Prospects for a (Semi-)Automated Papuan Comparative Linguistics and Reconstruction

Harald Hammarström

14 June 2019, Leiden
Papuan Comparative Linguistics

- Immense number of languages (∼ 869)
- Immense number of lineages (families + isolates)
  - 125 according to glottolog.org
  - 80 according to Palmer (2018:4-5)
  - 50 according to ethnologue.com
- Basic lexicon available in a published source for 767 lgs (88%)
Microgroups/Genera

- The Papuan language fall into perhaps 172 microlineages

  \[\text{Microlineage} \sim \text{genus (as in WALS)} \sim \text{a group of languages joined by at least 30\% lexicostatistical similarity}\]

- 104 of those microlineages have more than one member, i.e., microgroups

- To begin with, one would like to see a comparative-historical reconstruction of all these microgroups
Papuan microgroups with a published historical phonology with proto sound inventory and tracing to modern languages

**Nuclear Tirio:** Usher and Suter 2015
**Marind-Boazi-Yaqai:** Usher and Suter 2015
**Kamula-Elevala:** Suter and Usher 2017
**Greater Binanderean:** Smallhorn 2011
**Greater Awayu:** de Vries et al. 2012; Wester 2014
**West Inland Gulf of Papua:** Usher and Suter 2015
**Alor-Pantar:** Holton and Robinson 2014
**Far West Lakes Plain:** Clouse 1997:141-142
**East Tariku:** Clouse 1997:145-147
**West Tariku:** Clouse 1997:147-149
**Central Tariku:** Clouse 1997:149-151

**Lower Sepik:** Foley 1986:215-229; Foley 2018:213-220
**Bulaka River:** Usher 2014
**East Timor-Bunaq:** Schapper et al. 2014

**Mailuan:** Dutton 1982
**Asmat-Kamoro:** Voorhoeve 2005
**Ottilien:** Foley 2005:112-121
**Sogeram:** Daniels 2015
**Ndu:** Aikhenvald 2008:596-626; Laycock 1965:147-197
**Kainantu-Goroka:** Foley 1986:245-257; Xiao 1990
**Ok:** Healey 1964
**Koiarian:** Dutton 2010
**West Wapei:** Crowther 2001
**Sentanic:** Hartzler and Gregerson 1987
**Chimbu-Wahgi:** Capell 1962:105-128; Rarrick 2014
**Skou-Serra-Piore:** Donohue 2002; Donohue and Crowther 2005
**Enga-Kewa-Huli:** Franklin 1975
# Work on Language Relationships Across Areas: Raw

Raw number of “comparative” bibliographical items/year
Work on Language Relationships Across Areas: Per Ig

Number of “comparative” bibliographical items/year per language
I do not feel that the groundwork has yet been done to permit such wider groupings to be established on any large scale ... the evidence for larger groupings must be compelling, and the amassing of such evidence will be a slow process. There can be no short cuts to the classification of Papuan languages (Foley 1986:213).

- Available linguists prioritize documentation (for good reason)
- Are we inevitably looking at a slow process with no shortcuts?
- Perhaps computers can help?
Computer-Assisted Historical Linguistics for Papua?

  - Does not provide reconstructions
  - Intermediate steps not interpretable to a human

- Cognates-to-trees: Input cognate sets and obtain a historical tree with branch-lengths (Dunn 2014, Greenhill 2015, etc.)
  - The cognate sets have to be obtained from somewhere (usually a human)
  - Does not provide reconstructions
  - Intermediate steps and tree justification not interpretable to a human

- Wordlists-to-cognate-sets: Input wordlists and obtain cognate sets (List et al. 2017 etc.)
  - Let us dig deeper into this!
Cognate Detection

*Given meaning-aligned wordlists judge which word-forms are historically related*

<table>
<thead>
<tr>
<th>English</th>
<th>Turkish</th>
<th>Persian</th>
<th>Kurdish</th>
<th>Arabic</th>
<th>Hindi</th>
<th>Swedish</th>
</tr>
</thead>
<tbody>
<tr>
<td>wʌn</td>
<td>bir</td>
<td>yek</td>
<td>yek</td>
<td>wæːhed</td>
<td>ek</td>
<td>en</td>
</tr>
<tr>
<td>tuː</td>
<td>iki</td>
<td>do</td>
<td>dû</td>
<td>etneːn</td>
<td>doː</td>
<td>tvoː</td>
</tr>
<tr>
<td>əri</td>
<td>ɨtf̥</td>
<td>se</td>
<td>sê</td>
<td>tælæːtæ</td>
<td>t:i:n</td>
<td>treː</td>
</tr>
<tr>
<td>neim</td>
<td>isim/ad</td>
<td>esm</td>
<td>naw</td>
<td>?esm</td>
<td>naːm</td>
<td>namn</td>
</tr>
<tr>
<td>nous</td>
<td>burun</td>
<td>damaːgh</td>
<td>lût</td>
<td>mænæxiːr</td>
<td>naːk</td>
<td>nɛːsa</td>
</tr>
<tr>
<td>watər</td>
<td>su</td>
<td>aːb</td>
<td>aw</td>
<td>majja</td>
<td>pəniː</td>
<td>vaten</td>
</tr>
<tr>
<td>hed</td>
<td>baʃ/kafa</td>
<td>sar</td>
<td>ser</td>
<td>ræːs</td>
<td>sar</td>
<td>hʉːvʊd</td>
</tr>
<tr>
<td>nart</td>
<td>gedʒe</td>
<td>ʃab</td>
<td>ʃev</td>
<td>leːlæ</td>
<td>raːtriː</td>
<td>nat</td>
</tr>
<tr>
<td>boun</td>
<td>kemik</td>
<td>ostokhaːn</td>
<td>hestî</td>
<td>ʃadm</td>
<td>haḍḍiː</td>
<td>beːn</td>
</tr>
<tr>
<td>niuː</td>
<td>yeni</td>
<td>naw/taːze</td>
<td>nwːe</td>
<td>gediːd</td>
<td>nayaː</td>
<td>ny</td>
</tr>
<tr>
<td>wiː</td>
<td>biz</td>
<td>maː</td>
<td>ême</td>
<td>eðnæ</td>
<td>ham</td>
<td>viː</td>
</tr>
</tbody>
</table>

For today, let us conveniently ignore some complications

- Non-monomorphemic forms
- Meaning shift
- ...
Cognate Detection: State-of-the-Art

Nearly all past work in automated cognate detection (e.g., List et al. 2018, List 2014, Kondrak 2009, Steiner et al. 2011, List et al. 2017, St Arnaud et al. 2017 and references therein)

1. Align words phonetically
2. Compute similarity of aligned words
3. Group cognates that exceed a certain similarity threshold
Thresholds in Cognate Identification

Require tuning a threshold to cut a similarity-based score into a yes/no cognate decision

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Austronesian (Greenhill et al., 2008) [1]</td>
<td>4358</td>
<td>210</td>
<td>20</td>
<td>2864</td>
<td>0.64</td>
</tr>
<tr>
<td>Bai (Wang, 2006) [27]</td>
<td>1028</td>
<td>110</td>
<td>9</td>
<td>285</td>
<td>0.19</td>
</tr>
<tr>
<td>Chinese (Hóu, 2004) [28]</td>
<td>2789</td>
<td>140</td>
<td>15</td>
<td>1189</td>
<td>0.40</td>
</tr>
<tr>
<td>IndoEuropean (Dunn, 2012) [2]</td>
<td>4393</td>
<td>207</td>
<td>20</td>
<td>1777</td>
<td>0.38</td>
</tr>
<tr>
<td>Japanese (Hattori, 1973) [29]</td>
<td>1986</td>
<td>200</td>
<td>10</td>
<td>460</td>
<td>0.15</td>
</tr>
<tr>
<td>ObUgrian (Zhivlov, 2011) [30]</td>
<td>2055</td>
<td>110</td>
<td>21</td>
<td>242</td>
<td>0.07</td>
</tr>
<tr>
<td>TOTAL</td>
<td>16609</td>
<td>977</td>
<td>95</td>
<td>6817</td>
<td>0.30</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0170046.t002

“The key parameter we need to estimate is the best thresholds for cognate identification in some of the methods” (List et al. 2017:3)
The Threshold is the Problem

- The threshold can either be human-tuned or pre-trained with respect to some supervision/gold standard data set.
- Cognate detection and evaluation is typically done on data sets which include both shallow cognates and deep cognates.
  - Shallow cognate: German ’fünf’ vs English ’five’
  - Deep cognate: Prasuni ’wuču’ vs Sardinian ’chimbe’
- Dilemma
  - Strict threshold: Only shallow cognates are found
  - Loose threshold: Junk is found (along with shallow and deep cognates)
More Formally: First Step Cognate Detection

- Suppose you do not already know
  - The relevant sound-shifts
  - The classificatory tree of the input languages

  _Let’s call this variant First Step Cognate Detection_

- For a solution to be possible (whether for a human or machine cognate detector), one has to assume that cognates are more similar on average than non-cognates

\[
\sum_{x \neq y \in C_i} Sim(x, y) / |\{(x, y) | x \neq y \in C_i\}| > \sum_{x \neq y \notin C_i} Sim(x, y) / |\{(x, y) | x \neq y \notin C_i\}|
\]

  _Let’s call this property the Similarity Criterion_
I Propose

- **Shallow** first step cognate detection
  - Can be done
  - Can be done without a threshold
  - Shallow cognate = obeys the similarity criterion

- **Deep** first step cognate detection
  - Cannot be done
  - (Deep cognate detection must thus be done in several steps or with more information)
  - Deep cognate = does not obey the similarity criterion
Threshold-Free First Step Cognate Detection

- Thanks to the similarity criterion, there exists an optimization solution that maximizes

\[
\sum_{x \neq y \in C_i} \text{Sim}(x, y) \frac{1}{|\{(x, y) | x \neq y \in C_i\}|} - \sum_{x \neq y \notin C_i} \text{Sim}(x, y) \frac{1}{|\{(x, y) | x \neq y \notin C_i\}|}
\]

- The intuition is to contrast the cost of judging something cognate (penalty: dissimilarity) and judging something not cognate (penalty: similarity)

- Afaik, the only cognate detection paper in the literature that exploits this dichotomy is Ellison (2007)
  
  *This formulation is restricted to the case with exactly two input languages*

- Today we present a more transparent method for any number of input languages, based on the same optimization intuition
The Present Approach

1. **Input:** Set of $n$ word forms with the same meaning

2. **Pairwise Similarity:** Calculate the pairwise similarity between each pair of the $n$ words using a suitable similarity measure $S(x, y)$

3. **Significance Similarity:** Measure the significance $SS(x, y)$ of the similarity $S(x, y)$ by comparing $S(x, y)$ to $S(x, z)$ for all $z$ that have a different meaning.

4. **Divide the $n$ forms into subsets such that the average $SS(x, y)$ internally in a cognate set + average $1 - SS(x, y)$ between non-cognates is maximized (= correlation clustering)
Example: Pairwise Distances / Significance

\[
D(x, y) = 0.826 \quad \text{niwijcha poyo wanish wañi}
\]

\[
iwijcha \quad 0.000 \quad 1.000 \quad 0.750 \quad 0.875
\]

\[
poyo \quad 1.000 \quad 0.000 \quad 1.000 \quad 1.000
\]

\[
wanish \quad 0.750 \quad 1.000 \quad 0.000 \quad 0.333
\]

\[
wañi \quad 0.875 \quad 1.000 \quad 0.333 \quad 0.000
\]

\[
SS(x, y) = 0.146 \quad \text{niwijcha poyo wanish wañi}
\]

\[
iwijcha \quad 1.000 \quad 0.000 \quad 0.000 \quad 0.000
\]

\[
poyo \quad 0.000 \quad 1.000 \quad 0.000 \quad 0.000
\]

\[
wanish \quad 0.000 \quad 0.000 \quad 1.000 \quad 1.000
\]

\[
wañi \quad 0.000 \quad 0.000 \quad 0.750 \quad 1.000
\]

- \(D(x, y)\) in this example is simply normalized Levenstein distance as the (dis)similarity measure – anything more sophisticated is better.

- \(SS(x, y)\) is the proportion of words \(z\) with a different meaning such that \(D(x, y) \leq D(x, z)\).

- The conversion to \(SS\) is necessary to normalize the (dis)similarity so that a negative deviation can be pitted against a positive deviation.
Example: Significance Similarity

<table>
<thead>
<tr>
<th>Meaning</th>
<th>English</th>
<th>Swedish</th>
</tr>
</thead>
<tbody>
<tr>
<td>one</td>
<td>wʌn</td>
<td>naɪt</td>
</tr>
<tr>
<td>two</td>
<td>tuː</td>
<td>nat</td>
</tr>
<tr>
<td>three</td>
<td>θri</td>
<td></td>
</tr>
<tr>
<td>name</td>
<td>neɪm</td>
<td></td>
</tr>
<tr>
<td>night</td>
<td>nai̯t</td>
<td>nat</td>
</tr>
<tr>
<td>bone</td>
<td>boun</td>
<td>hɵnd</td>
</tr>
<tr>
<td>new</td>
<td>niuː</td>
<td></td>
</tr>
<tr>
<td>we</td>
<td>wiː</td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td>dɔɡ</td>
<td>hɵnd</td>
</tr>
<tr>
<td>nose</td>
<td>nɔus</td>
<td></td>
</tr>
<tr>
<td>water</td>
<td>watər</td>
<td></td>
</tr>
<tr>
<td>head</td>
<td>hed</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- $S(\text{nat, naɪt})$ will be higher than practically all of $S(\text{nat, wʌn}), S(\text{nat, tuː}), S(\text{nat, θri}), ... \rightarrow$ high significance
- $S(\text{dɔɡ, hɵnd})$ will not be higher than practically all of $S(\text{dɔɡ, wʌn}), S(\text{dɔɡ, tuː}), S(\text{dɔɡ, θri}), ... \rightarrow$ low significance
Significance Similarity

- SS wants to measure how non-random (“significant”) a certain form similarity is.
- SS(x, y) > 0.5 more similar than a chance pair of words.
- SS(x, y) < 0.5 less similar than a chance pair of words.
- Suppose two words x from $L_1$, y from $L_2$ with some meaning A.
- Presumably all, or nearly all, words from $L_2$ with a different meaning than A should be unrelated in form to x (Oswalt 1970).
- So we can compare $S(x, y)$ to a large array of $S(x, z)$ for z with a different meaning than x.
- SS(x, y) is the proportion of words for which $S(x, y) \geq S(x, z)$ (is more similar than a random pair of forms).
Clustering on Significance Similarity

- We now want to divide the $n$ forms into cognate subsets such that the average $SS(x, y)$ internally in a cognate set + average $1 - SS(x, y)$ between non-cognates is maximized.

- For every pair of words we have to choose between calling them
  1. cognate (and suffer a penalty if they have low SS) or non-cognate (and suffer a penalty if they have high SS)
  2. do this in a consistent way (so that cognacy preserves transitivity)

- This turns out to be a well-studied problem (called correlation clustering) for which there is an approximation algorithm (Demaine et al. 2006).

- For small $n$ an exhaustive search or a simple local search algorithm is sufficient in practice.
Let us look at an example

- 4 languages from the Torricelli family (North PNG)
- Data from a raw spreadsheet sent by Matthew Dryer (no cleaning/harmonization done)
- Most of us have never studied these languages and have few mature ideas on cognacy

<torricellimini1.html>
Towards Deep Cognate Detection

- Shallow cognates provide evidence for (shallow) subgrouping
  - Factor out the most recent subgroup
  - **Reconstruct** its proto-language via regular correspondences found in the shallow cognates
  - Redo (shallow) cognate detection, this time with the proto-language of the recognized subgroup instead of the surface forms

- Repeat

  *This way, deep cognates may be recognized iff surface divergent surface forms become similar by a series of nested regular correspondences*
The Three Pillars: Some Heuristic Approaches

● **Subgrouping:** A greedy solution
  - For every meaning, guess which cognate set is the oldest
  - The cognate set shared across the *deepest divide* is most likely the oldest
  - Thus this is the retention and the other cognate sets are innovations
  - Once innovations are distinguished from retentions, we can test for the subgroup best selected for by shared innovations

● **Reconstruction:** A greedy solution
  - In every cognate set, try one of the forms as ancestral
  - This gives equations to all modern forms
  - From such equations we can collect a set of potential sound changes
  - A potential sound change can be tested for significance across all cognate sets
  - Majority vote + play back of significant sound changes provide the reconstruction

● **Cognate detection:** (Just explained on previous slides)
Cognate Matrix to Most Demarcated Terminal Subgroup

1. For every meaning, guess which cognate set was present in the proto-language
   ▶ Heuristic: the value cognate set across the deepest divide is the most likely value for the proto-language

2. Throw away the retention & singleton isoglosses

3. Find the Most Demarcated Terminal (MDT) subgroup
   ▶ Heuristic: The MDT subgroup is the subset with the highest amount of supportive innovation isoglosses and the least amount of conflicting innovation isoglosses

4. Replace the languages of the MDT subgroup with its protolanguage
Retention vs Innovation

Which of A, B, C, D are innovations/retention?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>two</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>

Across **all** 184 meanings, the overall cognate distances between the languages are

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Agei [aif]</td>
<td>0.0</td>
<td>0.669</td>
<td>0.689</td>
<td>0.701</td>
<td>0.644</td>
</tr>
<tr>
<td>Aiku [ymo]</td>
<td>0.669</td>
<td>0.0</td>
<td>0.672</td>
<td>0.666</td>
<td>0.660</td>
</tr>
<tr>
<td>Aro [tei]</td>
<td>0.729</td>
<td>0.672</td>
<td>0.0</td>
<td>0.655</td>
<td>0.678</td>
</tr>
<tr>
<td>Bragat [aof]</td>
<td>0.701</td>
<td>0.666</td>
<td>0.655</td>
<td>0.0</td>
<td>0.685</td>
</tr>
<tr>
<td>Chinapeli [van]</td>
<td>0.644</td>
<td>0.660</td>
<td>0.678</td>
<td>0.685</td>
<td>0.0</td>
</tr>
</tbody>
</table>

The deepest divide (0.729) is between Aro and Agei which both share the A cognate.

Let us therefore guess that A is a retention in this case.

That makes B and C innovations.
Innovation Isoglosses to MDT Subgroup

- Throw away the retention isoglosses & singleton innovations
- We are now left with a list of innovation isoglosses that select various subsets of the languages at hand
- The MDT should be one which has the most unequivocal support isoglosses (the most supporting innovations and the least conflicting innovations)
- Heuristic: For each subset $S$ with at least one innovation
  - Do a Fisher Exact Test (FET) to measure how well each innovation $i$ selects $S$

$$\text{Subgroup}(S, I) = \prod_{i \in I} \text{FET}(S, i) = \prod_{i \in I} \sum_{k \geq |S \cap i|} \frac{\binom{|S|}{k} \binom{|L \setminus S|}{|i| - k}}{\binom{|S|}{|i|}}$$

  - Check if it beats what can be expected by random
  - Check that it doesn’t have a more recent subgroup within it (using the same test)
- If there is a $S$ that beats random and has no more recent subgroup within it, that is the Most Demarcated Terminal subgroup
Let us go back to that example

torricellimini1.html
Reconstruct the MDT Subgroup Proto-Language

- Suppose the Most Demarcated Terminal subgroup is $S = \{L_1, L_2, L_3\}$

<table>
<thead>
<tr>
<th></th>
<th>$L_1$</th>
<th>$L_2$</th>
<th>$L_3$</th>
<th>...</th>
<th>$L_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>...</td>
<td>B</td>
</tr>
<tr>
<td>M2</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>...</td>
<td>B</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- For each meaning
  - Determine which cognate to project to proto-$S$:
    - Project the most common (in $S$) cognate set to the proto-language, e.g., for meaning M1 project cognate set $A$ to proto-$S$
    - In case of a tie, e.g., M2, prefer the cognate set (here $B$) which is found outside $S$

  - Reconstruct the form for that cognate in proto-$S$

  *See next slides*
Form Reconstruction: Collecting Potential Sound Changes

- Given a set of cognate forms $x, y, z, \ldots$
- Assume the proto-sound and proto-condition for every sound change is preserved in at least one modern form
- Then the equations $*x \rightarrow x, *x \rightarrow y, *x \rightarrow z, *y \rightarrow x, *y \rightarrow z, \ldots$ etc encompass all relevant potential sound changes
- E.g. with $\{varm, worm, warm\}$, the equations

<table>
<thead>
<tr>
<th>Ancestral</th>
<th>Modern</th>
<th>Potential sound change(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>varm</td>
<td>worm</td>
<td>$v \succ w, a \succ o, v- \succ w-, Ca \succ Co, \ldots$</td>
</tr>
<tr>
<td>varm</td>
<td>warm</td>
<td>$v \succ w, v- \succ w-, \ldots$</td>
</tr>
<tr>
<td>worm</td>
<td>varm</td>
<td>$w \succ v, o \succ a$</td>
</tr>
<tr>
<td>worm</td>
<td>warm</td>
<td>$o \succ a$</td>
</tr>
</tbody>
</table>

- I experimented with extracting all uni- and bigram sound changes from such equations
Testing Potential Sound Changes

- Reverse-apply the sound change to all words
- Check how much the edit distance to its cognates improved/worsened ("gain")
- If the gain is better than random accept the sound sound change
  - Permutation tests (many variants) can represent the null hypothesis
  - Control for multiple testing of sound changes, e.g., if 560 potential sound changes are checked, an accepted sound change must be better than 560 random ones
Sound changes in the example

*torricellimini1.html*
Three Examples

18 Torricelli lgs (data from Matthew Dryer) torricelli1.html

10 random Austronesian lgs from Eastern Austronesia (data from Marian Klamer) marian10lgs1.html

10 random Trans New Guinea lgs (data from transnewguinea.org) tng101.html
Discussion

- In the present conceptualization
  - Subgrouping needs cognate information
  - Cognate detection is dependent on subgrouping
- In the present approach, this is done in a greedy see-saw manner (CD1, SG1, CD2, SG2, …)
- Why not go Bayesian?
Discussion

- In the present conceptualization
  - Subgrouping needs cognate information
  - Cognate detection is dependent on subgrouping

- In the present approach, this is done in a greedy see-saw manner (CD1, SG1, CD2, SG2, …)

- Why not go Bayesian?

  *Search space is prohibitive already with the tree topology, let alone with branch lengths, cognates judgment and regular sound changes intertwined. Heuristics needed to control the search space in Bayesian formulations. Preferable from a linguistic perspective to have more transparent heuristics than those.*
Conclusion

- Arguments to separate shallow and deep cognate detection
- Deep cognate detection addressed via iterative subgrouping and reconstruction
  - Heuristic subgroup detection
  - Heuristic discovery of sound changes
  - Heuristic iterated reconstruction
- All steps relatively transparent
Tell Me

- How to solve the meaning shift problem
- How to handle polymorphemic forms
- If this is what Papuan linguistics needs / does not need?
- ...


Crowther, Melissa. 2001. All the One language(s): comparing linguistic and ethnographic definitions of language in New Guinea. University of Sydney MA thesis.


Healey, Alan. 1964. The Ok Language Family in New Guinea. Australian National University doctoral dissertation. [Sometimes cited as A Survey of the Ok Family of Languages presumably because part of the thesis II-IV, which contains all linguistic data, carries this title.].


Steiner, Lydia, Peter F. Stadler & Michael Cysouw. 2011. A Pipeline for...


