Statistische Probleme der korpusbasierten Typologie

Balthasar Bickel

www.uni-leipzig.de/~bickel

Rückblick (April 2005)

Korpusbasierte Variablen in der Typologie:

1. Typologische Verteilungen zum Teil von Diskurspräferenzen voraussagbar

   … aber lokale/areale Geschichte bildet weitere Variablen

2. Typologische Varianz im Diskurs

   … aber soziologische/ethnologische Varianz bildet weitere Variablen

Heute (September 2005)

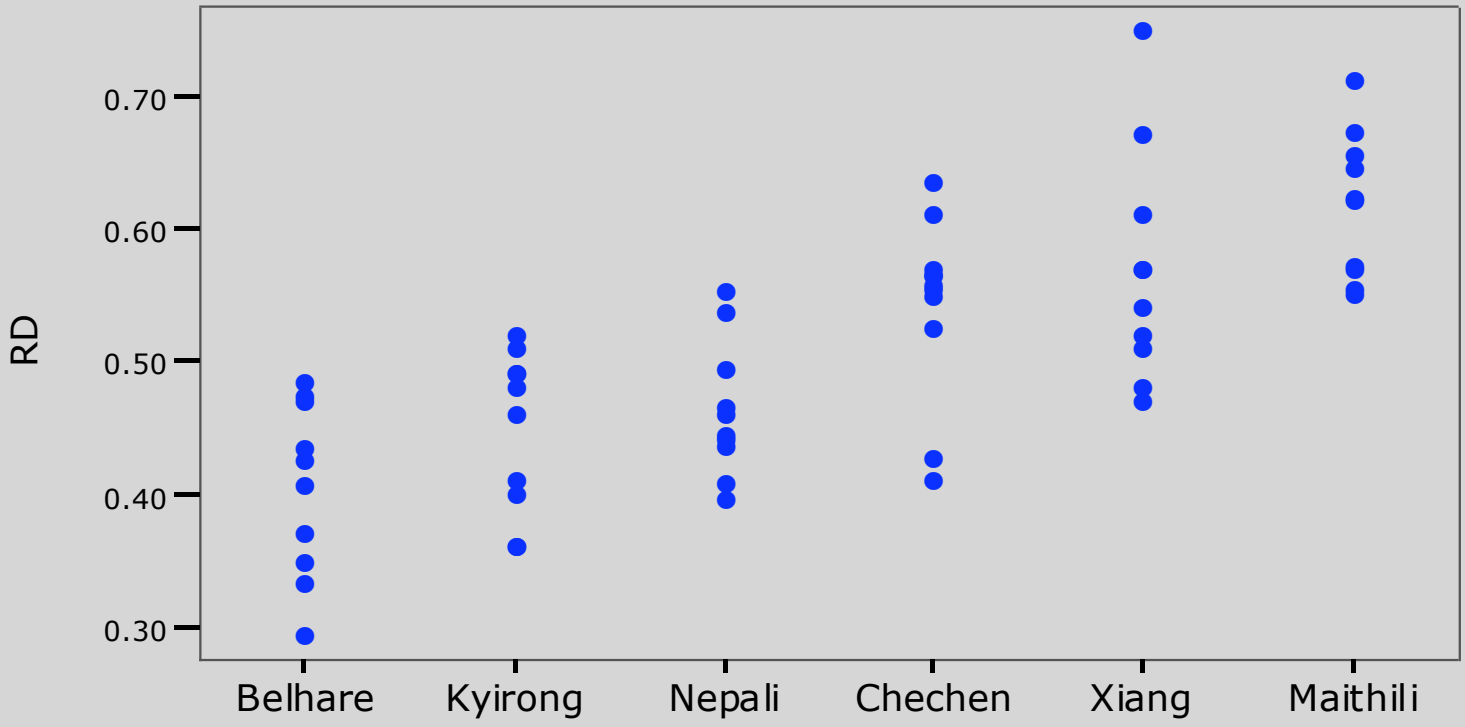
1. Natur typologischer Variablen und Folgen davon für korpusbasierte Typologie
2. Geltungsbereiche in der Typologie

Typologische Varianz, speziell im Diskurs

- Voraussetzung aller Typologie: die Varianz ist **strukturell** und **nicht individuell**
- Oft gemachte, aber auf einem Fehlschluss basierende Zusatzannahme: die Varianz ist notwendigerweise durch **"Sprachen"** (eine genealogische oder soziologische Einheit) bedingt = eine empirische, nicht apriorische Frage
- Beispiele zu strukturellen Variablen:
  - Grösse phonologischer Domänen
    - zwischensprachliche Varianz > innersprachliche Varianz
    - intergenealogische Varianz > intragenealogische Varianz
  - Affixkohärenz
    - Varianz nicht durch Sprache, sondern Kategorie bedingt (z.B. Tempus vs. Negation)
- Beispiel zu korpusbasierter Variable:
  - Referenzdichte

Beispiel: Referenzdichte (RD)

- Discourse variable: RD = *N* (overt arg) / *N* (poss. arg) (defined only for radical pro-drop languages)



Language vs. Individual: F (1,61) = 15.55, p < .001

Testdesign für RD-Typologie

- Aber: was an den Sprachen bedingt die Varianz? Welche Faktoren?
  - genealogische
  - areale
  - strukturelle
  - soziologische
  - NICHT: “Sprache”!
- Testdesign kann und soll “Sprachen” ignorieren!

Testdesign für RD-Typologie

- GLM design; interest in interactions
- DV = RD
- Structural factor: SYN (postulated in earlier work*)
  - agreement with NP (SYN = “case-based”)
  - agreement with argument (SYN = “none”)
- Sociological factors:
  - Social network (NET): loose, close
  - Literacy (LIT): literate, illiterate
- Unlikely: areal factors
- Not (yet) testable: genealogical factors

* Bickel, B. 2003. Referential density in discourse and syntactic typology. *Language* 79, 708-29.

Sample

	GEN	AREA	SYN	NET	RD predicted
Belhare	ST	Himalayas	–	close, rural	low
Kyirong			–	close, urban	low
Xiang		China	–	loose, urban	high
Nepali	IE	Himalayas	+	close, urban	high
Maithili			+	close, urban	high
Chechen	ND	Caucasus	+	loose, urban	high

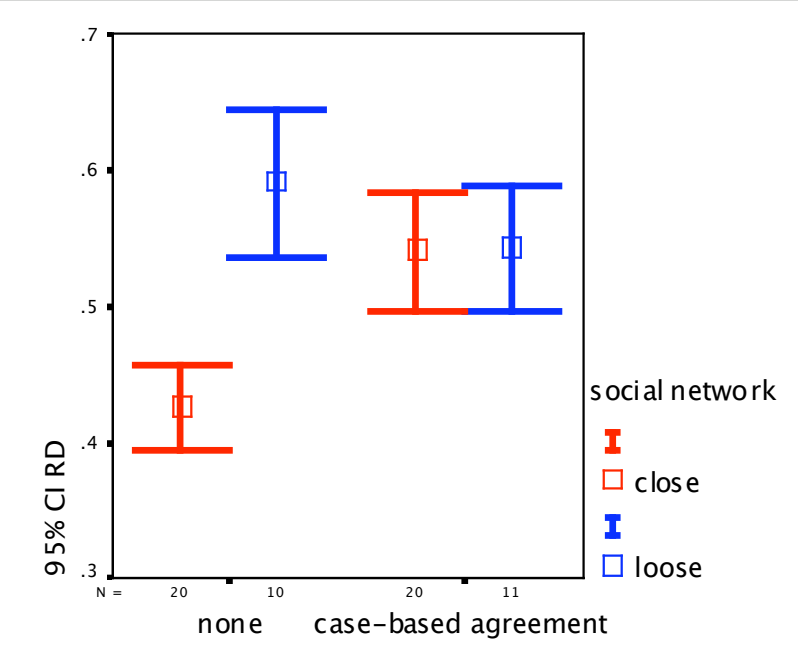
Cross-community variable:

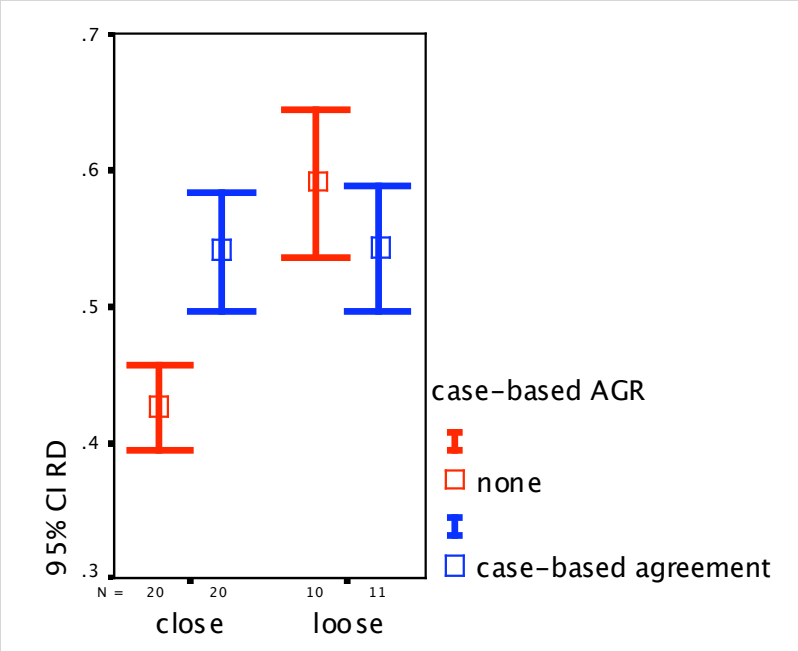
- Literacy (LIT)

Factorial analysis

3-way ANOVA (SYN * NET * LIT):

- main effect of SYN (F (55,1) = 10.75, p = .002)
- main effect of LIT (*F* (55,1) = 33.89, p < .001)
- interaction effect of SYN * NET (*F*(61,1) =11.22, p = .001)

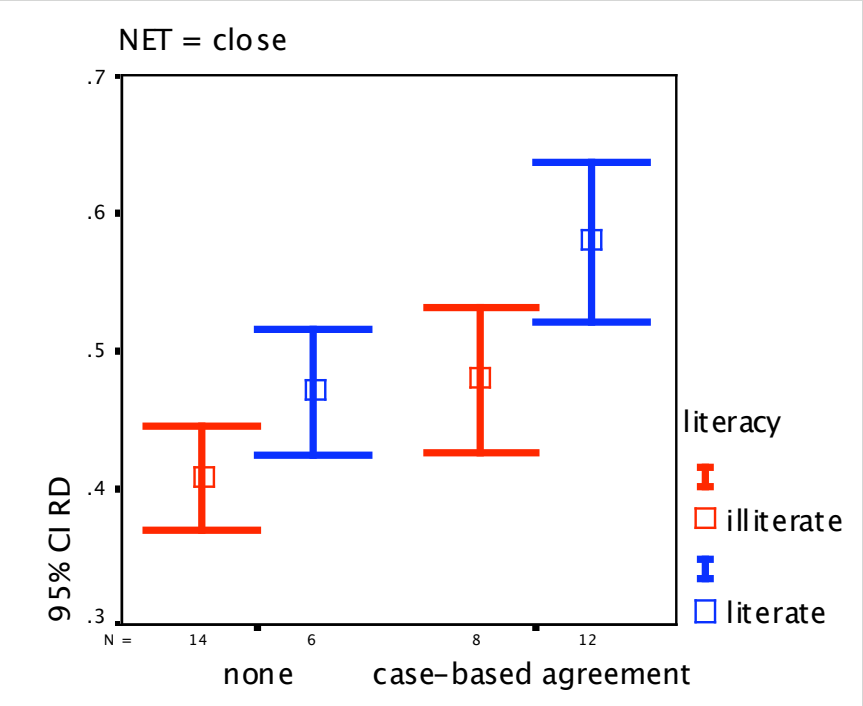




Factorial analysis (cont'd)

2-way ANOVA (SYN*LIT) under the NET = close condition

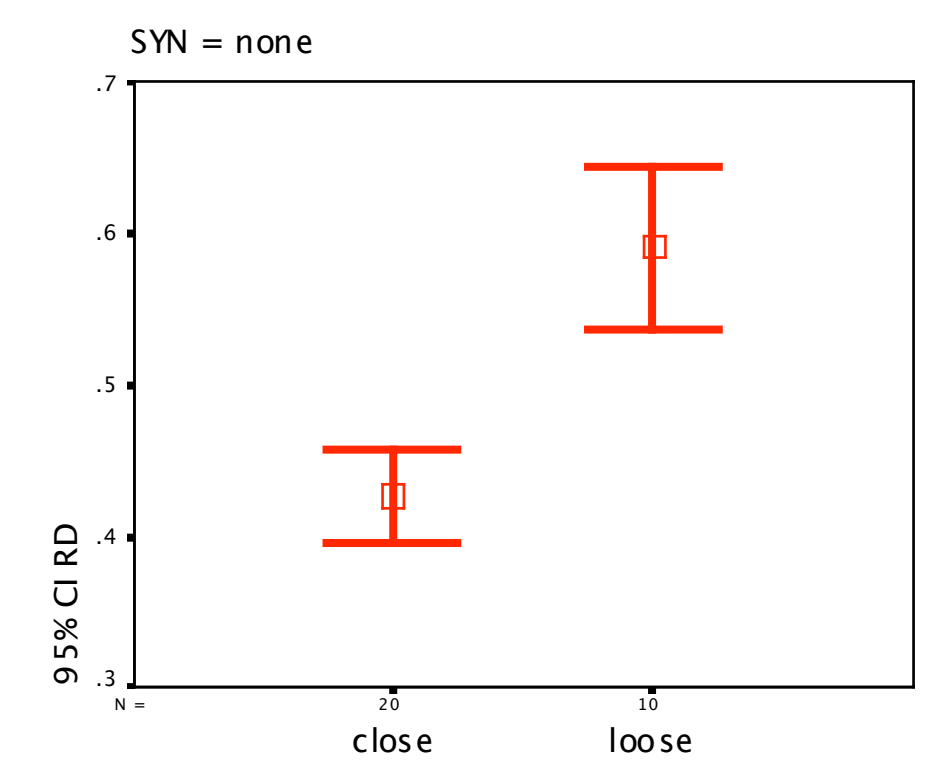
- main effect of SYN ($F(1,36) = 25.50$, $p < .001$, contrast = .08**)
- main effect of LIT ($F(1,40) = 12.14$, $p = .001$, contrast = .09*)
- no significant interaction



Factorial analysis (cont'd)

1-way ANOVA (NET) under the SYN = none condition (loose NET strictly implies LIT; hence LIT is excluded here)

- main effect of NET ($F(1,30) = 38.68$, $p < .001$, contrast = .16**)



Methodologische Konklusion

- “Sprachen” als genealogische und/oder soziologische Einheiten spielen in der korpusbasierten Typologie keine Rolle
- Alles GLM-Designs, v.a. ANOVA

Geltungsbereich

Wichtiger Störfaktor in der Typologie ist Genealogie (Verwandtschaft aufgrund eines gemeinsamen Vorfahren), aber wir können nicht genealogisch stratifizieren

- anstatt Zufallsproben, genealogisch balancierte Proben
- Probe = Population
- klassische Statistik nicht anwendbar
- exakte und Randomisierungsmethoden (Monte-Carlo)
- Randomisierungsbasierte ANOVA *

* Janssen, D., B. Bickel & F. Zúñiga 2005. Randomization tests in language typology. Ms.