Úvod do kvantitativní lingvistiky

ZS 2022

Tematická koncentrace textu (TC)

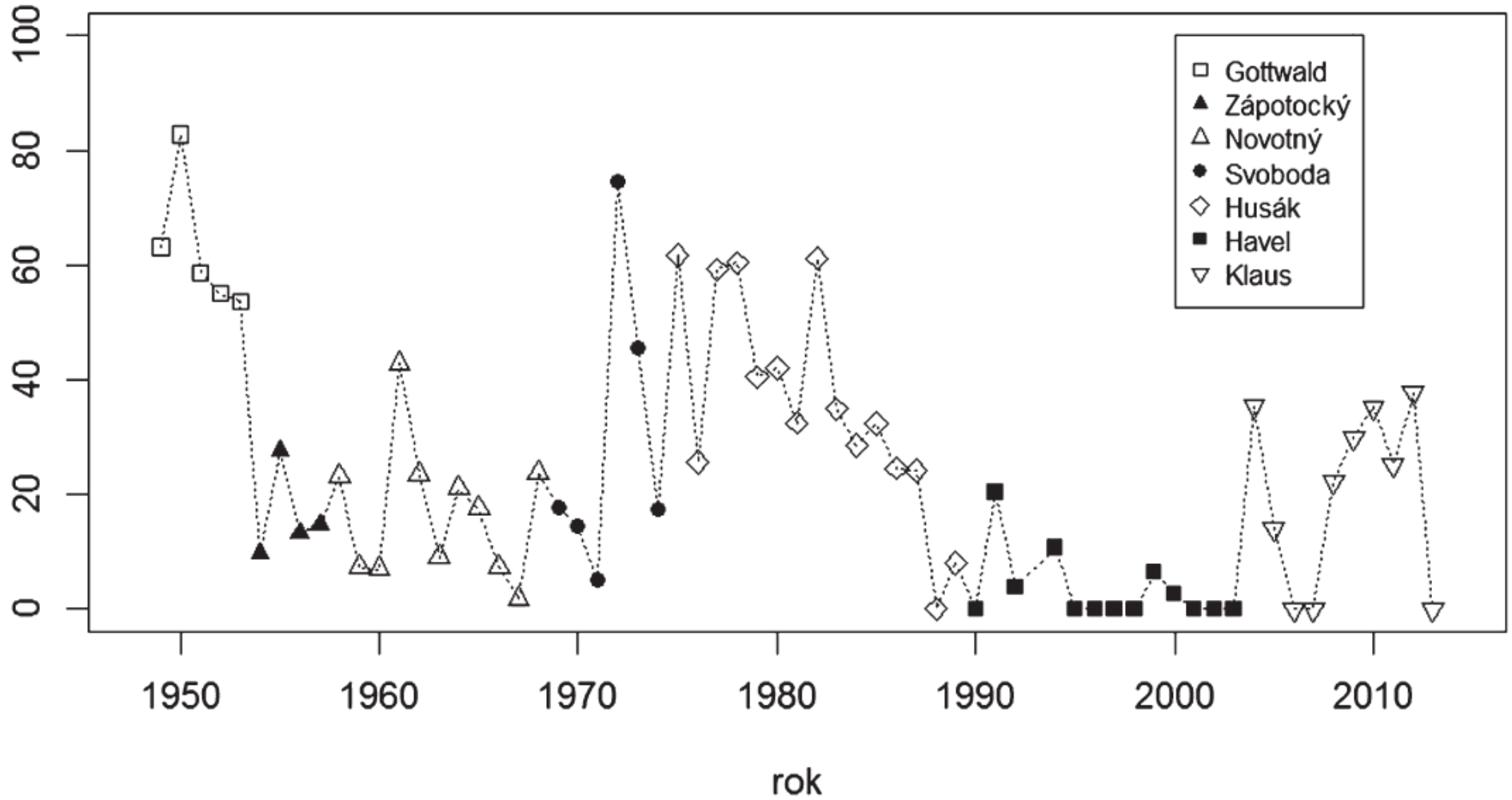
- vyjadřuje míru zaměřenosti textu na centrální téma/témata
- předpoklady
  - v různých textech se autor na dané téma či témata může zaměřovat s různou intenzitou;
  - lze identifikovat jazykové jednotky, které lze chápat jako nositele určitého tématu či témat;
  - míru zaměření se na dané téma či témata je možné detekovat analýzou frekvenčních charakteristik textu;
  - 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.

Opakování 1

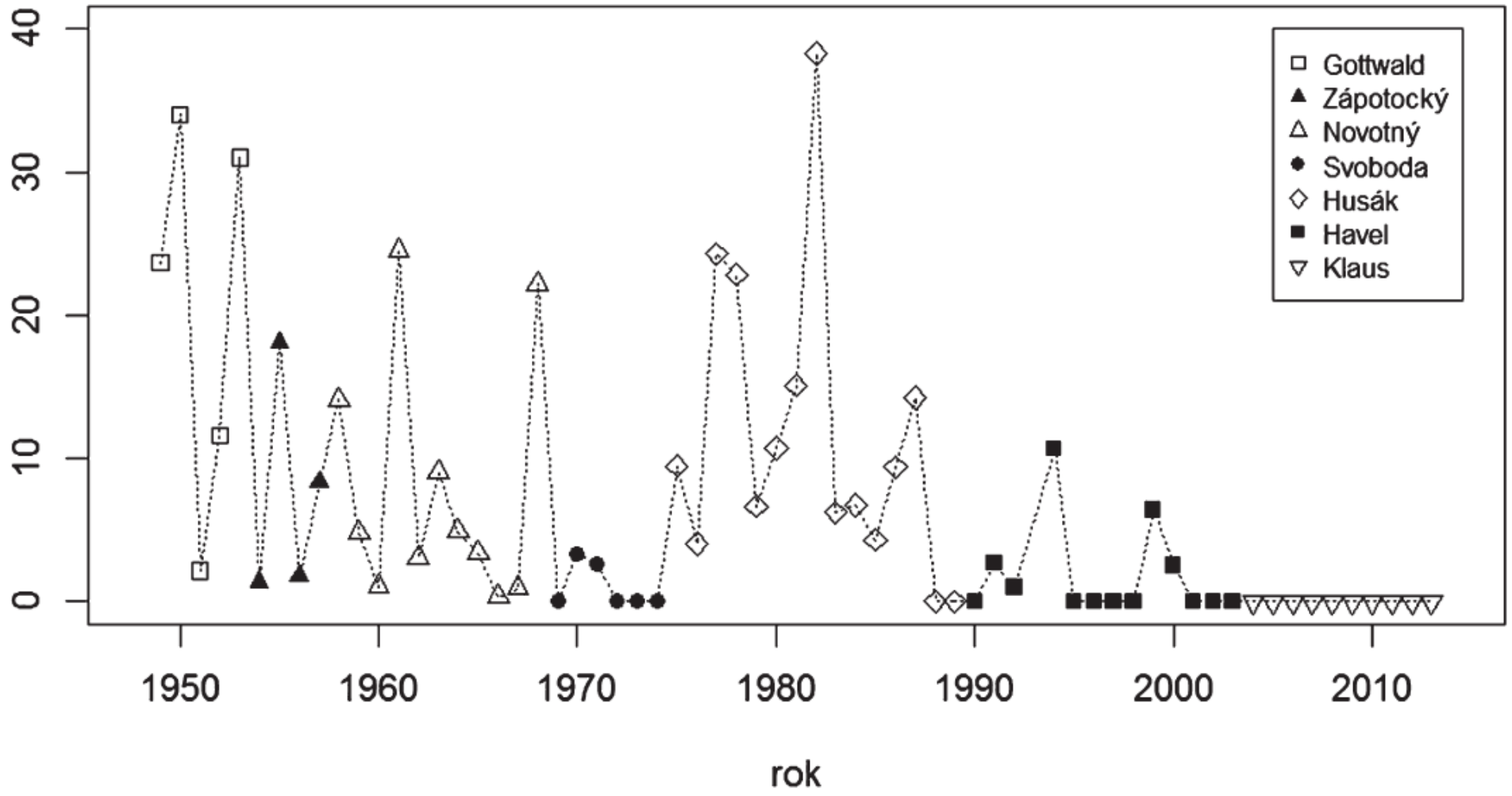
- z dat v souboru 221102_seminar_TC_pro vypocet.xlsx vypočítejte

  - h-bod
  - tematické váhy autosémantických slov
  - tematickou koncentraci daného textu

TK


Gottwald
Zápotocký
Novotný
Svoboda
Husák
Havel
Klaus
0
20
40
60
80
100
1950
1960
1970
1980
1990
2000
2010
rok

TK bez lemmatu *rok*

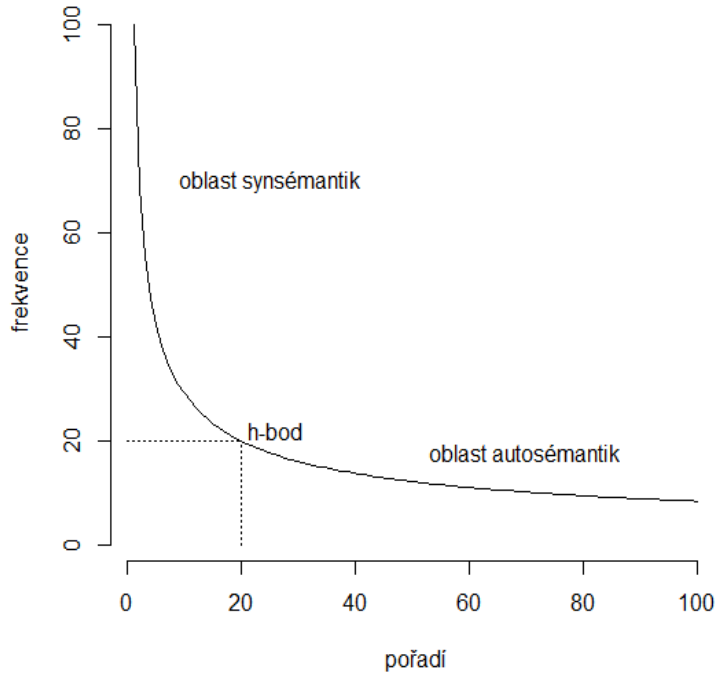


Jiné způsoby měření TK

- s. 28nn

Sekundární TK

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



Proporcionální TK

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

TK, STK, PTK, SPTK

- měří to samé?

- jak to zjistit?

- Čech, R., Garabík, R., Altmann, G. (2015). Testing the thematic concentration of text. Journal of Quantitative Linguistics, 22, 215-232.

Table 6. Correlation coefficients (r) between particular indicators.

	r
TC – PTC	**0.8584**
TC – STC	**0.7897**
STC – PTC	**0.7583**
SPTC – STC	**0.7044**
SPTC –TC	0.3308
SPTC –PTC	0.2450

- dále viz https://www.cechradek.cz/publ/2015_Cech_Garabik_Altmann_Testing_TC.pdf

1168 textů

	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

- více s. 33nn

Vztahy mezi TK, STK a PTK

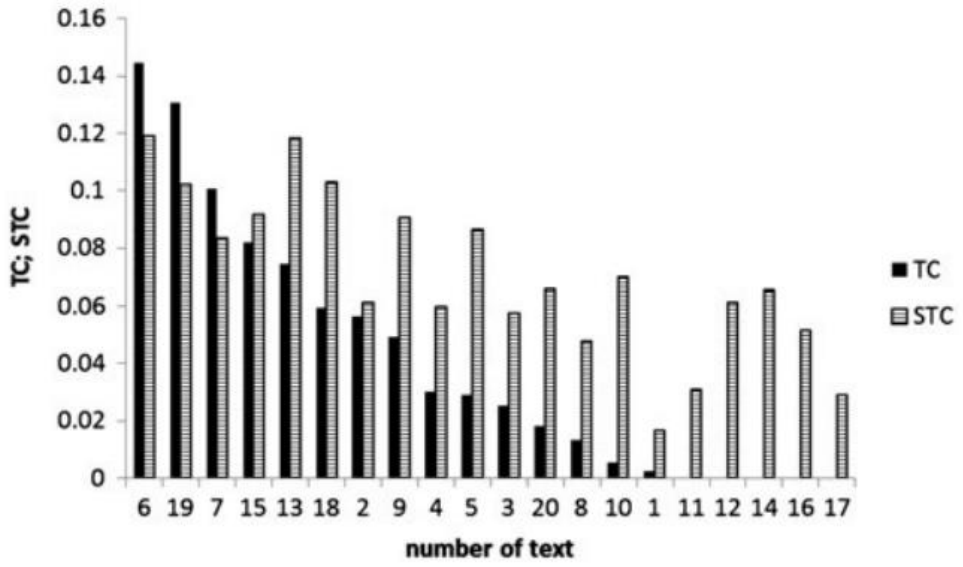


Fig. 8. *TC* and *STC* in particular texts. Texts are ranked (*x*-axis) in decreasing order in accordance to *TC*.

- interpretujte

Vztahy mezi TK, STK a PTK

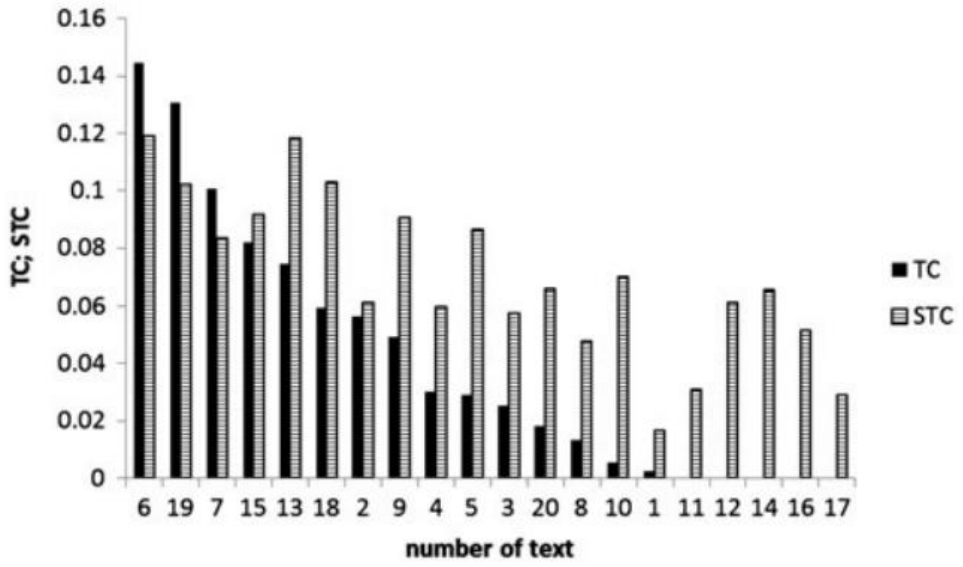


Fig. 8. *TC* and *STC* in particular texts. Texts are ranked (*x*-axis) in decreasing order in accordance to *TC*.

- …a specific tendency for the relationship between the TC and STC. Particularly, for texts with the highest TC, STC < TC, while for texts with the lower TC, STC > TC

Vztahy mezi TK, STK a PTK

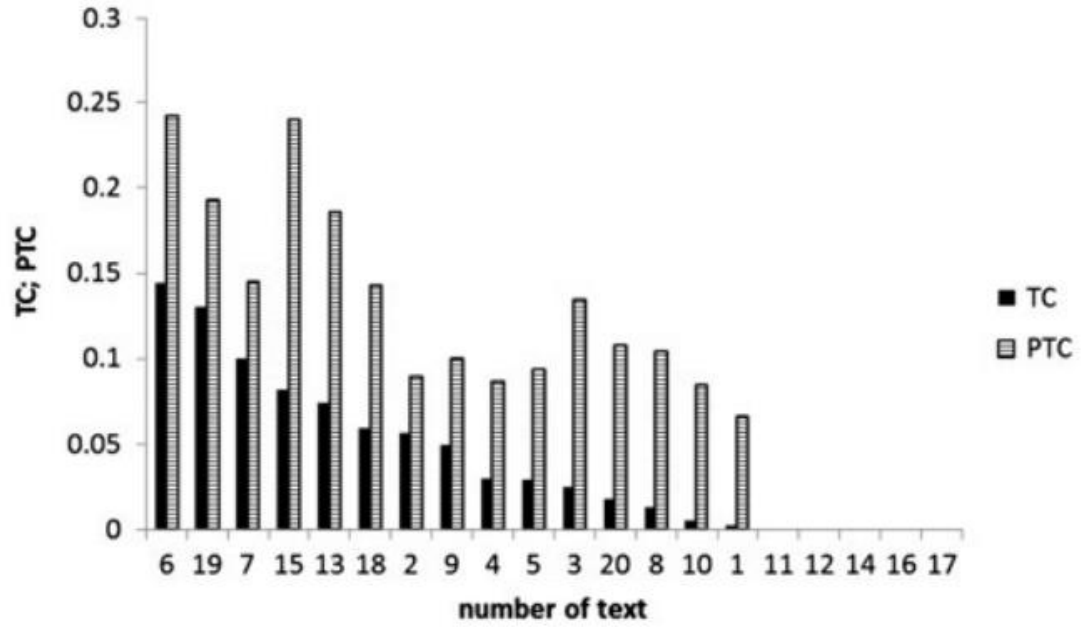


Fig. 9. *TC* and *PTC* in particular texts. Texts are ranked (*x*-axis) in decreasing order in accordance to *TC*.

Jazykové jednotky pro měření TK

- ???

Jazykové jednotky pro měření TK

- jak by se měla volba mezi slovním tvarem a lemmatem projevit na hodnotách TK, STK, PTK?

Jazykové jednotky pro měření TK

- jak by se měla volba mezi slovním tvarem a lemmatem projevit na hodnotách TK, STK, PTK?

Jazykové jednotky pro měření TK

- menší počtu textů s nulovou hodnotou tematické koncentrace,
- vyšší hodnota TK, STK a PTK lemmatizovaných textů.

Jazykové jednotky pro měření TK

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

	TK = 0	TK > 0
nelemmatizované	696	472
lemmatizované	486	682

- interpretujte

Jazykové jednotky pro měření TK

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

	TK = 0	TK > 0
nelemmatizované	696	472
lemmatizované	486	682

- interpretujte
- aplikujte vhodný statistický test

Jazykové jednotky pro měření TK

	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

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

- statisticky významné rozdíly

Jazykové jednotky pro měření TK

- když jsou mezi nelemmatizovanými a lemmatizovanými texty statisticky významné rozdíly, může mezi nimi být vztah?
- jak to ověřit?

Jazykové jednotky pro měření TK

- když jsou mezi nelemmatizovanými a lemmatizovanými texty statisticky významné rozdíly, může mezi nimi být vztah?
- jak to ověřit?

Jazykové jednotky pro měření TK

- když jsou mezi nelemmatizovanými a lemmatizovanými texty statisticky významné rozdíly, může mezi nimi být vztah?
- jak to ověřit?

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

- více viz s. 41nn

Koreferenční jednotka

- určete koreferenční jednotku označující Mariii z daného textu:

- Marie byla doma. Po hodině se učesala. Moc se jí to ale nelíbilo. Neměla ale čas, tak vyrazila. Na ulici na ni čekali studenti. „To jsme rádi, že už jste tady, paní učitelko.“ Všichni pak vyrazili cestou dolů. „Vy se máte,“ zaznělo po chvíli.

Koreferenční jednotka

- koreferenční jednotka a TK
  - očekávání

- více viz 43nn

Koreferenční jednotka

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