Stylistika III

ZS 2024

Stylistika

- "In this sense, analysing style means **looking systematically at the formal features of a text** and **determining their functional significance** for the interpretation of the text in question"
  (Jeffries & McIntyre 2010, p. 1)

Aktualizace (foregrounding)

- „**záměrná odchylka** od **standardního** užití jazykových výrazových prostředků“

  (Krčmová, 2017)

Aktualizace (foregrounding)

- „**záměrná odchylka** od **standardního** užití jazykových výrazových prostředků“

- pro analýzu potřebujeme
  - data, která vykazují vlastnosti standardního užití
    - např. referenční korpus, hodnocení uživatelů
  - analyzovaný jev v textu

- srovnáváme vlastnosti analyzovaného jevu v textu a v referenčních datech

Aktualizace (foregrounding)

- „**záměrná odchylka** od **standardního** užití jazykových výrazových prostředků“

- pro analýzu potřebujeme
  - **data, která vykazují vlastnosti standardního užití**
    - např. referenční korpus, hodnocení uživatelů
  - analyzovaný jev v textu

- srovnáváme vlastnosti analyzovaného jevu v textu a v referenčních datech

Aktualizace (foregrounding)

- „**záměrná odchylka** od **standardního** užití jazykových výrazových prostředků“

- pro analýzu potřebujeme
  - **data, která vykazují vlastnosti standardního užití**
    - např. referenční korpus, hodnocení uživatelů
  - **analyzovaný jev v textu**

- srovnáváme vlastnosti analyzovaného jevu v textu a v referenčních datech

Vlastnosti textu vs. referenční data

- text: N = 5000 slov; $f_{jenž}$ = 10; ipm = 2000
- SYN2020: N ≈ 100000 slov; $f_{jenž}$ = 15 673; ipm = 128,65

Klíčovost (keyness)

- jazykový jev, který se v daném textu vyskytuje **statisticky významně** častěji než v referenčním korpusu

- jevy: slovní tvar, lemma, fráze, slovní druh, syntaktická funkce…

Statistická významnost

Statistické testy významnosti

- 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

Statistické testy významnosti

- postulují se dvě hypotézy
  - **nulová hypotéza**:
    tvrzení, které obvykle deklaruje “žádný rozdíl”, tj. nalezený rozdíl je dán variabilitou dat, náhodou
    - (např. mince není falešná; frekvence daného slova se v daných textech/korpusech neliší)

Statistické testy významnosti

- postulují se dvě hypotézy
  - **nulová hypotéza**:
    tvrzení, které obvykle deklaruje “žádný rozdíl”, tj. nalezený rozdíl je dán variabilitou dat, náhodou
    - (např. mince není falešná; frekvence daného slova se v daných textech/korpusech neliší)
  - **alternativní hypotéza**:
    situace, kdy nulová hypotéza neplatí, tj. mezi proměnnými se předpokládá závislost; důležité je přitom nějaké teoretické zdůvodnění

Statistické testy významnosti

- testuje se platnost H0

- hladina významnosti
  - pravděpodobnost, že se zamítne nulová hypotéza, ačkoliv ona platí
  - obvykle 5 % (0,05) nebo 1 % (0,01)
  - p-hodnota (p-value)

- konvence
  - chyba 1. typu (neadekvátní zamítnutí H0, odpovídá hladině významnosti)
  - chyba 2. typu (neadekvátní nezamítnutí H0)

Statistické testy významnosti

- testuje se platnost H0
- zamítnutí H0 **neznamená, že H1 platí**

Statistické testy významnosti

- testuje se platnost H0
- zamítnutí H0 **neznamená, že H1 platí**
- zamítnutí H0 **znamená, že existuje určitá/vysoká pravděpodobnost toho, že naměřený rozdíl není možné vysvětlit vlivem náhody**
- H1 se nikdy **nepotvrzuje** (confirmation), vždy se jedná o **vyvracení** (rejection) H0

Statistické testy významnosti

- hod mincí (100x)
  - 50x panna a 50x orel → podvádí se?

Statistické testy významnosti

- hod mincí (100x)
  - 50x panna a 50x orel → podvádí se?
  - 52x panna a 48x orel → podvádí se?

Statistické testy významnosti

- hod mincí (100x)
  - 50x panna a 50x orel → podvádí se?
  - 52x panna a 48x orel → podvádí se?
  - 98x panna, 2x orel → podvádí se?

Statistické testy významnosti

- hod mincí (100x)
  - 50x panna a 50x orel → podvádí se?
  - 52x panna a 48x orel → podvádí se?
  - 98x panna, 2x orel → podvádí se?
  - 59x panna, 41 orel → podvádí se?

Statistické testy významnosti

- hod mincí (100x)
  - 50x panna a 50x orel → podvádí se?
  - 52x panna a 48x orel → podvádí se?
  - 98x panna, 2x orel → podvádí se?
  - 59x panna, 41 orel → podvádí se?
  - 60x panna, 40 orel → podvádí se?
  - …

Statistické testy významnosti

- pokud padne panna 61x, tak je větší než 95% pravděpodobnost, že jeden z hráčů podvádí
- jinými slovy: pravděpodobnost, že budeme neoprávněně tvrdit, že jeden z hráčů nepodvádí, je menší než 5%

Klíčovost (keyness)

- vyhodnocení → statistické testy, skóre

Klíčovost (keyness)

- vyhodnocení → statistické testy, skóre

- např. log-likelihood (*LL*)

$$LL = 2\left(f_{slovo\_text} \cdot log\frac{f_{slovo_{text}}}{f(o)_{slovo_{text}}} + f_{slovo\_korpus} \cdot log\frac{f_{slovo\_korpus}}{f(o)_{slovo\_korpus}}\right)$$

  - více viz McIntyre & Walker (2019, p. 154ff) Brezina (2018, p. 83ff)

Analýza klíčových slov (lemmat)

min. frekvence slova v textu = 3

nejfrekventovanější slova					
Klaus (2006)	f				
život	8				
rok	7				
volba	7				
politika	6				
země	6				
evropský	5				
občan	5				
přát	5				
velký	5				
člověk	5				

Analýza klíčových slov (lemmat)

min. frekvence slova v textu = 3

nejfrekventovanější slova		**klíčová slova**			
Klaus (2006)	f	Klaus (2006)	f	$f_{SYN2010}$	log likelihood
život	8	volba	7	23 529	36,41
rok	7	politika	6	18 866	31,99
volba	7	spoluobčan	3	717	31,29
politika	6	občan	5	14 679	27,33
země	6	přát	5	16 608	26,13
evropský	5	vážený	3	2 373	24,15
občan	5	život	8	92 237	23,00
přát	5	evropský	5	34 290	19,17
velký	5	volit	3	5 757	18,89
člověk	5	odpovědnost	3	6 066	18,59

Analýza klíčových slov (lemmat)

min. frekvence slova v textu = 3

nejfrekventovanější slova		**klíčová slova**			
Klaus (2006)	f	Klaus (2006)	f	$f_{SYN2010}$	log likelihood
život	8	volba	7	23 529	36,41
rok	7	politika	6	18 866	31,99
volba	7	spoluobčan	3	717	31,29
politika	6	občan	5	14 679	27,33
země	6	přát	5	16 608	26,13
evropský	5	vážený	3	2 373	24,15
občan	5	život	8	92 237	23,00
přát	5	evropský	5	34 290	19,17
velký	5	volit	3	5 757	18,89
člověk	5	odpovědnost	3	6 066	18,59

Analýza klíčových slov (lemmat)

min. frekvence slova v textu = 3

nejfrekventovanější slova		**klíčová slova**			
Klaus (2006)	f	Klaus (2006)	f	$f_{SYN2010}$	log likelihood
život	8	volba	7	23 529	36,41
rok	7	politika	6	18 866	31,99
volba	7	spoluobčan	3	717	31,29
politika	6	občan	5	14 679	27,33
země	6	přát	5	16 608	26,13
evropský	5	vážený	3	2 373	24,15
občan	5	život	8	92 237	23,00
přát	5	evropský	5	34 290	19,17
velký	5	volit	3	5 757	18,89
člověk	5	odpovědnost	3	6 066	18,59

Log-likelihood (LL)

$$LL = 2\left(f_{slovo\_text} \cdot log\frac{f_{slovo_{text}}}{f(o)_{slovo_{text}}} + f_{slovo_{ref.korpus}} \cdot log\frac{f_{slovo_{ref.korpus}}}{f(o)_{slovo_{ref.korpus}}}\right)$$

$$f(o)_{slovo_{text}} = \frac{N_{text} \cdot \left(f_{slovo_{text}} + f_{slovo_{ref.korpus}}\right)}{N_{text} + N_{ref.korpus}}$$

$$f(o)_{slovo_{korpus}} = \frac{N_{ref.text} \cdot \left(f_{slovo_{text}} + f_{slovo_{ref.korpus}}\right)}{N_{text} + N_{ref.korpus}}$$

Vlastnosti textu vs. referenční data

- text: N = 5000 slov; $f_{jenž}$ = 10; ipm = 2000
- SYN2020: N = 100000 slov; $f_{jenž}$ = 15 673; ipm = 128,65

Log-likelihood (*LL*)

$$f(o)_{slovo_{text}} = \frac{N_{text} \cdot \left(f_{slovo_{text}} + f_{slovo_{ref.\,korpus}}\right)}{N_{text} + N_{ref.\,korpus}} = \frac{5000(10 + 15673)}{100000000 + 5000} = 0.784$$

Log-likelihood (*LL*)

$$f(o)_{slovo_{text}} = \frac{N_{text} \cdot \left(f_{slovo_{text}} + f_{slovo_{ref.\,korpus}}\right)}{N_{text} + N_{ref.\,korpus}} = \frac{5000(10 + 15673)}{100000000 + 5000} = 0.784$$

$$f(o)_{slovo_{korpus}} = \frac{N_{ref.\,text} \cdot \left(f_{slovo_{text}} + f_{slovo_{ref.\,korpus}}\right)}{N_{text} + N_{ref.\,korpus}} = \frac{100000000(10+15673)}{100000000+5000} = 15682.22$$

Log-likelihood (LL)

$$f(o)_{slovo_{text}} = 0.784$$

$$f(o)_{slovo_{korpus}} = 15682.22$$

$$LL = 2\left(f_{slovo\_text} \cdot log\frac{f_{slovo_{text}}}{f(o)_{slovo_{text}}} + f_{slovo_{ref.\,korpus}} \cdot log\frac{f_{slovo_{ref.\,korpus}}}{f(o)_{slovo_{ref.\,korpus}}}\right) =$$

$$= 2\left(10 \cdot log\frac{10}{0.784} + 15673 \cdot log\frac{15673}{15682.22}\right) = 2(10 \cdot \log 12.75 + 15673 \cdot \log 0.999) =$$

$$= 2(10 \cdot 2.55 + 15673 \cdot (-0.00059)) = 2(25.5 - 9.21) = 2 \cdot 16.29 = 32.57$$

Log-likelihood (*LL*)

Table 5.4 Log-likelihood critical values and associated *p*-values for 1 degree of freedom (based on Rayson et al. 2004)

Log-likelihood critical value	*p*-value	Percentage level of confidence in a significant result	Percentage probability that result is due to chance
3.84	0.05	95	5
6.63	0.01	99	1
10.82	0.001	99.9	0.1
15.13	0.0001	99.99	0.01

McIntyre & Walker (2019, p. 158)

Klíčovost (keyness)

- postup
  - prostřednictvím statistického testu se vyberou klíčová slova
  - difference index (DIN) (https://wiki.korpus.cz/doku.php/manualy:kwords?s[]=din)

$$DIN = 100 \times \frac{RelFq(Ttxt) - RelFq(RefC)}{RelFq(Ttxt) + RelFq(RefC)}$$

  - DIN v intervalu <-100; 100>
    - hodnota -100 znamená, že daný jev se ve zkoumaném textu nevyskytuje, je pouze v referenčním korpusu (slovo tedy není ve zkoumaném textu prominentní)
    - hodnota 0 znamená, že daný jev má zhruba stejnou relativní frekvenci ve zkoumaném textu i v referenčním korpusu (slovo tedy není ve zkoumaném textu prominentní)
    - hodnota 100 značí, že slovo se vyskytuje pouze ve zkoumaném textu (může se tedy jednat o velmi prominentní slovo

DIN

$$DIN = 100\frac{\left(\frac{10}{5000}\right) - \left(\frac{15673}{100000000}\right)}{\left(\frac{10}{5000}\right) + \left(\frac{15673}{100000000}\right)} = 100\frac{0.002 - 0.000157}{0.002 + 0.000157} =$$

$$= 100\frac{0.001883}{0.002157} = 100 \cdot 0.8544 = 85.44$$

Klíčovost (keyness)

- aplikace KWords
  - https://kwords.korpus.cz/
  - https://wiki.korpus.cz/doku.php/manualy:kwords

Aktualizace (foregrounding) - příklad analýzy

- statistika a její interpretace
- ne vše, co je statisticky významné, musí být projevem aktualizace

- „That said, it is important to note that while **statistical significance and effect size** are indicators of what Baker (2006) terms *saliency*, **they are not necessarily indicators of foregrounding** (Mukařovsky [1932] 1964; Shklovsky [1917] 1965; van Peer 1986; see also McIntyre and Price 2018a for a summary).“

  (McIntyre & Walker, 2010, p. 164)

Domácí úkol na 16. 10. 2024

- vypočítejte hodnotu Log-likelihood a DIN z následujících hodnot

- počet slov v textu: $N_{text} = 2836$
- počet slov v referenčním korpusu: $N_{ref.\,korpus} = 100031037$
- frekvence slova v textu: $f_{slovo_{text}} = 14$
- frekvence slova v referenčním korpusu: $f_{slovo_{text}} = 10204$