Addressing Lexical Polymorphism in Linguistic Phylogenetics

Marc E. Canby

20 October 2022
• Ongoing, unpublished work

• Joint project with Tandy Warnow, Don Ringe, and Steve Evans

• **Goal:** Address polymorphism in linguistic phylogenetics

• Introduce new simulator for linguistic data
• Phonological, morphological, or lexical characters

• Each lexical character represents a different **semantic slot**, or meaning
  - **Examples**: “hand”, “mother”, “human being”, “male person”
  - Two languages have same state if words are cognate (shared linguistic ancestor): e.g. Spa. *mano* ~ Fr. *main* (“hand”)

• **Polymorphism** occurs when 2 words occur within the same meaning
  - **Examples**: “large” = {*big, large*}
    “human being” = {*human, man*}
  - Not as relevant for phonological / morphological characters
• Lots of lexical polymorphism in languages
  – We calculate 33% of lexical characters are polymorphic on Indo-European dataset (but may actually be higher…)

• Existing methods don’t typically handle polymorphism
  – Most linguists either throw out polymorphic characters, or code a polymorphism by using the most common state

• **Goal**: What is the best way to deal with polymorphism?
  – Compare methods in a simulation study
Causes of Polymorphism

- Polymorphism arises from **semantic shift** or **borrowing**

---

**“human being”**  
- **English**: human  
- **Old English**: mann  
- **Latin**: homō  
  - Composed of: **hum** (man) and **-ō** (being)

**“male person”**  
- **English**: man  
- **Old English**: wer  
- **Latin**: vir  
  - **Vir** (male) and **-us** (person)

---

- **human** borrowed from **French**
- **homme**
- **homaine**
- **humain**
- **homō**
- **vir**
Causes of Polymorphism

- Polymorphism as "shades of cognancy"

Modeling Polymorphism

• Substitution model is not sufficient, as it assumes that each character has only one state in a language
  – Need to have a “bag of words” for each character

• Polymorphism should be **bounded**:
  – A language should not have too many words for a single meaning
    • **Example:** A basic meaning like “human being” can be filled by *human* and *man* but not much else
  – The same word should not be used in too many meanings
    • **Example:** *man* can mean both “man” and “human being” or “man” and “husband” but not all three
Modeling Polymorphism

Each character is a separate birth-death process

Birth rate: $\lambda_i = \lambda$

Death rate: $\mu_i = i\mu$
Borrowing causes a state to be transmitted laterally

Birth rate: $\lambda_i = \lambda$
Death rate: $\mu_i = i\mu$
• Perform a simulation study to evaluate tree estimation methods

• Similar to Barbançon et al. (2013), but:
  – Different model of evolution (polymorphism model)
  – Modify maximum parsimony to address polymorphism
  – Do not look at distance-based methods
  – Determine best way to run Gray-Atkinson (Bayesian) method
• **Tree Generation:** Generate random tree

• **Network Generation:** Add $n$ contact edges randomly

• **Character Generation:** Create characters
  - Can control deviation from lexical clock, heterotachy, borrowing, amount of homoplasy, and rate of evolution.

• **Character Evolution:** Evolve characters under model
  - Any model (substitution, birth-death, etc.) can be plugged in, as long as certain functions are implemented
Experimental Study

- 32 trees with 30 languages each
- Simulate 4 datasets (replicates) down each tree
- 20 morphological characters (slow rate, moderate homoplasy, no polymorphism)
- 300 lexical characters (moderate homoplasy, equal numbers of slow, medium & fast rates)
  - Vary amounts of polymorphism between 33%, 50%, 66% of the characters (Low, Medium, High levels)
Estimation Methods

• 4 Methods based on **Maximum Parsimony**

• **MP 1:** Remove all polymorphic characters from dataset

• **MP 2:** Replace all polymorphic states with ?

• **MP 3:** Replace shared polymorphisms with same state

• **MP 4:** Replace each polymorphism with state that occurs most frequently

<table>
<thead>
<tr>
<th></th>
<th>Lat.</th>
<th>Lith.</th>
<th>Gk.</th>
<th>Skt.</th>
<th>OCS</th>
<th>Arm.</th>
<th>OIr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>0</td>
<td>1/2</td>
<td>3</td>
<td>1</td>
<td>1/2</td>
<td>3</td>
<td>0/4</td>
</tr>
<tr>
<td>MP 2</td>
<td>0</td>
<td>?</td>
<td>3</td>
<td>1</td>
<td>?</td>
<td>3</td>
<td>?</td>
</tr>
<tr>
<td>MP 3</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>?</td>
</tr>
<tr>
<td>MP 4</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>
Estimation Methods

- **GA**: Gray-Atkinson method under binary encoding

<table>
<thead>
<tr>
<th>Semantic Slot, Cognate Class</th>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>(“human being”, human)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(“human being”, man)</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(“male person”, human)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(“male person”, man)</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

- Tree is estimated by treating binary characters as evolving under CFN model
- **MCMC** sampling using MrBayes software
- May result in ancestral states being polymorphic, even if leaves are not
Want to obtain binary tree from estimation method to compare with reference tree

Use **RF rate** as error metric
- Fraction of edges in estimated tree not in reference tree
- Same as FN rate if trees are binary

MP methods may return multiple optimal trees
- Report average RF rate over all the trees

GA returns 200 trees **sampled** from posterior distribution
- Average RF rate over the trees
- Majority consensus (not binary)
- MAP tree (tree that occurs most often)
• GA-MAP is much better than GA-Avg or GA-Con
  – Going forward, we only report GA-MAP results
• Barbançon et al. (2012) did GA-Con, which is clearly sub-optimal
Relative Accuracy of Methods

- **GA**: Gray-Atkinson with MAP
- **MP 1**: Remove all polymorphic characters from dataset
- **MP 2**: Replace all polymorphic states with ?
- **MP 3**: Replace shared polymorphisms with same state
- **MP 4**: Replace each polymorphism with state that occurs most frequently

For low polymorphism, GA is worst, MP methods better

(a) Low Polymorphism
Relative Accuracy of Methods

- For low polymorphism, GA is worst, MP methods better
- MP 2-4 among the best for each polymorphism level
- MP 4 is best for every level, though GA is about the same under high polymorphism
• GA gets better as polymorphism level increases
• MP 1 (throw out polymorphic characters) gets worse with more polymorphism
• MP 2-4 more resistant to polymorphism level
• Taking MAP tree is best way to evaluate Gray-Atkinson

• Gray-Atkinson improves as polymorphism increases – why?

• Throwing out polymorphic characters (MP 1) is not good, especially under high polymorphism

• Replacing polymorphic states with most common state (MP 4) is best

• MP methods fairly resistant to polymorphism
Conclusions

- Novel parametric model of polymorphism
- Introduction of simulator that can be used with various models of evolution
- Helps linguists and practitioners:
  - Determine which estimation method is best with polymorphic characters
  - How to run Gray-Atkinson to get best results
  - Figure out how to best deal with polymorphic states when creating datasets
- What’s next?
  - More Bayesian methods, network reconstruction, likelihood methods, proper estimation methods for polymorphism...