Review

Grammaticality judgments unreliable

- vary with context
- sensitive to relative frequency
- affected by interactions of multiple conflicting constraints, including processing constraints

Usage data problematic

- unexamined confounds and correlations
- pooled data from different speakers
- lexical dependencies ignored
- cross-corpus differences

Data from controlled experiments

- experimental items = constructed sentences
- isolated from connected discourse
- artifactual default referents

Solutions

- use multiple sources of converging evidence: typological, usage-based, experimental, and introspective
- use modern data analysis: graphical visualization, descriptive statistics, multivariable modeling, qualitative interpretation of quantitative data

Documentation of *the problems from intuitions*:

Joan Bresnan. 2005. “A Few Lessons from Typology”.

Case studies of the *English dative alternation*:

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2005. “Predicting the Dative Alternation.” [corpus]

Joan Bresnan. 2006. “Is syntactic knowledge probabilistic? Experiments with the English dative alternation.” [experiments]

Case studies of the *English genitive alternation*:

Anette Rosenbach. 2003. “Iconicity and economy in the choice between the ’*s*-genitive and the *of*-genitive in English.” [experiments]

Lars Hinrichs and Benedikt Szmrecsányi. 2006. “Recent changes in the function and frequency of Standard English genitive constructions: a multivariate analysis of tagged corpora.” [corpus]

Hands-on quantitative data analysis with syntactic, semantic, and lexical data:

- R. Harald Baayen. 2006. *Practical Data Analysis for the Language Sciences with R* (forthcoming)
- class project with dative data from the CHILDES database

Methods of analysis of corpus and experimental linguistic data

- Install and learn to use R (open source statistical computing environment available for all platforms): dataframes, vector calculations
- Graphical data exploration – visualizing
  - *single random variables:* histograms, density plots, boxplots, ordered values, quantile plots
  - *two or more random variables:* barplots, mosaic plots, scatterplots, pairs plots, trellis graphics, smoothers

- Probability distributions
  - *Discrete distributions:* binomial (frequency of binary-valued variable in corpus), poisson (rate of occurrence of variable in a corpus)
  - *Continuous distributions:* normal distribution; t, F, χ^2

- # Basic statistical tests

Type of Data	Question?	If data are...	then do
1 numerical vector	normal distribution?		shapiro.test(), ks.test()
	equal probabilities?	counts	chisq.test()
	location of mean?	normal	t.test()
		non-normal	wilcox.test()
2 independent vectors	same distribution?		ks.test(), w jitter
	same means?	normal	t.test()
		non-normal	wilcox.test()
	same variances?	normal	var.test()
2 paired vectors	same means?	normal	t.test(...,paired = T)
		non-normal	wilcox.test(...,paired = T)
	functional relation?	normal	lm()
	correlated?	normal input	cor.test
		non-normal	cor.test(..., method = “spearman”)
1 numerical vector, 1 factor	different group means?	normal, same variances	lm(), anova(), aov()
		different variances	kruskal.test()
2 numerical vectors, 1 factor	different means? interactions?	normal	lm()
2 vectors of counts	different proportions?		chisq(), fisher.test()

Problems and pitfalls of linear regression: (i) outliers, (ii) nonlinear covariates

Snag of anova with factor levels > 2: multiple comparisons inflating chances of a significant result; use Bonferroni correction or Tukey’s H(onestly)S(ignificant)D(ifference)

- Clustering and Classification
  - principle components analysis (for tables of measurements)
  - classification trees
- Regression Modeling

  (to be continued on Thursday)