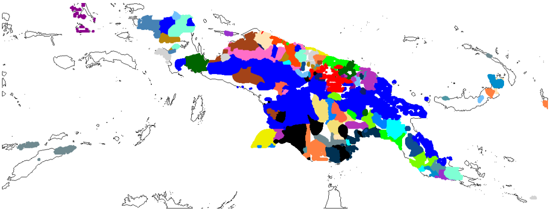
Prospects for a (Semi-)Automated Papuan Comparative Linguistics and Reconstruction

Harald Hammarström

14 June 2019, Leiden

Papuan Comparative Linguistics

- Immense number of languages ($\sim$ 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%)

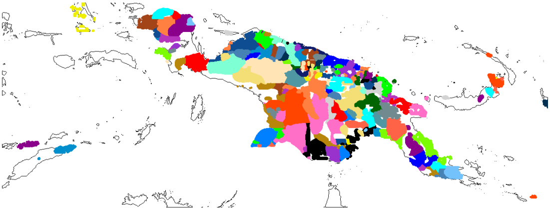


Microgroups/Genera

- The Papuan language fall into perhaps 172 microlineages

  Microlineage ∼ genus (as in WALS)∼ 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



∼ 27 Papuan Microgroups with Reconstructions

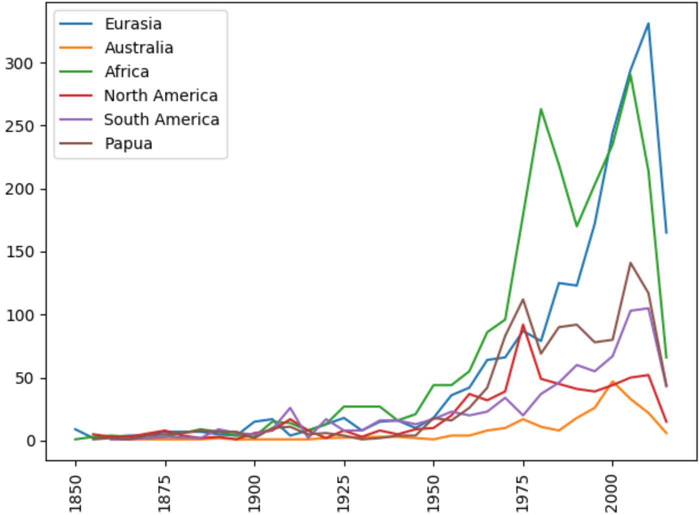
Papuan microgroups with a published historical phonology with proto sound inventory and tracing to modern languages

Nuclear Tirio: Usher and Suter 2015
Marind-Boazi-Yaqai: Usher and Suter 2015
Kamula-Elevala: Suter and Usher 2017
Greater Binanderean: Smallhorn 2011
Greater Awyu: 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

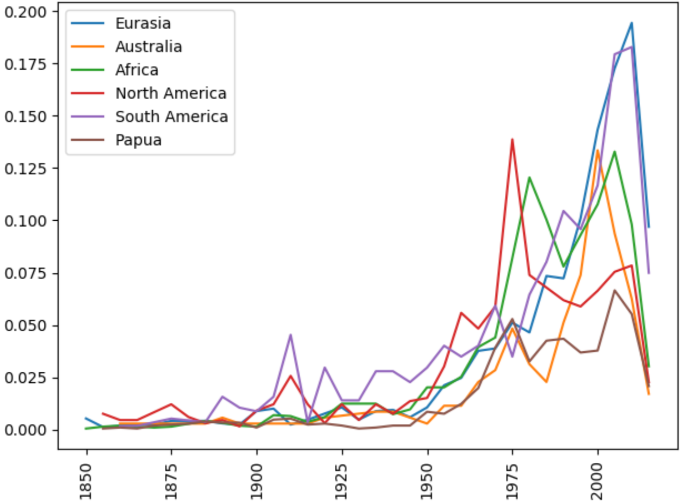
Work on Language Relationships Across Areas: **Raw**

Raw number of "comparative" bibliographical items/year



Work on Language Relationships Across Areas: **Per lg**

Number of "comparative" bibliographical items/year **per language**



Who is going to to do Papuan Historical Linguistics?

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 prioritze 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?

- Quick-and-dirty: Input ASJP 40 word-lists and obtain a similarity tree of languages (Wichmann 2012) (cf. Jäger 2019)
  - 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

English	Turkish	Persian	Kurdish	Arabic	Hindi	Swedish
wʌn	bir	yek	yek	wæːħed	ek	en
tuː	iki	do	dû	etneːn	doː	tvoː
θri	ytʃ	se	sê	tælæːtæ	tiːn	treː
neɪm	isim/ad	esm	naw	ʔesm	naːm	namn
noʊs	burun	damaːgh	lût	mænæxiːr	naːk	nɛːsa
watər	su	aːb	aw	mɑjjɑ	paːniː	vaten
hed	baʃ/kafa	sar	ser	rɑːs	sar	hʉːvʊd
naɪt	gedʒe	ʃab	ʃev	leːlæ	raːtriː	nat
boʊn	kemik	ostokhaːn	hestî	ʕɑdm	haḍḍiː	beːn
nɪuː	yeni	naw/taːze	nwê	gediːd	nayaː	ny
wiː	biz	maː	ême	eħnæ	ham	viː

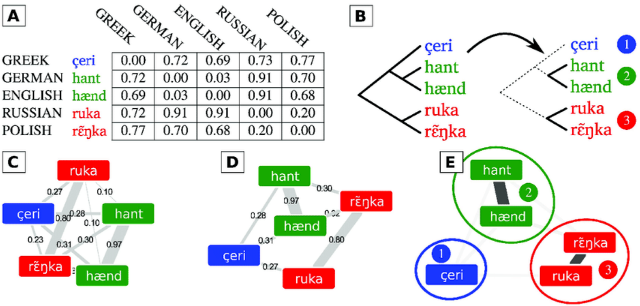
For today, let us conveniently ignore some complications

- 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*

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

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

$$\frac{\sum_{x \neq y \in C_i} Sim(x, y)}{|\{(x, y)|x \neq y \in C_i\}|} > \frac{\sum_{x \neq y \notin C_i} Sim(x, y)}{|\{(x, y)|x \neq y \notin C_i\}|}$$

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 similary 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 similary criterion

Threshold-Free First Step Cognate Detection

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

$$\frac{\sum_{x \neq y \in C_i} Sim(x,y)}{|\{(x,y)|x \neq y \in C_i\}|} - \frac{\sum_{x \neq y \notin C_i} Sim(x,y)}{|\{(x,y)|x \neq y \notin C_i\}|}$$

- 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. Signficance 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

0.826	niwijcha	poyo	wanish	wañi
niwijcha	0.000	1.000	0.750	0.875
poyo	1.000	0.000	1.000	1.000
wanish	0.750	1.000	0.000	0.333
wañi	0.875	1.000	0.333	0.000

$D(x, y)$

0.146	niwijcha	poyo	wanish	wañi
niwijcha	1.000	0.000	0.000	0.000
poyo	0.000	1.000	0.000	0.000
wanish	0.000	0.000	1.000	1.000
wañi	0.000	0.000	0.750	1.000

$SS(x, y)$

- $D(x, y)$ in this example is simply normalized Levenstein distance as the (dis)similarity measure – anything more sophistocated 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

Meaning	English	Swedish
one	wʌn	
two	tuː	
three	θri	
name	neɪm	
night	naɪt	nat
bone	boʊn	
new	nɪuː	
we	wiː	
dog	dɔg	hɵnd
nose	noʊs	
water	watər	
head	hed	
…	…	

- S(nat, naɪt) will be higher than practically all of S(nat, wʌn), S(nat, tuː), S(nat, θri), … → high significance
- S(dɔg, hɵnd) will not be higher than practically all of S(dɔg, wʌn), S(dɔg, tuː), S(dɔg, θri), … → 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 chose between calling them
    i. cognate (and suffer a penalty if they have low SS) or non-cognate (and suffer a penalty if they have high SS)
    ii. 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?

	Agei [aif]	Aiku [ymo]	Aro [tei]	Bragat [aof]	Chinapeli [van]	...
two	A	B	A	C	A	

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

	Agei [aif]	Aiku [ymo]	Aro [tei]	Bragat [aof]	Chinapeli [van]
Agei [aif]	0.0	0.669	0.689	0.701	0.644
Aiku [ymo]	0.669	0.0	0.672	0.666	0.660
Aro [tei]	0.729	0.672	0.0	0.655	0.678
Bragat [aof]	0.701	0.666	0.655	0.0	0.685
Chinapeli [van]	0.644	0.660	0.678	0.685	0.0

- 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

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

  - 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

`torricelliminil.html`

Reconstruct the MDT Subgroup Proto-Language

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

	$\overbrace{L1 \quad L2 \quad L3}^{S}$			...	L10
M1	A	A	B	...	B
M2	A	B	C	...	B
...					

- 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*, ...
- 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

Ancestral		Modern	Potential sound change(s)
varm	$\rightarrow$	worm	$v > w$, $a > o$, $v- > w-$, $Ca > Co$, ...
varm	$\rightarrow$	warm	$v > w$, $v- > w-$, ...
worm	$\rightarrow$	varm	$w > v$, $o > a$
worm	$\rightarrow$	warm	$o > a$
...			

- I experimented with extracting all uni- and bigram sound changes from such equations

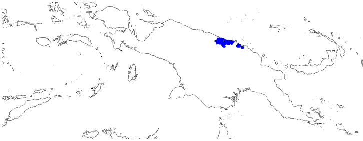
Testing Potential Sound Changes

- 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

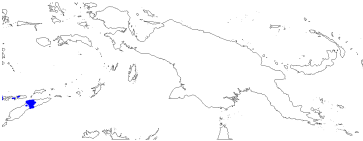
Sound changes in the example

`torricellimini1.html`

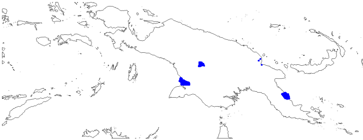
Three Examples



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 adressed 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?
- ...

Aikhenvald, Alexandra Y. 2008. *The Manambu language of East Sepik, Papua New Guinea*. Oxford: Oxford University Press.

Capell, Arthur. 1962. *Linguistic Survey of the South-Western Pacific (New and revised edition)* (South Pacific Commission Technical Paper 136). Noumea: South Pacific Commission.

Clouse, Duane A. 1997. Toward a reconstruction and reclassification of the Lakes Plain languages of Irian Jaya. In Karl J. Franklin (ed.), *Papers in Papuan linguistics No. 2* (Pacific Linguistics: Series A 85), 133-236. Canberra: Research School of Pacific and Asian Studies, Australian National University.

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

Daniels, Don. 2015. A Reconstruction of Proto-Sogeram: Phonology, Lexicon, and Morphosyntax. University of California at Santa Barbara doctoral dissertation.

Demaine, Erik D., Dotan Emanuel, Amos Fiat & Nicole Immorlica. 2006. Correlation Clustering in General Weighted Graphs. *Theoretical Computer Science* 361(2-3). 172–187.

Donohue, Mark & Melissa Crowther. 2005. Meeting in the middle: interaction in North-Central New Guinea. In Andrew Pawley, Robert Attenborough, Jack Golson & Robin Hide (eds.), *Papuan Pasts: Studies in the Cultural, Linguistic and Biological History of the Papuan-speaking Peoples* (Pacific Linguistics 572), 167-184. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Donohue, Mark. 2002. Which Sounds Change: Descent and Borrowing in the Skou Family. *Oceanic Linguistics* 41(1). 171–221.

Dunn, Michael. 2014. Language phylogenies. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge Handbook of Historical Linguistics*, 190-211. New York: Routledge.

Dutton, Tom E. 1982. Borrowing in Austronesian and Non-Austronesian languages of coastal South-East Mainland Papua New Guinea. In Amran Halim, Lois Carrington & Stephen A. Wurm (eds.), *Papers from the third international conference on Austronesian linguistics, Vol 1: Currents in Oceanic* (Pacific Linguistics: Series C 74), 109-177. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Dutton, Tom. 2010. *Reconstructing Proto Koiarian: The history of a Papuan language family* (Pacific Linguistics 610). Canberra: Research School of Pacific and Asian Studies, Australian National University.

Ellison, T. Mark. 2007. Bayesian Identification of Cognates and Correspondences. In *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology* (SigMorPhon '07), 15–22. Stroudsburg, PA, USA: Association for Computational Linguistics.

Foley, William A. 1986. *The Papuan languages of New Guinea* (Cambridge language surveys). Cambridge: Cambridge University Press.

Foley, William A. 2005. Linguistic prehistory in the Sepik-Ramu Basin. In Andrew Pawley, Robert Attenborough, Jack Golson & Robin Hide (eds.), *Papuan Pasts: Studies in the Cultural, Linguistic and Biological History of the Papuan-speaking Peoples* (Pacific Linguistics 572), 109-144. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Foley, William A. 2018. The Languages of the Sepik-Ramu Basin and Environs. In Bill Palmer (ed.), *Papuan Languages and Linguistics*, 197-432. Berlin: Mouton.

Franklin, Karl J. 1975. Comments on Proto-Engan. In Stephen A. Wurm (ed.), *New Guinea Area Languages and Language Study Vol 1: Papuan Languages and the New Guinea linguistic scene* (Pacific Linguistics: Series C 38), 263-276. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Greenhill, Simon J. 2015. TransNewGuinea.org: An Online Database of New Guinea Languages. *PLoS ONE* 10(10). 1–17.

Hartzler, Margaret & Kenneth J. Gregerson. 1987. Towards a reconstruction of Proto Tabla-Sentani phonology. *Oceanic Linguistics* 26. 1–29.

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.].

Holton, Gary & Laura C. Robinson. 2014. The internal history of the Alor-Pantar language family. In Marian Klamer (ed.), *Alor-Pantar languages: History and typology*, 55-97. Berlin: Language Science Press.

Jäger, Gerhard. 2019. Computational historical linguistics. *Theoretical Linguistics* to appear. to appear.

Kondrak, Grzegorz. 2009. Identification of Cognates and Recurrent Sound Correspondences in Word Lists. *Traitement Automatique des Langues* 50(2). 201–235.

Laycock, Donald C. 1965. *The Ndu language family (Sepik District, New Guinea)* (Linguistic Circle of Canberra Publications: Series C, Books 1). Canberra: Australian National University.

List, Johann-Mattis, Mary Walworth, Simon J. Greenhill, Tiago Tresoldi & Robert Forkel. 2018. Sequence comparison in computational historical linguistics. *Journal of Language Evolution* 3(2). 130.

List, Johann-Mattis, Simon J. Greenhill & Russell D. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLOS ONE* 12(1). 1–18.

List, Johann-Mattis. 2014. Sequence comparison in historical linguistics. Düsseldorf: Heinrich Heine University doctoral dissertation.

Oswalt, Robert L. 1970. The detection of remote linguistic relationships. *Computer Studies in the Humanities and Verbal Behavior* 3. 117–129.

Palmer, Bill. 2018. Language families of the New Guinea Area. In Bill Palmer (ed.), *Papuan Languages and Linguistics*, 1-20. Berlin: Mouton.

Rarrick, Samantha Carol. 2014. Evaluating the Chimbu-Wahgi Subgrouping. In Priscila Leal & Gordon West (eds.), *Proceedings 2014: Selected Papers from the Eighteenth College-Wide Conference for Students in Languages, Linguistics & Literature*, 94-106. Honolulu: University of Hawaiʻi at Mānoa.

Schapper, Antoinette, Juliette Huber & Aone van Engelenhoven. 2014. The relatedness of Timor-Kisar and Alor-Pantar languages: A preliminary demonstration. In Marian Klamer (ed.), *Alor-Pantar languages: History and typology*, 99-154. Berlin: Language Science Press.

Smallhorn, Jacinta. 2011. *The Binanderean languages of Papua New Guinea* (Pacific Linguistics 625). Canberra: Research School of Pacific and Asian Studies, Australian National University.

St Arnaud, Adam. 2017. Identifying cognate sets across dictionaries of related languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2519–2528. Association for Computational Linguistics.

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

Computational Historical Linguistics. *Language Dynamics & Change* 1. 89–127.

Suter, Edgar & Timothy Usher. 2017. The Kamula-Elevala Language Family. *Language and Linguistics in Melanesia* 35. 105–131.

Usher, Timothy & Edgar Suter. 2015. The Anim Languages of Southern New Guinea. *Oceanic Linguistics* 54(1). 110–142.

Usher, Timothy. 2014. Bulaka River Consonants. *Journal of Language Relationship* 12. 31–50.

Voorhoeve, Bert. 2005. Asmat-Kamoro, Awyu-Dumut and Ok: An enquiry into their linguistic relationship. In Andrew Pawley, Robert Attenborough, Jack Golson & Robin Hide (eds.), *Papuan Pasts: Studies in the Cultural, Linguistic and Biological History of the Papuan-speaking Peoples* (Pacific Linguistics 572), 145-166. Canberra: Research School of Pacific and Asian Studies, Australian National University.

de Vries, Lourens, Ruth Wester & Wilco van den Heuvel. 2012. The Greater Awyu language family of West Papua. In Harald Hammarström & Wilco van den Heuvel (eds.), *History, contact and classification of Papuan languages* (LLM Special Issue 2012), 269-312. Port Moresby: Linguistic Society of Papua New Guinea.

Wester, Ruth. 2014. A linguistic history of Awyu-Dumut: Morphological Study and Reconstruction of a Papuan Language Family. Vrije Universiteit Amsterdam doctoral dissertation.

Wichmann, Søren. 2012. A classification of Papuan languages. In Harald Hammarström & Wilco van den Heuvel (eds.), *History, contact and classification of Papuan languages* (LLM Special Issue 2012), 313-386. Port Moresby: Linguistic Society of Papua New Guinea.

Xiao, Hong. 1990. A Genetic Comparison of Hua, Awa and Binumarien. *Language and Linguistics in Melanesia* 21. 143–166.