Moving beyond Pāṇini: causal theories in linguistics

Balthasar Bickel
a VERY brief history of linguistics
or: why linguistics has a problem with causal theories
The origin of grammatical analysis

Pāṇini’s *Aṣṭādhyāyī* (fl. 4th c. BCE)

3,959 rules of Sanskrit

An example:

"2.3.1 if not already expressed,

2.3.2 for goal: case 2 (ACC)

2.3.46 for gender and number only (i.e. no role specs): case 1 (NOM)

3.4.69 for agent, goal or intransitive: *laḥ* (finite verb endings)"

We get can accusative on goals *because* it’s the law.
The origin of grammatical analysis

Pāṇini’s *Aṣṭādhyāyī* (fl. 4th c. BCE)

2.१.१ समर्थः पदविधः ।
2.१.२ सुभाषितवितः परासः ।
2.१.३ प्राकः कलारः समासः ।
2.१.४ सह मृपः ।
2.१.५ अव्ययभावः ।
2.१.६ अव्ययः सिभकिसमीपसमृद्धः-
व्यृद्धिभावः सम्प्रवायः-
शब्दप्रादभविस्म्यः सम्प्रवायः-
सम्प्रवायः ।
2.१.७ यथासः ।
2.१.८ यावदव्यः ।
2.१.९ सुप्रवन्ति ।
2.१.१० अक्षशालाकासः परिणः ।
2.१.११ विभाषः ।
2.१.१२ अपरिवलहिरः पश्चः ।
2.१.१३ आङ्ग्यावृत्तिविधः ।
Linguistics is engineering, even now

• Formulate the most concise, most parsimonious, most elegant description, like Pāṇini!

• Mostly a goal in itself: “pure linguistics” (Lazard 2012*)

• But perhaps not so interesting for other disciplines:
  • The most elegant and concise description may not capture
    • the generalizations by which children learn
    • the components that fit with the phylogeny of language
    • the units that brains process
  • Still, linguists adopt the Pāṇinian style even for cross-linguistic work...
• Fomulate a law and explain away any counter-examples!

• And so the law causes the facts!

• Illustration: The Final-Over-Final-Constraint (a modern version of Greenberg Universal #2; Biberauer et al. 2014*)

\[
\begin{align*}
\text{e.g. } & [\text{PP } [\text{NP YP N} ] \text{ P}] \\
\text{e.g. } & * [\text{PP } [\text{NP N YP} ] \text{ P}] 
\end{align*}
\]
Counterexample in Harar Oromo (Kushitic, Owens 1985)

\[
[PP [NP \text{maná} [NP \text{obbolesá} \text{ xiyyá} ] ] = tt]
\]

\begin{array}{cccc}
\text{house} & \text{brother} & \text{my} & \text{in} \\
N & NP & P
\end{array}

Solution: Explain the example away, e.g. limit the FOFC to complements with the same category features (Biberauer et al. 2014*) and argue that Oromo postpositions are [+V], or indeed not postposition at all.
Why not?

• Nothing is guaranteed to be exceptionless, not even “exceptionless \((p<.05)\)” (Piantadosi & Gibson 2014*)

• No idea what survived the human population bottlenecks 20-60kya!

• So pick generalizations that are justified (Chomsky 1964ff), but this leaves us in the end perhaps only with very abstract generalizations like
  
  • simple composition \((\alpha & \beta)\), as shared with other species (e.g. mongooses, Janssen et al. 2012+)
  
  • supra-regularity, as shared with other cognitive domains (e.g. action, Fitch 2014†)
  
  • recursion, as shared with other species when limited to regular grammars (e.g. Tamarin monkeys; Fitch & Hauser 2004‡)
  
  • asymmetry (categories), as shared with other species (e.g. Campbell monkeys; Ouattara et al. 2009§)

*Cogn Sci, †BMC Biol, ‡Phys Life Rev, ††Science, ‡‡PLOS One
A cheap way out

• Plough through databases, find soft constraints (correlations). Then explain them *post hoc*...

• **but this is the very problem that brings us here!**
  
  • sample?
  
  • missing data
  
  • unclear stochastic process
  
  • *causality?*
Perhaps after nearly 2500 years, it’s time to move on!
A more expensive way out: a normal science approach

• How is the (evolutionary, diachronic, ontogenetic) development of specific parts of languages *caused* by the natural and social ecology of language?

• For this, we need:

1. **Theories** on how natural and social conditions cause specific patterns in language evolution, change and development so that structures end up with the distributions we observe.

2. Fine-grained variables for *measuring* these distributions. Adequate iff
   - descriptively correct
   - cross-linguistically applicable
   - in sync with what we know about processing, acquisition

3. Statistical **models** for testing (1) against (2)
Theories

- General framework (cf. talks by Dan Dediu, Morten Christiansen, Florian Jaeger, Jasmeen Kanwal, Christian Bentz)
Causal theories — some examples

- **Event-based theories:** contact effects limited to concrete, *localized and historical* events, with no functional motivation, e.g. events in Eurasia in the least 14ky:

![Geographical distribution of NO clade](image)

Figure 2: Geographical distribution of NO clade. (a–g) Spatial frequency distributions of the NO clade: NO*, N (overall distribution of hg N), O (overall distribution of hg O), N*, N1, N2, N3. Maps are based on data from Supplementary Table 1. We label various panels following the YCC ‘by mutation’ format by adding the relevant mutation suffix.
Causal theories — some examples

- **Functional theories:** processing and communication principles cause certain directions in language change, e.g.
  - High cost of voicing in word-final position favors development and maintenance of final devoicing (Blevins 2004*)
  - Low humidity disfavors development and maintenance of rich tonal distinctions (Everett et al. 2015+; also Coupé’s talk)
  - Signal transmission in verb-final structures is safer with case makers (Hall et al. 2013†, Gibson et al. 2013‡)
  - Informative communication prefers certain lexical patterns (Regier’s talk)
  - Priming trends cause differences in NP frequency (Bickel 2003§)
  - **Perhaps:** supra-regular computation favors the development and maintenance of embedded phrase structures (“Dendrophilia”, Fitch 2014#)
• Signals may be weak

• Causes trigger *possible* change, but actualization requires many opportunities for change (many speakers, many generations) because:

  • uncertainty of social propagation (but once there, we get amplification through feedback loop in the next generation; cf Dediu’s talk)

• competing forces: e.g. contact events can enhance or suppress a principled trigger of change

• In fact, a causal trigger must not be too strong: it might harm communication and acquisition!
**Methodological challenge**

- must pick up signals of change: diachronic transition probabilities (Maslova 2000 etc.)
- even when languages don’t belong to a family (44-47% of all families have only 1 known member*)
Traditional approaches

- Family relations are a confound (Galton’s Problem, Simpson’s Paradox), so control for them by:
  - strategic sampling (Dryer 1989*), or re-sampling (Everett et al. 2015+)
  - modeling them as fixed (Dediu & Ladd 2007†, Bickel et al. 2009‡) or random (Jaeger et al. 2011§, Bentz & Winter 2013#) factors

- but...
  - even after controlling for confounds,
  - synchronic frequency estimates $\not\Rightarrow$ transition probabilities:
    - the process may not have reached stationarity (Maslova 2000¶)
    - indeed sometimes has not reached stationarity (Cysouw 2011‖),
    - especially when it is driven by local contact events!

*Stud. Lang, †PNAS, ‡PNAS, §Phon. Domains, ¶Ling Typ, #Lang Dyn Change, ‑Ling Typ, ‖Ling Typ
and more problems..

- also, shared inheritance or parallel development within a family can be the very signal we seek to pick up!
- E.g. DOM in Romance (e.g. Spanish *a*, Romanian *pe*) or Indo-Iranian (e.g. Hindi *-ko*, Nepali *-lāi*, Persian *râ*)
The Family Bias Method (or the Family of Family Bias Methods)

Core ideas:

1. Families are not a confound but demonstrated families are the very basis on which we can estimate transition probabilities (Greenberg 1978*, Maslova 2000+ etc.)

→ estimate difference in transition probabilities, eg. $P(A>B) > P(A<B)$: “family biases”

2. We can estimate family biases even for isolates and small families via extrapolation (Bickel 2013§)

*Univ. of Lang. 1, +Ling Typ., §Ling Typ Hist Cont
Step 1: estimating family biases in sufficiently large families

Set-based approach:

- Infer a family bias if A “dominates”, using e.g. a binomial test. (If nothing dominates, we don’t know.)

Tree-based approach:

- Estimate the best-fitting transition rate matrix $Q$ in a Continuous-Time Markov chain
- Infer a family bias if $q_{AE} \neq q_{EA}$ fits the data better than $q_{AE} = q_{EA}$ (LR or BF)
Step 1: estimating family biases in sufficiently large families

Assumptions

<table>
<thead>
<tr>
<th></th>
<th>set-based</th>
<th>tree-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>family model</td>
<td>tree, wave, linkage, network</td>
<td>tree (strict)</td>
</tr>
<tr>
<td>stochastic process of</td>
<td>independent multinomial trial</td>
<td>Continuous-Time Markov or Wiener process</td>
</tr>
<tr>
<td>diachronic event</td>
<td></td>
<td></td>
</tr>
<tr>
<td>data requirement</td>
<td>none</td>
<td>non-constant</td>
</tr>
<tr>
<td>family requirement</td>
<td>none</td>
<td>topology; branch lengths*</td>
</tr>
</tbody>
</table>

*e.g. length 1 between each node, assuming that anagenetic change in, say, the lexicon, is irrelevant for type change, especially if caused by contact (Thomason & Kaufman 1988)*
Step 2: estimate bias probabilities behind small families and isolates

- Use the mean probability of bias in large families for estimating the *probability that a small family is what survives of a large family with a bias* (in whatever direction). E.g. Laplace estimates on biases with 95%CI:

<table>
<thead>
<tr>
<th>Region</th>
<th>Africa</th>
<th>Eurasia</th>
<th>Pacific</th>
<th>N/C America</th>
<th>S America</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autotyp</td>
<td>.0</td>
<td>.027</td>
<td>.034</td>
<td>.0002</td>
<td>0.01</td>
</tr>
</tbody>
</table>

- if estimated to be biased, estimate direction of bias value (e.g. E) based on what they have, allowing for deviations with a probability based on deviations in large families, and resolving ties at random, e.g.

- take the mean across many extrapolations (e.g. 10,000)
Simulation of a discrete-time Markov process, where language varieties can (within steps of ca. 100 years ~ 3 generations)

- **give birth**: Poisson process with birth rate $\lambda = [.7, .8]$
- **die or stay alive**: Bernoulli process with survival prob. $\pi = [.1, .2]$

![Simulation Diagram](image)
Performance of methods in simulations (preliminary!)

• add a binomial variable with a family bias
• and see what we can recover, varying the definition of ‘small family’ and the rejection level of binomial test for inferring a bias in a family:

\[ \Delta \text{Pr(bias direction)} \]

\[ \text{large family threshold} \]

\[ \alpha = .05 \]
\[ \alpha = .1 \]
\[ \alpha = .15 \]

Mean \( \text{Pr(bias direction)} \) estimated lower than built in
Mean \( \text{Pr(bias direction)} \) estimated higher than built in
So, we have framework and a method
→ apply in two case studies focusing on methods
Case Study #1: the Trans-Pacific Hypothesis

- Causal theory grounded in the peopling of the Pacific and the Americas vs. the younger spreads in Eurasia 20-1kya and Africa in the past 2ky: contact triggers change towards similar properties

- **Hypothesis:** families show different diachronic biases in the Trans-Pacific area vs. elsewhere, keeping many diverse properties that were swept away through contact elsewhere
Case Study #1: the Trans-Pacific Hypothesis

- Data from AUTOTYP and (re-coded) WALS, $N \geq 250$, $k < 10$
- 354 multinomial variables coded for $N=[250, 1370]$ languages
- Set-based family bias estimates of large ($N \geq 5$) families with, $a=.1$
- Tree-based family bias estimates of non-constant large families, $BF>2$
- Extrapolations, then Fisher Exact Test of MEAN BIASES IN VARIABLE $\times$ AREA

**Methods:**
- set-based
- tree-based (MCMC)
- tree-based (ML)
Case Study #1: the Trans-Pacific Hypothesis

- False Discovery Rate ($q$) estimates (using Dabney & Storey’s 2014 bootstrap method):

<table>
<thead>
<tr>
<th>Method</th>
<th>Significant at $\alpha &lt; .05$</th>
<th>$q$ at that level</th>
<th>Significant at $q &lt; .1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set-based</td>
<td>73</td>
<td>0.16</td>
<td>32</td>
</tr>
<tr>
<td>Tree-based (MCMC)</td>
<td>71</td>
<td>0.15</td>
<td>26</td>
</tr>
<tr>
<td>Tree-based (ML)</td>
<td>43</td>
<td>0.27</td>
<td>17</td>
</tr>
</tbody>
</table>

- From this, subtract variants of variables, e.g re voicing distinctions in WALS:
  - MADVOI: \{\text{none, in\_plos\_\&\_fric, in\_plos\_only, in\_fric\_only}\}
  - MADVOI2: \{\text{none, some}\}

→ **30 true discoveries** (mean, set-based and MCMC-based estimates)
Case Study #1: the Trans-Pacific Hypothesis

- **Top 15:**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Source</th>
<th>N(lgs)</th>
<th>p (sets)</th>
<th>p (MCMC)</th>
<th>p (ML)</th>
<th>Trans-Pacific</th>
<th>Other</th>
<th>Variant of</th>
</tr>
</thead>
<tbody>
<tr>
<td>MADVOI2</td>
<td>WALS</td>
<td>565</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0001</td>
<td>-voicing</td>
<td>+voicing</td>
<td></td>
</tr>
<tr>
<td>DRYPOS</td>
<td>WALS</td>
<td>794</td>
<td>0.0000</td>
<td>0.0007</td>
<td>0.0069</td>
<td>+poss pref</td>
<td>-poss pref; +poss suff</td>
<td></td>
</tr>
<tr>
<td>MADVOI</td>
<td>WALS</td>
<td>565</td>
<td>0.0000</td>
<td>0.0018</td>
<td>0.0079</td>
<td>-voicing in plos/fric</td>
<td>+voicing in plos/fric</td>
<td>MADVOI2</td>
</tr>
<tr>
<td>DRYPOS0</td>
<td>WALS</td>
<td>591</td>
<td>0.0000</td>
<td>0.0003</td>
<td>0.0000</td>
<td>+poss pref; poss suff</td>
<td>-poss pref; +poss suff; -both</td>
<td>DRYPOS0</td>
</tr>
<tr>
<td>MADLAT2</td>
<td>WALS</td>
<td>565</td>
<td>0.0001</td>
<td>0.0002</td>
<td>0.0002</td>
<td>-latterals</td>
<td>+latterals</td>
<td></td>
</tr>
<tr>
<td>BAKADP2</td>
<td>WALS</td>
<td>377</td>
<td>0.0002</td>
<td>0.0002</td>
<td>0.0009</td>
<td>-adp</td>
<td>+adp</td>
<td></td>
</tr>
<tr>
<td>DRYGEN</td>
<td>WALS</td>
<td>1102</td>
<td>0.0002</td>
<td>0.0024</td>
<td>0.0009</td>
<td>-NGen</td>
<td>+NGen</td>
<td></td>
</tr>
<tr>
<td>MADLAT</td>
<td>WALS</td>
<td>565</td>
<td>0.0002</td>
<td>0.0031</td>
<td>0.0046</td>
<td>-non-obstr lat</td>
<td>+non-obstr lat</td>
<td>MADLAT2</td>
</tr>
<tr>
<td>DRYGEN0</td>
<td>WALS</td>
<td>1020</td>
<td>0.0002</td>
<td>0.0002</td>
<td>0.0001</td>
<td>-Nnp</td>
<td>-nNp; +nNp</td>
<td></td>
</tr>
<tr>
<td>POLYAGR</td>
<td>AUTOTYP</td>
<td>331</td>
<td>0.0004</td>
<td>0.0001</td>
<td>0.0018</td>
<td>-without; +POLYAGR</td>
<td>+without; -POLYAGR</td>
<td></td>
</tr>
<tr>
<td>DRYDEM0</td>
<td>WALS</td>
<td>1011</td>
<td>0.0004</td>
<td>0.0004</td>
<td>0.0017</td>
<td>+DemN; -NDem</td>
<td>-DemN; +NDem</td>
<td></td>
</tr>
<tr>
<td>MADPRS</td>
<td>WALS</td>
<td>565</td>
<td>0.0006</td>
<td>0.0000</td>
<td>0.0019</td>
<td>+Labial-velars</td>
<td>+Labial-velars</td>
<td></td>
</tr>
<tr>
<td>LOCUS.POSS</td>
<td>AUTOTYP</td>
<td>270</td>
<td>0.0008</td>
<td>0.0376</td>
<td>0.3543</td>
<td>-H</td>
<td>-H</td>
<td></td>
</tr>
<tr>
<td>MADTONO2</td>
<td>WALS</td>
<td>525</td>
<td>0.0008</td>
<td>0.0009</td>
<td>0.0029</td>
<td>+atonal; -tonal</td>
<td>-atonal; +tonal</td>
<td></td>
</tr>
<tr>
<td>HASWAN03</td>
<td>WALS</td>
<td>269</td>
<td>0.0011</td>
<td>0.0011</td>
<td>0.0055</td>
<td>+desid aff</td>
<td>+implicit subj; -desid aff</td>
<td></td>
</tr>
<tr>
<td>LOCUS.POSS.S</td>
<td>AUTOTYP</td>
<td>276</td>
<td>0.0013</td>
<td>0.0025</td>
<td>0.3346</td>
<td>-H</td>
<td>-H</td>
<td>LOCUS.POSS</td>
</tr>
</tbody>
</table>

- **Pearson Residual Analysis:**
  - 83% positive for outside Trans-Pacific (mean across methods)
  - 28% positive inside Trans-Pacific (mean across methods)
Case Study #2: The Anti-Ergative Hypothesis

• Primacy of A arguments in processing:

\[
\begin{align*}
\text{dass} & \quad \text{Peter} \quad \text{Lehrerinnen} \\
\text{that} & \quad \text{Peter: X/A/P? teachers: A/P?}
\end{align*}
\]

\[
\begin{aligned}
\text{mag} & \quad \text{[NP1 was A!]} \\
\text{likes} & \\
\text{mögen} & \quad \text{[NP1 was P!]} \\
\text{like} & 
\end{aligned}
\]

• The comprehension system tends to first assume that an unmarked initial NP is S or A, but not P

• If this NP later turns out to be P, this triggers an N400 (+ LPS):

→ ERP effect ("Anti-Ergative Effect")
Case Study #2: The Anti-Ergative Hypothesis

The Anti-Ergative Effect is independent of:

- **Frequency**: because of frequent A drop, initial NPs in Turkish tend to be P arguments, but the effect is still there (Demiral et al. 2008*)

- **Animacy**: initial NPs in Turkish tend to be inanimate, but the effect is still there (Demiral et al. 2008*)

- **Topicality**: initial NPs in Chinese show the effect regardless of whether the context makes them topical or not (Wang et al. 2010+)

- **The role played by {S,A} vs {P} alignment in grammar**: very restricted relevance in Chinese but the effect is there nevertheless (Wang et al. 2009#)
Case Study #2: The Anti-Ergative Hypothesis

And it even shows up in languages with ergative case, such as Hindi:

\[
\begin{align*}
kitāb & \quad \text{book(FEM)[NOM]} \\
bec-ī & \quad \text{sell-PP.FEM} \\
(Rām-ne) & \quad \text{Ram-ERG} \\
kitāb-ko & \quad \text{book(FEM)-ACC} \\
bec-ā & \quad \text{sell-PP.MASC} \\
(Rām-ne) & \quad \text{R-ERG}
\end{align*}
\]

Although Hindi NOM structurally includes and often prefers a P-reading, the processing system first interprets it as S or A!
Case Study #2: The Anti-Ergative Hypothesis

Hypothesis:

• If the Anti-Ergative Effect indeed applies universally to every unmarked initial NP, and if systems adapt to their processing environment, expect them
  ‣ to attempt to reanalyze initial NPs as covering \{S,A\}
  ‣ to avoid reanalyzing initial NPs as covering \{S,P\}
Case Study #2: The Anti-Ergative Hypothesis

• Tested on 617 languages, 712 subsystems (e.g. past vs. nonpast); excluding V-initial structures

• Controlling for possible event-based areal diffusion effects

\[ E(S \neq A) \quad A (S = A) \]

(means per language, across all NP types, clause types, and valency classes)
Bias for ergatives vs. against ergatives is determined both by:

- contact histories ($\text{AREA} \times \text{BIAS DIRECTION}$, LR $p<.01$)
- Anti-Ergative Effect: more ergative biases than anti-ergative biases across all areas (binomial $ps<.05$)

Results are the same across methods and genealogical data (set-based vs tree-based estimates, AUTOTYP vs. GLOTTOLOG trees etc.)
Conclusions

• Causal theories are tricky in traditional, Pāṇinian linguistics

• Alternative: theories of historical contact events and functional constraints → causes for biases in language change

• Now testable (though we obviously still need better methods, e.g. sensitive to partial tree or network structures in families)

• Describe language so we can test theories: descriptions need to become even more typologically informed than in the past