Beyond universals: exploring the conditions of language

Balthasar Bickel
‘God particles’ of human language?

How about the elements that are found in all and only human languages?
The classical typological answer: search for exceptionless universals


“But what an achievement it would be were we able to confront a language and say to it: ‘you have such and such a specific property and hence, also such and such further properties and such and such an overall character’ – were we able, as daring botanists have indeed tried, to construct the entire lime tree from its leaf. If one were allowed to baptize an unborn child, I would choose the name typology.”
“Universal laws”
Problems with the classical typological answer

- Problem #1: exceptionless in a sample ≠ impossible
- OK, but what if “exceptionless” = never observing in a sample with \( p < .05 \)?
- Piantadosi & Gibson (2013*):

\[N\] independent languages sampled

212 language families in WALS; many less are really independent

*Cogn Science
And of course we tend to find counterexamples:

• Claim: *Linear order is fixed within grammatical words, or depends on scope.*
  ‣ Counterexample: Chintang (Sino-Tibetan; Bickel et al 2007+)
    
    \[
    a-ma-ap-yokt-u-c-e \sim ma-a-ap-yokt-u-c-e \sim ma-ap-a-yokt-u-c-e \text{ etc.}
    \]

    2sA-NEG-shoot-NEG-3P-3ns-PST \quad \text{NEG-2sA-shoot-NEG-3P-3ns-PST} \quad \text{NEG-shoot-2sg-NEG-3P-3ns-PST}

    ‘You didn’t shoot them.’

• Claim: *Syntactic ergativity requires morphological ergativity*
  ‣ Counterexample: Oirata (Timor-Alar-Pantar; Donohue & Brown 1999+)
    
    a. \[inte \ [ihar \ [mara-n]] \text{ asi.}\]
        \quad 1peNOM \quad \text{dog.NOM} \quad \text{go-REL} \quad \text{see}

        ‘We saw the dog that had left.’

    b. \[ihar \ [ante \ \text{asi-n}] \text{ mara.}\]
        \quad \text{dog.NOM} \quad 1sNOM \quad \text{see-REL} \quad \text{go}

        ‘The dog that I saw.’

    c. \*[ihar \ [ani \ \text{asi-n}] \text{ mara.}\]
        \quad \text{dog} \quad 1sACC \quad \text{see-REL} \quad \text{go}

        ‘The dog that saw me left.’

*Language, +Austr J of Ling*
Worse: what do our samples represent?

- Unclear which structures survived the population bottlenecks in hour history (cf. Evans & Levinson 2009*, Dediu & Levinson 2013+)

Li & Durbin (2011#):

Autosomes Africa

Autosomes Out-of-Africa

(Estimates based on individual whole-genome sequences)

*BBS, +Frontiers in Psych, #Nature
Another problem: potentially spurious correlations

- If agglutinative, then verb-final and simple syllables
  (Konstanz Universals Archive #11, #372)
- If there is case, then there is number (KUA#116)
- If SHAPE as Adj, then COLOR and SIZE as Adj (KUA#141)
- If nouns inflect for case, verbs also inflect (for something) (KUA#228)
- If VO and atonal, then NRel.

(Roberts & Winters 2013*)
So, if exploring samples is not a safe route to exceptionless universals, what to do?
The classical structuralist answer: *guarantee* exceptionless universals!

Pāṇini’s Cakra: Formulate a generalization and then explain away counterexamples.

“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)”

› Apply Pāṇini’s Cakra to a universal, and you win!
How to guarantee universals with Pāṇini’s Cakra

- **Problem**: Lack of phonological syllables in Gokana (Hyman 1983*)
  - $C_2 \subset C_1$: $C_2$ only \{b, l, g\}
  - but differences are defined by words, not syllables:
    - $C_1V$, $C_1VV$, $C_1VC_2$, $C_1VVV$, $C_1VVVV$
    - $C_1VC_2V$, $C_1VVC_2V$, $C_1VC_2VV$, $C_1VVC_2VV$

- **Solution**: Assume syllables as universals nevertheless but add specific constraints: $C_2$ only in weak (second) syllables; and derive $C_1VC_2$ from $[\sigma C_1V][\sigma C_2V]$ (Hyman 2011+)

But why not an analysis without syllables?
$$bCV([blg]|V\{1,3}\|(V?([blg]V\{1,2}\}))\ b$$

*Curr Appr Afr Ling, +Phonology*
How to guarantee universals with Pāṇini’s Cakra

- **Problem:** Lack of grammatical or phonological words in Vietnamese (Schiering, Hildebrandt & Bickel 2010*):

  1. Engl. red → redd-ish, not reddish, but: *red-not-ish
  2. Vietnamese ṭọ-ṭọ, không ṭọ-ṭọ, oder: ṭọ-không-ṭọ

      or cà phê (from French café): cà với phê ‘coffee and the like’

- **Solution:** Assume words as a universal nevertheless (Vogel 2009†), but allow them to be interruptable under specific circumstances.

But why not assume a variable here?
Languages with vs. languages without words?

---

*J of Ling, †Universals of Language Today*
How to guarantee universals with Pāṇini’s Cakra

• **Problem:** Violations of Greenberg Universal #2 and of the Final-Over-Final Constraint in Harar Oromo (Kushitic, Owens 1985)

\[
[PP [NP maná [NP obbolesá xiyyá ] ] = tt]
\]

- house
- brother
- my
- in

\[ N \quad P \]

• **Solution:** Limit the FOFC to complements with the same category features (Biberauer et al. 2008*) and argue that Oromo postpositions are [-N], or indeed not postposition at all.

But why not assume a variable here? So that disharmonic head-final structures are dispreferred but not excluded?

*WCCFL*
How to guarantee universals with Pāṇini’s Cakra

• **Problem:** lack of nested phrase structures in Pirahã (Everett 2005*, 2009+)

• **Solution:** Assume nested phrase structures as a universal nevertheless and limit embedding to 1 level under specific circumstances (Nevins et al. 2009+).

But why not assume a variable here? Especially since we know how nested structures can come and go (e.g. [X [GEN-Y] ] < “X belongs to Y“ in Tok Pisin)

*Curr Anthr, +Language*
So, if throwing Pāṇini’s Cakra is not a safe route to exceptionless universals either, what next?
Justifying universals by first principles

• Criteria of Learnability (Chomsky 1964*ff): a universal is justified if we need it for explaining the fact that language is learnable.

  **But:** phonotactics, word structures, postpositions, non-nested NPs etc. are all learnable from the input even without assuming syllables, words, FOFC, obligatorily nested NPs!

• Perhaps, unlike this kind of stuff, we need at least hierarchical phrase structure, with labels and dependencies (MERGE), for learnability

  **But:** even CFG grammars (with strong generative capacity and structure dependence), turn out to be learnable from the input (Ambridge et al. 2008++, Perfors et al. 2011##)!
Justifying universals by first principles

• Considerations of Evolution (Chomsky 2004*ff): a universal is justified if we need it for explaining the fact that language evolved

  But: no evidence that syllables, words, FOFC, obligatorily nested NPs are needed, and such things would have unclear selectional advantages anyway

• Perhaps, unlike this stuff, we need at least MERGE for explaining language evolution because this directly captures the supra-regular capacity that humans have, unlike other species (cf. Fitch’s talk)

  But: MERGE is only one of many ways of computing supra-regular syntax; e.g. model-theoretic syntax (Pullum & Scholz 2001+), Construction Grammar etc. → no help for arbitrating universals

*in Structures and Beyond, in +Logical Aspects of Computational Linguistics
So, if not even considerations of “explanatory adequacy” are a safe route to exceptionless universals either, what now?
Two options

A. Keep on throwing Pāṇini’s Cakra anyway: keep universals as “working hypotheses”, or “programs”, and fight for them at all costs!

B. Give up on universals!
An alternative: a normal science, post-Pāṇinian approach

• How do specific parts of languages arise and develop over time (evolutionary, historical, or over the lifespan), given their natural and social ecology?

• For this, we need:

  1. Causal theories on how natural and social factors drive language evolution, change and development so that structures end up with the distributions we observe

  2. Fine-grained variables for measuring these distributions, formulated in sync with what we know about processing, acquisition etc

  3. Statistical models for testing (1) against (2)
Causal theories — some examples

- **Event-based theories**: once-off spreads, limited to concrete historical events, e.g. in Eurasia, over a period of at least the past 14ky

Rootsi et al. 2007 in *Europ J Hum Gen*, Maddieson 2005 in *WALS*
Causal theories — some examples

- **Functional theories:** cognitive/physiological and social/communicative principles cause certain directions in language evolution and change so that languages better fit their environment, e.g.

- High cost of voicing in word-final position favors development and maintenance of final devoicing (Blevins 2004*)

- Communicative need for distinguishing questions from statements causes development and maintenance of interrogative vs. declarative form (Dryer 2005+)

- *Perhaps:* certain kinship systems favor development and maintenance of special “kintax” morphology (Evans 2003 for review#)

- *Perhaps:* supra-regular computation in pattern recognition favors the development and maintenance of embedded phrase structures (cf. Fitch’s Dendrophilia Hypothesis)

*Evolutionary Phonology, +WALS, #Ann Rev Anth
Case study: a causal theory

• Joint work with Ina Bornessel-Schlesewsky demonstrates cognitive primacy of A arguments:

\[
\text{dass Peter Lehrerinnen that Peter: } \exists / A / P? \text{ teachers: } A / P? \quad \begin{cases} 
\text{mögen } [\text{NP1 was P!}] \\
\text{like } \\
\text{mag } [\text{NP1 was A!}] \\
\text{likes}
\end{cases}
\]

• 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 costs something:

\[\rightarrow \text{ERP effect ("Anti-Ergative Effect")}\]
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#)
And it even shows up in languages with ergative case, such as Hindi:

<table>
<thead>
<tr>
<th>Hindi Word</th>
<th>English Translation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>kitāb</td>
<td>book (FEM) [NOM]</td>
<td>kitāb</td>
</tr>
<tr>
<td>bec-ī</td>
<td>sell-PP. FEM</td>
<td>bec-ī</td>
</tr>
<tr>
<td>(Rām-ne)</td>
<td>Ram-ERG</td>
<td>(Rām-ne)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hindi Word</th>
<th>English Translation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>kitāb-ko</td>
<td>book (FEM) - ACC</td>
<td>kitāb-ko</td>
</tr>
<tr>
<td>bec-ā</td>
<td>sell-PP. MASC</td>
<td>bec-ā</td>
</tr>
<tr>
<td>(Rām)</td>
<td>R [NOM]</td>
<td>(Rām)</td>
</tr>
</tbody>
</table>

Although Hindi NOM structurally includes and often prefers a P-reading, the processor first interprets it as S or A!
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\}

• But expect actual signals in diachrony to be weak:
  • the costs are low and so ergative systems can be happily processed and transmitted over generations
  • actualization requires many opportunities for change (many speakers, many generations)
  • there are many counter-acting forces, e.g. conservatism, areal spread, new developments of ergatives, e.g. from focus markers highlighting the special saliency of agents, spread of special valency classes etc.
Testing the 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≠A)

A (S=A)

(means per language, across all NP types, clause types, and valency classes)
Need a method that

- captures effects over time, not simply synchronic distribution (because there is no guarantee of stationarity, Maslova 2000*)
- yet also picks up signals from isolates and small families

* and picks up signals from innovating as much as from maintaining a preferred structure
- and allows assessing confounding effects such as those from areal diffusion, other processing factors — and interactions between all these
The Family Bias Method

• Step 1: estimate biases in diachrony in large families (N > 5).

Several options, *two of which are used here*:

A. **Set-based methods** (ignoring tree topologies)

Observations in *demonstrably* related languages: Possible diachronic interpretations:

![Diagram of set-based methods]

Inference (under all interpretations):

\[ P(E > A) > P(A > E) \]  

(“Family Bias”)

\[ P(E > A) \approx P(E > A) \]  

(“no bias”, “diverse”, “neutral”)

→ Conclude bias if there are more A than E, as decided by a binomial test

Bickel 2011 in *Ling Typ*, 2013 in *Lang Typol and Hist Cont*; Software: familybias (comparativelinguistics.uzh.ch)
The Family Bias Method

**Step 1: estimate biases in diachrony in large families (N \( \geq 5 \)).**

Several options, *two of which are used here:*

**B. Tree-based methods**

(as used e.g. by Dunn et al. 2011* )

- Estimate the transition rate matrix of a continuous-time Markov model so that it maximizes the likelihood, e.g.

\[
L(D|T) = \sum_{i \in \{ A, E \}} P_{S_6}(i) P_{iA}(t) P_{iE}(t)
\]

- Or, approximate Bayesian marginal likelihoods via MCMC sampling over trees

- Compare these likelihoods to infer biases, 

\[
P(E \succ A) > P(A \succ E)
\]

(Pagel 1999*, 2004*; Felsenstein 2004#)

---

\*Nature, +Syst Biol, Inferring Phylogenies; Software: BayesTraits and R:geiger
### The Family Bias Method

- Set-based vs. tree-based estimates have both advantages and disadvantages:

<table>
<thead>
<tr>
<th></th>
<th>set-based</th>
<th>tree-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>need branch lengths (known or estimated)</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>need tree topology (known or estimated)</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>can handle invariant data</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

- Use both when possible and compare results.
- Same results in our dataset, except for Indo-European...
The Family Bias Method: Indo-European

Set-based: \( P(E > A) > P(A > E) \), \( p < .001 \)

Tree-based, \( P(E > A) \approx P(A > E) \)
ML logBF=.08 (\( p = .77 \))

(Topology and branch lengths based on nodes in Glottolog)

glottolog.org (Hammarström & Nordhoff 2014)
The Family Bias Method: Indo-European

- But no difference when based on estimated instead of fixed tree (BayesPhylogenies based on cognate replacements; Dunn et al. 2011*):

\[ P(E > A) > P(A > E) \]

MCMC logBF = 6.49

*Nature, Thanks to Michael Dunn for sharing the trees
The Family Bias Method

- 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: S=A or S≠A)
  - if estimated to be biased, estimate direction of bias value (e.g. S=A) based on what they have, allowing for deviations with a probability based on deviations in large families, and resolving ties at random
  - take the mean across many extrapolations (e.g. 2,000)

Simulation study shows that this method is very conservative:
- overestimation of biases and bias direction ≤ .05
- underestimation ≤ .21 for biases, ≤ .07 for bias directions
Bias for ergatives vs. against ergatives is determined both
• by contact histories (\(\text{AREA} \times \text{BIAS DIRECTION}, \ p = .003\))
• by Anti-Ergative Effect: proportion of ergative biases smaller than proportion of anti-ergative biases across all areas (all \(ps < .05\))

Results are independent of method for large family estimates (set-based, tree-based, ML, MCMC, AUTOTYP vs. GLOTTOLOG trees etc.)
Diversification strongly depends on area ($p < .001$)
Conclusions

• Results do not depend on
  • individual datapoints ("counterexamples") and fights on what is the "right" analysis (throwing Pāṇini’s Cakra), but on general, quantifiable patterns
  • sampling choices, since with methods like the Family Bias Method we can use exhaustive samples (unlike in classical, sampling-based typology)

• Approach in line with the normal science triad — causal theory, data, statistical modeling

• and in line with the old insight that nothing in linguistics makes sense expect in the light of history (cf. Dobzhansky re biology)

http://www.spw.uzh.ch/distributionaltypology
Thanks to my colleagues in the Anti-Ergative study

- Ina Bornkessel-Schlesewsky (Marburg)
- Kamal Choudhary (Ropar, India)
- Alena Witzlack-Makarevich (Kiel)