Tematická koncentrace textu

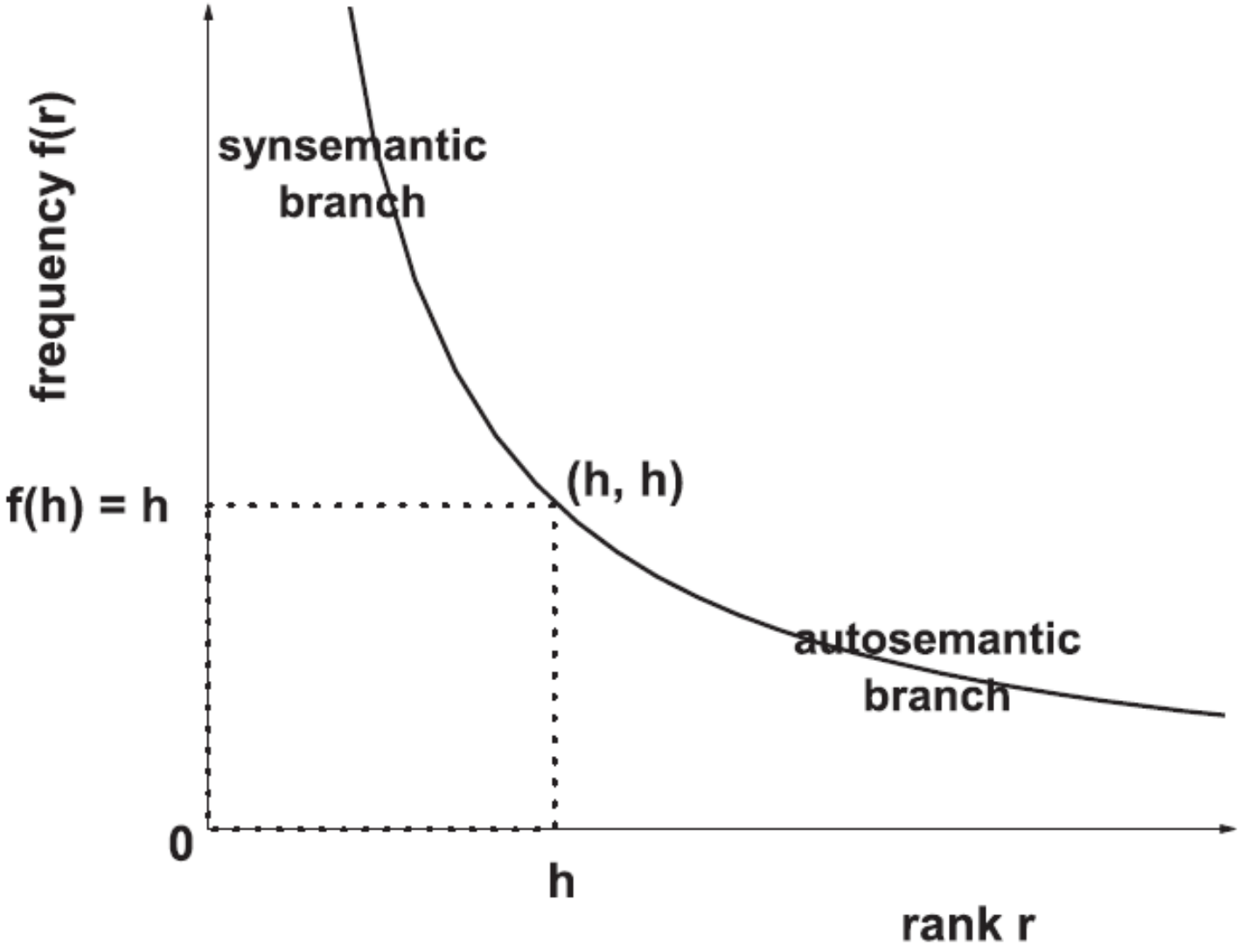
Radek Čech

Cíle

- pomocí kvantitativní textové analýzy
  - odhalit ústřední téma textu (pokud existuje)
  - zjistit tematickou „sílu“ jednotlivých slov reprezentujících témata textu
  - zjistit tematickou koncentraci celého textu

Metodologie

- Popescu et al. (2009)
  - tematická slova → substantiva, verba, adjektiva
  - východiskem frekvenční charakteristika textu
    - h-bod → určuje fuzzy hranici mezi synsémantickými a autosémantickými slovy

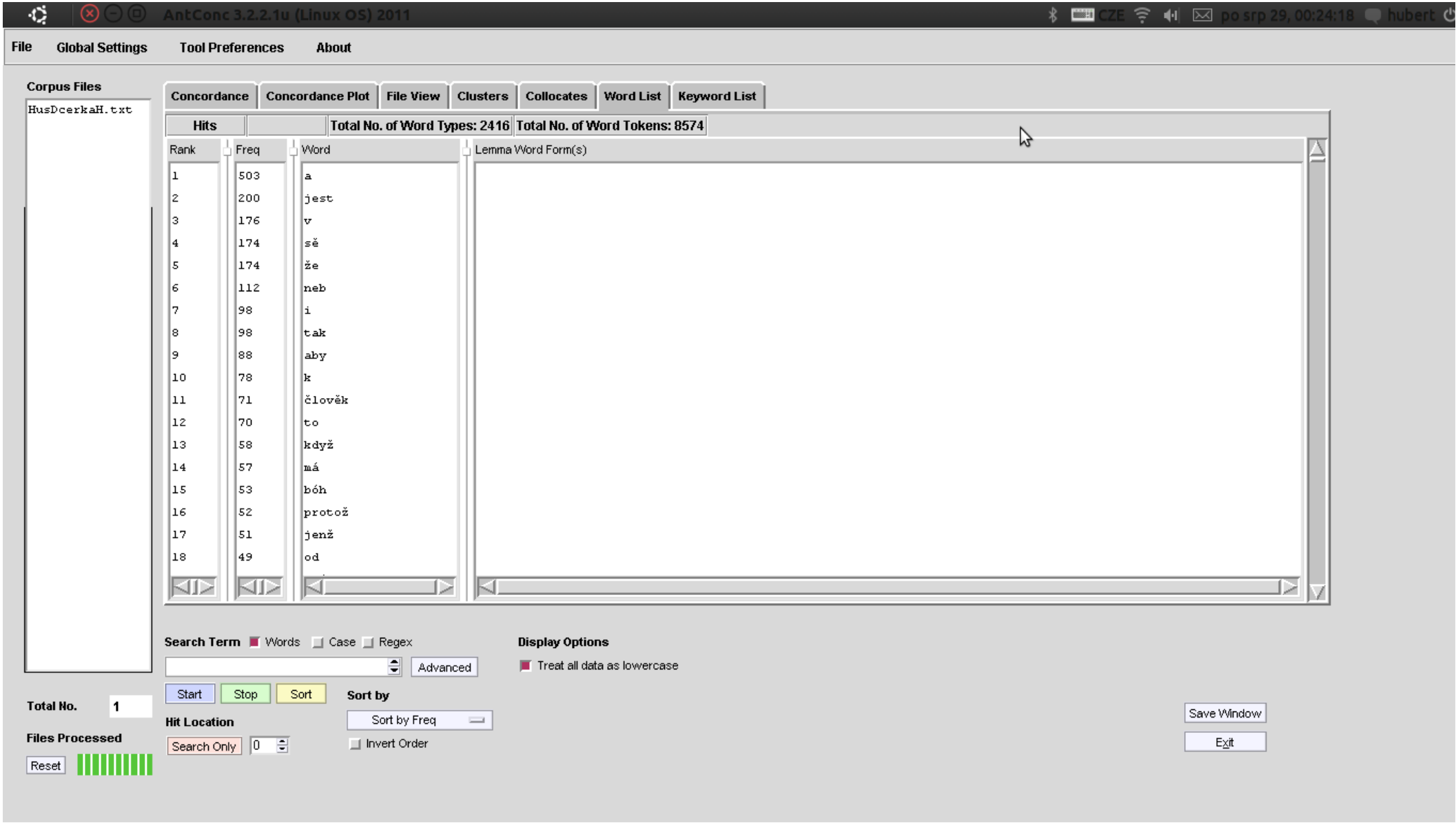

frequency f(r)
synsemantic
branch
f(h) = h
(h, h)
autosemantic
branch
0
h
rank r

Metodologie

- Popescu et al. (2009)
  - tematická slova → substantiva, verba, adjektiva
  - východiskem frekvenční charakteristika textu
    - h-bod → určuje fuzzy hranici mezi synsémantickými a autosémantickými slovy
  - tematická slova na h-bodem vyjadřují hlavní téma textu (lepší výsledky u lemmatizovaných textů)

Postup

- text → J. Hus *Dcerka*
  - (text poskytlo Oddělení vývoje jazyka UJČ)
- vytvoření frekvenční distribuce slovních tvarů
  - AntConc
    - http://www.antlab.sci.waseda.ac.jp/antconc_index.html


AntConc 3.2.2.1u (Linux OS) 2011
CZE
po srp 29, 00:24:18
hubert
File
Global Settings
Tool Preferences
About
Corpus Files
HusDcerkaH.txt
Concordance
Concordance Plot
File View
Clusters
Collocates
Word List
Keyword List
Hits
Total No. of Word Types: 2416
Total No. of Word Tokens: 8574
Rank
Freq
Word
Lemma Word Form(s)
1 503 a
2 200 jest
3 176 v
4 174 sě
5 174 že
6 112 neb
7 98 i
8 98 tak
9 88 aby
10 78 k
11 71 člověk
12 70 to
13 58 když
14 57 má
15 53 bóh
16 52 protož
17 51 jenž
18 49 od
Search Term
Words
Case
Regex
Advanced
Start
Stop
Sort
Sort by
Sort by Freq
Invert Order
Hit Location
Search Only
0
Display Options
Treat all data as lowercase
Total No.
1
Files Processed
Reset
Save Window
Exit

Frekvenční distribuce

pořadí	f	slovo	pořadí	f	slovo	pořadí	f	slovo
1	503	a	12	70	to	**23**	**41**	**die**
2	200	jest	13	58	když	24	41	ho
3	176	v	14	57	má	25	41	na
4	174	sě	**15**	**53**	**bóh**	26	39	ale
5	174	že	16	52	protož	27	38	jeho
6	112	neb	17	51	jenž	28	37	z
7	98	i	18	49	od	**29**	**36**	**boha**
8	98	tak	19	48	své	30	35	jako
9	88	aby	20	47	by	31	33	jakož
10	78	k	21	47	s	32	33	ty
11	**71**	**člověk**	22	43	li	**33**	**30**	**ďáble**

H-bod

$$h = f(h)$$

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

$$h_{Dcerka} = \frac{33 \cdot 33 - 30 \cdot 32}{33 - 32 + 33 - 30} = 32.25$$

Index tematické koncentrace

- *r´*.... pořadí slova před h-bodem
- tematická váha: *h-r´*
  - slova s nejnižším rankem mají největší váhu
- relativizace
  - tematická váha daného slova vynásobená jeho vlastní frekvencí je podělena součtem všech vah a nejvyšší možnou frekvencí → index tematické koncentrace slova

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

Postup

- výpočet normalizační konstanty *C*

$$\frac{2}{h(h-1)\,f(1)} = \frac{2}{32.25(31.25)503} = 0.00000394532 = C$$

- *člověk* (*r‘ = 11, f = 71*)

$$(h-r')\,f(r')C = (32.25-11)71 \cdot C = 0.005952502$$

Tematická koncentrace textu

slovo	*r*	*f*	*(h-r')f(r')C*
člověk	11	71	0.00592
bóh	15	53	0.00361
die	23	41	0.00150
boha	29	36	0.00046
TC			0.01152

$$tcu = 1000(TC)$$

$$tcu_{Dcerka} = 1000(0.01152) = 11.52$$

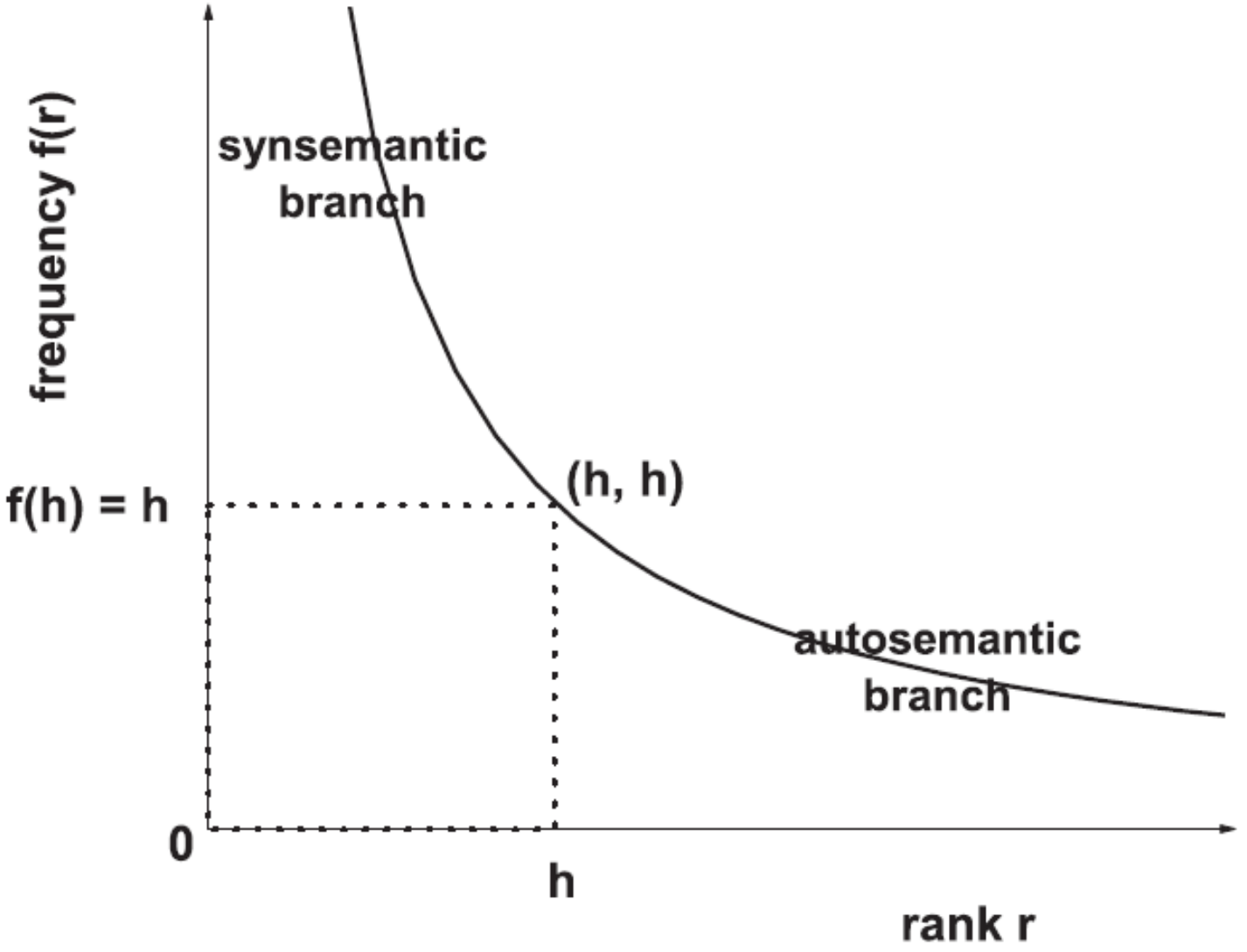
Tematická koncentrace textu

- k čemu je to dobré?
  - jednoduchá metoda umožňující zjistit, o čem daný text je a „jak moc“ se týká centrálního tématu (případně témat)
  - kvantifikace umožňuje porovnávat texty (za pomocí statistických testů)

- nedostatek
  - pokud žádné tematické slovo nad h-bodem, TC=0 → všechny texty s TC=0 jsou si rovny

Index negativní TC

- východisko
  - h-bod fuzzy hranice
  - pokud text tematicky „roztříštěn“ → tematická slova mají relativně malou frekvenci
  - čím je text tematicky „roztříštěnější“, tím větší vzdálenost mezi h-bodem a prvním tematickým slovem (pořádí dáno frekvencí)


frequency f(r)
synsemantic
branch
f(h) = h
(h, h)
autosemantic
branch
0
h
rank r

Index negativní TC

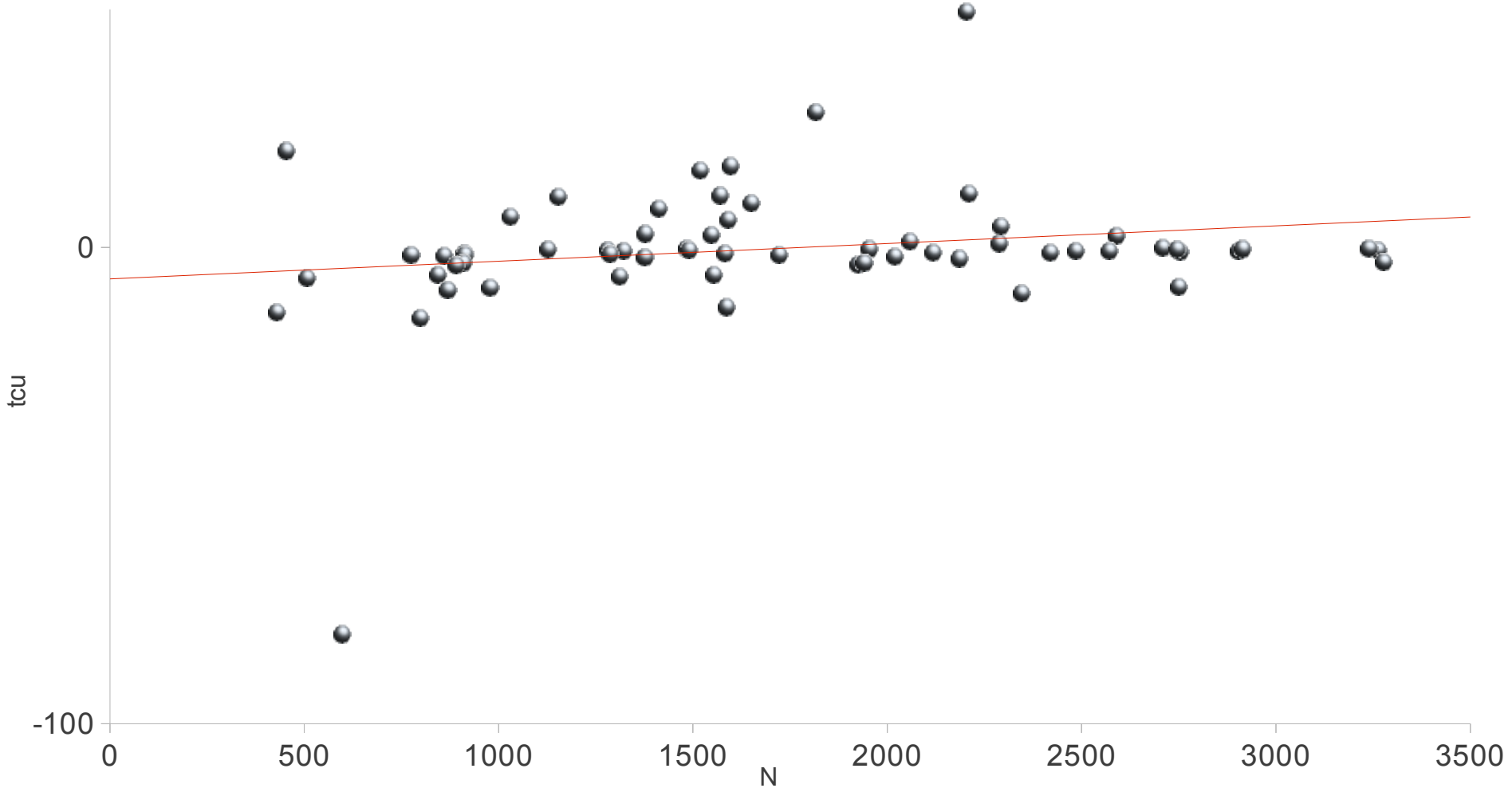
- postup
  - u textů *TC*=0 hledáno první tematické slovo za h-bodem

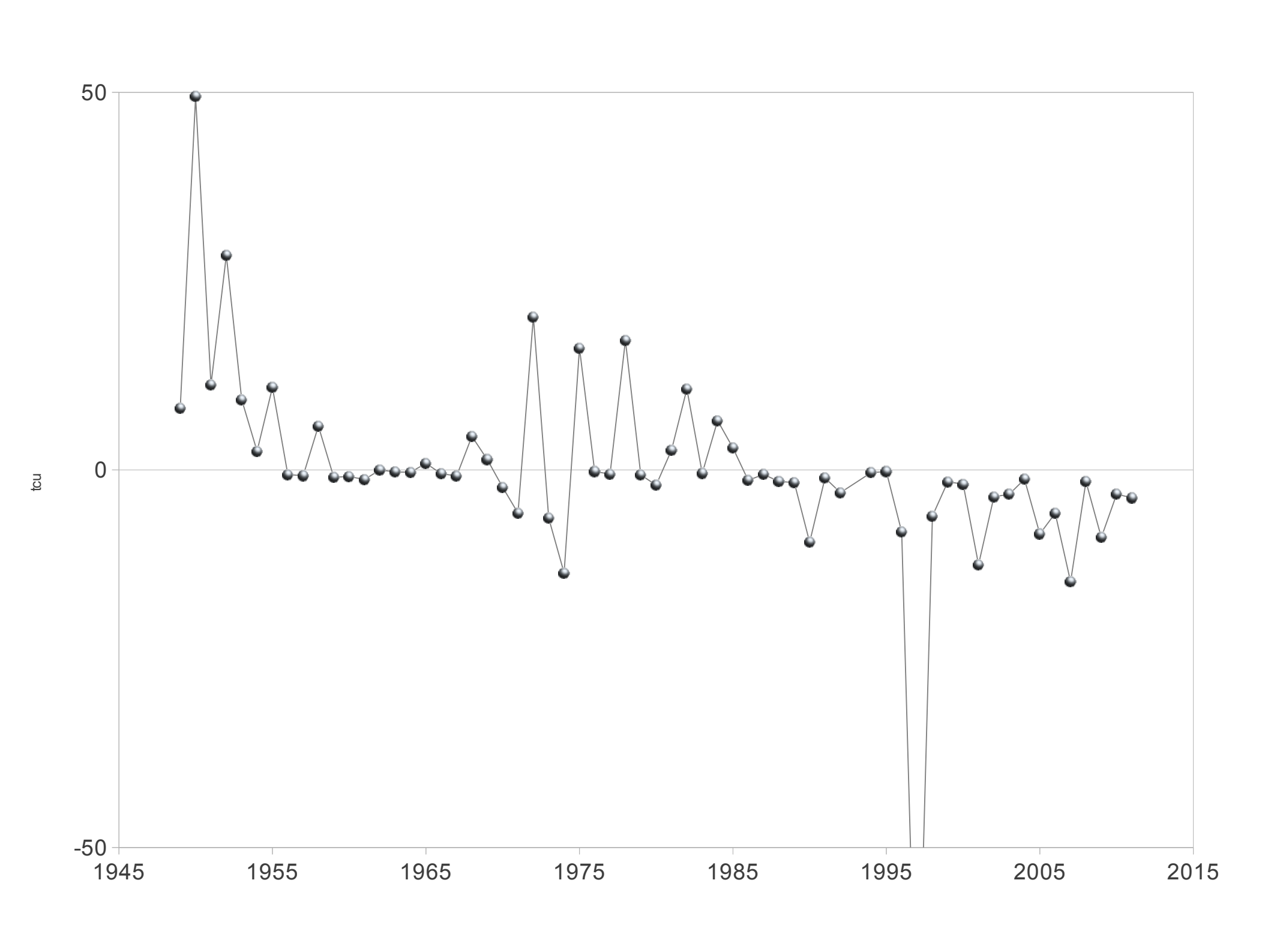
$$TC_{\text{neg}} = (h - r'') \frac{h}{f(r'')} C$$

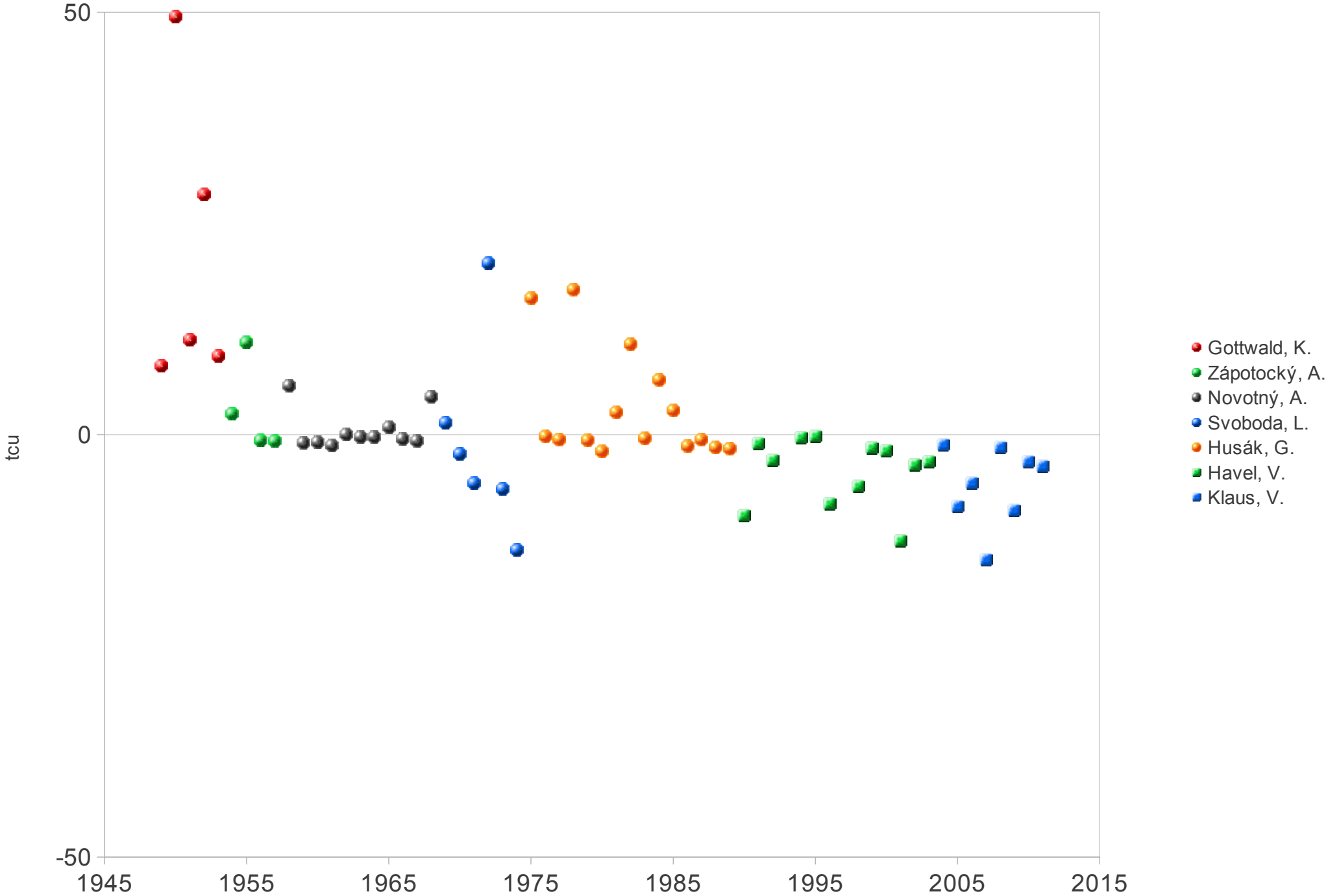
Novoroční projevy československých a českých prezidentů

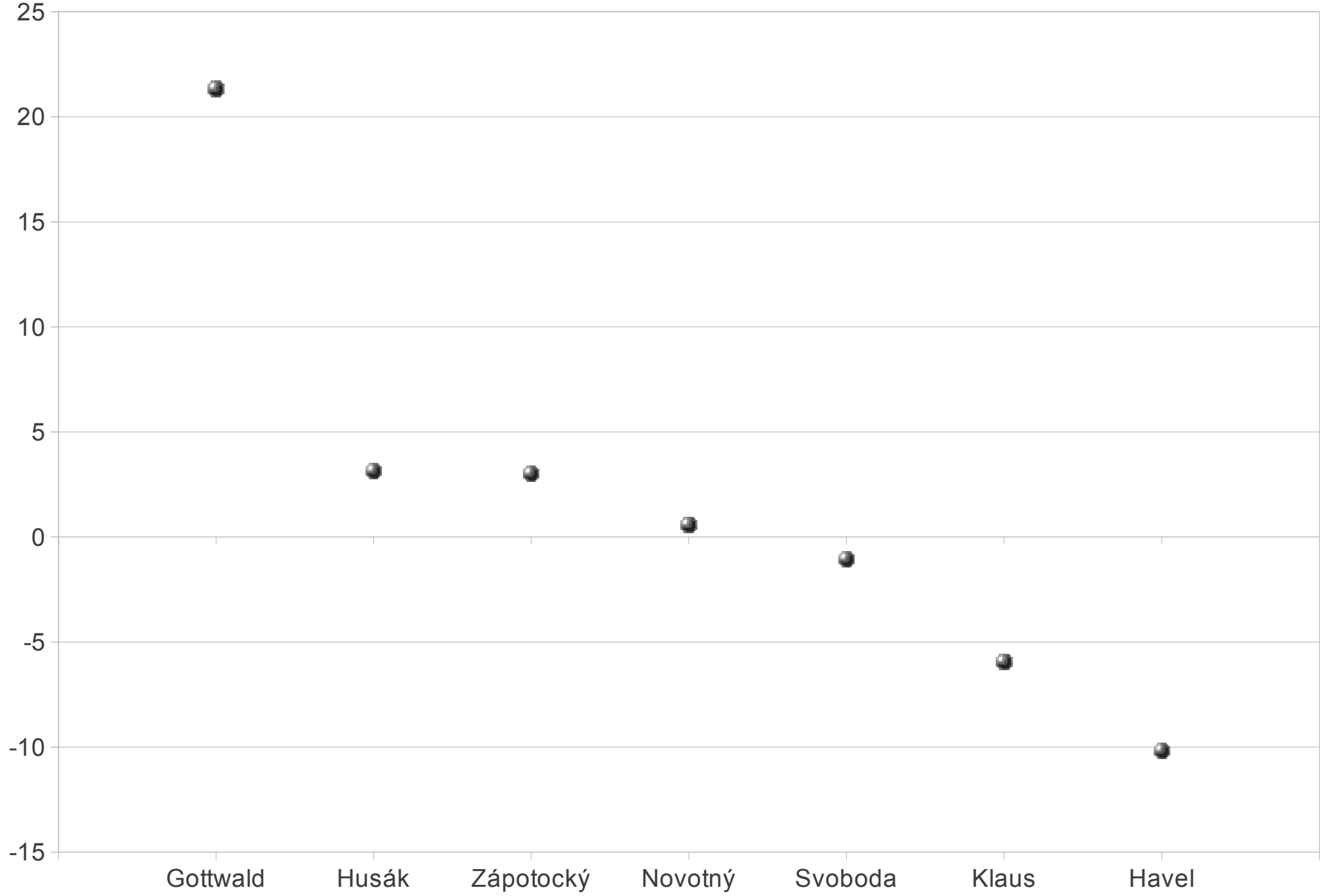
- 1949 – 2011 (kromě r. 1993)
- relativně stabilní žánr

Délka textu versus TC

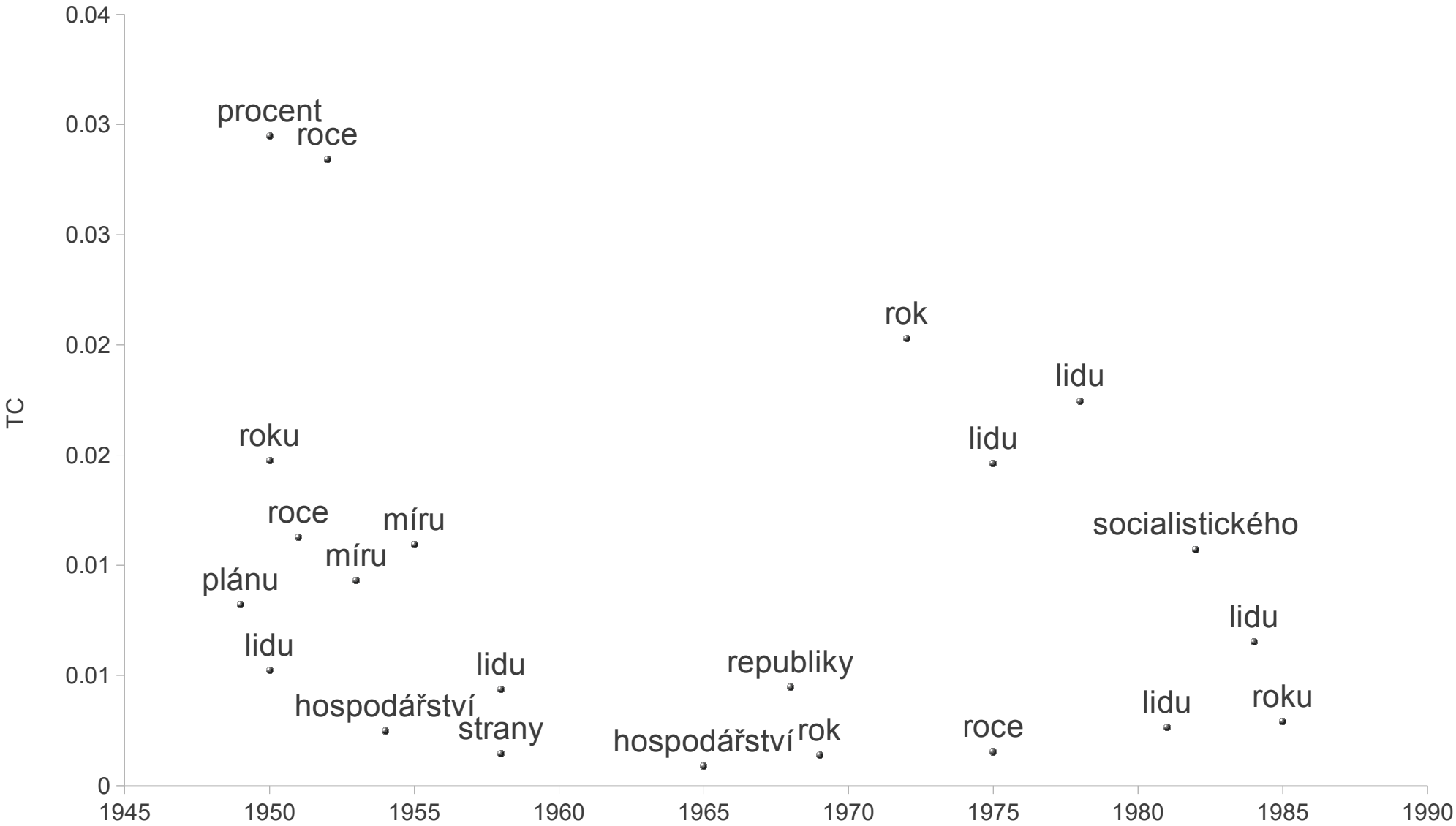


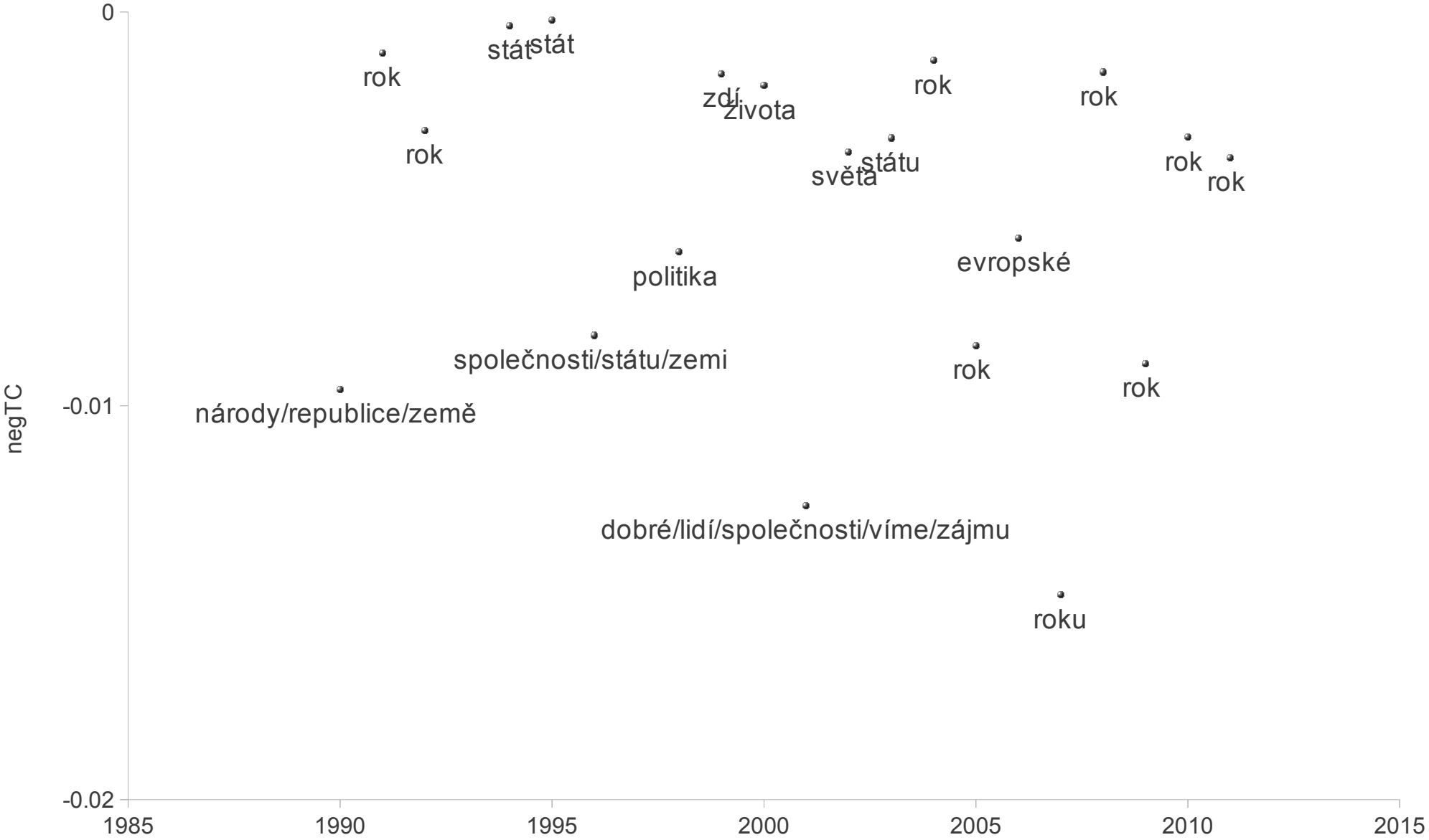

50
0
-50
tcu
1945
1955
1965
1975
1985
1995
2005
2015


50
0
-50
tcu
1945
1955
1965
1975
1985
1995
2005
2015
Gottwald, K.
Zápotocký, A.
Novotný, A.
Svoboda, L.
Husák, G.
Havel, V.
Klaus, V.


25
20
15
10
5
0
-5
-10
-15
Gottwald
Husák
Zápotocký
Novotný
Svoboda
Klaus
Havel

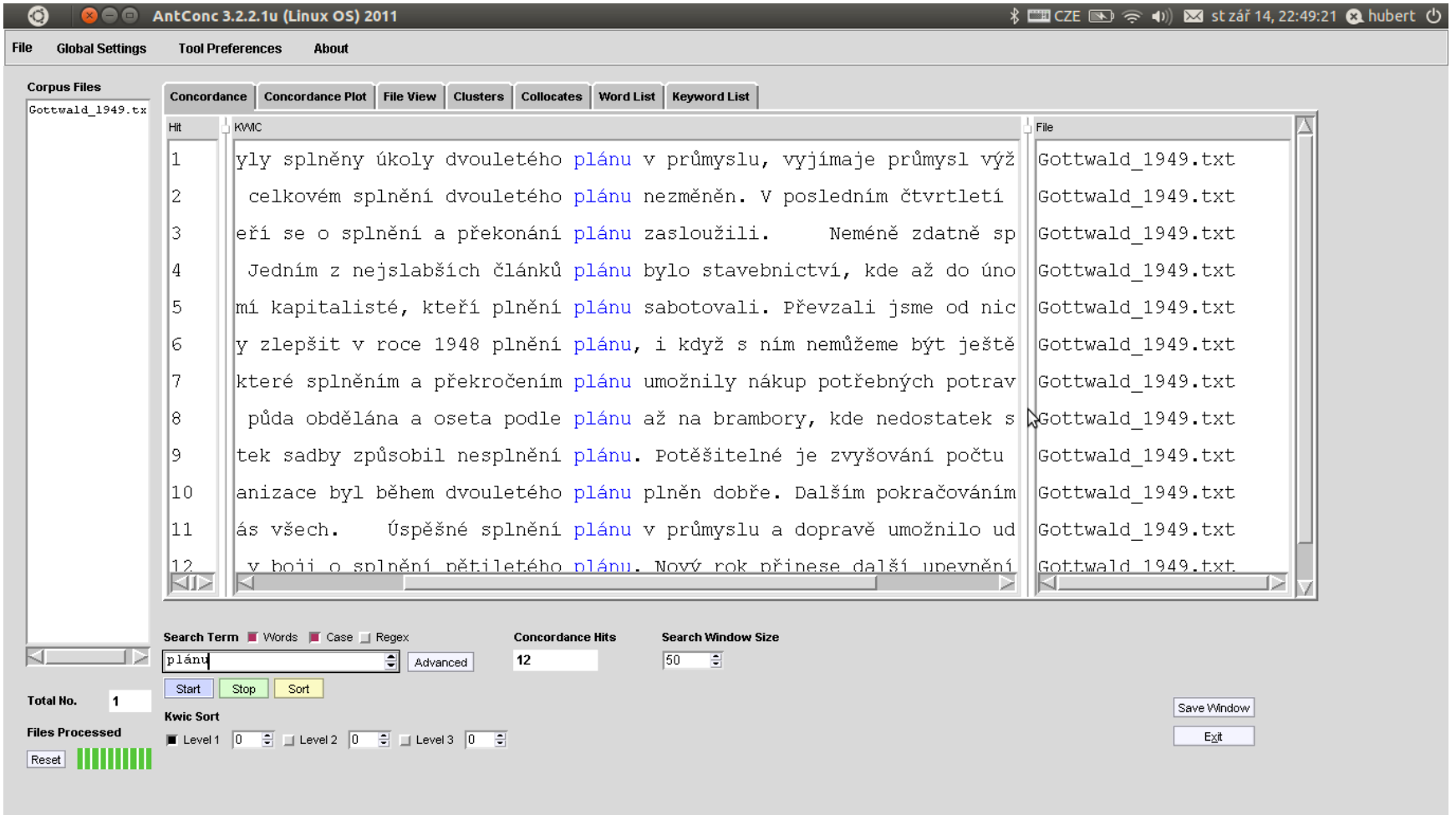
Tematická slova




0
-0.01
-0.02
negTC
1985
1990
1995
2000
2005
2010
2015
stát
stát
rok
rok
zdí
života
rok
rok
světa
státu
rok
rok
politika
evropské
společnosti/státu/zemi
rok
rok
národy/republice/země
dobré/lidí/společnosti/víme/zájmu
roku

Tematická slova

- AntConc
  - možnost detailní analýzy textu



Výhledy

- navrhnout statistické testy pro testování rozdílu mezi
  - jednotlivými texty
  - obdobími (např. totalita x demokracie)
  - autory

Děkuji za pozornost!

radek.cech@osu.cz
www.cechradek.cz