



Environment prototypicality effects on syntactic alternation

Gabriel Doyle and Roger Levy
UCSD Linguistics

gdoyle@ling.ucsd.edu

<http://idiom.ucsd.edu/~gdoyle/papers/bls08-envproto.pdf>

Berkeley Linguistics Society

February 9, 2008



A Big Question

How do speakers choose among the various ways an idea can be expressed?



A Big Question

How do speakers choose among the various ways an idea can be expressed?

I'm hungry



A Big Question

How do speakers choose among the various ways an idea can be expressed?

I'm hungry

I feel hungry



A Big Question

How do speakers choose among the various ways an idea can be expressed?

I'm hungry

I feel hungry

I hunger

I want to eat



A Big Question

How do speakers choose among the various ways an idea can be expressed?

I'm hungry

I feel hungry

I hunger

I'd like some food

I desire food

I need to eat

I want to eat

My stomach is growling

I am peckish

Food, I want it!

Eating would be wise

I'm going to grab some food

I'm starving!

Feed me!

Give me food now

My tummy is rumbling

I need something to eat

I could stand to eat



A Big Question

How do speakers choose among the various ways an idea can be expressed?

- Very large (infinite?) space of possible expressions
- Different purposes, meanings, focuses
- What makes one variant preferred by speakers?



Approaching the Question

- Recent work on syntactic alternations
 - Dative (Bresnan et al 2007)
 - Genitive (O'Connor et al 2004)
 - Passives (Weiner and Labov 1983)
 - *That*-omission (Jaeger 2005, 2006)
 - Topicalization (Snider & Zaenen 2006)
- Narrowly construed alternations
 - Always two possible alternants
 - Still sheds light on the mechanics of speaker choice
 - Gradient constraints



Approaching the Question

- Gradient constraints

That movie gave me the creeps [NP NP]

*That movie gave the creeps to me [NP PP]



Approaching the Question

- Gradient constraints

That movie gave **me** **the creeps** [NP NP]

*That movie gave **the creeps** **to me** [NP PP]

This story is designed to give **the creeps** **to people who hate spiders** (Bresnan et al 2007)



Approaching the Question

- Gradient constraints
 - Semantics
 - Accessibility
 - Processing
 - Syntactic Parallelism
- Outstanding question: how does change of syntactic category affect alternations?
 - Many alternations use different syntactic categories
 - e.g., NP **NP** v. NP **PP** in dative
 - Previous studies haven't considered this as a factor
 - Key in understanding sentence-level variation



Outline

- The *needs doing* alternation
 - The dog needs to be fed ~ The dog needs feeding
- 4 steps to investigate the alternation
 - Establish mixed-category status/environment prototypicality
 - Show lack of categorical semantic constraints
 - Determine gradient factors with regression model
 - Investigate environment prototypicality



The *needs doing* alternation

needs to be done ~ **needs doing**

the *to be* form

the *ing* form

In many cases, both sound equally good:

The couch **needs to be cleaned**

The couch **needs cleaning**



The *needs doing* alternation

needs to be done ~ needs doing

the *to be* form

the *ing* form

The couch needs to be cleaned

The couch needs cleaning

past participle:
verbal (& adjectival?)

gerund:
verbal & nominal properties



Gerunds as a mixed category

English gerunds: mixed verbal/nominal categories
(Malouf 2000)

verbal: gerunds can govern NPs

verbal: adverbs, not adjectives, modify gerunds

verbal/nominal: genitive or accusative subject

nominal: NP-like external distributions



Gerunds as a mixed category

English gerunds: mixed verbal/nominal categories
(Malouf 2000)

verbal: gerunds can govern NPs

verbal: adverbs, not adjectives, modify gerunds

verbal/nominal: genitive or accusative subject

nominal: NP-like external distributions

His seeing Mike angered me



Environment Prototypicality

*The couch needs a to be cleaned.
The couch needs a cleaning.



Environment Prototypicality

Preceding
determiner



*The couch needs **a** to be cleaned.

The couch needs **a** cleaning.



Environment Prototypicality

Preceding
determiner

Prototypical nominal environment;
verbal element: **MISMATCH**

*The couch needs **a** to be **cleaned**.

The couch needs **a** cleaning.



Environment Prototypicality

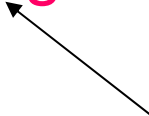
*The couch needs **a** to be **cleaned**.

The couch needs **a cleaning**.

Preceding
determiner



Prototypical nominal environment;
nominal/verbal element: **MATCH**





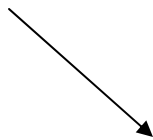
Environment Prototypicality

The paper needs to be completely rewritten.
?The paper needs completely rewriting.



Environment Prototypicality

Preceding
adverb



The paper needs to be **completely** rewritten.
?The paper needs **completely** rewriting.



Environment Prototypicality

Preceding
adverb

Prototypically verbal position,
verbal element: **MATCH**

The paper needs to be **completely rewritten**.

?The paper needs **completely rewriting**.

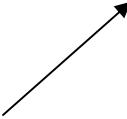


Environment Prototypicality


The paper needs to be **completely** rewritten.

?The paper needs **completely** **rewriting**.

Preceding
adverb



Prototypically verbal position,
nominal/verbal element: **MISMATCH**





Environment Prototypicality Hypothesis

- This will have a gradient effect on speaker choice
- Nominal properties of gerund makes *ing* form preferred in more nominal environments
 - and dispreferred in more verbal ones



Semantic (Near) Equivalence

- Can investigate this as a lexico-syntactic variable only if no categorical meaning-based constraints exist (Weiner and Labov 1983)
- Investigate 4 possible categorical semantic restrictions
 - Proposals 1-3, from Murphy (referenced by Murray Frazer & Simon 1996)
 - Proposal 4, our own



Semantic (Near) Equivalence

- Murphy's Proposals
 - *to be* form implies subject's possessor is agent
 - John's car needs to be washed
 - ⇒ John will wash the car
 - *ing* form implies pre-existing subject
 - (*?) My paper needs writing
 - *ing* form implies benefit to subject
 - (*?) My car needs selling
 - Abundant counter-examples in BNC, WWW



Semantic (Near) Equivalence

- Our proposal
 - achievement/state verbs can't use *ing* form
 - but in our subset of the BNC:
 - 39 achievement *ing*, 47 achievement *to be*
 - 5 state *ing*, 28 state *to be*



Moving past categorical constraints

- The alternation cannot be reduced to a set of categorical constraints
- Instead we consider a set of gradient constraints
- Strength and significance of gradient factors estimated by logistic regression model



Mixed-effects logistic regression

- estimates probability of the *to be* form
- considers discrete and continuous factors
- constraint strength based on coefficient magnitudes
- domination, ganging-up effects possible
 - like stochastic OT
- “random effect of verb”
 - each verb can have an idiosyncratic preference for one alternant
 - other factors controlled for



Dataset

- Examples extracted using a `tgrep` search on automatically parsed British National Corpus (BNC)
- Retrospectively sampled (Agresti 2002) random subset
 - 502 examples of each alternant (1004 total)
 - Original breakdown: 4436 *to be*, 1321 *ing*
- Hand-annotated for control factors



Control factors

Factors associated with “raised” subject:

- animacy, concreteness, definiteness, pronominality, relativization

Other factors:

- tense, inflection of *need*, modality, aspect, negation, modals, verb particle

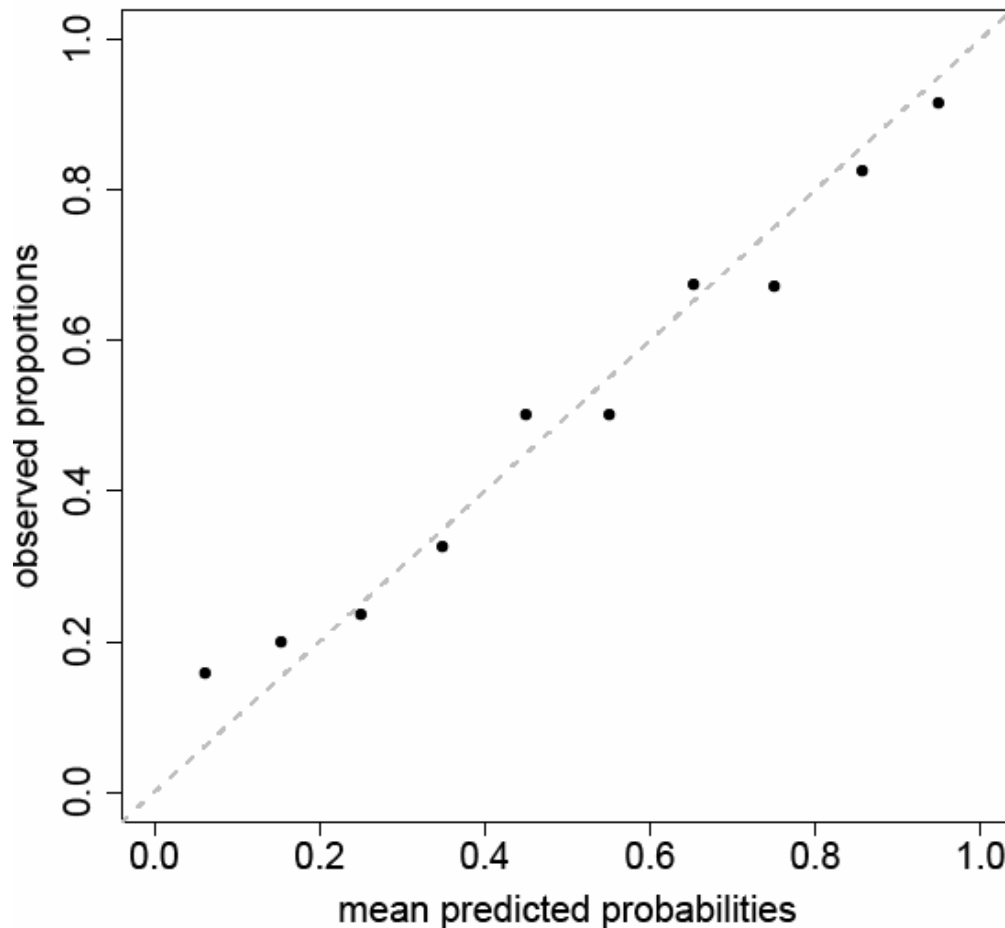
Quantitative predictors:

- verb length, verb frequency
- subject length, post-verbal dependent length, ambiguously-attached phrase length



Model results: overall accuracy

Quantiles of Predicted & Observed Probabilities of *to be*

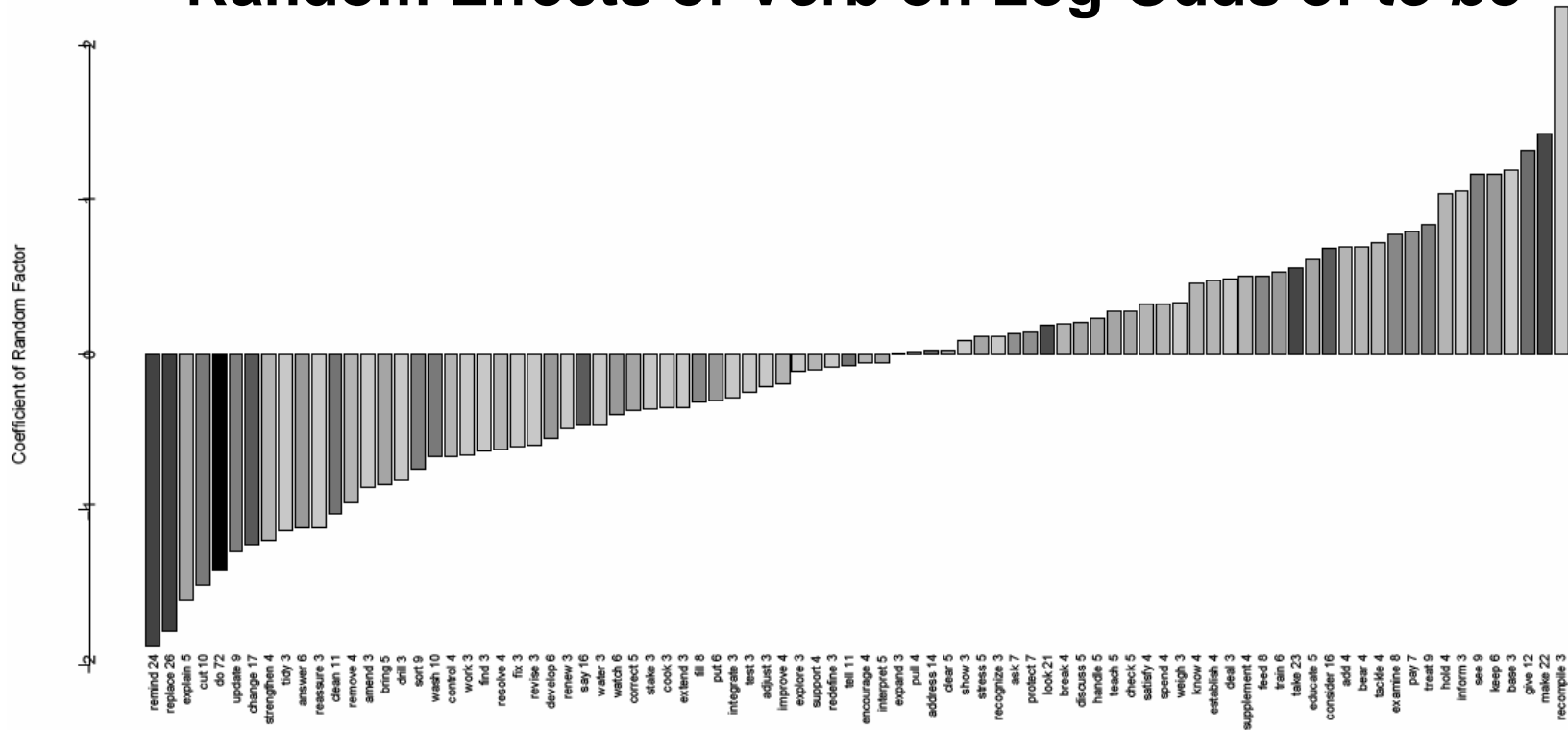


- regression model classifies 74% correctly in 5-way cross-validation (baseline = 50%)
- predicted and observed probabilities tightly correlated ($R^2=.98$)



Model results: verb-specific preferences

Random Effects of Verb on Log-Odds of *to be*

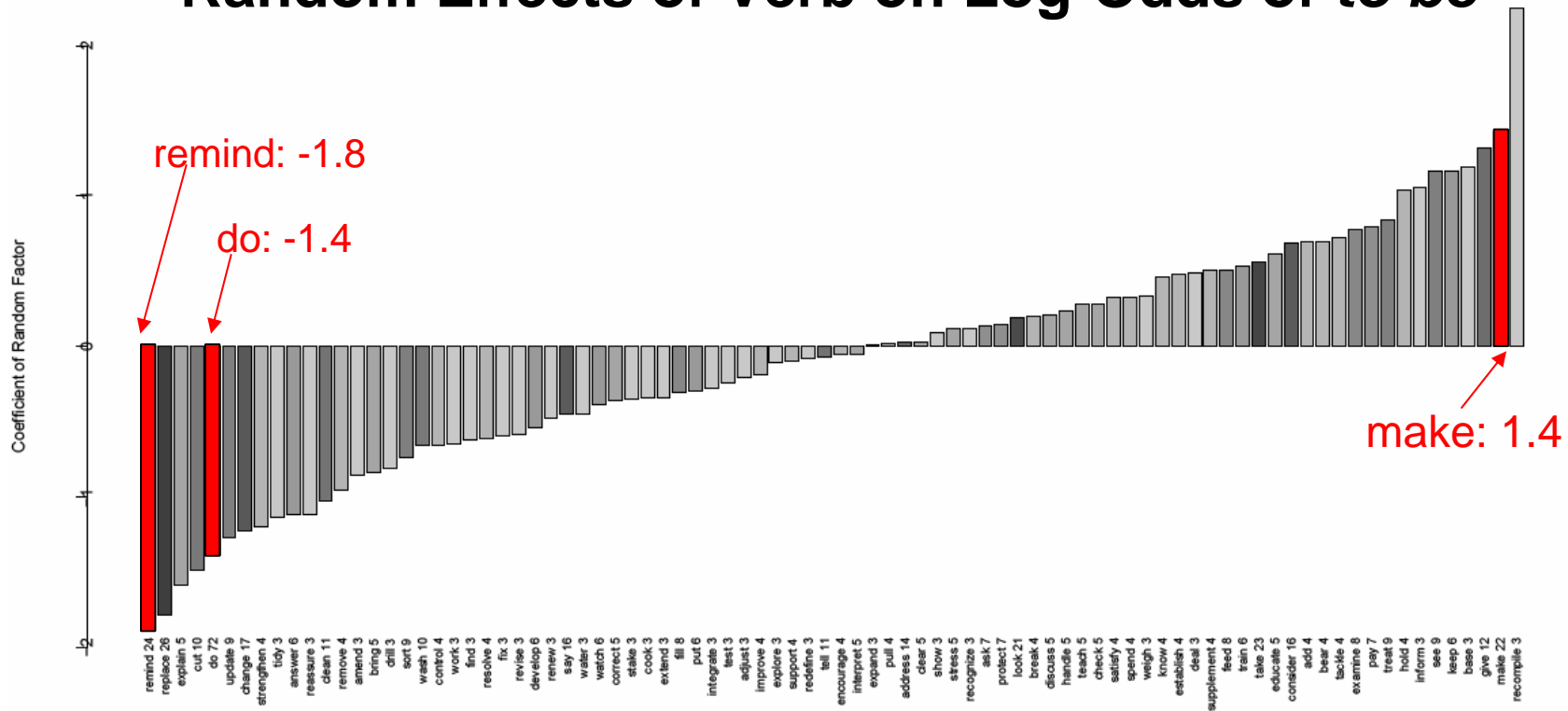


Positive favors *to be*, negative favors *ing*
darker bars mean more verb attestations

Random effect of verb $\sim N(0,0.546)$

Model results: verb-specific preferences

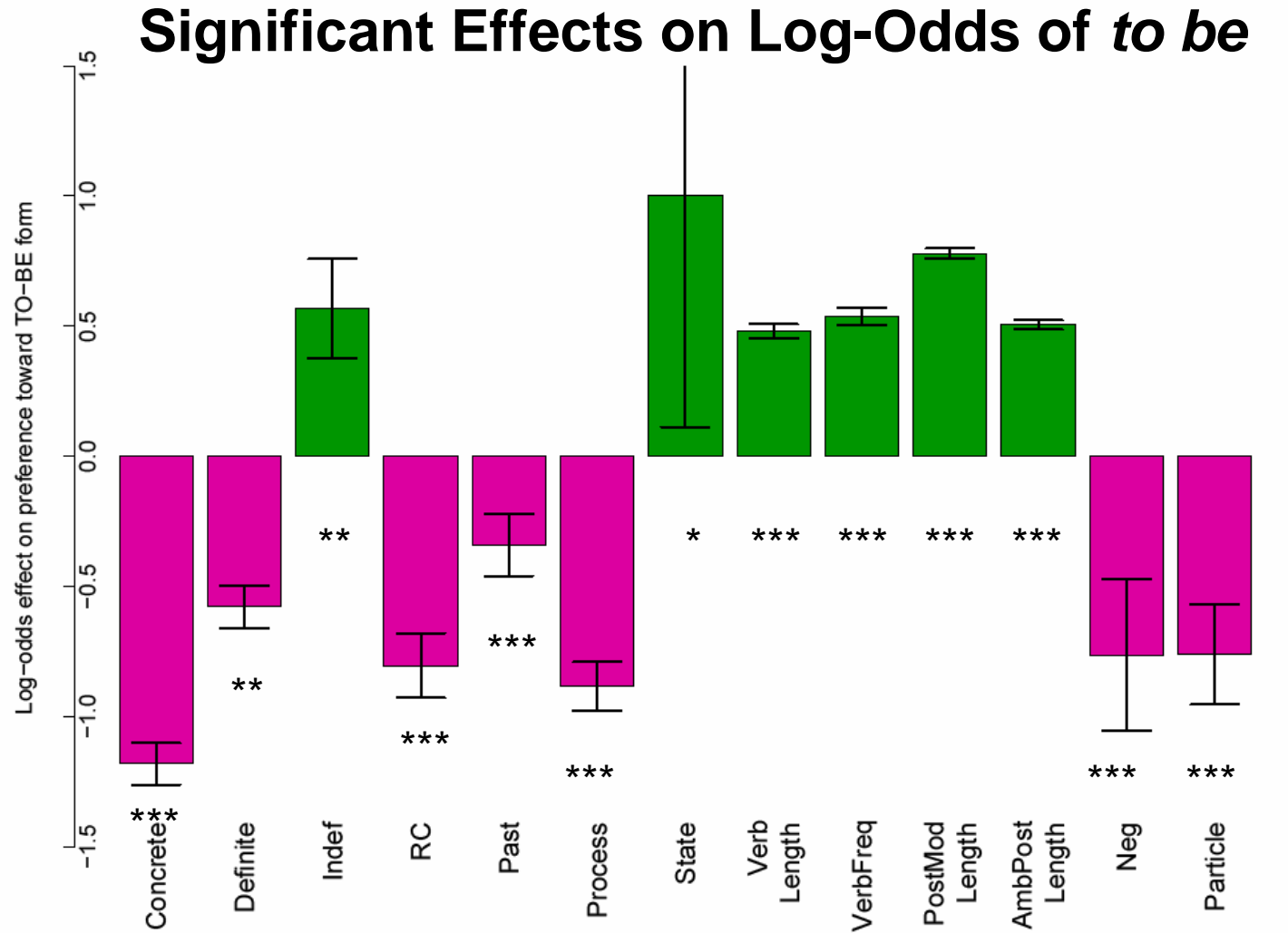
Random Effects of Verb on Log-Odds of *to be*



Positive favors *to be*, negative favors *ing*
darker bars mean more verb attestations

Random effect of verb $\sim N(0,0.546)$

Model results: control factors



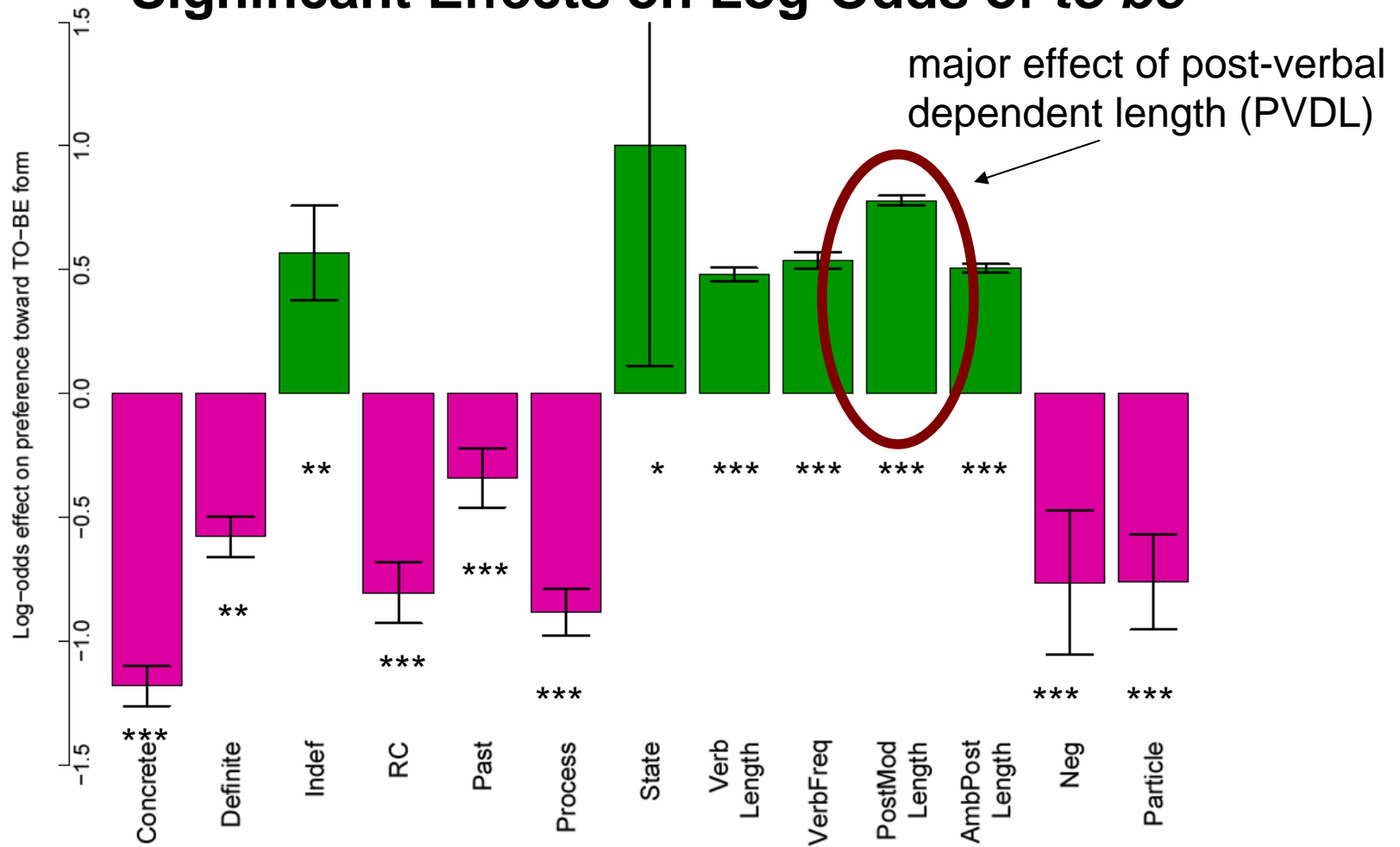
significance levels: * $p < .05$

** $p < .01$

*** $p < .001$

Model results: control factors

Significant Effects on Log-Odds of *to be*



significance levels: * $p < .05$

** $p < .01$

*** $p < .001$



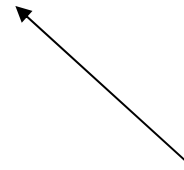
Structural Bias

- One distinguishing part of the environment is post-modifier type.
- Verb phrases and noun phrases have different prototypical post-modifiers
 - VPs: adverbs, by-phrases, complementizers
 - NPs: locative PPs, relative clauses
- Post-modifiers after the alternants tend to be more prototypically verbal post-modifiers
 - despite balanced dataset



Structural Bias

I need (tell) that he eats candy




CP post-modifier
prototypical verbal
environment



Structural Bias

I need to be told that he eats candy

Strictly verbal
alternant preferred



CP post-modifier
prototypical verbal
environment





Structural Bias

- Various measures of environment prototypicality available
- Structural bias: $\frac{P(\textit{Environment}|\textit{NP})}{P(\textit{Environment}|\textit{VP})}$
- Our approximation: $\frac{P(W_1|\textit{NP})}{P(W_1|\textit{VP})}$
- High structural bias implies more nominal environment
 - So high structural bias should favor *ing*



Including Structural Bias

- high structural bias corresponds to increased probability of *ing* form
 - as EPH predicts
- structural bias has marginal significance ($p=0.07$) if PVDL included
- significance of PVDL drops to $p=0.03$ (from $p<10^{-6}$), strength cut in half
 - no other factors change significantly
- structural bias and post-verbal dependent length highly correlated ($\rho=-0.91$)
- Both are good predictors, but is one or the other the cause?



Conclusions

- needs to be done ~ needs doing is driven by gradient, not categorical factors
 - similar to dative alternation, that-omission
- logistic regression model with 14 factors yields 74% cross-validated accuracy on classification
 - verbs have strong idiosyncratic effects on speaker choice
- environment prototypicality effects appear to be important in speaker choice
 - mixed-category membership is salient to speakers during production



Future directions

- Better probabilistic measures of environment prototypicality
 - disentangling PVDL and EP
- Acceptability rating study
 - does a non-prototypical environment make a sentence awkward?
- Self-paced reading time study
 - does a non-prototypical environment make a sentence harder to process?



Thanks!

- UCSD Computational Psycholinguistics Lab
 - Klinton Bicknell
 - Adam Bickett
 - Albert Park
 - Nathaniel Smith
- UCSD Center for Research in Language
- Joan Bresnan
- T. Florian Jaeger
- Tatiana Nikitina
- Tom Wasow
- Arnold Zwicky