,078431	ibid.
921	Italské listy 19	Čapek, K.	603	0	0	0	0,008205	0,016786	0,066667	ibid.
922	Italské listy 20	Čapek, K.	646	0	0	0	0	0,001003	0	ibid.
923	Italské listy 21	Čapek, K.	428	0	0,008913	0	0	0,017927	0	ibid.
924	Italské listy 22	Čapek, K.	590	0	0,006433	0	0	0,006853	0	ibid.
925	Italské listy 23	Čapek, K.	643	0	0,000624	0	0,025415	0,028314	0,081761	ibid.
926	Italské listy 24	Čapek, K.	578	0	0,009615	0	0	0,01207	0	ibid.
927	Italské listy 25	Čapek, K.	814	0	0,000425	0	0	0	0	ibid.
928	Obrázky z Holandska 01	Čapek, K.	577	0	0,003309	0	0	0,02427	0	ibid.
929	Obrázky z Holandska 02	Čapek, K.	133	0	0	0	0	0,114286	0	ibid.
930	Obrázky z Holandska 03	Čapek, K.	610	0	0,012037	0	0,037037	0,064849	0,10119	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
931	Obrázky z Holandska 04	Čapek, K.	416	0,094276	0,061728	0,166667	0,092593	0,088523	0,148515	ibid.
932	Obrázky z Holandska 05	Čapek, K.	507	0	0	0	0	0,003551	0	ibid.
933	Obrázky z Holandska 06	Čapek, K.	270	0,301684	0,212121	0,4	0,41958	0,286859	0,507937	ibid.
934	Obrázky z Holandska 07	Čapek, K.	373	0,051809	0,043985	0,126761	0,035714	0,053571	0,087379	ibid.
935	Obrázky z Holandska 08	Čapek, K.	341	0,072	0,049091	0,140625	0,033566	0,06	0,123457	ibid.
936	Obrázky z Holandska 09	Čapek, K.	343	0	0	0	0	0,033381	0	ibid.
937	Obrázky z Holandska 10	Čapek, K.	405	0	0,003556	0	0	0,030909	0	ibid.
938	Obrázky z Holandska 11	Čapek, K.	93	0,190476	0,209524	0,363636	0,047619	0,130102	0,2	ibid.
939	Obrázky z Holandska 12	Čapek, K.	331	0,067227	0,075953	0,111111	0,067227	0,111183	0,225352	ibid.
940	Obrázky z Holandska 13	Čapek, K.	405	0,1	0,087374	0,160714	0,103704	0,09418	0,141304	ibid.
941	Obrázky z Holandska 14	Čapek, K.	422	0	0,016619	0	0,006173	0,042735	0,074468	ibid.
942	Obrázky z Holandska 15	Čapek, K.	1113	0	0,000496	0	0,002792	0,020524	0,040741	ibid.
943	Obrázky z Holandska 16	Čapek, K.	1227	0	0	0	0	0,006588	0	ibid.
944	Obrázky z Holandska 17	Čapek, K.	1302	0	0	0	0,025641	0,03139	0,052198	ibid.
945	Výlet do Španěl 01	Čapek, K.	739	0	0,001603	0	0	0,013889	0	ibid.
946	Výlet do Španěl 02	Čapek, K.	389	0	0,023188	0	0,008696	0,033597	0,089552	ibid.
947	Výlet do Španěl 03	Čapek, K.	495	0	0,001083	0	0	0,004141	0	ibid.
948	Výlet do Španěl 04	Čapek, K.	671	0	0,024564	0	0	0,018131	0	ibid.
949	Výlet do Španěl 05	Čapek, K.	899	0	0	0	0	0,005445	0	ibid.
950	Výlet do Španěl 06	Čapek, K.	370	0	0,011364	0	0	0,017756	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
951	Výlet do Španěl 07	Čapek, K.	338	0	0	0	0	0,00947	0	ibid.
952	Výlet do Španěl 08	Čapek, K.	524	0	0,013302	0	0	0,014323	0	ibid.
953	Výlet do Španěl 09	Čapek, K.	598	0,035714	0,034203	0,108696	0,011905	0,043132	0,088496	ibid.
954	Výlet do Španěl 10	Čapek, K.	447	0,069124	0,040766	0,12987	0,043969	0,04281	0,108911	ibid.
955	Výlet do Španěl 11	Čapek, K.	253	0	0,013333	0	0	0,011111	0	ibid.
956	Výlet do Španěl 12	Čapek, K.	688	0	0	0	0	0	0	ibid.
957	Výlet do Španěl 13	Čapek, K.	547	0	0,001289	0	0,015198	0,026684	0,086207	ibid.
958	Výlet do Španěl 14	Čapek, K.	819	0	0	0	0	0,003676	0	ibid.
959	Výlet do Španěl 15	Čapek, K.	640	0	0	0	0	0,006808	0	ibid.
960	Výlet do Španěl 16	Čapek, K.	511	0,050505	0,048951	0,188235	0,047242	0,062915	0,181818	ibid.
961	Výlet do Španěl 17	Čapek, K.	574	0	0,00133	0	0,008163	0,036667	0,06015	ibid.
962	Výlet do Španěl 18	Čapek, K.	489	0	0	0	0	0,012903	0	ibid.
963	Výlet do Španěl 19	Čapek, K.	1787	0	0,018191	0	0,054329	0,048589	0,154167	ibid.
964	Výlet do Španěl 20	Čapek, K.	1695	0,019576	0,020039	0,060172	0,040334	0,043423	0,078125	ibid.
965	Výlet do Španěl 21	Čapek, K.	1341	0,006547	0,014416	0,04811	0,023583	0,026783	0,066474	ibid.
966	Výlet do Španěl 22	Čapek, K.	336	0	0,039773	0	0,028571	0,050325	0,129032	ibid.
967	Výlet do Španěl 23	Čapek, K.	576	0	0	0	0	0	0	ibid.
968	Výlet do Španěl 24	Čapek, K.	519	0,016279	0,018499	0,090909	0,007633	0,030897	0,07619	ibid.
969	Výlet do Španěl 25	Čapek, K.	450	0	0,012685	0	0,007955	0,020928	0,079646	ibid.
970	Výlet do Španěl 26	Čapek, K.	283	0	0	0	0	0	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
971	Výlet do Španěl 27	Čapek, K.	616	0,012097	0,02256	0,076271	0,019713	0,025511	0,07971	ibid.
972	Výlet do Španěl 28	Čapek, K.	553	0	0,001998	0	0	0,016162	0	ibid.
973	Výlet do Španěl 29	Čapek, K.	510	0	0,010786	0	0,011328	0,056211	0,072464	ibid.
974	Jak se co dělá: Jak se dělají noviny	Čapek, K.	759	0	0,0172	0	0,019006	0,037465	0,07027	ibid.
975	Jak se co dělá: Z čeho se skládají noviny	Čapek, K.	272	0,1875	0,139815	0,25	0,083916	0,097756	0,155172	ibid.
976	Jak se co dělá: O redakci	Čapek, K.	3400	0	0,005705	0	0,011031	0,020599	0,05489	ibid.
977	Jak se co dělá: Jak vzniká číslo ranních novin	Čapek, K.	1091	0	0,007565	0	0	0,009983	0	ibid.
978	Jak se co dělá: Další činitelé	Čapek, K.	727	0,008027	0,057268	0,076923	0,038499	0,07185	0,13615	ibid.
979	Jak se co dělá: Jak se dělá: film	Čapek, K.	278	0,06213	0,070375	0,162791	0,246154	0,153846	0,235294	ibid.
980	Jak se co dělá: Krátký, ale nutný výklad o lidech	Čapek, K.	215	0,25625	0,209722	0,323529	0,409091	0,295756	0,428571	ibid.
981	Jak se co dělá: Honba za námětem	Čapek, K.	480	0,02381	0,043956	0,121212	0,166667	0,139589	0,233871	ibid.
982	Jak se co dělá: Čtyři filmové náměty	Čapek, K.	2580	0,03994	0,043649	0,124236	0,033605	0,03526	0,111702	ibid.
983	Jak se co dělá: Od námětu k scenáriu	Čapek, K.	957	0	0,010913	0	0,020219	0,044635	0,102564	ibid.
984	Jak se co dělá: Stavíme	Čapek, K.	708	0,003571	0,014167	0,06015	0,008627	0,01921	0,06962	ibid.
985	Jak se co dělá: Točíme	Čapek, K.	2061	0,036467	0,038906	0,111111	0,047309	0,047579	0,130795	ibid.
986	Jak se co dělá: Jak se tedy dělá: film	Čapek, K.	359	0,066667	0,070076	0,153846	0,022409	0,106658	0,105263	ibid.
987	Jak se co dělá: V dílnách a laboratořích	Čapek, K.	610	0	0,007857	0	0	0,016667	0	ibid.
988	Jak se co dělá: Premiéra	Čapek, K.	202	0,166667	0,161728	0,193548	0,383754	0,258697	0,4	ibid.
989	Jak se co dělá: Jak vzniká divadelní hra	Čapek, K.	399	0	0,008242	0	0	0,024451	0	ibid.
990	Jak se co dělá: První počátky	Čapek, K.	405	0,072751	0,097375	0,214286	0,199248	0,186184	0,411765	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
991	Jak se co dělá: Obsazení	Čapek, K.	331	0,343434	0,227273	0,367347	0,251196	0,171188	0,2	ibid.
992	Jak se co dělá: Režie	Čapek, K.	334	0,043174	0,084868	0,254545	0,062169	0,07967	0,185185	ibid.
993	Jak se co dělá: Čtená zkouška	Čapek, K.	522	0,126984	0,115995	0,164384	0,085781	0,112121	0,204082	ibid.
994	Jak se co dělá: Ve zkušebním sále	Čapek, K.	527	0,105324	0,135893	0,268519	0,071296	0,119935	0,258993	ibid.
995	Jak se co dělá: Další zkoušky	Čapek, K.	380	0,020833	0,077953	0,097222	0	0,076634	0	ibid.
996	Jak se co dělá: Kus dozrává	Čapek, K.	871	0,14902	0,127096	0,286667	0,159659	0,125541	0,223958	ibid.
997	Jak se co dělá: Generální zkouška	Čapek, K.	752	0,047792	0,077248	0,166667	0,096138	0,100646	0,188172	ibid.
998	Jak se co dělá: Další průběh	Čapek, K.	1064	0,04607	0,044476	0,092391	0,043975	0,060986	0,13913	ibid.
999	Jak se co dělá: Premiréra	Čapek, K.	1965	0,02967	0,024914	0,061086	0,036607	0,031226	0,090573	ibid.
1000	Jak se co dělá: Po premiéře	Čapek, K.	301	0	0,035579	0	0	0,038462	0	ibid.
1001	Jak se co dělá: Průvodce po zákulisí	Čapek, K.	3252	0	0,002633	0	0,004313	0,014016	0,02625	ibid.
1002	Lidové noviny: Novoroční datel	Čapek, K.	638	0	0,002646	0	0	0,013889	0	ibid.
1003	Lidové noviny: Slyšel jsem, že	Čapek, K.	646	0,013228	0,042593	0,084034	0,027132	0,038836	0,075676	ibid.
1004	Lidové noviny: O berních fasích	Čapek, K.	483	0	0,011722	0	0,028831	0,046869	0,142857	ibid.
1005	Lidové noviny: Děti a válka	Čapek, K.	473	0,440476	0,309066	0,386667	0,496753	0,29697	0,491228	ibid.
1006	Lidové noviny: Zvrácené poměry	Čapek, K.	161	0,050794	0,100529	0,190476	0	0,1	0	ibid.
1007	Lidové noviny: O městském dítěti	Čapek, K.	588	0	0,01049	0	0,066358	0,072743	0,165517	ibid.
1008	Lidové noviny: V tísni těchto dnů	Čapek, K.	222	0	0	0	0	0,089091	0	ibid.
1009	Lidové noviny: Alej fáborů	Čapek, K.	193	0	0,019481	0	0	0,048485	0	ibid.
1010	Lidové noviny: Ženy a děti	Čapek, K.	145	0	0,02381	0	0	0,040741	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
1011	Lidové noviny: Mladý sedlák	Čapek, K.	198	0	0	0	0	0	0	ibid.
1012	Lidové noviny: Kurzíva militaristická	Čapek, K.	174	0	0	0	0,12963	0,110269	0,162791	ibid.
1013	Lidové noviny: Slovákům	Čapek, K.	207	0	0	0	0,088889	0,141414	0,26087	ibid.
1014	Lidové noviny: Propříště	Čapek, K.	270	0	0	0	0	0,024476	0	ibid.
1015	Lidové noviny: Aréna	Čapek, K.	567	0	0,012407	0	0,02765	0,037634	0,103448	ibid.
1016	Lidové noviny: Dobře to dopadlo	Čapek, K.	496	0	0	0	0,026273	0,040925	0,073394	ibid.
1017	Lidové noviny: V jásotu národa	Čapek, K.	245	0,081818	0,082828	0,1875	0,036364	0,071625	0,12	ibid.
1018	Lidové noviny: Indiskrece z Ženevy	Čapek, K.	187	0	0,047619	0	0	0	0	ibid.
1019	Lidové noviny: Karlovy Vary	Čapek, K.	545	0,020513	0,041667	0,094118	0,044643	0,086285	0,178571	ibid.
1020	Lidové noviny: Byla to revoluce?	Čapek, K.	565	0,019565	0,045717	0,096774	0,006988	0,032699	0,061224	ibid.
1021	Lidové noviny: Tíseň	Čapek, K.	665	0	0	0	0	0,002251	0	ibid.
1022	Lidové noviny: Nouzové práce zdarma	Čapek, K.	194	0,12	0,106667	0,181818	0,094276	0,103788	0,166667	ibid.
1023	Lidové noviny: Jako každého roku	Čapek, K.	410	0	0,001894	0	0,026853	0,032821	0,105263	ibid.
1024	Lidové noviny: Tvář doby	Čapek, K.	460	0	0,013333	0	0	0,021415	0	ibid.
1025	Lidové noviny: Letadla nad Prahou	Čapek, K.	198	0	0	0	0	0	0	ibid.
1026	Lidové noviny: Novostavba nad Prahou	Čapek, K.	200	0	0	0	0	0	0	ibid.
1027	Lidové noviny: Na karlovarském letišti	Čapek, K.	1510	0	0,008	0	0,003699	0,011263	0,036017	ibid.
1028	Lidové noviny: Neštěstí a technika	Čapek, K.	710	0	0,019835	0	0,004874	0,032069	0,05641	ibid.
1029	Lidové noviny: Chudí těchto dnů	Čapek, K.	663	0	0,001344	0	0,022523	0,035506	0,075	ibid.
1030	Lidové noviny: Hospodářská tíseň	Čapek, K.	273	0	0	0	0	0,006324	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
1031	Lidové noviny: Dobrá práce	Čapek, K.	488	0	0	0	0	0,031173	0	ibid.
1032	Lidové noviny: Regionalismus	Čapek, K.	547	0	0,019231	0	0,043651	0,050187	0,145038	ibid.
1033	Lidové noviny: Spoluvina	Čapek, K.	539	0	0	0	0	0,029777	0	ibid.
1034	Lidové noviny: Národ nás nepotřebuje	Čapek, K.	674	0,035256	0,045238	0,104762	0,085586	0,090469	0,173913	ibid.
1035	Lidové noviny: Dvě stě tisíc	Čapek, K.	754	0	0,035954	0	0	0,029086	0	ibid.
1036	Lidové noviny: Pětasedmdesát let	Čapek, K.	204	0	0,042857	0	0	0,02277	0	ibid.
1037	Lidové noviny: Do dvacátého roku	Čapek, K.	764	0	0	0	0	0,020363	0	ibid.
1038	Lidové noviny: Pumy nad světem	Čapek, K.	528	0	0	0	0	0,006667	0	ibid.
1039	Lidové noviny: Budeme žít!	Čapek, K.	262	0	0	0	0,035333	0,049933	0,111111	ibid.
1040	Zahradníkův rok 01	Čapek, K.	617	0	0,005784	0	0,022105	0,032385	0,070064	ibid.
1041	Zahradníkův rok 02	Čapek, K.	643	0	0	0	0	0,015238	0	ibid.
1042	Zahradníkův rok 03	Čapek, K.	975	0	0,012392	0	0	0,029666	0	ibid.
1043	Zahradníkův rok 04	Čapek, K.	656	0	0,016512	0	0,036877	0,056344	0,087302	ibid.
1044	Zahradníkův rok 05	Čapek, K.	787	0,00974	0,020455	0,081818	0,039562	0,067538	0,166667	ibid.
1045	Zahradníkův rok 06	Čapek, K.	747	0	0,019963	0	0	0,017241	0	ibid.
1046	Zahradníkův rok 07	Čapek, K.	880	0	0,012342	0	0	0,027987	0	ibid.
1047	Zahradníkův rok 08	Čapek, K.	515	0,051282	0,05711	0,114286	0,036946	0,064229	0,091837	ibid.
1048	Zahradníkův rok 09	Čapek, K.	943	0	0,006972	0	0,014481	0,019735	0,054852	ibid.
1049	Zahradníkův rok 10	Čapek, K.	642	0	0,045848	0	0,086096	0,082442	0,160622	ibid.
1050	Zahradníkův rok 11	Čapek, K.	1134	0	0,00581	0	0	0,011738	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
1051	Zahradníkův rok 12	Čapek, K.	561	0	0,01061	0	0,011084	0,036782	0,07563	ibid.
1052	Zahradníkův rok 13	Čapek, K.	1103	0,024951	0,020683	0,073059	0,030303	0,028746	0,07037	ibid.
1053	Zahradníkův rok 14	Čapek, K.	467	0	0,005051	0	0	0,032086	0	ibid.
1054	Zahradníkův rok 15	Čapek, K.	923	0,01548	0,017544	0,070064	0,017772	0,025595	0,064356	ibid.
1055	Zahradníkův rok 16	Čapek, K.	584	0,05153	0,040943	0,125	0,049784	0,068939	0,178862	ibid.
1056	Zahradníkův rok 17	Čapek, K.	1139	0	0,001835	0	0,023264	0,024481	0,059233	ibid.
1057	Zahradníkův rok 18	Čapek, K.	573	0	0,009696	0	0	0,030139	0	ibid.
1058	Zahradníkův rok 19	Čapek, K.	859	0	0	0	0	0,009897	0	ibid.
1059	Zahradníkův rok 20	Čapek, K.	655	0	0,005831	0	0,009354	0,028211	0,06875	ibid.
1060	Zahradníkův rok 21	Čapek, K.	989	0	0,006688	0	0	0,012957	0	ibid.
1061	Zahradníkův rok 22	Čapek, K.	521	0	0,00571	0	0	0,036623	0	ibid.
1062	Zahradníkův rok 23	Čapek, K.	1000	0	0,004855	0	0,01697	0,026159	0,053846	ibid.
1063	Zahradníkův rok 24	Čapek, K.	702	0	0,012955	0	0	0,023525	0	ibid.
1064	Zahradníkův rok 25	Čapek, K.	1090	0	0,004496	0	0	0,019337	0	ibid.
1065	Zahradníkův rok 26	Čapek, K.	641	0	0,020261	0	0,03125	0,031863	0,082759	ibid.
1066	Dopis Anně Nešporové 01	Čapek, K.	675	0	0,003151	0	0,029577	0,048795	0,113445	ibid.
1067	Dopis Anně Nešporové 02	Čapek, K.	260	0	0	0	0	0	0	ibid.
1068	Dopis Anně Nešporové 03	Čapek, K.	1452	0	0,000966	0	0	0,000769	0	ibid.
1069	Dopis Anně Nešporové 04	Čapek, K.	508	0	0,033333	0	0	0,020945	0	ibid.
1070	Dopis Anně Nešporové 05	Čapek, K.	182	0	0	0	0	0	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
1071	Dopis Anně Nešporové 06	Čapek, K.	603	0	0	0	0	0,013818	0	ibid.
1072	Dopis Anně Nešporové 07	Čapek, K.	431	0	0	0	0,018462	0,028889	0,084112	ibid.
1073	Dopis Anně Nešporové 08	Čapek, K.	442	0	0,006968	0	0	0,016722	0	ibid.
1074	Dopis Anně Nešporové 09	Čapek, K.	480	0,034965	0,028846	0,102564	0,035714	0,04533	0,095238	ibid.
1075	Dopis Anně Nešporové 10	Čapek, K.	282	0	0	0	0	0,013935	0	ibid.
1076	Dopis Anně Nešporové 11	Čapek, K.	498	0	0	0	0	0,006617	0	ibid.
1077	Dopis Anně Nešporové 12	Čapek, K.	893	0	0,004487	0	0,005128	0,023988	0,042705	ibid.
1078	Dopis Anně Nešporové 13	Čapek, K.	1092	0	0,005797	0	0,010256	0,013846	0,039409	ibid.
1079	Dopis Anně Nešporové 14	Čapek, K.	536	0	0,019698	0	0	0,022917	0	ibid.
1080	Dopis Anně Nešporové 15	Čapek, K.	338	0,016317	0,073718	0,12069	0,006366	0,040066	0,081081	ibid.
1081	Dopis Anně Nešporové 16	Čapek, K.	751	0	0,005544	0	0	0,006345	0	ibid.
1082	Dopis Heleně 01	Čapek, K.	413	0	0	0	0,006993	0,022436	0,068627	ibid.
1083	Dopis Heleně 02	Čapek, K.	318	0	0	0	0	0,002755	0	ibid.
1084	Dopis Heleně 03	Čapek, K.	388	0	0	0	0	0,016222	0	ibid.
1085	Dopis Heleně 04	Čapek, K.	139	0,133333	0,207143	0,235294	0	0,078571	0	ibid.
1086	Dopis Heleně 05	Čapek, K.	725	0	0,006603	0	0	0,011628	0	ibid.
1087	Dopis Heleně 06	Čapek, K.	572	0,019481	0,035964	0,095745	0,014286	0,026374	0,066667	ibid.
1088	Dopis Heleně 07	Čapek, K.	378	0	0	0	0	0,001374	0	ibid.
1089	Dopis Heleně 08	Čapek, K.	277	0	0	0	0	0,015515	0	ibid.
1090	Dopis S. K. Neumannovi 01	Čapek, K.	544	0	0,011312	0	0	0,037778	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
1091	Dopis S. K. Neumannovi 02	Čapek, K.	749	0	0	0	0	0,025817	0	ibid.
1092	Dopis S. K. Neumannovi 03	Čapek, K.	770	0	0	0	0	0,002346	0	ibid.
1093	Dopis S. K. Neumannovi 04	Čapek, K.	612	0,04914	0,040154	0,098214	0,038889	0,035526	0,08982	ibid.
1094	Dopis S. K. Neumannovi 05	Čapek, K.	770	0	0	0	0	0	0	ibid.
1095	Dopis S. K. Neumannovi 06	Čapek, K.	3342	0	0,003596	0	0,006773	0,019354	0,045764	ibid.
1096	Dopis S. K. Neumannovi 07	Čapek, K.	536	0	0	0	0	0,003501	0	ibid.
1097	Dopis S. K. Neumannovi 08	Čapek, K.	867	0	0	0	0	0,005682	0	ibid.
1098	Dopis S. K. Neumannovi 09	Čapek, K.	573	0	0	0	0	0,02161	0	ibid.
1099	Dopis O. Scheinpflugové 01	Čapek, K.	986	0	0	0	0	0,001335	0	ibid.
1100	Dopis O. Scheinpflugové 02	Čapek, K.	373	0	0,024351	0	0	0,010989	0	ibid.
1101	Dopis O. Scheinpflugové 03	Čapek, K.	676	0	0	0	0,008076	0,013984	0,05848	ibid.
1102	Dopis O. Scheinpflugové 04	Čapek, K.	425	0	0	0	0	0,016254	0	ibid.
1103	Dopis O. Scheinpflugové 05	Čapek, K.	437	0	0,017483	0	0	0,014423	0	ibid.
1104	Dopis O. Scheinpflugové 06	Čapek, K.	725	0,025621	0,030066	0,1	0,008429	0,041107	0,0625	ibid.
1105	Dopis O. Scheinpflugové 07	Čapek, K.	166	0	0,045918	0	0,119048	0,127551	0,277778	ibid.
1106	Dopis O. Scheinpflugové 08	Čapek, K.	183	0	0	0	0	0,008889	0	ibid.
1107	Dopis O. Scheinpflugové 09	Čapek, K.	466	0	0,000289	0	0	0	0	ibid.
1108	Dopis O. Scheinpflugové 10	Čapek, K.	528	0	0,014286	0	0	0,005808	0	ibid.
1109	Dopis O. Scheinpflugové 11	Čapek, K.	574	0	0	0	0	0,011111	0	ibid.
1110	Dopis O. Scheinpflugové 12	Čapek, K.	536	0	0,009146	0	0	0,024985	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
1111	Dopis O. Scheinpflugové 13	Čapek, K.	549	0	0,025641	0	0	0,016329	0	ibid.
1112	Dopis O. Scheinpflugové 14	Čapek, K.	488	0	0	0	0	0	0	ibid.
1113	Dopis O. Scheinpflugové 15	Čapek, K.	818	0	0	0	0	0,001773	0	ibid.
1114	Dopis O. Scheinpflugové 16	Čapek, K.	693	0	0	0	0	0	0	ibid.
1115	Dopis O. Scheinpflugové 17	Čapek, K.	701	0	0	0	0	0,00117	0	ibid.
1116	Dopis O. Scheinpflugové 18	Čapek, K.	310	0	0,032407	0	0,033333	0,074675	0,127273	ibid.
1117	Dopis O. Scheinpflugové 19	Čapek, K.	727	0	0	0	0	0,005887	0	ibid.
1118	Dopis O. Scheinpflugové 20	Čapek, K.	482	0	0	0	0	0,011541	0	ibid.
1119	Dopis O. Scheinpflugové 21	Čapek, K.	588	0	0	0	0	0,013442	0	ibid.
1120	Dopis TGM 01	Čapek, K.	53	0	0,222222	0	0	0	0	ibid.
1121	Dopis TGM 02	Čapek, K.	185	0	0	0	0	0	0	ibid.
1122	Dopis TGM 03	Čapek, K.	137	0	0,057143	0	0,3	0,266667	0,214286	ibid.
1123	Dopis TGM 04	Čapek, K.	503	0	0	0	0,00625	0,035294	0,072	ibid.
1124	Dopis TGM 05	Čapek, K.	265	0	0	0	0	0,013774	0	ibid.
1125	Dopis TGM 06	Čapek, K.	168	0	0,015306	0	0	0,020833	0	ibid.
1126	Dopis TGM 07	Čapek, K.	200	0	0	0	0	0	0	ibid.
1127	Dopis TGM 08	Čapek, K.	549	0	0,008403	0	0,042857	0,036667	0,096774	ibid.
1128	Dopis TGM 09	Čapek, K.	230	0	0,047619	0	0	0,030303	0	ibid.
1129	Dopis TGM 10	Čapek, K.	250	0	0	0	0	0,009877	0	ibid.
1130	Dopis TGM 11	Čapek, K.	382	0	0	0	0	0,003991	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
1131	Dopis TGM 12	Čapek, K.	76	0	0	0	0	0,06	0	ibid.
1132	Dopis TGM 13	Čapek, K.	279	0	0	0	0	0,003828	0	ibid.
1133	Dopis TGM 14	Čapek, K.	141	0	0	0	0	0	0	ibid.
1134	Dopis TGM 15	Čapek, K.	126	0	0	0	0	0	0	ibid.
1135	Dopis TGM 16	Čapek, K.	298	0	0	0	0	0	0	ibid.
1136	Dopis TGM 17	Čapek, K.	241	0	0	0	0	0,027446	0	ibid.
1137	Dopis TGM 18	Čapek, K.	234	0	0	0	0	0,02139	0	ibid.
1138	Dopis TGM 19	Čapek, K.	345	0,133333	0,103704	0,272727	0,055944	0,078526	0,137931	ibid.
1139	Dopis TGM 20	Čapek, K.	111	0	0	0	0	0	0	ibid.
1140	Dopis TGM 21	Čapek, K.	208	0,104167	0,111607	0,263158	0,035273	0,101852	0,166667	ibid.
1141	Dopis TGM 22	Čapek, K.	104	0	0,071429	0	0	0	0	ibid.
1142	Dopis TGM 23	Čapek, K.	269	0	0	0	0	0	0	ibid.
1143	Dopis TGM 24	Čapek, K.	555	0	0,010417	0	0	0,004085	0	ibid.
1144	Dopis Věře Hrůzové 01	Čapek, K.	311	0	0,034091	0	0	0,00947	0	ibid.
1145	Dopis Věře Hrůzové 02	Čapek, K.	365	0	0	0	0	0	0	ibid.
1146	Dopis Věře Hrůzové 03	Čapek, K.	385	0	0	0	0	0,00519	0	ibid.
1147	Dopis Věře Hrůzové 04	Čapek, K.	450	0	0,025179	0	0,024545	0,041649	0,098901	ibid.
1148	Dopis Věře Hrůzové 05	Čapek, K.	392	0	0	0	0	0	0	ibid.
1149	Dopis Věře Hrůzové 06	Čapek, K.	243	0	0,020513	0	0	0,021883	0	ibid.
1150	Dopis Věře Hrůzové 07	Čapek, K.	439	0	0	0	0	0	0	ibid.

číslo textu	název	autor	N	TK slovní tvary	STK slovní tvary	PTK slovní tvary	TK lemmata	STK lemmata	PTK lemmata	zdroj
1151	Dopis Věře Hrůzové 08	Čapek, K.	502	0,068323	0,053936	0,106557	0,082621	0,050079	0,131737	ibid.
1152	Dopis Věře Hrůzové 09	Čapek, K.	313	0	0	0	0	0	0	ibid.
1153	Dopis Věře Hrůzové 10	Čapek, K.	272	0	0	0	0	0,008547	0	ibid.
1154	Dopis Věře Hrůzové 11	Čapek, K.	353	0	0	0	0	0,014652	0	ibid.
1155	Dopis Věře Hrůzové 12	Čapek, K.	265	0	0	0	0	0	0	ibid.
1156	Dopis Věře Hrůzové 13	Čapek, K.	454	0	0	0	0	0	0	ibid.
1157	Dopis Věře Hrůzové 14	Čapek, K.	347	0,115385	0,079882	0,157895	0,021429	0,028022	0,109756	ibid.
1158	Dopis Věře Hrůzové 15	Čapek, K.	380	0	0	0	0	0,001972	0	ibid.
1159	Celní unie v ohrožení		429	0,053333	0,115152	0,170213	0,034632	0,116883	0,135593	Pražský závislostní korpus 2.0.
1160	Stát – podnikatelé – nezaměstnanost		610	0,032634	0,081731	0,181818	0,15561	0,206435	0,357143	ibid.
1161	Na život a na smrt – nejlépe po americku		458	0	0,032967	0	0,034161	0,102355	0,101852	ibid.
1162	Voda a teplo = peníze		602	0,030612	0,100733	0,121622	0,139037	0,152741	0,225225	ibid.
1163	Podnikání v éteru		424	0,16431	0,166061	0,319149	0,324444	0,30202	0,518519	ibid.
1164	Poklidné kompetence		369	0,061765	0,147154	0,175	0,190927	0,172894	0,15942	ibid.
1165	Je-li vypovídání smluv legální , je nutné novelizovat zákony		258	0	0,031746	0	0,123589	0,138503	0,304348	ibid.
1166	Podnikatelská banka nabírá dech		433	0,080808	0,185859	0,176471	0,219608	0,148841	0,259259	ibid.
1167	Rusko zve zahraniční investory		590	0,021164	0,063899	0,101266	0,126254	0,124843	0,218045	ibid.
1168	Jak statistický úřad počítá míru inflace		850	0,141429	0,146167	0,344086	0,229032	0,229228	0,525714	ibid.

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