Úvod do kvantitativní lingvistiky

ZS 2024

Hypotéza - opakování

- která tvrzení *jsou/nejsou* testovatelnými hypotézami?
  1. *delší klauze (měřeno vpočtu slov) mají v průměru kratší slova (měřeno v počtu slabik) než klauze kratší*
  2. *v odborných textech je hodně dlouhých vět*
  3. *pokud je slovo syntakticky závislé na substantivu, je to přívlastek*
  4. *auxiliáry jsou v průměru kratší něž autosémantika*
  5. *mezi délkou slova měřenou v počtu hlásek a v počtu slabik je lineární závislost*
  6. *děti z měst mají bohatou slovní zásobu*
  7. *čeština je jeden z nejkomplikovanějších jazyků na světě*
  8. *čím je slovo delší, tím má více hlásek*
  9. *čím je člověk starší, tím v průměru používá více zájmen*

Metodologie

(teorie)

↓

verbální formulace hypotézy + operacionalizace

↓

matematická formalizace

↓

experiment

↓

matematické vyhodnocení experimentu

↓

lingvistická explanace

Operacionalizace vs. klasifikace

- Způsoby klasifikace jevů
  - $V(A) = V(B)$, nebo $V(A) \neq V(B)$
  - $V(A) > V(B)$, nebo $V(A) = V(B)$, nebo $V(A) < V(B)$
  - $V(A) - V(B) = d$

Operacionalizace vs. klasifikace

- Způsoby klasifikace jevů
  - $V(A) = V(B)$, nebo $V(A) \neq V(B)$
    - V = substantivum: *dům* = *stůl*; *dům* ≠ *spát*
  - $V(A) > V(B)$, nebo $V(A) = V(B)$, nebo $V(A) < V(B)$
  - $V(A) - V(B) = d$

Operacionalizace vs. klasifikace

- Způsoby klasifikace jevů
  - $V(A) = V(B)$, nebo $V(A) \neq V(B)$
    - V = substantivum: *dům = stůl*; *dům ≠ spát*
  - $V(A) > V(B)$, nebo $V(A) = V(B)$, nebo $V(A) < V(B)$
    - V = synt. objekt:
      Vidím <u>Marii</u> > Myslím <u>na Marii</u> > Dívám se <u>na Marii</u> > > Zabil Marii <u>sekyrou</u>
  - $V(A) - V(B) = d$

Operacionalizace vs. klasifikace

- Způsoby klasifikace jevů
  - $V(A) = V(B)$, nebo $V(A) \neq V(B)$
    - *V* = substantivum: *dům = stůl*; *dům ≠ spát*
  - $V(A) > V(B)$, nebo $V(A) = V(B)$, nebo $V(A) < V(B)$
    - *V* = synt. objekt:
      Vidím Marii > Myslím na Marii > Dívám se na Marii > > Zabil Marii sekyrou
  - $V(A) - V(B) = d$
    - *V* = délka slova (ve slabikách):
      Trojanovice vs. *Ostrava*: 5 – 3 = 2
      Trojanovice vs. *Praha*: 5 – 2 = 3
      Trojanovice vs. *Malenovice*: 5 – 5 = 0

Operacionalizace vs. klasifikace

- už „pouhá“ kvantifikace klasifikace přináší hlubší pohled na dané jevy
- srov. korpusová lingvistika

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

? intuice

- jaký bude rozdíl ve frekvenci substantiv v PUB a ODB textech

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.

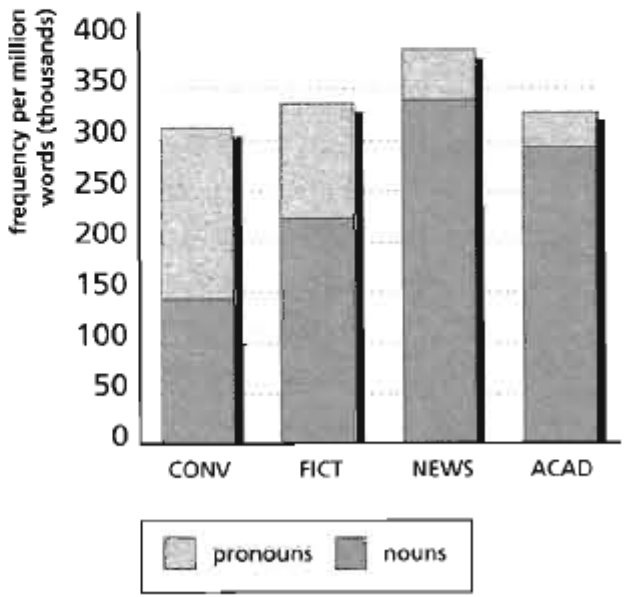
- The relative frequency of nouns is much higher in object position and as a complement or object of a preposition than in subject position.

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



Operacionalizace

- je třeba jasně a jednoznačně definovat proměnné, mezi kterými se předpokládá závislost
- H: čím je slovo frekventovanější, tím je vyšší jeho synonymie
  - v čem může být problém?

Operacionalizace

- je třeba jasně a jednoznačně definovat proměnné, mezi kterými se předpokládá závislost
- H: čím je slovo frekventovanější, tím je vyšší jeho synonymie
  - frekvence: je třeba jasně uvést
    - co se myslí *slovem*
    - jak a kde se budou počítat frekvence (srov. různé subkorpusy výše)

Operacionalizace

- je třeba jasně a jednoznačně definovat proměnné, mezi kterými se předpokládá závislost
- H: čím je slovo frekventovanější, tím je vyšší jeho synonymie
  - frekvence: je třeba jasně uvést
    - co se myslí *slovem*
    - jak a kde se budou počítat frekvence (srov. různé subkorpusy výše)
  - synonymie: je třeba uvést
    - jak se bude polysémie kvantifikovat

Operacionalizace

- H: ženy mají větší pasivní slovní zásobu než muži
  - problém?

Operacionalizace

- H: ženy mají větší pasivní slovní zásobu než muži
  - jak definovat „pasivní slovní zásobu“?

Operacionalizace

- špatná operacionalizace znehodnocuje celou analýzu

- H: subjekt je v češtině v průměru delší než objekt

  - je to empiricky testovatelná hypotéza?

Operacionalizace

- špatná operacionalizace znehodnocuje celou analýzu

- H: subjekt je v češtině v průměru delší než objekt

  Muž v triku veze naše dopisy

  - je tato věta v souladu s hypotézou?

Operacionalizace

- špatná operacionalizace znehodnocuje celou analýzu
- H: subjekt je v češtině v průměru delší než objekt

Muž v triku veze naše dopisy

  - pokuste se formulovat některé problémy s určováním délky

Operacionalizace

- špatná operacionalizace znehodnocuje celou analýzu
- H: subjekt je v češtině v průměru delší než objekt

  Muž v triku veze naše dopisy

  - některé problémy s určováním délky:
    - rozvitý vs. nerozvitý subjekt/objekt?
    - počet slov?
    - počet slabik?
    - počet morfémů?
    - je neslabičná předložka samostatným slovem?

- způsob měření ovlivňuje podobu výsledku!!!

Operacionalizace

- jasné vymezení
- dokumentace
  - ideálně technická zpráva
- replikovatelnost

Úkol

- ověřte hypotézu o vztahu frekvence a délky slova
- formulujte hypotézu
- proveďte analýzu
  - můžete použít i jen jeden krátký text, bude snazší to zpracovat…
  - https://ezcalc.me/word-frequency-counter/
  - Excel: fce DÉLKA
- vytvořte dokument, kde bude
  - formulována hypotéza,
  - operacionalizace popsána tak, aby byl experiment replikovatelný (definice jednotek, popis jazykového materiálu)
  - prezentovány výsledky
- pokud chcete, pracujte ve dvojicích

Úkol

délka	průměrná frekvence	počet slov o dané délce
1		
2		
3		
4		
…		