In Defense of Corpus Data

(continued)

Summary from Week 2:

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 of the dative alternation

In the case of the dative alternation:

- effects of givenness, animacy *cannot* be explained by simpler theories (e.g. Hawkins 1994, Snyder 2003)
- pooled data from different speakers does *not* invalidate grammatical inference (contra Newmeyer 2003)

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.

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

And,

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.

3. The problem of lexical biases

What really drives the dative alternation *still* remains unclear.

We have assumed that NPs can be drawn out of the database and examined independently for their properties of discourse accessibility, animacy, pronominality, and the like.

But these NPs come from different verbs and different senses of the same verb!

Question 3:

Do the apparent effects of givenness and animacy on the choice of dative syntax hold, when they are conditioned on specific verb senses?

38 verbs $\times$ 5 semantic classes

Examples:

give.t = transfer: *give you an armband*

give.c = communication: *give me this cock and bull story . . .*

give.a = abstract: *give that a lot of thought*

pay.t = transfer: *pay somebody good money*

pay.a = abstract: *pay attention to cats*

55 verb senses in use in dataset

Use a multilevel model to condition the binary response on the verb sense:

Model B: Response $\sim$

fixed effects: 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) $-$ 1

random effect: verb sense

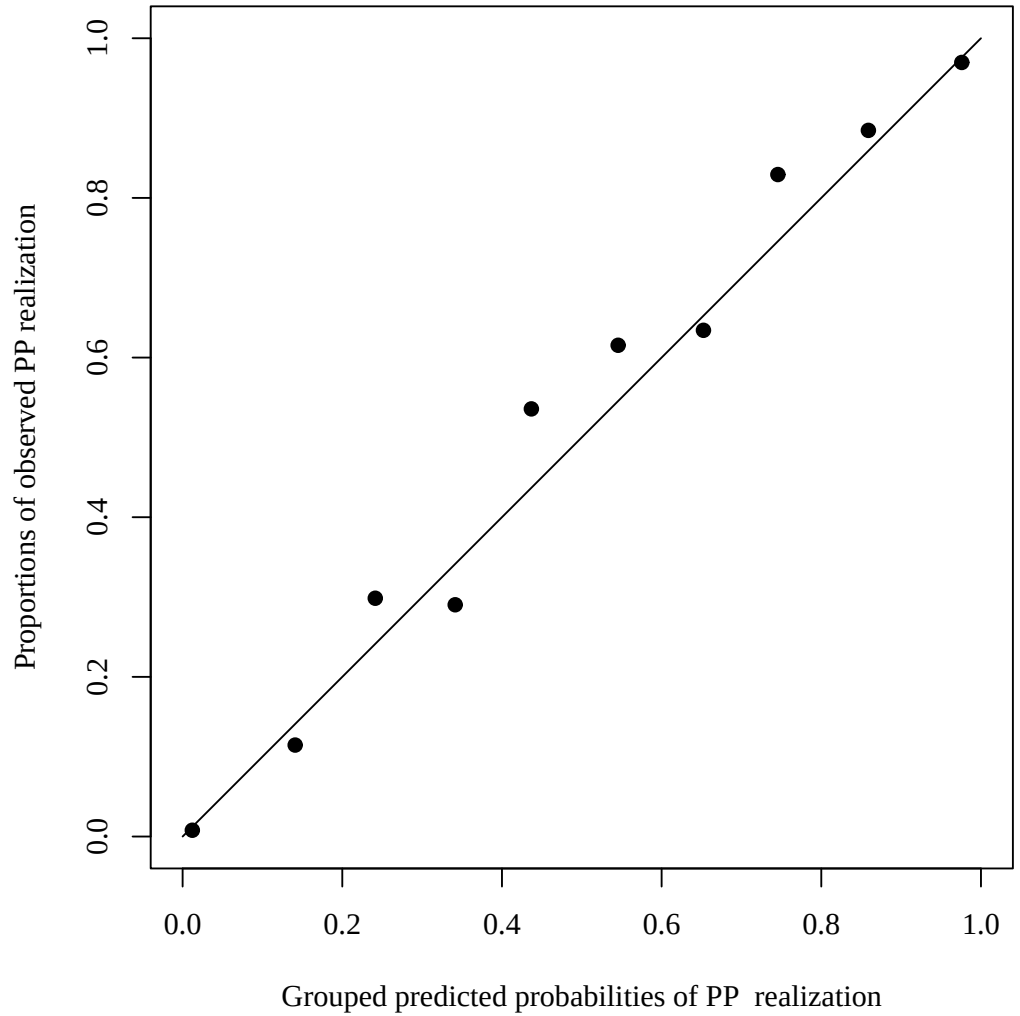
A Generalized Linear Model with a Single Random Intercept

$$logit[Pr(\boldsymbol{Y_{ij}} = y_{ij}|u_i)] = \boldsymbol{X_{ij}}\boldsymbol{\beta} + u_i$$

The conditional probability of a response given a cluster i is systematically linked to a linear combination of fixed cross-cluster explanatory variables $\boldsymbol{X_{ij}}$ and a randomly varying normally distributed cluster effect.

% Classification Table for Model B (1 = PP; cut value = 0.50)				
		Predicted:		% Correct
		0	1	
Observed:	0	1809	50	97%
	1	68	433	86%
			Overall:	95%

Model B plot of observed against predicted responses



How well does Model B 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 B parameters on each training set and score its predictions on the unseen testing set.

Mean overall score (average % correct predictions on unseen data) = 94%. Very good!

The model formula showing harmonic alignment:

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

$X\hat{\beta} =$

$1.5\{\text{a}\} + 0.58\{\text{c}\} + 0.96\{\text{f}\} - 3.28\{\text{p}\} + 2.7\{\text{t}\}$

$+1.5\{\text{accessibility of recipient} = \text{nongiven}\}$

$-1.2\{\text{accessibility of theme} = \text{nongiven}\}$

$+1.7\{\text{pronominality of recipient} = \text{nonpronoun}\}$

$-2.2\{\text{pronominality of theme} = \text{nonpronoun}\}$

$+0.7\{\text{definiteness of recipient} = \text{indefinite}\}$

$-1.7\{\text{definiteness of theme} = \text{indefinite}\}$

$+1.5\{\text{animacy of recipient} = \text{inanimate}\}$

$+0.4\{\text{person of recipient} = \text{nonlocal}\}$

$-0.2\{\text{number of recipient} = \text{plural}\}$

$+0.7\{\text{number of theme} = \text{plural}\}$

$+0.35\{\text{concreteness of theme} = \text{nonconcrete}\}$

$-1.1\{\text{parallelism} = 1\} - 1.2 \cdot \text{length difference (log scale)}$

and $\hat{u} \sim N(0, 2.27)$.

Relative magnitudes of significant effects in Model B

	Coefficient	Odds Ratio PP	95% C.I.
nonpronominality of recipient	1.73	5.67	3.25–9.89
inanimacy of recipient	1.53	5.62	2.08–10.29
nongivenness of recipient	1.45	4.28	2.42–7.59
indefiniteness of recipient	0.72	2.05	1.20–3.5
plural number of theme	0.72	2.06	1.37–3.11
structural parallelism in dialogue	-1.13	0.32	0.23–0.46
nongivenness of theme	-1.17	0.31	0.18–0.54
length difference (log scale)	-1.16	0.31	0.25–0.4
indefiniteness of theme	-1.74	0.18	0.11–0.28
nonpronominality of theme	-2.17	0.11	0.07–0.19

Answer to Question 3:

The influence of givenness and animacy and the other variables on the choice of dative syntax remain significant when they are conditioned on specific verb senses.

4. The problem of cross-corpus differences

Question 4:

Does it make sense to relate frequencies of usage to grammar? (Keller and Asudeh 2002: 240)

After all, unlike the grammaticality of a linguistic form, which is an idealization over usage, the actual frequency of usage of a form is a function of both grammatical structure and extra-grammatical factors such as memory limitations, processing load, and the context.

In fact it is true that

the frequencies of double-object constructions in the Switchboard collection of recordings of telephone conversations $\neq$

frequencies in the Treebank Wall Street Journal collection of news and financial reportage

V NP NP's $=$ 79% of total Switchboard datives ($n = 2360$)

V NP NP's $=$ 62% of total Wall Street Journal datives ($n = 905$)

Fit the same model to the combined data from two different corpora and compare fits to the components.

Model C: Response $\sim$

fixed effects: semantic class + accessibility of recipient + accessibility of theme + pronominality of recipient + pronominality of theme + definiteness of recipient + definiteness of theme + animacy of recipient + concreteness of theme + length difference (log scale) $-$ 1

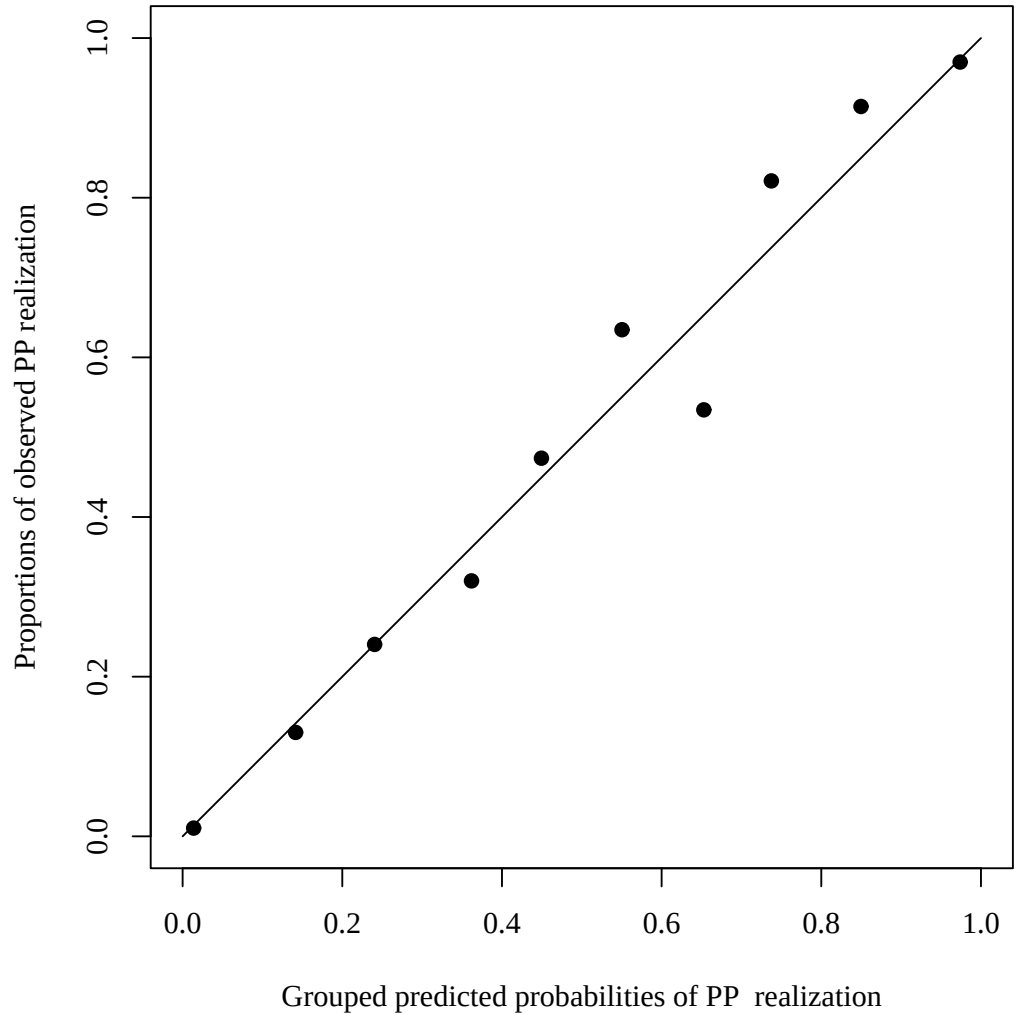
random effect: verb sense

Model C = Model B minus three factors (person, number, and parallelism) not marked in our Wall Street Journal dative dataset

Model C data has 110 different verb senses

Model C Classification Table (1 = PP; cut value = 0.50)				
		Predicted:		% Correct
		0	1	
Observed:	0	2320	96	96%
	1	119	730	86%
			Overall:	93%

Model C plot of observed against predicted responses



How well does Model C 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 $(n = 1265)$ and score its predictions on the unseen testing set.

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

Model C on component corpora

% NP NP's	Switchboard	Wall Street Journal
predicted	79%	63%
actual	79%	62%

How can this be?

Inputs vary. For example:

Wall Street Journal recipients: nouns outnumber pronouns 5 to 1

Switchboard recipients: pronouns outnumber nouns almost 4 to 1

The pressure for pronominal recipients to appear in the NP object position is about the same across the two corpora. *There are more double object constructions in the Switchboard corpus in part because there are simply more recipient pronouns.*

Setting pronouns aside, the proportion of dative NP NP constructions is higher in the Wall Street Journal data than in the Switchboard data, and Model C captures this difference between the corpora:

Model C on component corpora

% NP NP's (non-pronouns)	Switchboard	Wall St. Journal
predicted	49%	58%
actual	49%	55%

Again, how can this be?

Again, inputs vary. For example, *among non-pronoun complements to dative verbs:*

Wall Street Journal median length differential (log scale) = 1.1

Switchboard median length differential (log scale) = 0.69

The pressure for longer themes to appear at the end, favoring the V NP NP construction, is about the same in both of the two corpora. *There are more double object constructions in the Wall Street Journal corpus when we set pronouns aside in part because there are simply longer theme noun phrases.*

Answer to Question 4:

Some striking differences between different corpora can be explained as the response of the same model to quantitatively different inputs.

The statistical structure embedded in the model has generality and captures significant structural properties of language beyond the contingencies of a particular corpus.

But is there really *no* difference between the two corpora with respect to how strong the predictors are?

We investigated this question by adding to Model C an additional factor ‘modality’. whose value is ‘s’ for the Switchboard data and ‘w’ for the Wall Street Journal data, and then developing further models to study all interactions with modality.

There is a small but significant higher probability of using the V NP PP structure in the Wall Street Journal data, but there is no indication whatsoever that the other parameters of the model are different for data from the two corpora.

The simplest model, which treats modality as a simple main effect, is also the most accurate.

We conclude that the model for spoken English transfers beautifully to written, except that in written English, there is a slightly higher probability of using the prepositional dative structure.

Of course, it is always possible that in other registers and corpora and other regional varieties of English, further changes are required…

Conclusions:

The kinds of questions that have been raised about usage data are *empirical questions*.

- correlated factors seeming to support reductive theories
- pooled data invalidating grammatical inference
- apparent generalizations stemming from lexical biases
- cross-corpus differences undermining corpus grammar

Answers can be found by using modern statistical theory and modeling strategies widely used in other fields (biology, medicine).

Along with formal syntactic and semantic properties and relative structural complexity, the properties of animacy and discourse accessibility have a stable effect on dative syntax across written and spoken modalities, across verb senses, and across speakers.