Kvantitativní analýza textu

(její nástrahy, meze, perspektivy)

Radek Čech

http://www.cechradek.cz

Kvantitativní analýza textu

- kvantifikace a její důsledky
- kvantifikace a operacionalizace – případ volby jazykových jednotek aneb „každá volba je špatná“
- sekundární tematická koncentrace textu
- QUITA – software pro kvantitativní analýzu textů

Význam kvantifikace

- nekvantifikovaný popis
  - $V(A) = V(B)$, nebo $V(A) \neq V(B)$
  - V … vlastnost
  - A, B … pozorované jevy
- co znamená kvantifikovat?
  - označení vlastnosti číslem
  - $V(A) - V(B) = d$
  - d … rozdíl hodnot
- jak interpretovat d?

Význam kvantifikace v textologii

Detailnější poznatky můžeme vyčíst ze starší knihy Frekvence slov, slovních druhů a tvarů v češtině (1961), kde jsou zvlášť pojednány texty popularizující a vědecké. Tak např. v odborných textech je silný podíl substantiv (32,9 % – novější údaje z jiného souboru textů, 1983, mluví dokonce o 44,35 % různých lexémů) a adjektiv je dokonce relativně více než v textech jiných (16,25 %), sloves naopak méně (14,15 %, v materiálu z r. 1983 14,96 %), dosti vysokou frekvenci mají předložky (10,73 %). Ve vlastních vědeckých textech je překvapivě o něco méně substantiv i adjektiv a sloves, zato přibývá zájmen. Ve srovnání s jinými typy textů je významný především nižší podíl sloves, která jsou kromě toho poměrně stereotypní. V materiálu Českého národního korpusu bylo zachyceno (podle přednášky M. Kopřivové z r. 2005) v odborných textech z takřka 20 milionů zpracovaných jednotek 5,5 mil. substantiv, 2,3 mil. adjektiv, 2,2 mil. sloves, 1,2 mil. spojek a zájmen a 1,7 mil. předložek. I tu jde o orientační údaj, který může být dále upřesňován prací s celým již zpracovaným materiálem s přihlédnutím k jednotlivým použitým textům.

(Čechová et al. 2008, s. 218)

SYN 2010

ODB (SYN2010)					SYN2010 (bez ODB)			
pořadí	**POS**	**f**	**%**		**pořadí**	**POS**	**f**	**%**
1	subst.	8908919	32.94		1	subst.	20899938	28.73
2	verb.	4074532	15.07		2	verb.	13753983	18.90
3	adj.	3782195	13.99		3	pron.	8837590	12.15
4	prep.	2907295	10.75		4	prep.	7785085	10.70
5	pron.	2431471	8.99		5	adj.	7301172	10.03
6	konj.	2079657	7.69		6	konj.	5733385	7.88
7	adv.	1667110	6.16		7	adv.	5442020	7.48
8	num.	821434	3.04		8	num.	1825676	2.51
9	part.	365034	1.35		9	part.	1117393	1.54
10	inter.	6118	0.02		10	inter.	61697	0.08
		27043765	100				72757939	100

SYN 2010

ODB (SYN2010)					PUB (SYN2010)			
pořadí	**POS**	**f**	**%**		**pořadí**	**POS**	**f**	**%**
1	subst.	8908919	32.94		1	subst.	11322190	34.17
2	verb.	4074532	15.07		2	verb.	5328752	16.08
3	adj.	3782195	13.99		3	prep.	3879399	11.71
4	prep.	2907295	10.75		4	adj.	3870151	11.68
5	pron.	2431471	8.99		5	pron.	2861782	8.64
6	konj.	2079657	7.69		6	konj.	2199028	6.64
7	adv.	1667110	6.16		7	adv.	2044801	6.17
8	num.	821434	3.04		8	num.	1170565	3.53
9	part.	365034	1.35		9	part.	449945	1.36
10	inter.	6118	0.02		10	inter.	5528	0.02
		27043765	100				33132141	100.00

Biber et al. (1999): Longman Grammar of Spoken and Written English

The distribution of nouns and pronouns varies greatly depending upon register (2.3.5, 2.4.14). It further turns out that the use of pronouns v. full noun phrases varies in relation to syntactic role.

CORPUS FINDINGS [3,16]

Pronouns are slightly more common than nouns in conversation.

At the other extreme, nouns are many times more common than pronouns in news and academic prose.

The noun-pronoun ratio varies greatly depending upon syntactic role.

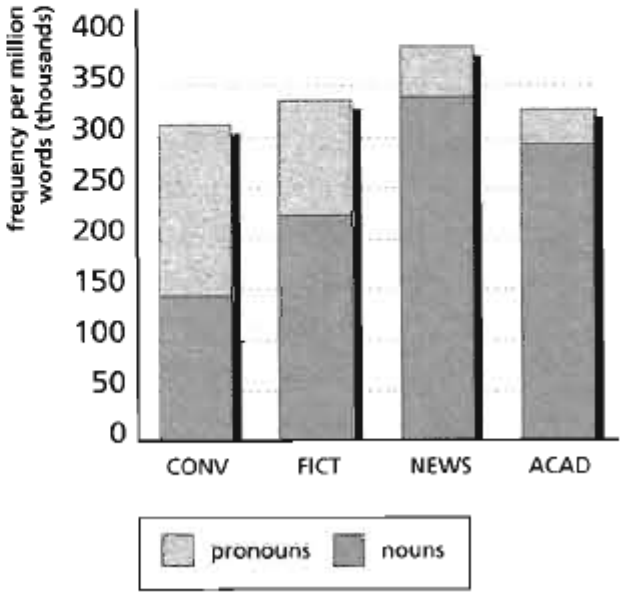
- The relative frequency of nouns is much higher in object position and as a complement or object of a preposition than in subject position.

DISCUSSION OF FINDINGS

As illustrated in 4.1.1, there are important differences in the reliance on nouns v. pronouns across registers. In

Figure 4.1

Distribution of nouns v. pronouns across registers



Význam kvantifikace v textologii

- každá smysluplná kvantifikace je založena na předpokladu, že rozdíly hodnot u sledované vlastnosti jsou ve vztahu k jiné vlastnosti, např.
  - délka slova vs. polysémie
  - typ textu vs. výskyt slovních druhů
- důsledky
  - vlastnosti jazyka, které jsou mnohdy popisovány izolovaně, jsou interpretovány ve vzájemných vztazích
  - kvantifikace stanovuje velikost rozdílu
  - jak zjištěné velikosti interpretovat?

Interpretace kvantifikace

- převažuje procentuální vyjádření rozdílů
- většinou doplněné komentáři typu: pozorovaný rozdíl je „malý“, „větší“, „téměř stejný“, že „se velmi liší“ apod.
- jak velký rozdíl je „malý“, „velký“?

Proporce subst., adj. a verb v různých typech textu (SYN 2010)

	ODB	PUB	BEL
substantiva	32,94 %	34,17 %	24,17 %
verba	15,07 %	16,08 %	21,26 %
adjektiva	13,99 %	11,68 %	8,66 %

Od popisu k hypotéze

- předpokládaný vztah mezi dvěma vlastnostmi = působení mechanismu
- zaměříme-li na vztahy mezi vlastnostmi, překračujeme hranice popisu jednotek směrem k odhalování toho, proč jsou sledované vlastnosti oněch jednotek takové, jaké jsou
- nástrojem analýzy tohoto typu je empiricky testovatelná hypotéza

Hypotéza

- hypotéza (Greis 2009, s. 11)
  - tvrzení, které se týká více než jednoho jevu či případu;
  - má alespoň implicitně strukturu podmínkového souvětí, tj. *„jestliže…, pak…“*, případně *„čím…, tím…“* (např. čím je slovo frekventovanější, tím je kratší);
  - je falzifikovatelné (tj. vyvratitelné) prostřednictvím experimentu, který dovoluje rozhodnout, zda predikce formulovaná prostřednictvím hypotézy je vyvrácena, či ne (vyhodnocení se většinou experimentu pomocí statistických testů).

Hypotéza

- která tvrzení *jsou*/*nejsou* testovatelnými hypotézami?
  1. *hodně mužů má pleš*
  2. *pokud se v knize vyskytují biblický příběh, je to apokryf*
  3. *jestli se zavedou řidičáky „na zkoušku“, může se snížit nehodovost mladých řidičů a řidiček*
  4. *muži mají častěji pleš než ženy*
  5. *jestliže se je sloveso dokonavé, častěji se na něj váže přímý akuzativní předmět než na sloveso nedokonavé*
  6. *ženy jsou citlivé*
  7. *čím je slovo frekventovanější, tím je větší jeho polysémie*
  8. *jestli se zavedou řidičáky „na zkoušku“, sníží se nehodovost mladých řidičů a řidiček*
  9. *nářečí často ovlivňují podobu mluveného jazyka obyvatel dané nářeční oblasti*

Od popisu k hypotéze

klasifikace jevů

↓

klasifikace + kvantifikace + verbální interpretace pozorovaných rozdílů

↓

snaha po „hlubší“ interpretaci pozorovaných rozdílů =
= hlavní pozornost zaměřena již nikoliv na jevy samotné,
ale na vztahy mezi nimi (mechanismy)

↓

hypotéza

↓

testování hypotéz (experiment) + statistická interpretace výsledků

↓

(?teorie?)

Experimentální přístup

- výhody
  - replikovatelnost
  - pokud testujeme stejnou hypotézu na různých jazycích můžeme dospět k jazykovému zákonu
  - vyvrácení hypotézy nemusí vždy znamenat neplatnost celé hypotézy, ale jen odhalení tzv. hraničních podmínek
  - intersubjektivita
- nevýhody
  - redukcionismus
  - specifické nároky (technické prostředky, aplikace statistiky atp.), se kterými se „běžně“ školený lingvista musí vyrovnat

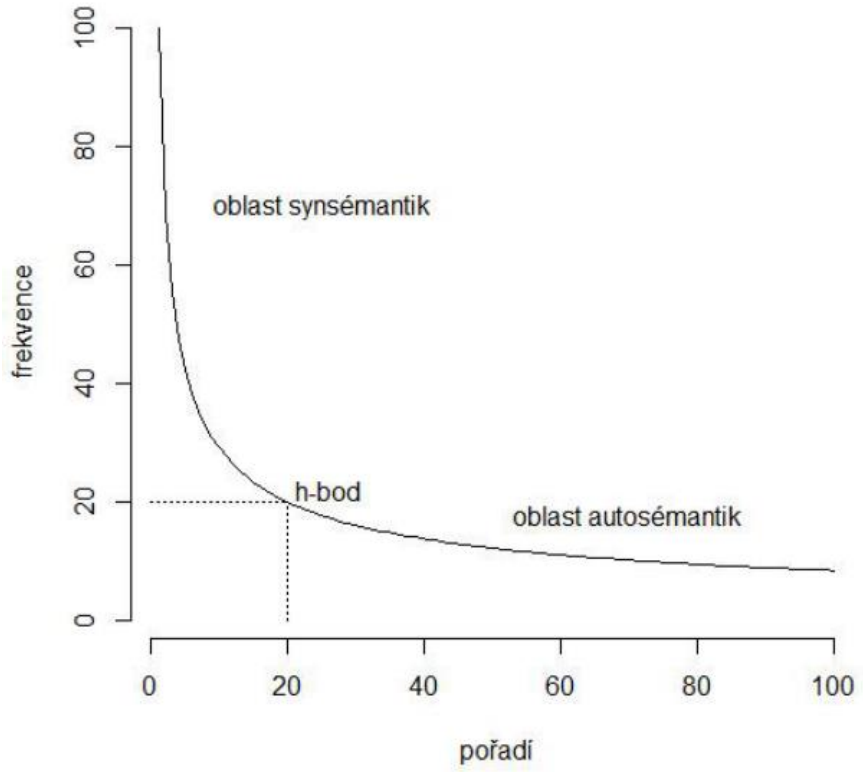
Operacionalizace a její problémy

- H_0: subjekt není v češtině v průměru delší než objekt
- H_1 : subjekt je v češtině v průměru delší než objekt

Muž v triku veze naše dopisy

Operacionalizace a její problémy v kvantitativní analýze textů

- volba jazykových jednotek při analýzy tematické koncentrace textů



Volba jazykových jednotek

- slovní tvar
- lemma
  - „škola“
- jednotka zahrnující synonyma
- koreferenční jednotka
- hreb (sémantický agregát)

Slovní tvary vs. lemmata

text	slovní tvary	lemmata
Často… (poezie)	**0,0294**	0
Filosofové… (poezie)	**0,1905**	0,0476
Hlubokým… (poezie)	0	**0,0457**
Dopis Olze 01	0	0
Dopis Olze 11	**0,0048**	0,0022
Dopis Olze 19	0	**0,0217**
Role českého prezidenta (článek)	0,0188	**0,0823**
Průchod spravedlnosti (článek)	**0,1125**	0,1105
Šifra socialismus (esej)	0,0063	**0,0583**

Hreb

hreb	elemnty
ja	{počíta**m**, so**m**, stáva**m** sa, bojí**m** sa, obaja, vie**me**, nás, ma, ja, cíti**m**, mne, mi, ma, ja, cíti**m**, dúfa**m**, dúfa**me**, so**m**, so**m**, nemôže**me**, ľúbi**m**, nevie**m**, nevie**m**, nevie**m**, bude**m**, bude**me**, plače**m**, smeje**m** sa, dúfa**m**, naša}
vedieť	{vieme, neviem, neviem, neviem}
objatie	{to, objatie, ktoré, ktoré}

Operacionalizace a její problémy v kvantitativní analýze textů

- je nezbytné prezentovat detailní metodologický postup (v ideálním případě přiložit technickou zprávu)
- často/většinou nemáme kritéria pro „správnou“ volbu

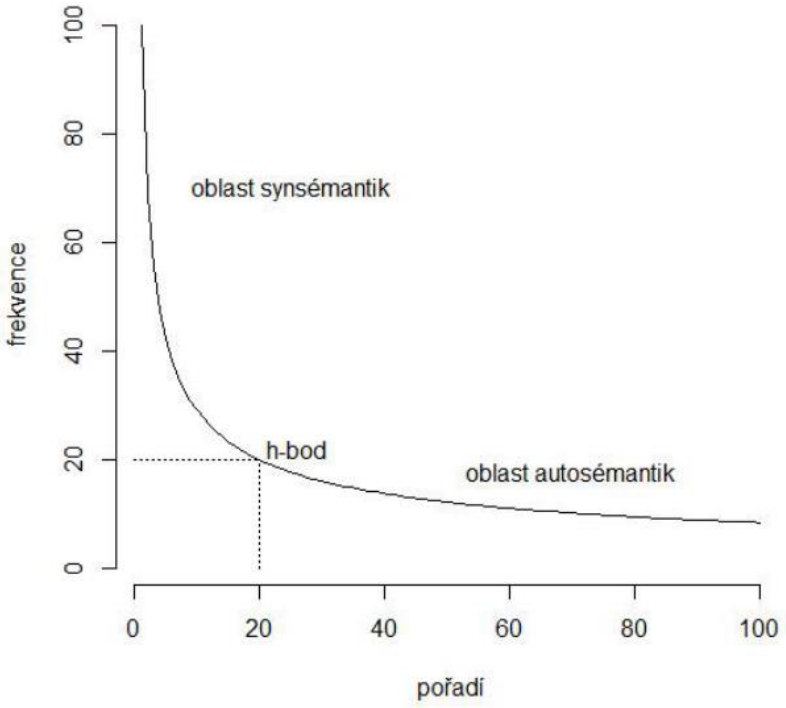
Sekundární tematická koncentrace

- nulové hodnoty tematické koncentrace
  - z teoretického hlediska nejsou problém
  - omezují použití tohoto typu analýzy

text (lemmatizováno)	TK
Bělorusko, náš nový východní soused	0
Co si opravdu myslím o stíhačkách?	0
Diktátoři porozumí jen síle	0
Dozrál čas	0
Historická šance pro naši zemi - nepromarněme ji!	0,0271
Litomyšlské znaky	0
Nepodléhejme rétorice diktátora Castra	0
Svět bez Zdeňka	0,0359
Ztráta paměti?	0,0693
Co je to Hrad? Prostě jímka!	0,0688

Sekundární tematická koncentrace

$$STK = \sum_{r'=1}^{2h} \frac{(2h - r')f(r')}{h(2h-1)f(1)}$$



Sekundární tematická koncentrace

text (lemmatizováno)	TK	STK
Bělorusko, náš nový východní soused	0	0,0515
Co si opravdu myslím o stíhačkách?	0	0,0113
Diktátoři porozumí jen síle	0	0,0568
Dozrál čas	0	0,0171
Historická šance pro naši zemi - nepromarněme ji!	0,0271	**0,0416**
Litomyšlské znaky	0	0,0189
Nepodléhejme rétorice diktátora Castra	0	0,0127
Svět bez Zdeňka	**0,0359**	0,0299
Ztráta paměti?	**0,0693**	0,0552
Co je to Hrad? Prostě jímka!	0,0688	**0,0753**

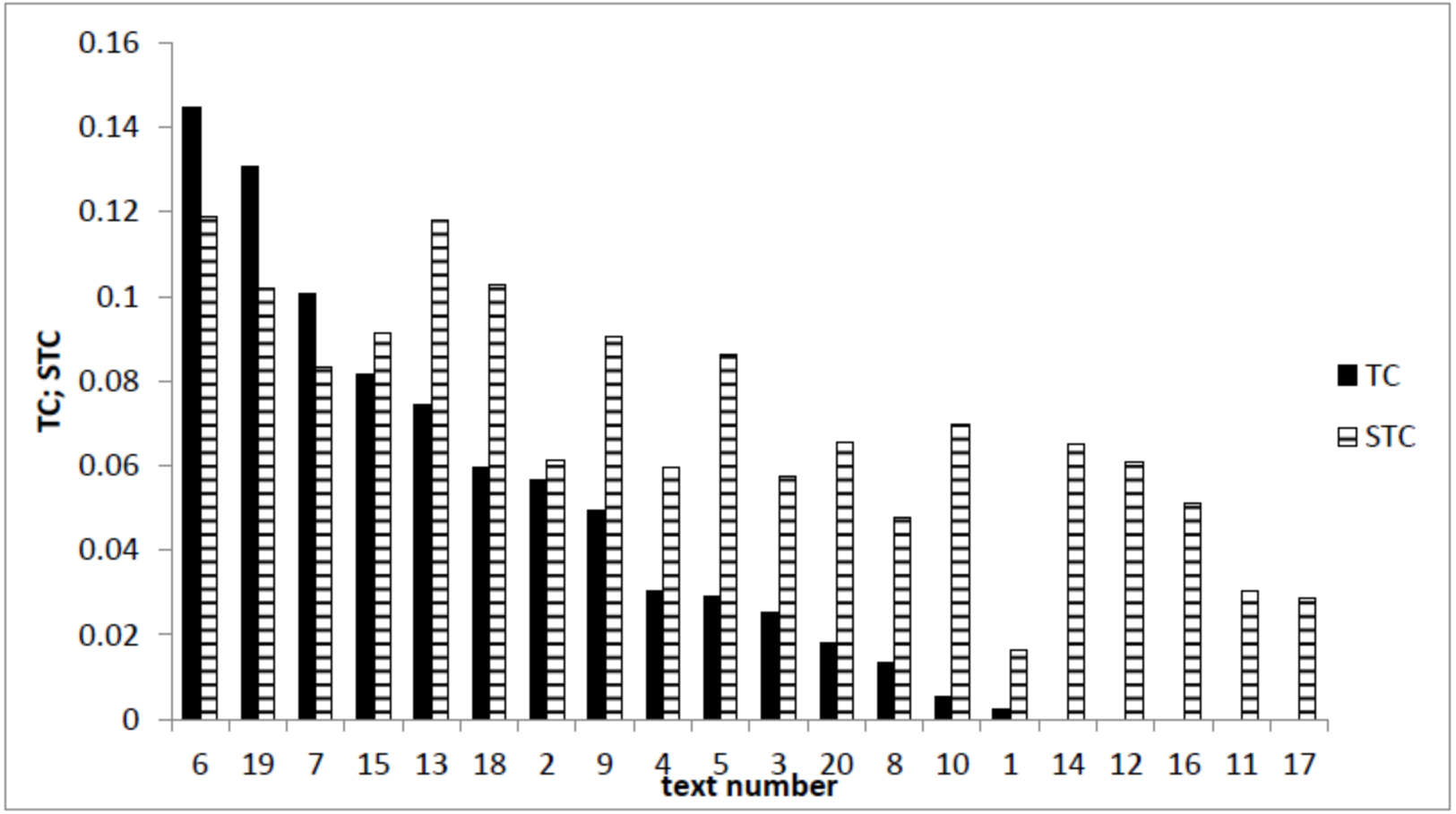


Figure 3. *TC* and *STC* in particular texts. Texts are ranked (*x*-axis) in the decreasing order in accordance to *TC*.

QUITA

- software pro kvant. analýzu textů
- dostupný na
  - https://code.google.com/p/oltk/
- stručné charakteristika zde:
  - http://www.cechradek.cz/publ/2014_Kubat_etal_QUITA.pdf