Exploring similarities across linguistic structures: phylogenetic methods beyond phylogeny

Balthasar Bickel
University of Zurich
Goals

• Show that comparative linguistics has severely suffered from a categorical non-quantitative approach to structure (the Pāṇinian Approach)

• Urgent need for quantitative similarity analysis

• Phylogenetics offers extremely useful tools for this

• Case studies

• Can perhaps be generalized to other areas in the humanities that face problems with categorical concepts (cf. the constant appeal of Wittgenstein’s Family Resemblance idea)
Similarities between linguistic structures

• An old problem in comparative linguistics: how to develop cross-linguistically applicable notions for analysis?

Capturing languages “in their own terms”  ↔  Generalizing about human language(s)
Example: a cross-linguistic notions of “adverbial clause”?

• Standard approach (not only in linguistics but more generally in the humanities): “the Insightful Definition Approach”
  • Chapter 1: come up with a good definition
  • Chapter 2: apply it
  • Chapter 3: gain insights

• So let’s try: “an adverbial clause is a clause that is:
  • dependent on a main clause,
  • but not functioning as an argument of the main clause predicate”
Example: a cross-linguistic notions of “adverbial clause”?

• But somehow different from “chaining”:

Taunya (MacDonald 1990)

wawiya matafo sa?ama ?afa fei-mene-a-te ate-ni yau-fe-o. yau-fe-pa fofe si?i mango young small INDEF ready-STAT-3s-DS old.woman-ERG see-PERF-DEP see-PERF-SS come mark
tefe-a-te mopamo ne-pi-sou ai ne-pi-sou fofe-pa ?e si?i ne-pi tei-?ora?aro-ti-pa wai put-3s-DS yo.sib 3s-GEN-COM e.sib 3s-GEN-COM come-SS DEM mark 3s-GEN touch-erase-PERF-SS again
nen-sa teme si?i tefe-i-nu.
3p-CONTR other mark put-3p-NARR

‘Only a young small mango was ready, and the old woman saw it. She saw it and came and put a mark on it. Then the younger brother and the elder brother came and erased that mark of hers and they themselves put another mark (on the mango).’
Example: a cross-linguistic notions of “adverbial clause”?

• In response to this, Foley & Van Valin (1984) added more defining properties:
  • disjunct scope
  • no WH

Taiya (MacDonald 1990)

a. * nen mei momune-i-nani=ra pofei-ti nen=tu-e=nae?  
   3p here sit-3p-ASS=TOP talk-CONJ 3p=give-2=POLAR.Q
   ‘They sat here and/but did you talk to them?’
   or ‘When they sat here, did you talk to them?’
   or ‘Did you talk to them when they sat here?’

b. * we mei fofe-a-nani=ra ...
   who[NOM] here come-3s-ASS=TOP
   Intended: ‘*if who did come here...’
Example: a cross-linguistic notions of “adverbial clause”?

• Nice contrast with chaining:

   Tauya (MacDonald 1990)
   
   a. *tepau-fe-pa  yate fitau-a=nae?*
      break-PRF-SS go  throw-2=POLAR.Q
      ‘Did you break it and go away?’
      or ‘Did you go away after breaking it?’ (presupposing either ‘you went away’ or ‘you broke it’)
   
   b. *ne-ni  we  tu-a-te  yau-i=ne?*
      3s-ERG who[NOM] [3sP]give-3s-DS [3sP]see-3p=PARAMETRICAL.Q
      ‘Who did he give it to? and they saw him’ (‘Who did he give it to when they saw him?’)
Example: a cross-linguistic notions of “adverbial clause”?

• BUT there is “adverbial subordination” with disjunct scope and WH:

Belhare (Bickel 1993)

a. ne-e yuŋ-a=naa mundhupt-he i?
   DEM-LOC [3sS]sit-SBJV.PST=TOP [3sS]chat-PST Q
   ‘When he was here, did he say something?’ (or was he silent?)
   or ‘Did he say something when he was here?’ (or later only?)
   but not ‘Was he here, and did he say something?’

b. sa-a ya=m-phekt-a-k=naa n-κhatd-at-ni-gak=phe?
   who-[s]ERG call=3nsA-call-SBJV.PST-2=TOP NEG-go-PST-NEG-2=IRR
   ‘You hadn’t gone if who had called you?’
The heart of the problem

- Tauya and Belhare both have “adverbial subordination”
- but they can’t be instances of the same universal concept because they don’t share all defining properties:
  - Tauya: \{dependent, non-argumental, disjunct, no WH\}
  - Belhare: \{dependent, non-argumental, disjunct, WH ok\}
- i.e. there is **hidden diversity** behind such notions as “adverbial subordination”
  - just like behind notions such as “agent”, “voiced”, “phonol. word”, “/a/”
  - and even more behind such notions as “perfective”, “participle” etc.
The heart of the problem

• What to do? What’s the right definition? Who says?

• Is a ban on WH a critical property of subordination?
  • if yes, what about Belhare? Why this apparent exception?
  • if not, the concept and its definition are no longer insightful:

  We could no longer say: “you can’t form WH questions because the clause is subordinate”

  or: “it’s subordinate; therefore, we predict that a ban on WH”

  and, vice versa, we could no longer say: “there is is ban on WH; therefore, it’s subordinate”

  or: “a ban on WH is a test/diagnostic for subordination”
The heart of the problem

• On close inspection, all **insightful concepts** are **absolute universals** in disguise:
  
  • subordination: \{no WH\} ↔ \{disjunct scope\}
  
  • or even: \{no WH\} ↔ \{disjunct scope\} ↔ \{focusability\}
  
• But, then, what to do with exceptions?
The classical response

• The universal is true. Let’s explain away the exception:

• There must be something special in Belhare that overrides the constraint against WH, or what looks like WH, isn’t really WH

(This follows from the general Pāṇinian Approach that has dominated research on grammar since its beginnings: language as a categorical system)

fl. 5th/6th century BCE
One current alternative response: retreat!

- The exceptions are just language-specific noise; for comparison, use “comparative concepts” that gloss over the noise (Haspelmath 2010)

- so, define adverbial subordination as a comparative concept only via disjunct scope; or via ‘non-assertion’ or whatever

- But this creates the illusion of more uniformity than is empirically warranted (cf. Evans & Levinson 2009):
  - “How many languages have adverbial subordination?” (which kind do you mean?)
  - “How does it develop over time?” (which kind?)
Another attempt: Multivariate Typology

- Structures are mostly similar and hardly ever identical.
- Similarity is a simple concept: identity in some, difference in other variables.
  - For studying similarities, we need large systems of fine-grained variables that fully capture the range of known variation: Multivariate Typology.
  - And with this, we can describe and measure the variation, instead of reducing it — i.e. do what most other disciplines would do when confronted with variation.
An alternative: Multivariate Typology

- How many structures? — As many as are distinct in $V_1 \ldots V_k$

- How many variables? — As many as are of interest to the research question, e.g. all variables needed to capture cross-linguistic differences in the syntax of clause linkage; or in morphological coding.

- Which variables? — Developed as needed for distinguishing structures during data collection (Bickel & Nichols 2002), or pre-determined by the research question.

  - structural variables (finite list of choices): morphosyntactic or semantic properties in which structures are alike or differ.

  - denotation variables (non-finite list of choices): denotations (stimuli, contexts, functions) in which structures are alike or differ
A multivariate typology of clause linkage
Some variables

**IIL-scope:** The scope of illocutionary operators in the main clause is
- **conjunct:** extends to the main clause and the dependent clause
- **disjunct:** extends to either the main or the dependent clause but never to both
- **local:** is limited to the main clause
- **extensible:** extends to either the main clause alone or to both the main clause and the dependent clause, but never to the dependent clause alone
- **constraint-free:** is not regulated by the clause linkage type

**T-scope:** The scope of tense or status operators in the main clause is
- **conjunct:** extends to the main clause and the dependent clause
- **local:** is limited to the main clause
- **extensible:** extends to either the main clause alone or to both the main clause and the dependent clause, but never to the dependent clause alone

**Finiteness:** The dependent clause is headed by a verb form that is
- **finite:** at least as many categories must be marked as in main clauses
- **nonfinite:** only fewer categories are allowed
- **any:** either the same range or less categories can be marked

**IIL-mark:** Marking of illocutionary force operators in the dependent clause is
- **ok:** allowed
- **banned:** not allowed
- **harmonic:** allowed but only if it matches the marking on the main clause

**T-mark:** Marking of tense or status operators in the dependent clause is
- **ok:** allowed
- **banned:** not allowed
- **harmonic:** allowed but subject to constraints based on the tense or status choice in the main clause

**Symmetry:** The range of categories that can be expressed on linked clauses is
- **symmetrical:** must match
- **asymmetrical:** can be different
- **free:** can be different and can even include elements of different type (different parts of speech, clauses and NPs, etc.)
• Are there constructions that are more like each other than others?

• Any trends in certain areas or families (“Papuan-style subordination”, “Slavic coordination”)?

• Any universal trends (“structures supporting narrative use tend to be more similar to symmetrical and-like structures than to topic-related structures”)?

• If there are universal trends, what properties are responsible for them?

  ▸ Need for formal similarity analyses as a heuristic
Similarity analysis

• Similarities could pattern in terms
  • clusters, possibly centered on prototypes
  • scales/continua
  • taxonomies (trees)

• No reason to assume any of these choices!
  • classical clustering algorithms not suitable since
  • they make assumptions on the kind of expected pattern:
    • underlying dimensions/scales,
    • clusters/partitions
    • trees
Similarity analysis

- Use Split graphs, such as NeighborNets, for similarity analysis!

- split graphs are
  - really just a 2D-visualization of a distance matrix, with no implicit evolutionary interpretation (split graphs are “implicit” in this regard: Nichols & Warnow 2008)
  - designed to assess patterns in cluster shapes, specifically treeness
- well suited as general heuristics
### Data: some suitable distance measure (here, relative Hamming distance)

<table>
<thead>
<tr>
<th>Language:but</th>
<th>Marker</th>
<th>ILL-scope</th>
<th>T-scope</th>
<th>Finiteness</th>
<th>ILL-mark</th>
<th>T-mark</th>
<th>Symmetry</th>
<th>WH</th>
<th>Extraction</th>
<th>FOC</th>
<th>Position</th>
<th>Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amele:but</td>
<td>gba</td>
<td>local</td>
<td>local</td>
<td>fin.</td>
<td>*</td>
<td>✓</td>
<td>symm.</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>fix.:pre</td>
<td>ad-S</td>
</tr>
<tr>
<td>Amele:chain</td>
<td>-me, -PV</td>
<td>conj.</td>
<td>conj.</td>
<td>nonfin.</td>
<td>*</td>
<td>*</td>
<td>asymm.</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>fix.:pre</td>
<td>ad-S</td>
</tr>
<tr>
<td>Amele:or</td>
<td>fo ~o</td>
<td>conj.</td>
<td>local</td>
<td>fin.</td>
<td>✓</td>
<td>✓</td>
<td>symm.</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>fix.:pre</td>
<td>ad-S</td>
</tr>
<tr>
<td>Amele:cond</td>
<td>fi</td>
<td>disj.</td>
<td>extens.</td>
<td>any</td>
<td>✓</td>
<td>✓</td>
<td>flex</td>
<td>✓</td>
<td>NA</td>
<td>NA</td>
<td>flex-rel.</td>
<td>ad-S</td>
</tr>
<tr>
<td>Greek:p.c.</td>
<td>-men/-nt</td>
<td>disj.</td>
<td>extens.</td>
<td>nonfin.</td>
<td>*</td>
<td>*</td>
<td>asymm.</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>flex-rel.</td>
<td>ad-S</td>
</tr>
<tr>
<td>Belhare:and</td>
<td>=cha...=cha</td>
<td>flex</td>
<td>local</td>
<td>fin.</td>
<td>✓</td>
<td>✓</td>
<td>symm.</td>
<td>NA</td>
<td>*</td>
<td>✓</td>
<td>flex-adj.</td>
<td>ad-S</td>
</tr>
<tr>
<td>Belhare:chain</td>
<td>ki(na)(huŋ)</td>
<td>flex</td>
<td>extens.</td>
<td>fin.</td>
<td>harm.</td>
<td>✓</td>
<td>asymm.</td>
<td>✓</td>
<td>possible</td>
<td>✓</td>
<td>flex-adj.</td>
<td>ad-S</td>
</tr>
<tr>
<td>Belhare:inc</td>
<td>-kone</td>
<td>local</td>
<td>extens.</td>
<td>nonfin.</td>
<td>✓</td>
<td>*</td>
<td>asymm.</td>
<td>✓</td>
<td>*</td>
<td>✓</td>
<td>flex-adj.</td>
<td>ad-S</td>
</tr>
<tr>
<td>Belhare:sub</td>
<td>-naa...</td>
<td>disj.</td>
<td>extens.</td>
<td>fin.</td>
<td>*</td>
<td>✓</td>
<td>flex</td>
<td>✓</td>
<td>*</td>
<td>✓</td>
<td>flex-adj.</td>
<td>ad-S</td>
</tr>
<tr>
<td>Burúshaski:chain</td>
<td>n(V)-�-Σ-(i)n</td>
<td>flex</td>
<td>extens.</td>
<td>nonfin.</td>
<td>NA</td>
<td>*</td>
<td>asymm.</td>
<td>NA</td>
<td>NA</td>
<td>*</td>
<td>flex-adj.</td>
<td>ad-S</td>
</tr>
<tr>
<td>Chantyal:chain</td>
<td>-si, -rə</td>
<td>NA</td>
<td>extens.</td>
<td>nonfin.</td>
<td>NA</td>
<td>*</td>
<td>asymm.</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>flex-adj.</td>
<td>ad-S</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Amele:but</td>
<td>0.62</td>
<td>0.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amele:or</td>
<td>0.25</td>
<td>0.62</td>
<td>0.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amele:cond</td>
<td>0.62</td>
<td>0.88</td>
<td>0.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greek:part.coni.</td>
<td>0.62</td>
<td>0.38</td>
<td>0.75</td>
<td>0.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Belhare:and</td>
<td>0.38</td>
<td>0.88</td>
<td>0.25</td>
<td>0.50</td>
<td>0.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Belhare:chain</td>
<td>0.50</td>
<td>0.75</td>
<td>0.38</td>
<td>0.44</td>
<td>0.62</td>
<td>0.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Belhare:inconseq</td>
<td>0.62</td>
<td>0.50</td>
<td>0.62</td>
<td>0.56</td>
<td>0.38</td>
<td>0.50</td>
<td>0.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Belhare:sub</td>
<td>0.38</td>
<td>0.75</td>
<td>0.50</td>
<td>0.22</td>
<td>0.38</td>
<td>0.50</td>
<td>0.45</td>
<td>0.55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burushaski:chain</td>
<td>0.71</td>
<td>0.43</td>
<td>0.71</td>
<td>0.71</td>
<td>0.29</td>
<td>0.38</td>
<td>0.38</td>
<td>0.25</td>
<td>0.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chantyal:chain</td>
<td>0.67</td>
<td>0.33</td>
<td>0.67</td>
<td>0.67</td>
<td>0.17</td>
<td>0.50</td>
<td>0.33</td>
<td>0.00</td>
<td>0.67</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Interim conclusion

- Split graphs are a useful addition to existing similarity analysis tools
- On more case study...
Another case study: cross-lexemic semantic roles

- Joint work with A. Witzlack-Makarevich, T. Zakharko & L. Bierkandt:
  - Evaluate evidence for cross-lexemic ("molecular") semantic roles ("experiencer", "agent", "theme", "unaccusative S" etc.),
  - i.e. between lexeme-specific ("atomic") roles ("killer", "said words" etc.) and generalized roles ("S", "A", "P" etc.)
  - So far mostly a debate between theories on the "right" definition of cross-lexemic roles
  - Turn into an empirical issue, by using a multivariate typology of how languages group predicates in morphosyntax
  - specifically groupings by non-canonical case assignments
Case study: cross-lexemic semantic roles

Non-canonical case assignments:

<table>
<thead>
<tr>
<th></th>
<th>Djambarrpuyŋu</th>
<th>Tsez</th>
<th>Nias</th>
<th>Chechen</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NOM</td>
<td>DAT</td>
<td>ABS</td>
<td>ABS</td>
</tr>
<tr>
<td>A of ‘love’</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>A of ‘hate’</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>A of ‘fear’</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A of ‘wait for’</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Jaccard distances:

<table>
<thead>
<tr>
<th></th>
<th>A of ‘love’</th>
<th>A of ‘hate’</th>
<th>A of ‘fear’</th>
</tr>
</thead>
<tbody>
<tr>
<td>A of ‘love’</td>
<td>1/3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A of ‘hate’</td>
<td></td>
<td>1/2</td>
<td>3/4</td>
</tr>
<tr>
<td>A of ‘fear’</td>
<td></td>
<td></td>
<td>1/2</td>
</tr>
<tr>
<td>A of ‘wait’</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

152 languages
42 cross-linguistically recurrent meanings for S
93 cross-linguistically recurrent meanings for A
177 cross-linguistically recurrent meanings for P
etc.
Results for S (sole arguments of one-argument predicates):

sensation

experience

activity

δ=.22
Results for S (sole arguments of one-argument predicates):

Fuzzy Cluster Analysis
optimized at $D=.48$ with $k=5$ and $r=1.2$
Conclusion

- Old-standing problem in comparative linguistics: how to develop cross-linguistically applicable notions for analysis?

- Answer: Multivariate Typology, registering all regards in which structures differ vs. are alike → variables

- Resulting matrices need similarity analysis.

- One method from phylogenetics, split graphs, is a useful tool for similarity analysis, even without stakes in evolutionary explanations (which I don’t have here!)

- Perhaps the approach of Multivariate Typology can be generalized to other cases where the “Insight Definition Approach” fails in the humanities because of too much diversity