Úvod do kvantitativní lingvistiky

Radek Čech

Historie KL

- G. K. Zipf (1902-1950)
- PLK – B. Trnka (problematika těsnopisu)
- M. Těšitelová a kol.
- G. Altmann, R. Köhler, L. Hřebíček

Místo KL v lingvistice

- cíle lingvistiky (obecně)
- metody lingvistiky (obecně)
- předmět lingvistiky (obecně)

Základní rysy KL

- cílem nikoliv popis (klasifikace) jednotek jazykového systému, ale vytvoření modelu, jehož vlastnosti jsou založeny na **empiricky testovatelných hypotézách**
- lingvistika jako **experimentální věda**

KL

(teorie)

↓

verbální formulace hypotézy + operacionalizace

↓

matematická formalizace

↓

experiment

↓

matematické vyhodnocení experimentu

↓

lingvistická explanace

Kvantitativní lingvistika

- Co je teorie?
- Jak poznat, která teorie je lepší než jiná?
- Plyne něco z teorie?
- Lingvistické teorie?

KL

(teorie)

↓

verbální formulace hypotézy + operacionalizace

↓

matematická formalizace

↓

experiment

↓

matematické vyhodnocení experimentu

↓

lingvistická explanace

Kvantitativní lingvistika

- Co je hypotéza?
- Formální vlastností hypotézy?
- Lingvistické hypotézy…

Hypotéza

- která tvrzení *jsou*/*nejsou* testovatelnými hypotézami?
  1. *hodně mužů má pleš*
  2. *pokud se v knize vyskytují biblický příběh, je to apokryf*
  3. *jestli se zavedou řidičáky „na zkoušku“, může se snížit nehodovost mladých řidičů a řidiček*
  4. *muži mají častěji pleš než ženy*
  5. *jestliže se je sloveso dokonavé, častěji se na něj váže přímý akuzativní předmět než na sloveso nedokonavé*
  6. *ženy jsou citlivé*
  7. *čím je slovo frekventovanější, tím je větší jeho polysémie*
  8. *jestli se zavedou řidičáky „na zkoušku“, sníží se nehodovost mladých řidičů a řidiček*
  9. *nářečí často ovlivňují podobu mluveného jazyka obyvatel dané nářeční oblasti*

KL

(teorie)

↓

verbální formulace hypotézy + **operacionalizace**

↓

matematická formalizace

↓

experiment

↓

matematické vyhodnocení experimentu

↓

lingvistická explanace

Kvantitativní lingvistika

- 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

- je třeba jasně a jednoznačně definovat proměnné, mezi kterými se předpokládá závislost
- H: čím je slovo frekventovanější, tím je polysémnější
  - 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 polysémnější
  - frekvence: je třeba jasně uvést
    - co se myslí *slovem*
    - jak se budou počítat frekvence
  - polysémie: je třeba uvést
    - jak se bude polysémie kvantifikovat

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

  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!!!

KL

(teorie)

↓

verbální formulace hypotézy + operacionalizace

↓

matematická formalizace

↓

experiment

↓

matematické vyhodnocení experimentu

↓

lingvistická explanace

Kvantitativní lingvistika

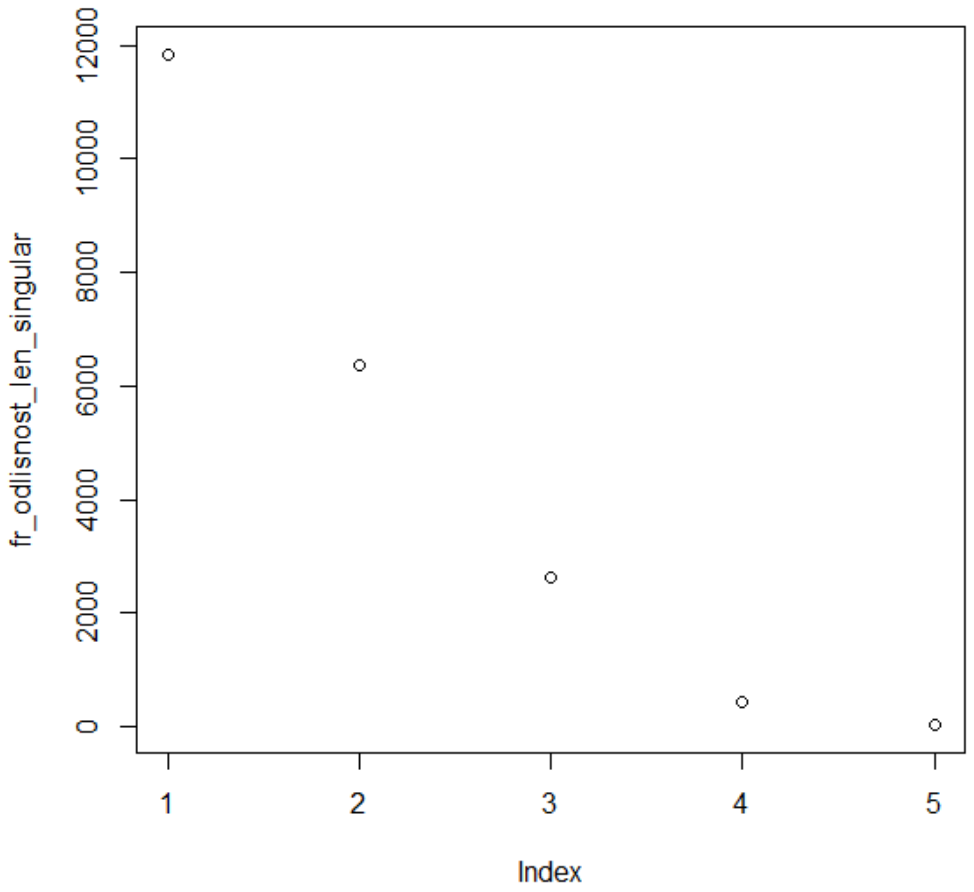
- Matematická formalizace
  - vyjádření sledovaných vlastností čísly
  - např. frekvenční distribuce, vztah mezi dvěma vlastnostmi atp.

Matematická formalizace

- hypotéza
  - čím více morfo-fonetických změn se v daném tvaru slova projevuje (vzhledem k nominativu), tím menší je jeho frekvence
- zdůvodnění
  - Minimisation of producing, encoding, memory effort (speaker)
  - Minimisation of decoding, (hearer)
  - the conserving effect of token frequency

Matematická formalizace

počet změn	***f***
0	11847
1	6356
2	2616
3	447
4	25



KL

(teorie)

↓

verbální formulace hypotézy + operacionalizace

↓

matematická formalizace

↓

experiment

↓

matematické vyhodnocení experimentu

↓

lingvistická explanace

Kvantitativní lingvistika

- Co je experiment?
  - vlastnosti
  - chyby
  - replikace

KL

(teorie)

↓

verbální formulace hypotézy + operacionalizace

↓

matematická formalizace

↓

experiment

↓

matematické vyhodnocení experimentu

↓

lingvistická explanace

Kvantitativní lingvistika

- Matematické vyhodnocení experimentu
- náhoda vs. mechanismus
- statistický test

Statistické testy významnosti

- testuje se pravděpodobnost toho, zda je pozorovaná závislost způsobena náhodou, či ne
- hod mincí (100x)
  - pokud padne 50x panna a 50x orel, nepředpokládá se, že by někdo podváděl
  - pokud padne 52x panna a 48x orel, nepředpokládá se, že by někdo podváděl
  - padne-li však 98x panna, dá se předpokládat, že někdo podvádí
  - a co když padne panna 59x nebo 60x nebo 61x nebo 62x…?

Statistické testy významnosti

- dá se spočítat (srov. např. Greis 2010, s. 33nn), že pokud padne panna 61x, tak je větší než 95procentní pravděpodobnost, že jeden z hráčů podvádí
- jinými slovy: pravděpodobnost, že budeme neoprávněně tvrdit, že jeden z hráčů nepodvádí, je menší než 5procentní

Statistické testy významnosti

- porovnávají se dvě hypotézy
  - nulová hypotéza: tvrzení, které obvykle deklaruje “žádný rozdíl”, tj. nalezený rozdíl je dán variablitou dat, náhodou (např. mince není falešná; mezi formou jazyka a četností užívaní *bychom*/*bysme* není rozdíl)
  - alternativní hypotéza: situace, kdy nulová hypotéza neplatí, tj. mezi proměnnými se předpokládá závislost; důležité je přitom nějaké teoretické zdůvodnění

Statistické testy významnosti

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

Chí-kvadrát test dobré shody

- příklad: předpokládáme, že v románech se bude častěji používat nespisovná varianta slova “bychom” než v publicistických textech
  - proměnnými jsou: a) typ textu; b) varianta slova

  H_0: mezi typem textu a používáním nespisovné varianty slova “bychom” není žádný vztah

  H_1: mezi typem textu a používáním nespisovné varianty slova “bychom” je vztah, tj. tato forma se častěji vyskytuje v próze

Chí-kvadrát test dobré shody

	SYN2005nov (romány)	SYN2005pub (publicistika)
bychom	5260	6679
bysme	714	39
% bysme	13,6	0,6

	SYN2005nov (romány)	SYN2005col (povídky)
bychom	5260	1660
bysme	714	136
% bysme	13,6	8,2

Chí-kvadrát test dobré shody

	SYN2005nov (romány)	SYN2005pub (publicistika)
bychom	5260	6679
bysme	714	39
% bysme	13,6	0,6
p = 0,00000000000000022		

	SYN2005nov (romány)	SYN2005col (povídky)
bychom	5260	1660
bysme	714	136
% bysme	13,6	8,2

Chí-kvadrát test dobré shody

	SYN2005nov (romány)	SYN2005pub (publicistika)
bychom	5260	6679
bysme	714	39
% bysme	13,6	0,6
p = 0,00000000000000022		

	SYN2005nov (romány)	SYN2005col (povídky)
bychom	5260	1660
bysme	714	136
% bysme	13,6	8,2
p = 0,0000001851		

Statistické testy významnosti

- někteří lingvisté (vesměs korpusoví) jsou přesvědčeni, že čím více dat máme k dispozici, tím lépe
  - Sinclair: "*The more, the better!*"

- z hlediska použitelnosti statistických testů to však neplatí,
  - srov. Rietvield T. et al.: Pitfalls in Corpus Research. *Computers and the Humanities* 38: 343–362, 2004.

Statistické testy významnosti

	romány	novely	Σ	% novely
konstrukce *A*	500000	501800	1001800	50,09%
konstrukce *B*	501500	500000	1001500	49,93%
Σ	1001500	1001800	2003300	
chi^2 = 5.43, p=0,020				

KL

(teorie)

↓

verbální formulace hypotézy + operacionalizace

↓

matematická formalizace

↓

experiment

↓

matematické vyhodnocení experimentu

↓

<u>lingvistická explanace</u>

Kvantitativní lingvistika

- popis vs. interpretace vs. explanace

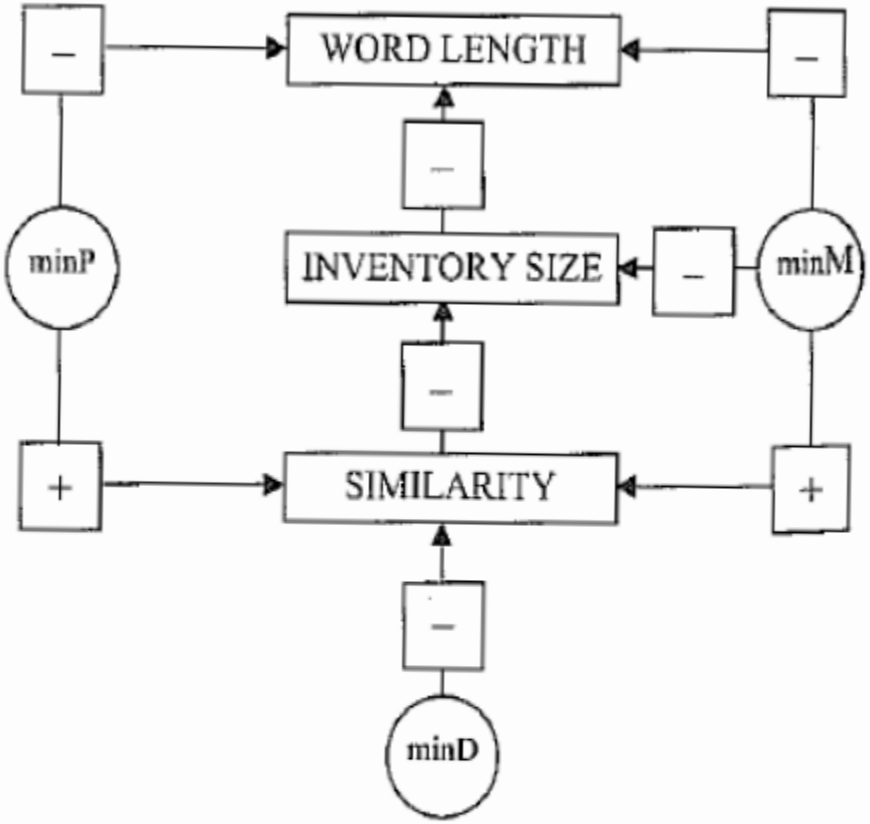


Fig. 53.1: Control circuit on the phoneme/word level, consisting of three requirements and three system variables. The squares represent proportionality operators and give the sign of their numerical value; requirements are symbolised by circles.

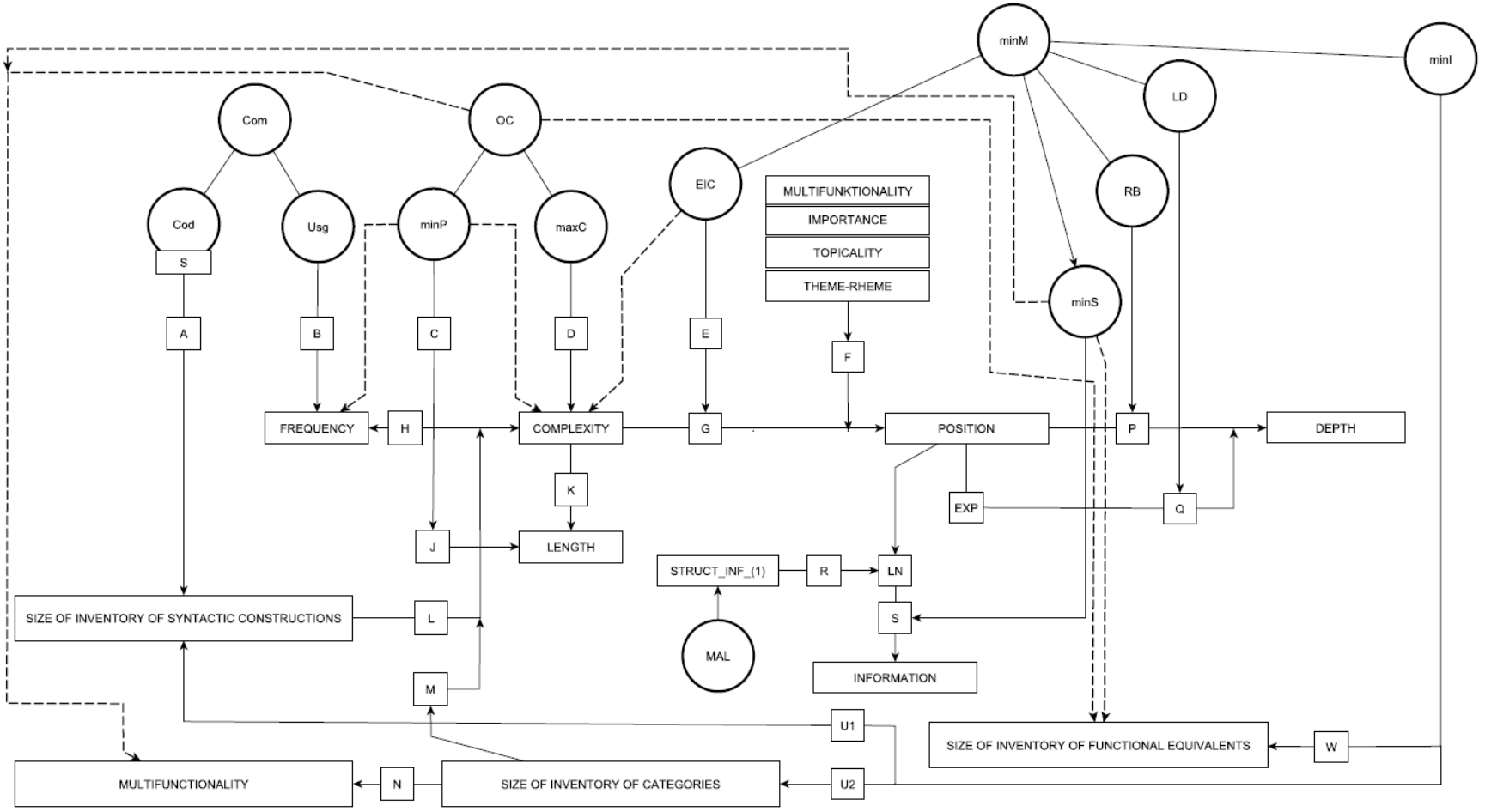


Figure 4.33: The structure of the syntactic subsystem as presented in this volume

Kvantitativní lingvistika

- International Quantitative Linguistics Association (http://www.iqla.org/)
- journals:
  - *Journal of Quantitative Linguistics* (http://www.tandfonline.com/action/journalInformation?show=aimsScope&journalCode=njql20#.UytfjfI5OM4)
  - *Glottotheory* (http://www.degruyter.com/view/j/glot)
  - *Glottometrics* (http://www.ram-verlag.eu/journals-e-journals/glottometrics/)
- book series
  - *Quantitative Linguistics* (de Gruyter)
  - *Studies in Quantitative Linguistics* (RAM-Verlag)

Kvantitativní lingvistika

Děkuji za pozornost!

(www.cechradek.cz)