(Super small) Data or Super (Small Data)? Using (Inter)Individual Estimates to Get the Most from the Least

a talk about data with surprisingly little data

Phillip M. Alday Matthias Schlesewsky
Ina Bornkessel-Schlesewsky

EXAL+, 16 January 2016

Introduction

A few small assumptions (making my bias explicit)

0. Languages differ, but there is one human brain, i.e. core computational mechanism, processing them all.

A few small assumptions (making my bias explicit)

0. Languages differ, but there is one human brain, i.e. core computational mechanism, processing them all.
1. The goal of psycho- and neurolinguistic research is to develop models of human language processing, both in perception and production.
2. Quantitative models are generally more specific and precise and thus more desireable than qualitative models.

A few small assumptions (making my bias explicit)

0. Languages differ, but there is one human brain, i.e. core computational mechanism, processing them all.
1. The goal of psycho- and neurolinguistic research is to develop models of human language processing, both in perception and production.
2. Quantitative models are generally more specific and precise and thus more desireable than qualitative models.

Research in practice entails developing and testing models both in their general shape and in their specific parameter values.

Why are we here?

Why are we here?

Because of (Indo-)Germanic hegemony

Why are some languages less studied?

- lack of interest
- lack of money
- lack of appropriate research infrastructure
- lack of "accessibility"
- dominance of Western European (esp. Germanic) and Chinese institutions

Arabic and other understudied languages . . .

Why is Arabic understudied?

Arabic and other understudied languages . . .

Why is Arabic understudied?

- widely spoken
- geopolitically important
- extensive written tradition
- interesting for a number of reasons

Arabic and other understudied languages . . .

Why is Arabic understudied?

- widely spoken
- geopolitically important
- extensive written tradition
- interesting for a number of reasons

We don't have much neurocognitive data on Arabic, but we do have a lot of other data.

I can't believe I'm saying this but maybe more traditional linguistics has something to teach us ...

Prior information

Traditional linguistics

Typology and "grammars"

- understand, acknowledge and model language variation and diversity (that's why we're here today)
- deep relationship between processing strategies and structures (*neurotypology*)
- genetic relationships and language change provide insights into local optima (Zipf's Law – coming to Evolang!)


Unless you're using evidence-based procedures, I can't hear a word you're saying.
someecards.com

(Slightly less) Traditional linguistics

Offline ratings

- acceptability judgments
  - use subject's own behavioral response in predicting neuro-response, not just as trial-rejection criterium
  - norming and pre-test data
- interpretations
- used e.g. in research on Competition Model from MacWhinney and Bates
- qualitative rankings can be made quantitative

Corpus linguistics

Frequency

- frequency plays a crucial role in many models of language processing and change
- accurate estimate and not just subjective impression allows for testing the desired hypothesis
- important whether
  - frequency influences processing strategy
  - processing constraints drive frequency
  - all of the above

Corpus linguistics

Existence

- null evidence always a problem
- as well as existence of speaker-produced errors

http://www.tofugu.com/wp-content/uploads/2015/05/cat-cartoon-grammar.jpg

Computer linguistics

Asking whether a machine can think is like asking whether a submarine can swim – E. Dijkstra

Computer linguistics

Asking whether a machine can think is like asking whether a submarine can swim – E. Dijkstra

Nonetheless, there are some common principles at work (some form of computation; buoyancy).

Computer linguistics

Stochastic incrementality

- modern research parsers have a stochastic component trained on large corpora
- this provides an ideal way to combine models and corpus data
- incremental parsers (e.g. shift-reduce parsers) additionally provide a model of the effects of incrementality

Natural language processing

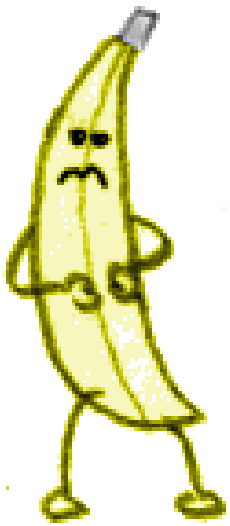
- focus on extraction and expression of meaning
- instead of mapping to a platonic abstraction

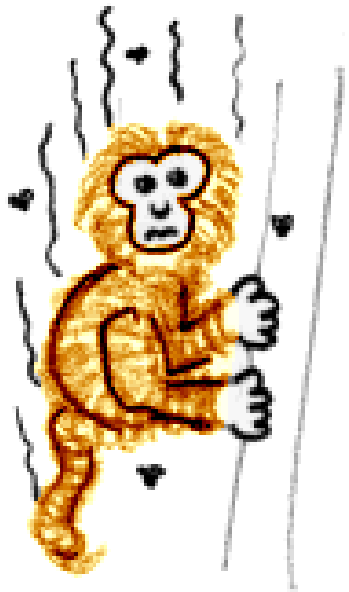
Computer linguistics

We gave the monkeys$_1$ the bananas$_2$

... because they$_1$ were hungry.

... because they$_2$ were ripe.





http://specgram.com/CLIII.4/08.phlogiston.cartoon.zhe.html

Iterative updates

- incorporating prior information – as model parameters/covariates, in structure of the model, or even as a prior in the Bayesian sense – is an iterative process.
- the basic logic of scientific research is based on successive approximations to reality.
- approximations are improved by removing error, and so we should focus on the things we can't yet explain.

Iterative updates in practice

- using *a priori* estimates from previous research can provide good model fit
- rapid quantitative empirical estimates of model free parameters possible
- both methods yield similar fit
- but the latter is preferred (less sensitive to researcher whim and post-hoc massaging)

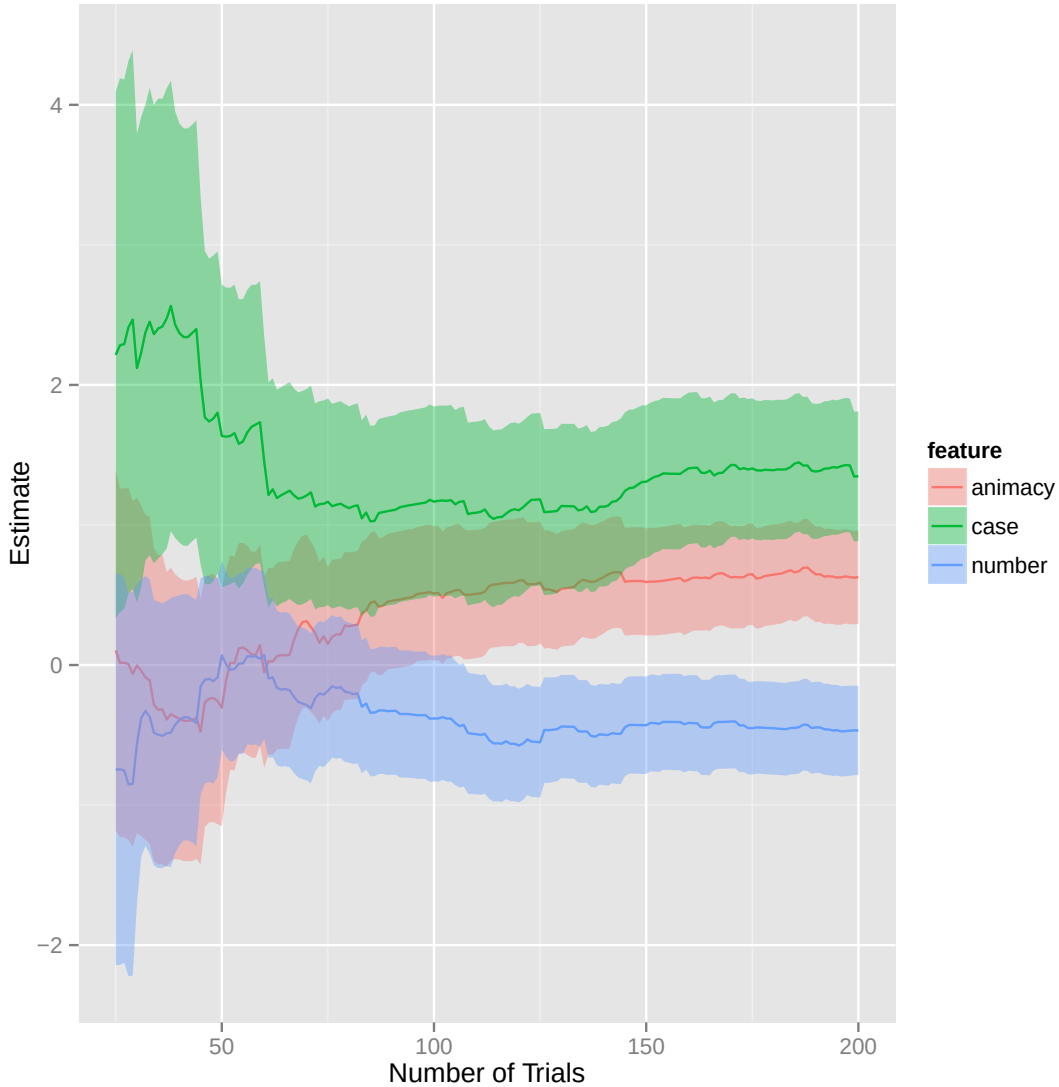
	Df	AIC	logLik
quantized from previous research	11	395190	-197584
estimates from a single-subject	11	395223	-197600
pooled estimates from four subjects	11	395252	-197615

Alday et al (2014, Neuroinformatics), Alday et al (2015, Ling. Vanguard)

Rapid acquisition of parameter values

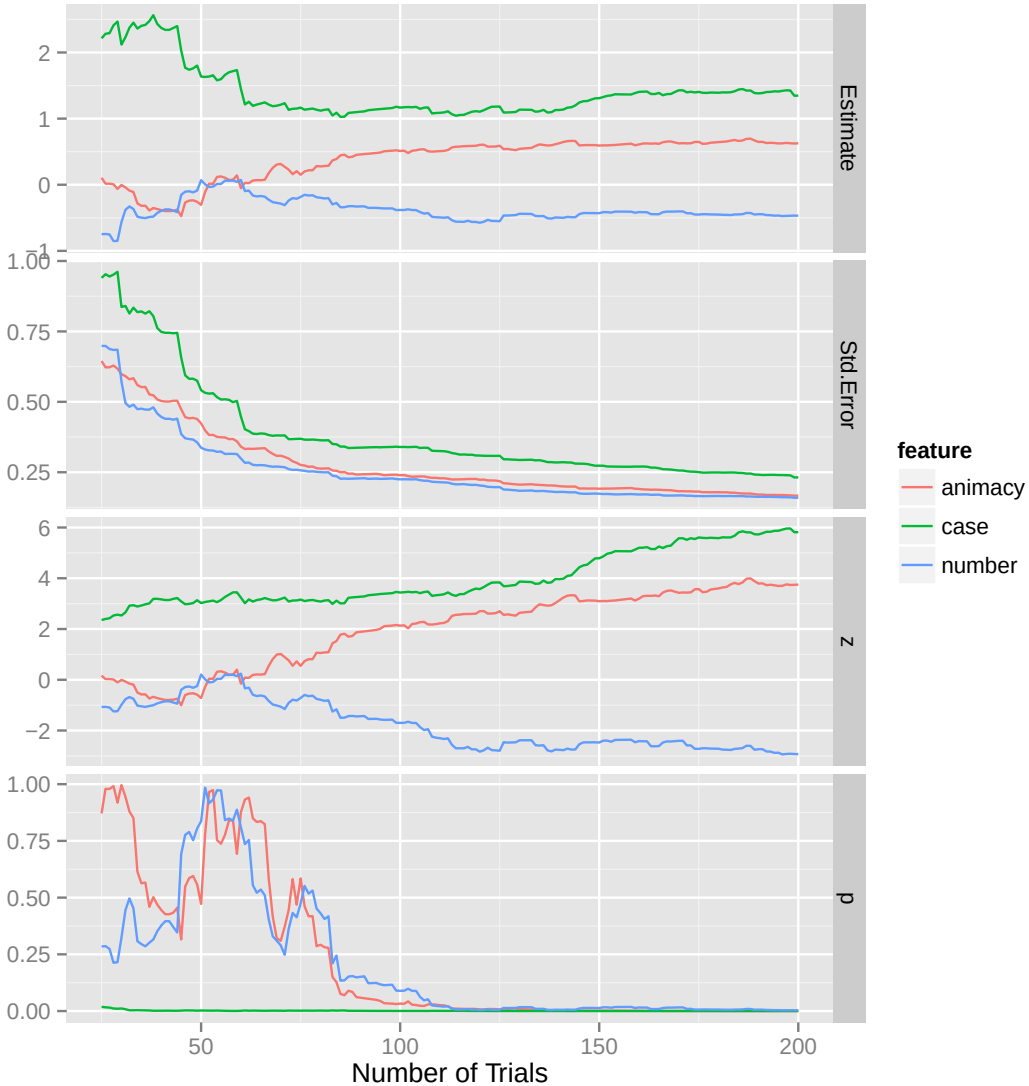
- simple behavorial design
- verb-final constructions in German: *dass* NP1 NP2 V
- NPs: random selection of fully crossed design for CASE x ANIMACY x NUMBER
- V: randomly assigned number, all transitive
- canonical, non-canonical and ungrammatical sentences (doubel accusatives, no agreement)
- task: pick the actor

Rapid acquisition of parameter values



Alday et al (2015), Linguistic Vanguard

Rapid acquisition of parameter values



Alday et al (2015), Linguistic Vanguard

Rapid acquisition of parameter values

- We managed to estimate three parameters in a model in less than 30 minutes.
- But for those parameters, we reached a point of diminishing returns well before the end of the experiment.
- We could have used our time better.

I can't believe I'm saying this but maybe more spam has something to teach us . . .

Adaptive testing

"Unbiased" Sampling

- purely random sampling is the foundation of experimentation

"Unbiased" Sampling

- purely random sampling is the foundation of experimentation
- except when it's not:
  - inclusion and exclusion criteria for participants and trials
  - constructed, controlled stimuli
  - etc.
- all of this done either before or after the experiment, but not during
- (ideally the criteria are determined before, even if applied after!)
- we choose our stimuli, participants and trials so that we can focus on hypotheses of interest

"Unbiased" Sampling

- purely random sampling is the foundation of experimentation
- except when it's not:
  - inclusion and exclusion criteria for participants and trials
  - constructed, controlled stimuli
  - etc.
- all of this done either before or after the experiment, but not during
- (ideally the criteria are determined before, even if applied after!)
- we choose our stimuli, participants and trials so that we can focus on hypotheses of interest
- with modern computing, we can focus in real time!

Bias isn't always a bad thing in statistics

- we can exclude conditions dynamically in an experiment
- we prioritize conditions based on variance and convergence rate
- priorities may change during an experiment
  - sudden improvement in one parameter
  - complex interaction leads to a previous estimate becoming worse
- priority has an upper bound to prevent "false" dominance via high variance / low convergence
  - poorly chosen model
  - trivial or non-existent effect
  - free variation (both inter- and intra-individual)

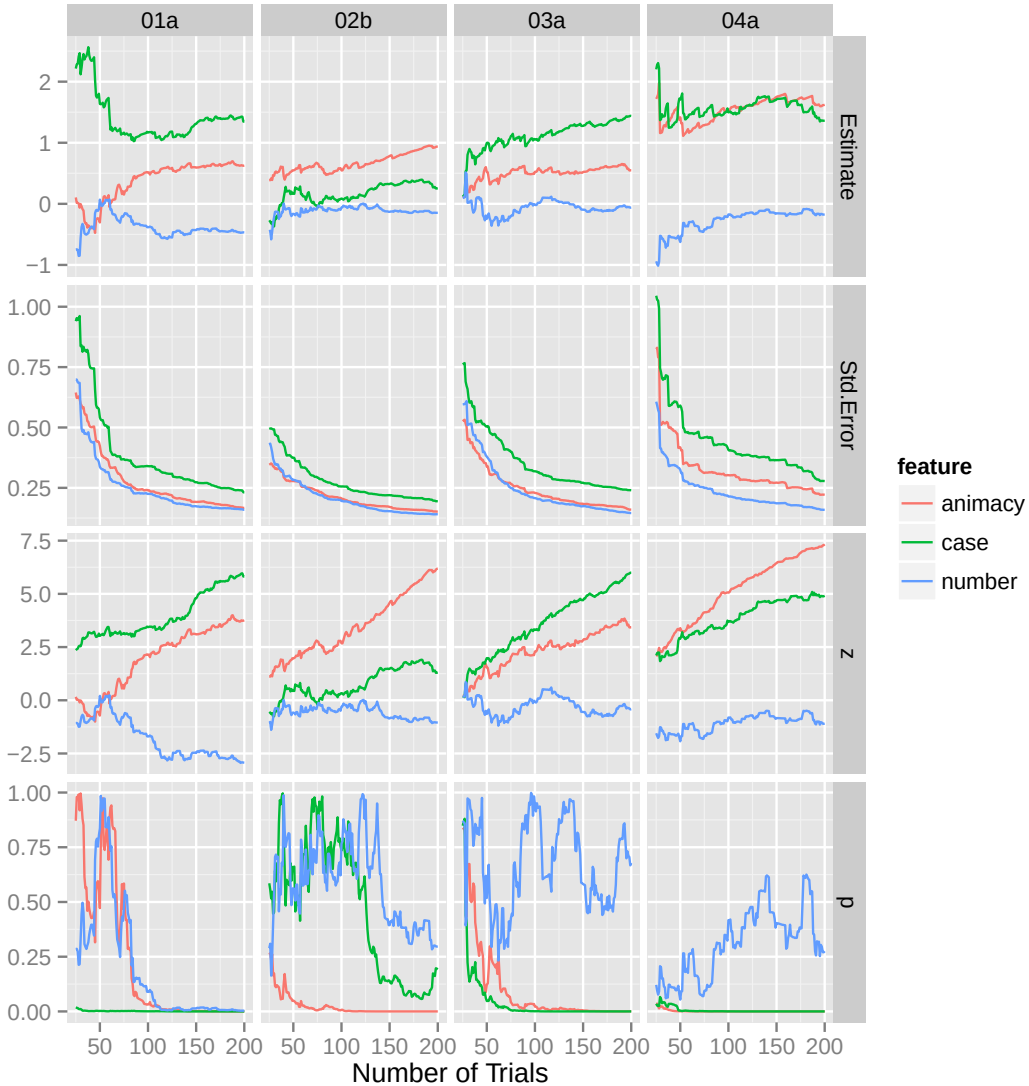
Bias isn't always a bad thing in statistics

- we can exclude conditions dynamically in an experiment
- we prioritize conditions based on variance and convergence rate
- priorities may change during an experiment
  - sudden improvement in one parameter
  - complex interaction leads to a previous estimate becoming worse
- priority has an upper bound to prevent "false" dominance via high variance / low convergence
  - poorly chosen model
  - trivial or non-existent effect
  - free variation (both inter- and intra-individual)
- biased estimators can outpeform unbiased estimators!

Precedent and prior formulation

- such adaptive techniques are widely used in spam filtering
- medical research has special rules for clinical trials with strong effects apparant early on
- philosophical relationship to MCMC, especially considerations of ideal chain length
- relationship to filtering – remove distracting, "less interesting" portions of the signal
- indeed this can be viewed as an example of Kalman filtering or recursive Bayesian estimation

Rapid acquisition, again



Alday et al (2015), Linguistic Vanguard

Principles and Parameters

(No, I'm not going to stop the bad word play any time soon.)

Principles I

Use the data we already have available

- previous research
- the Internet

Principles I

Use the data we already have available

- previous research
- the Internet

Focus on model specification and comparison

- reduce free parameters
- maximize predictive power
- strive for parsimony
- use modern statistical tools (mixed-effects models, Bayesian methods)

Principles II

Maximize entropy and optimize data collection

- between and within participants
- between and within experiments

Principles II

Maximize entropy and optimize data collection

- between and within participants
- between and within experiments

Hire a programmer and statistician

- (I have reasonable rates, but a very long wait list . . .)

From (super small) data to super (small data)

Stand on the shoulders of ~~giants~~ giant data sets

Big data provides an empirical, quantitative starting point.

Nucleation point for explosive, iterative growth

Recursion is not just something to study in language, but something for studying language! Research, participants and trials are not context-free.

In other words...

Use your computer and your statistics like it's 2016!

Questions?

http://www.economist.com/blogs/prospero/2015/05/johnson-polyglots

Slides, all published code, and some code in progress available at:
`palday.bitbucket.org`