Analýza klíčových slov
versus
tematická koncentrace textu

Radek Čech
(KČJ FF OU)

ÚČNK FF UK
Praha
3. 12. 2013

Analýza klíčových slov

- cíl: detekovat slova, která lze díky relativně vysoké/nízké frekvenci považovat za slova vyjadřující hlavní téma/témata textu
- klíčové slovo – slovo, které se v daném textu používá se signifikantně vyšší nebo nižší frekvencí, než je tomu v referenčním korpusu
- testovým kritériem
  - chí-kvadrát test
  - log-likelihood test

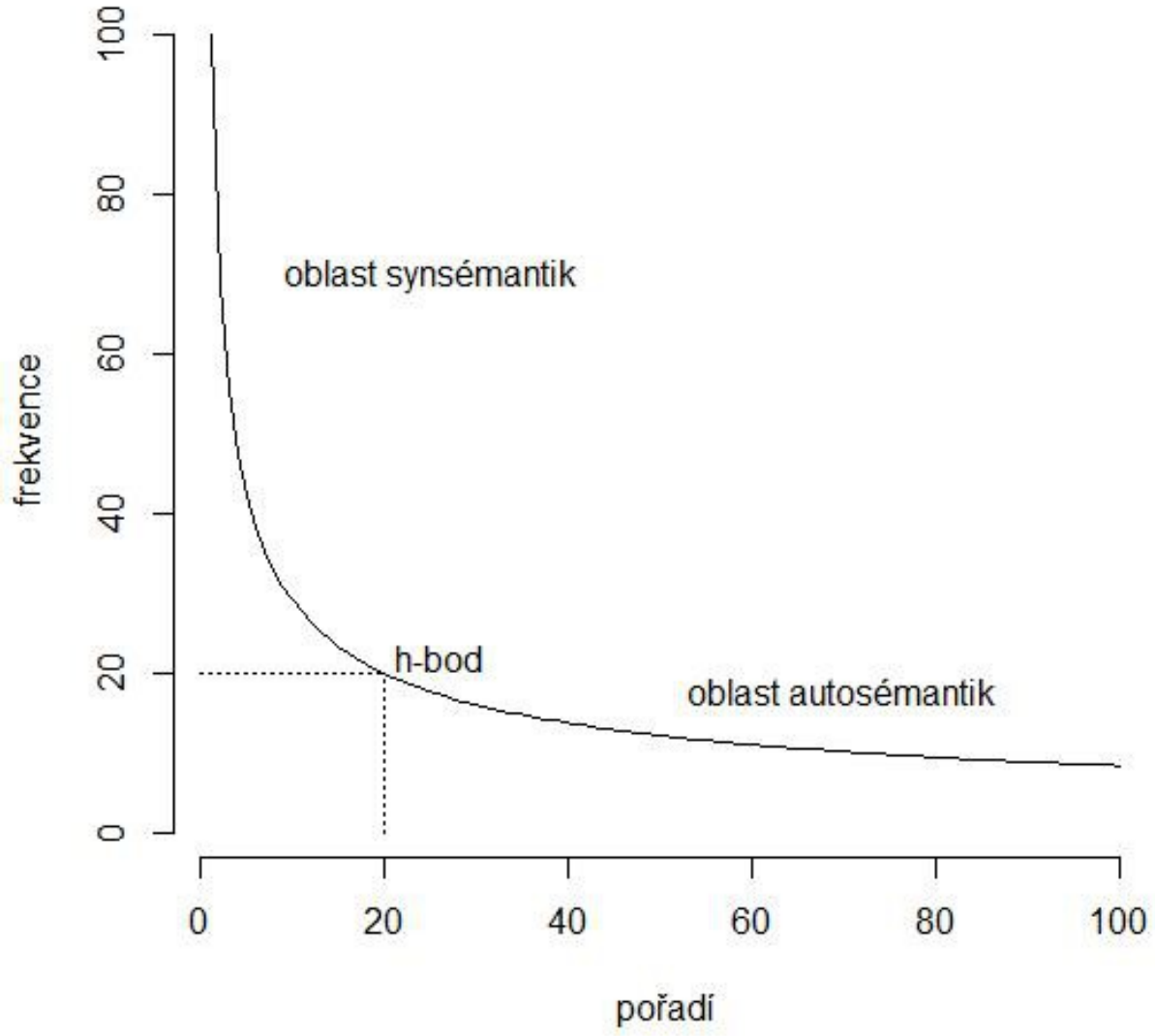
- L. Jehlička *Mládí a poezie* (log-likelihood; p = 0,001; min(f) = 3; all signif. types; stop-list: pron., prep, conj., number; ignore case; ref. corpus: SYN2010)

Form	LL	Fq(text)	Fq(ref)	Form	LL	Fq(text)	Fq(ref)
hrdinstvím	85.144	5	38	*jest*	31.067	4	3346
hrdinství	77.212	6	386	*poezii*	27.974	3	1146
mládí	60.701	7	3713	*ducha*	21.622	3	3328
poezie	46.29	5	1975	*mladí*	19.363	3	4873
katolictví	41.943	3	110	*člověku*	18.805	3	5356
otroctví	33.223	3	476	*lásky*	18.012	3	6128
nejen	31.731	7	30472	*především*	17.619	5	37073
oběti	31.342	4	3232	*pak*	17.489	9	161980
životu	31.223	4	3281	*není*	16.93	8	129953

Tematická koncentrace textu

- cíl: kvantifikovat celkovou charakteristiku textu, přičemž kvantifikace se chápe jako prostředek, který umožní testovat rozdíly mezi jednotlivými texty (příp. skupinami textů); tyto rozdíly mohou být důsledkem vliv žánrů, autorství atd.
- detekce jednotlivých slov a kvantifikace jejich tzv. tematické váhy „jen“ dílčím krokem
- TK založena na frekvenční struktuře textu a tzv. h-bodu

Tematická koncentrace textu



- L. Jehlička *Mládí a poezie; h*-bod = 9; sekundární TK: 2*h* = 18

pořadí	slovo	frekvence	pořadí	slovo	frekvence
1	*a*	53	11	*na*	8
2	*v*	21	12	*li*	8
3	*je*	19	13	*není*	8
4	*k*	15	14	*však*	8
5	*se*	15	15	*nejen*	7
6	*ale*	15	16	***mládí***	7
7	*že*	13	17	*s*	7
8	*aby*	10	18	*hrdinství*	6
9	*pak*	9	19	*bez*	6
10	*i*	9	20	*bylo*	6

- L. Jehlička *Mládí a poezie; h*-bod = 10,5

pořadí	lemma	frekvence	pořadí	lemma	frekvence
1	*být*	55	11	*ten*	10
2	*a*	53	12	*jenž*	10
3	*v*	24	13	*tento*	10
4	*se*	21	14	*aby*	10
5	*k*	15	15	*pak*	9
6	*ale*	15	16	*i*	9
7	*že*	13	17	*poezie*	8
8	*on*	13	18	*však*	8
9	***člověk***	13	19	*na*	8
10	***hrdinství***	11	20	*život*	8

Tematická váha slova

- tematická váha slova

$$(h-r')\cdot f(r')$$

- normalizace

$$C=\sum_{r=1}^{h}(h-r)\cdot f(1)$$

- index tematické váhy slova

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

Tematická váha slova

- index tematické váhy lemmatu *člověk*
  - $h = 10{,}5$; $r = 8$; $f = 13$

$$TV_{člověk} = 2\frac{(h - r')\cdot f(r')}{h(h-1)\cdot f(1)} = 2\frac{(10{,}5-8)\cdot 13}{10{,}5(10{,}5-1)\cdot 55} = 0{,}011848$$

KW vs. TK

- společné znaky
  - analýza lexikálních charakteristik textu
  - intersubjektivita
  - automatická detekce jednotek reprezentujících hlavní téma/témata textu

Problémy KW & TK

- KW & TK
  - volba jazykových jednotek
    - slovní forma, lemma, koreferenční jednotka, hreb...
  - vliv délky textu
- KW
  - vliv referenčního korpusu
  - volba hladiny významnosti
- TK
  - v mnohých případech se nevyskytuje nad *h*-bodem žádné autosémantické slovo

TK

- cíl: kvantifikovat celkovou charakteristiku textu, přičemž kvantifikace se chápe jako prostředek, který umožní testovat rozdíly mezi jednotlivými texty (příp. skupinami textů); tyto rozdíly mohou být důsledkem vliv žánrů, autorství atd.

$$TK_{text} = \sum_{r'=1}^{T} 2\frac{(h - r') \cdot f(r')}{h(h-1) \cdot f(1)}$$

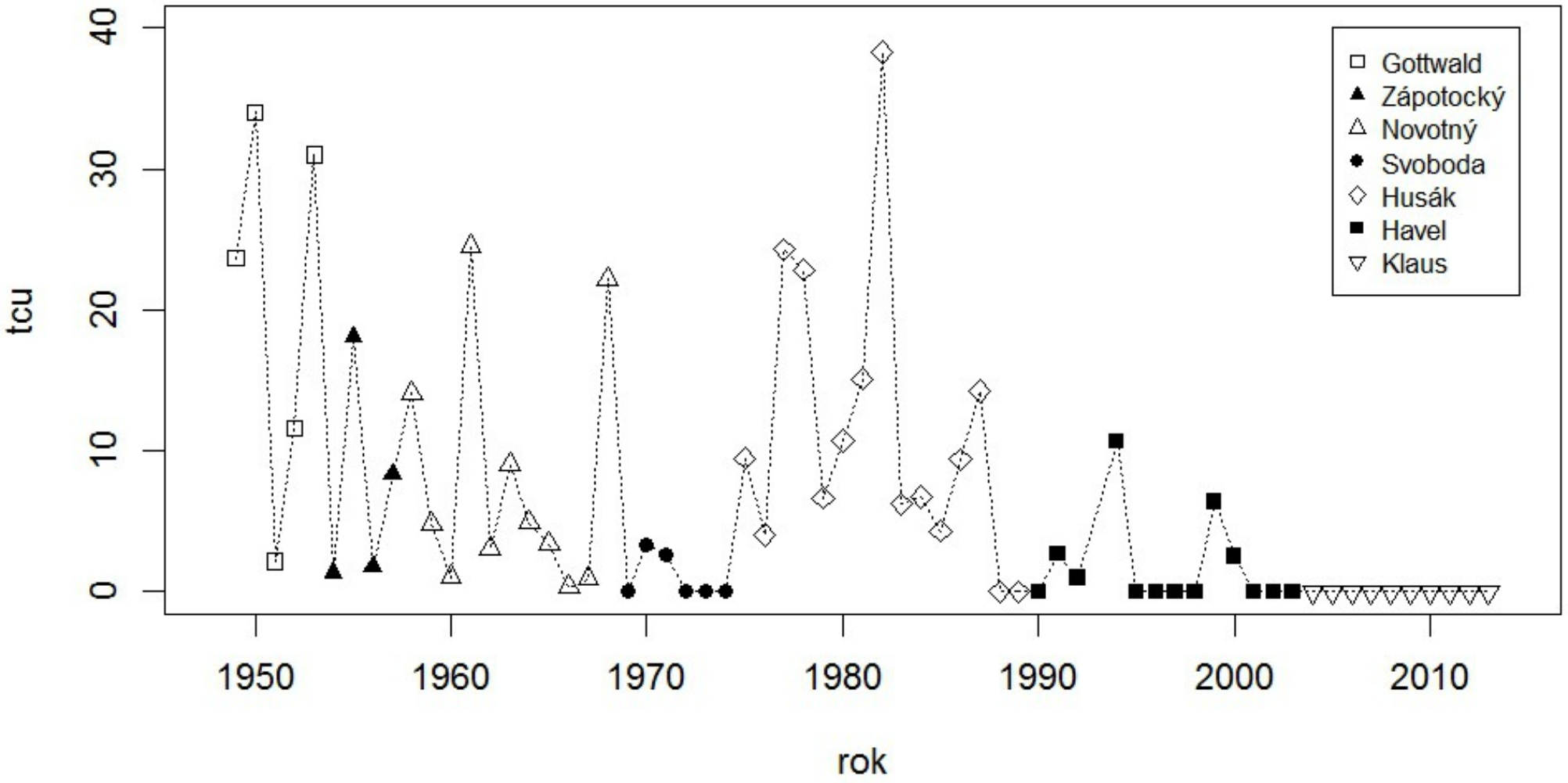
TK

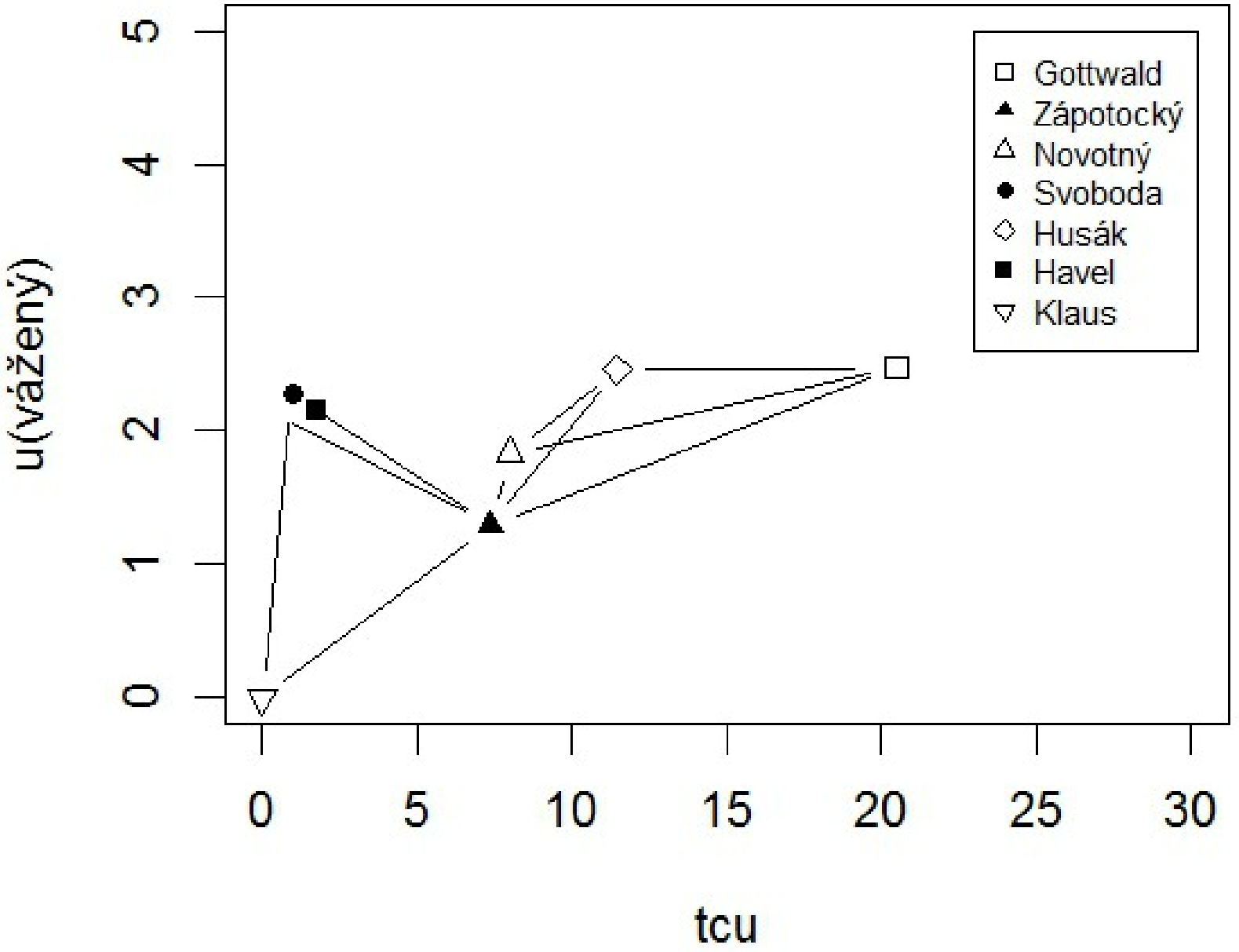
$$VAR(TK) = \left(\frac{2}{h(h-1)f(1)}\right)^2 \cdot \left(\sum_{r'=1}^{T} f(r')\right) \cdot m_{2,r'}$$

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

$$m_{1,r'} = \frac{\sum r' \cdot f(r')}{f(r')}$$

$$|u| = \frac{TK_1 - TK_2}{\sqrt{VAR(TK_1) + VAR(TK_2)}}$$


tcu
0
10
20
30
40
1950
1960
1970
1980
1990
2000
2010
rok
Gottwald
Zápotocký
Novotný
Svoboda
Husák
Havel
Klaus


u(vážený)
0
1
2
3
4
5
0
5
10
15
20
25
30
tcu
Gottwald
Zápotocký
Novotný
Svoboda
Husák
Havel
Klaus

KW vs. TK

KW

kvantifikace

↓

matematická formalizace

↓

statistická interpretace

↓

popis

↓

explanace

TK

hypotéza

↓

kvantifikace

↓

matematická formalizace

↓

experiment

↓

statistická interpretace

↓

explanace

Korpusová vs. kvantitativní lingvistika

KL

kvantifikace

↓

matematická formalizace

↓

statistická interpretace

↓

popis

↓

explanace

QL

(teorie)

↓

hypotéza

↓

kvantifikace

↓

matematická formalizace

↓

experiment

↓

statistická interpretace

↓

explanace

Korpusová vs. kvantitativní lingvistika

KL

kvantifikace

↓

matematická formalizace

↓

statistická interpretace

↓

popis

↓

explanace

QL

(teorie)

↓

hypotéza

↓

kvantifikace

↓

matematická formalizace

↓

experiment

↓

statistická interpretace

↓

explanace

Děkuji za pozornost!

(www.cechradek.cz)