Modelling sound change in relation to time-depth and geography: a case study on the Indo-European and Tupían language families

GERD CARLING, SANDRA CRONHAMN, NIKLAS JOHANSSON, JOOST VAN DE WEIJER

SCIHS, UC Berkeley, May 31, 2014
Why analyse sound change?

• Many current computational cladistic analyses are based on the method of **basic vocabulary comparison** (BVC, see Hammarström 2014, 60f.) or **typology** (e.g., Muysken & O’Connor 2014, 1ff.).

• Traditional comparative method rests on three legs (e.g., Meillet 1924, 11; Campbell 2014, 107ff.).
  – **Vocabulary**
  – **Phonology** (Morphophonology)
  – **Morphology** (Morphosyntax)

• Conventional subgrouping (tree-models, wave-models, combinations) based on common innovations or isoglosses (Gamkrelidze & Ivanov 1984, Garrett 2006, Josephson 2013)
Why Indo-European and Tupí?

• Indo-European language family
  = sound change well attested and investigated
  (Kümmel 2007)

• Tupí (like all South American languages families)
  = very little data
  – No ancient sources
  – Sound change → not well investigated
  – Vocabularies → sparse
Background: project LUNDIC (Lund University) Database including geographic data

Collection of data from 400 languages (14 families):

- Basic vocabulary data (200 Swadesh lists)
- Culture-specific vocabulary data (pre-defined sets, focus on subsistence, religion, kinship, 2-300 generic terms).
- Cognate analysis (loans/inherited words distinguished)
- Full vocabularies (2000) for some languages.
- Sound change charts (for comparative linguistic analysis).
- Typological data (focus on basic typology, alignment, information structure).
Basic questions

• How does sound change relate to:
  – Language diversity and change
  – Time depth
  – Geographic spread

• Aims:
  1) Testing quantification of sound change against basic vocabulary
  2) Testing the results against time depth and geography
# Tools

- For clustering and subgrouping:

<table>
<thead>
<tr>
<th>Program</th>
<th>Analysis</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>R</strong></td>
<td>Biplots (Principal Component Analysis)</td>
<td>Demonstrates the amount of variance of samples based on a relative similarity between objects</td>
</tr>
<tr>
<td><strong>MrBayes</strong></td>
<td>Cladograms</td>
<td>Creates cladistic trees based on a hierarchy of individual elements</td>
</tr>
</tbody>
</table>

- For geographic analysis:

<table>
<thead>
<tr>
<th>Program</th>
<th>Analysis</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ArcGIS</strong></td>
<td>(Layered) maps</td>
<td>Demonstrates the geographic spread of variance and clusters</td>
</tr>
</tbody>
</table>
Theoretical framework: sound change, time depth and geography

• Critical changes are parallel and independent (Meillet 1924:380).

• The reconstructed state is a linguistic structural unity, which we should expect to correspond to a linguistic unity reminiscent of a spoken language.

• The connection of a reconstructed unity to an exact time and place (“nation indo-européenne”) in history must remain uncertain (Meillet 1924: 375).
Basic model of quantification, sound change

• Sound change charts, model of, e.g., Meillet (1924), Krahe (1966).

• Focus on critical and common innovations, as derived from comparative method.

• Reconstructed phoneme systems used as a basis.
Model of quantification, sound change

<table>
<thead>
<tr>
<th></th>
<th>*ū</th>
<th>ā-ŷ[i-mut]</th>
<th>ų-ŷ[i-mut]-i:</th>
<th>ų-u:</th>
<th>ų-au</th>
<th>ų-i:</th>
<th>ų-i:-I</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PIE</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Proto-Germanic</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Old Norse</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Swedish</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Old English</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>English</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Gothic</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Proto-Celtic</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Old Irish</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Irish</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Old Welsh</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong>Welsh</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Gaulish</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Proto-Italic</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Latin</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Old Italian</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Italian</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Old French</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>French</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Oscan</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Our analysis

Meillet (1924)

Notes:
1° Skr. e et o sont des longues issues d'anciennes diphtongues indo-iraniciennes ai, au conservées en vieux perse ; le fait qu'elles représentent des diphtongues est reconnaisable en sanskrit même et a été vu par les grammairiens indigènes. — Les diphtongues
Model of quantification: time and space

• Language states (including reconstructed) quantified as layers, connected to time periods and coordinates.

• The greater time-depth, the more unstable the basis for a reconstruction, both in time and space (space-time model of Meid 1975).

![Diagram of space-time model]

- Raum (space)
- Zeit (time)

\[ ?'_1 \rightarrow ?'_2 \]
Methodology: sound change data

• Point of departure: established reconstructed phoneme inventory of Proto-Indo-European (5000-8000 BP) and Proto-Tupí (5000-6000 BP), e.g. (simplified),

Proto-Indo-European (PIE) *kʷ in *kʷo- ’what’

* kʷo-

what    kí-    quis    tís

English    Sanskrit    Latin    Greek
Methodology: sound change data

- The evolution of the reconstructed phonemes is followed to the reconstructed, historically attested and contemporary language states, e.g. (simplified),

<table>
<thead>
<tr>
<th>Time</th>
<th>PIE</th>
<th>Proto-Germanic</th>
<th>Old Norse</th>
<th>Swedish</th>
<th>Old English</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>*k^w</td>
<td>*x^w</td>
<td>hv</td>
<td>v</td>
<td>x^w</td>
<td>w (ʍ)</td>
</tr>
</tbody>
</table>

* PIE: Proto-Indo-European
* PIE: Proto-Germanic
* Old Norse
* Swedish
* Old English
* English
Methodology

• Each change is quantified by value 1
• Earlier stages of change are inherited by their descendants, e.g., Modern English has inherited all values from Old English, Proto-Germanic, etc.
  → Indirect relative chronology
• *<i>k</i><small>wo</small> → Eng. <i>what</i> & Swe. <i>vad</i>
  → Hittite <i>kui</i>-
Data availability and comparability

<table>
<thead>
<tr>
<th></th>
<th>Indo-European</th>
<th>Tupí</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary branches</td>
<td>~10</td>
<td>~10</td>
</tr>
<tr>
<td>Reconstructed languages</td>
<td>12</td>
<td>19</td>
</tr>
<tr>
<td>Historical languages</td>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>Contemporary languages</td>
<td>13</td>
<td>40</td>
</tr>
<tr>
<td><strong>Total languages</strong></td>
<td><strong>43</strong></td>
<td><strong>62</strong></td>
</tr>
</tbody>
</table>
Selection of changes in current set

• **Indo-European** (sound changes well established) → Focus on critical innovations

  – Unconditioned changes:
    » Regular sound changes
    » Complete phoneme deletions

  – Conditioned changes: selection of changes generally affecting more than one language and/or more than one phoneme
    » If no unconditioned change exists for a phoneme
      → all conditioned changes included
    » Relevant sound laws; Grimm’s, Osthoff's, Ruki-rule etc.
    » Important conditioned changes, such as palatalizations, velarizations, labializations
Selection of changes in current set

- **Tupí** (sound changes less well established)
  - Based on the reconstructed phonological system by Rodrigues & Cabral (2012)
  - Sparse information about sound changes → everything is included
  - Different branches are differently well-documented → possible effect on the results

- **Total number of changes in data sets:**
  - Tupí 237
  - Indo-European 891
Biplot: Indo-European sound change

Numbers: relative distance to 0.0 = Proto-Indo-European (all sounds value 1)
Colours: traditional subgroupings
Comparison: sound change vs. basic vocabulary (Indo-European)

- Basic vocabulary → clustered subgrouping
- Sound change → gradient subgrouping
Biplot: Tupí sound change

Numbers: relative distance to 0.0 = Proto-Tupí (all sounds value 1)
Colours: traditional subgroupings
Comparison: Indo-European vs. Tupí sound change

- Indo-European = more gradient
- Tupí = more clustered
- Possibly due to the difference in data granularity
- Matches traditional subgrouping
Geographic Spread vs. Innovativeness – Tupí

- Colour coding: distance from zero = Proto-Tupí
- Blue/Green/Red/Yellow = Directions
- Grading = Distance
Cladogram: Tupí

Scale: relative distance from zero

Colour coding: traditional subgrouping
Geographic Spread vs. Innovativeness – Indo-European

- Colour coding: distance from zero = Proto-Indo-European
- Blue/Green/Red/Yellow = Directions
- Grading = Distance

Legend

Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>☑️</td>
<td>Modern</td>
</tr>
<tr>
<td>□</td>
<td>Old</td>
</tr>
<tr>
<td>🟢</td>
<td>Most conservative</td>
</tr>
<tr>
<td>🟡</td>
<td>Most innovative</td>
</tr>
</tbody>
</table>

Grading

- Distance
Cladogram: Indo-European

Scale: relative distance from zero  Colour coding: traditional subgrouping
Conclusion, method in general

Advantages

• Results correspond to conventional subgroupings.
• Methodology embraces common innovations (independent change) as well as shared similarities (parallel evolution)
• Methodology suitable for measuring innovativeness (branch length)

Limitations

• Requires a solid preparatory work of comparative linguistic analysis (disadvantage for undescribed language families)
• Data granularity plays some role, but not for the overall results
Results and perspectives

Results:

• Phylogenetic models should be based on mixed data sets.

• Areal effects (with results from both trees and biplots), both in subgrouping and linguistic distance

• Supports a geographical diffusion model, related to convergence.
Results and perspectives

Perspectives:

- Can possibly be useful for the discussion on proto-language time depth?
- Data sets can be used for, e.g., measuring directionality of sound change?
Thank You!

Funding from:
Marcus and Amalia Wallenberg Foundation
Crafoord Foundation

Special thanks to:
Anne Goergens for creating maps
Love Eriksen for providing geographic data for Tupí
Ana Suelly Cabral, Niklas Burenhult, Rob Verhoeven & Filip Larsson for valuable remarks
Key References


Primary Sources

**Indo-European**


**Tupí**
