In Defense of Corpus Data

Summary from Week 1:

Introspective judgments about decontextualized, constructed examples…

- may underestimate the space of grammatical possibility because of absence of context
- may reflect relative frequency within the space of grammatical possibility
- may fail to reflect the interactions of multiple conflicting constraints, including processing constraints

An alternative source of data:

the spontaneous use of language in natural settings

But

surprisingly, many syntacticians believe that such ‘usage data’ (corpora) are irrelevant to the theory of grammar.

Summary from Week 1:

Corpus data are problematic because…

- correlated variables can be explained by simpler theories (e.g. Hawkins 1994, Snyder 2003)
- pooled data from different speakers may invalidate grammatical inference
- lexical biases are not accounted for
- cross-corpus differences undermine the relevance of corpus studies to grammatical theory

Bresnan, Cueni, Nikitina, and Baayen (in press):

—the four problems in the critique of usage data are *empirical issues*

—can be resolved by using modern statistical theory and modelling strategies widely used in other fields.

Case study: the dative alternation

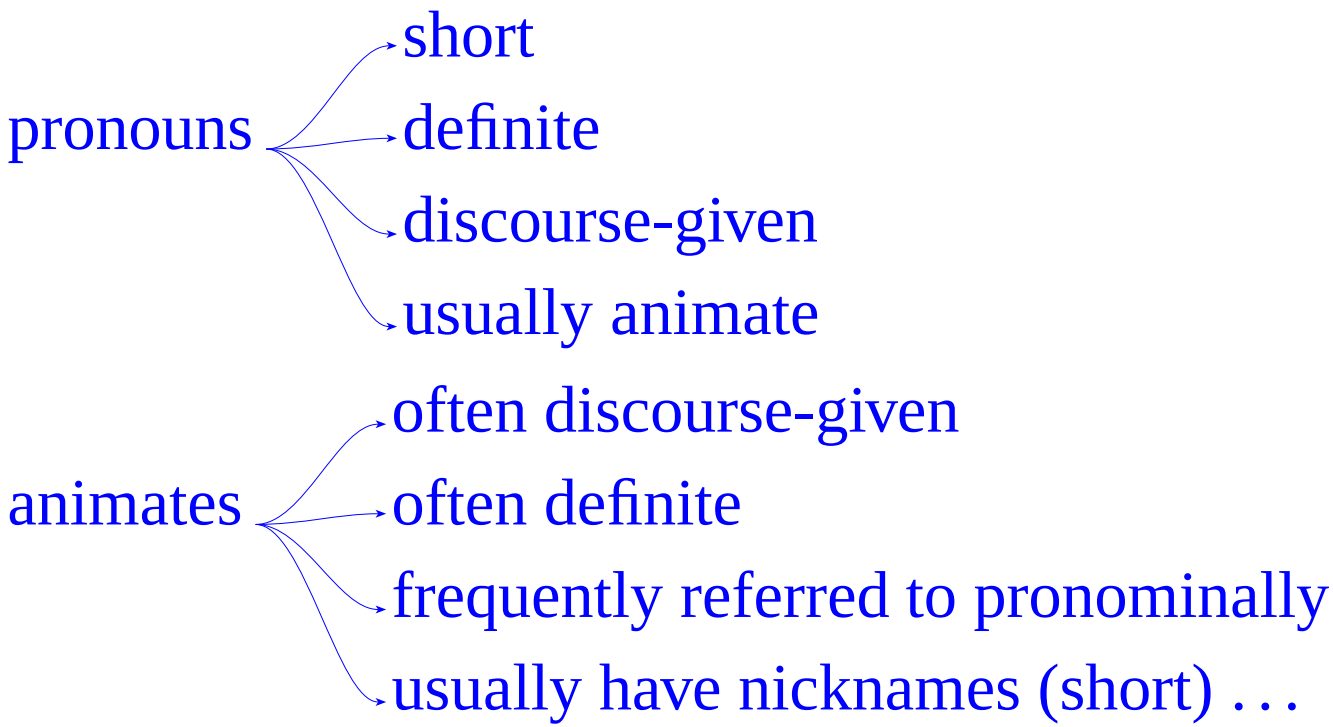
Corpus studies of English have found that various properties of the recipient and theme have a quantitative influence on dative syntax (Thompson 1990, Collins 1995, Snyder 2003, Gries 2003, ao):

discourse accessibility
relative length
pronominality
definiteness
animacy

$\Rightarrow$

dative construction choice

Yet what really drives the dative alternation remains unclear because of pervasive correlations in the data:



Correlations tempt us into reductive theories that explain effects in terms of just one or two variables (e.g. Hawkins 1994, Snyder 2003)

A beautifully simple theory:

1. Givenness correlates with shorter, less complex expressions (less description needed to identify)
2. Shorter expressions occur earlier in order to facilitate parsing (more complex after less)

Apparent effects of givenness (and correlated properties like animacy) could reduce to the preference to process syntactically complex phrases later than simple ones (Hawkins 1994).

Question 1:

Are these effects of discourse accessibility, animacy, and the like the epiphenomena of syntactic complexity effects in parsing?

Use logistic regression to control simultaneously for multiple variables related to a binary response.[a]

Use large samples of richly annotated data: 2360 dative observations from the three-million-word Switchboard collection of recorded telephone conversations.

[a] Williams 1994; Arnold, Wasow, Losongco, and Ginstrom 2000; cf. Gries 2003

explanatory variables:

- discourse accessibility, definiteness, pronominality, animacy (Thompson 1990, Collins 1995)
- differential length in words of recipient and theme (Arnold et al. 2000, Wasow 2002, Szmrecsanyi 2004b)
- structural parallelism in dialogue (Weiner and Labov 1983, Bock 1986, Szmrecsanyi 2004a)
- number, person (Aissen 1999, 2003; Haspelmath 2004; Bresnan and Nikitina 2003)
- concreteness of theme

plus 5 broad semantic classes of uses of verbs which participate in the dative alternation:

- abstract (abbreviated ‘a’): *give it some thought*
- transfer of possession (‘t’): *give an armband*, *send*
- future transfer of possession (‘f’): *owe*, *promise*
- prevention of possession (‘p’): *cost*, *deny*)
- and communication (‘c’): *tell*, *give me your name*, said on a telephone

Model A:

Response $\sim$

semantic class + accessibility of recipient + accessibility of theme + pronominality of recipient + pronominality of theme + definiteness of recipient + definiteness of theme + animacy of recipient + person of recipient + number of recipient + number of theme + concreteness of theme + structural parallelism in dialogue + length difference (log scale)

The Logistic Regression Model

$$\text{logit}[\text{Probability}(\text{Response} = 1)] = \boldsymbol{X}\boldsymbol{\beta}$$

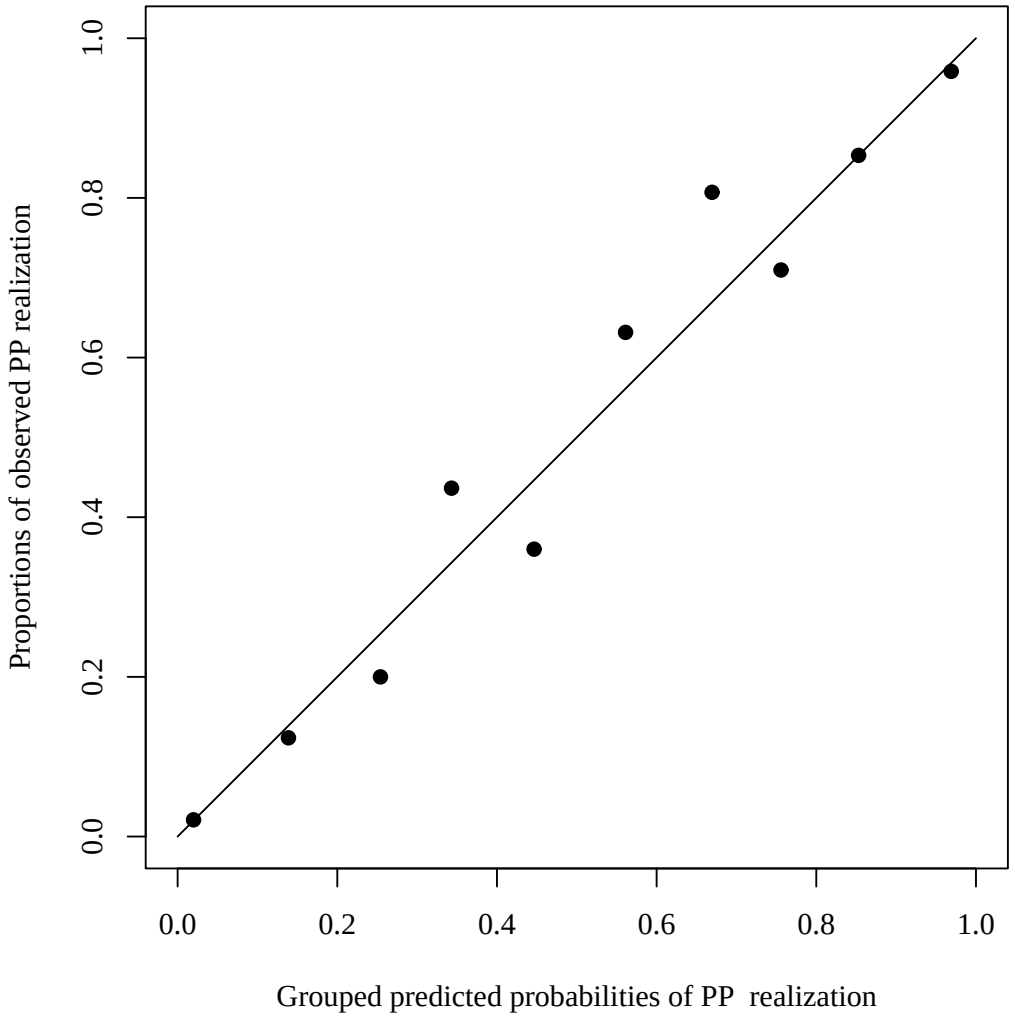
or

$$\text{Probability}(\text{Response} = 1) = \frac{1}{1 + \exp(-\boldsymbol{X}\boldsymbol{\beta})}$$

Classification Table for Model A (1 = PP; cut value = 0.50)				
		Predicted:		% Correct
		0	1	
Observed:	0	1796	63	97%
	1	115	386	77%
			Overall:	92%

% Correct from always guessing NP NP (=0): 79%

Model A plot of observed against predicted responses



How well does the model generalize to new data?

Divide the data randomly 100 times into a training set of sufficient size for the model parameters $(n = 2000)$ and a testing set $(n = 360)$.

Fit the Model A parameters on each training set and score its predictions on the unseen testing set.

Mean overall score (average % correct predictions on unseen data) = 92%.

All of the model predictors except for number of recipient are significant.

All, $p < 0.001$ except person of recipient, number of theme, and concreteness of theme, $p < 0.05$.

What Model A shows.

Harmonic alignment of prominence scales with syntactic position:

discourse given $\succ$ not given

animate $\succ$ inanimate

definite $\succ$ indefinite

pronoun $\succ$ non-pronoun

recipient shorter $\succ$ recipient longer

V NP NP

V NP PP

The model formula:

$$\text{Probability}\{\text{Response} = 1\} = \frac{1}{1+\exp(-X\beta)}, \text{ where}$$

$$\begin{aligned} X\hat{\beta} = \; & 0.95 \\ & -1.34\{\text{c}\} + 0.53\{\text{f}\} - 3.90\{\text{p}\} + 0.96\{\text{t}\} \\ & +0.99\{\text{accessibility of recipient} = \text{nongiven}\} \\ & -1.1\{\text{accessibility of theme} = \text{nongiven}\} \\ & +1.2\{\text{pronominality of recipient} = \text{nonpronoun}\} \\ & -1.2\{\text{pronominality of theme} = \text{nonpronoun}\} \\ & +0.85\{\text{definiteness of recipient} = \text{indefinite}\} \\ & -1.4\{\text{definiteness of theme} = \text{indefinite}\} \\ & +2.5\{\text{animacy of recipient} = \text{inanimate}\} \\ & +0.48\{\text{person of recipient} = \text{nonlocal}\} \\ & -0.03\{\text{number of recipient} = \text{plural}\} \\ & +0.5\{\text{number of theme} = \text{plural}\} \\ & -0.46\{\text{concreteness of theme} = \text{nonconcrete}\} \\ & -1.1\{\text{parallelism} = 1\} - 1.2 \cdot \text{length difference (log scale)} \end{aligned}$$

and $\{c\} = 1$ if subject is in group c, 0 otherwise.

Positive coefficients favor PP dative, negative favor NP:

$$+0.99\{\text{accessibility of recipient} = \text{nongiven}\}$$
$$-1.1\{\text{accessibility of theme} = \text{nongiven}\}$$
$$+1.2\{\text{pronominality of recipient} = \text{nonpronoun}\}$$
$$-1.2\{\text{pronominality of theme} = \text{nonpronoun}\}$$
$$+0.85\{\text{definiteness of recipient} = \text{indefinite}\}$$
$$-1.4\{\text{definiteness of theme} = \text{indefinite}\}$$
$$+2.5\{\text{animacy of recipient} = \text{inanimate}\}$$
$$+0.48\{\text{person of rec} = \text{nonlocal}\}$$
$$-1.2 \cdot \text{length difference (log scale)} < 0 \quad [\text{len(rec)} > \text{len(th)}]$$
$$-1.2 \cdot \text{length difference (log scale)} > 0 \quad [\text{len(rec)} < \text{len(th)}]$$

This is harmonic alignment with syntactic position

Answer to Question 1:

The Harmonic Alignment effects on syntactic choice cannot be reduced to one single predictor.

In particular, the syntactic complexity in parsing hypothesis does not explain the influence of given-ness (and animacy, etc.) on the choice of dative syntax.

Question 2

A persistent question about corpus studies of grammar . . .

in Newmeyer's (2003: 696) words:

> "The Switchboard Corpus explicitly encompasses conversations from a wide variety of speech communities. But how could usage facts from a speech community to which one does not belong have any relevance whatsoever to the nature of one's grammar? There is no way that one can draw conclusions about the grammar of an individual from usage facts about communities, particularly communities from which the individual receives no speech input."

This is an empirical question:

What the speakers share in their choices of dative syntax might outweigh their differences.

The Switchboard Corpus is annotated for speaker identity.

424 total speakers $\Rightarrow$ total of 2360 instances of dative constructions

- 228 speakers $\Rightarrow 4-7$ each
- 106 speakers $\Rightarrow 8-12$ each
- 42 speakers $\Rightarrow 13-19$ each
- 11 speakers $\Rightarrow 20+$ each

The data are extremely unbalanced.

Speaker identity is a source of unknown dependencies in the data.

The effect of these unknown dependencies on the reliability of the estimates can be estimated from the observed data using modern statistical techniques:[a]

> When data dependencies fall into many small clusters (each speaker defines a 'cluster'), assume a 'working independence model' (our Model A) and revise the covariance estimates using bootstrap sampling with replacement of entire clusters.

[a] Efron and Tibshirani (1986, 1993); Feng, McLerran, Grizzle (1996); Harrell (2001)

in other words…

Create multiple copies of the data by resampling *from the speakers*. The same speakers' data can randomly occur many times in each copy.

Repeatedly re-fit the model to these copies of the data and use the average regression coefficients of the re-fits to correct the original estimates for intra-speaker correlations.

If the differences among speakers are large, they will outweigh the common responses and the findings of Model A will no longer be significant.

Result: the model coefficients are the same; the confidence intervals of the odds ratios[a] are wider, reflecting the reduction of independent observations in our data caused by the presence of clusters of speaker dependencies.

An odds ratio of 1 means that the odds of a dative PP and a dative NP are the same, so the outcome is 50%–50%.

We want the confidence intervals to stay nicely away from 1!

[a]—the intervals in which you can be confident that the chance of error stays below threshold (<5%)

Model A

Relative magnitudes of significant effects with corrected error estimates

	Coefficient	Odds Ratio PP	95% C.I.
inanimacy of recipient	2.54	12.67	5.56–28.87
nonpronominality of recipient	1.17	3.22	1.70–6.09
nongivenness of recipient	0.99	2.69	1.37–5.3
transfer semantic class	0.96	2.61	1.44–4.69
indefiniteness of recipient	0.85	2.35	1.25–4.43
plural number of theme	0.50	1.65	1.05–2.59
person of recipient	0.48	1.62	1.06–2.46
nongivenness of theme	-1.05	0.35	0.19–0.63
structural parallelism in dialogue	-1.13	0.32	0.22–0.47
nonpronominality of theme	-1.18	0.31	0.19–0.50
length difference (log scale)	-1.21	0.3	0.22–0.4
communication semantic class	-1.34	0.26	0.13–0.55
indefiniteness of theme	-1.37	0.25	0.15–0.44

Answer to Question 2:

The influence of discourse accessibility, animacy, and the like on dative syntax remain significant when differences in speaker identity are taken into account.

What the speakers share in the choice of dative syntax outweighs their differences.

To be continued . . .

Reading assignment for Thursday:

Read pages 1 to 16 (to line 8) of Bresnan, Cueni, Nikitina, and Baayen.

Is it at all convincing? What is your view?