Trees, Waves and Friends - Group Work

Maria Zielenbach

May 2023

Tree Tasks

1. Read through the short summary of the Tree Model. Have a look at the examples. How does the figure in example 1 depict the basic assumptions of the Tree Model?

2. Trick question: Try to subgroup the dialects in the table of example 2. (Nein = no, Ja = yes, teils = partially) Dutch (Niederländisch) represents the older stage, all Ja and teils fields are innovations.

3. Read through the short summary of traditional Lexicostatistics. Have a look at example 3. How did Grimes & Grimes get from the similarities in fig. 1 to the classification in fig. 3?

4. Trick question: Compare the sound correspondences proposed by Voorhoeve (1994) with the Tree infered by Grimes & Grimes (1984). Can you find any matches between isoglosses in Voorhoeve's correspondences and the subgroupings in the Tree?

5. Read through the short summary of Bayesian Phylogenetics and have a look at example 4.

6. Compare the consensus tree and the densitree in the example (note: they show different languages!). What are the advantages and disadvantages of the two trees?

7. Imagine you encounter a language tree in the literature. What information would you need in order to know on which method it is based? Which steps would you take to evaluate the tree?

8. Try to fill out the first three rows of the table on p. 9. Which similarities and differences do you notice between the models?

9. What do you like or don't like about the Tree Model and the methods used to infer Trees? Can you think of cases for which it is not applicable? Which problems did you encounter in the trick questions?

The Traditional Tree Model

The Tree Model (Germ. *Stammbaummodell*) is *the* model in historical linguistics. It is usually attributed to German philologist August Schleicher (1821-1868), however, there are language trees that precede Schleicher's (1853) paper (cf. List et al. (2016)).

> "Aus der Art und Weise, wie sämmtliche indogermanische Sprachen unter einander verwandt sind, schloss man nun mit Recht, dass sie aus einer Urspache entsprungen seien, dass eine Nation, das indogermanische Urvolk, sich mit der Zeit in jene acht Völker getheilt habe, von denen jedes in ähnlicher Weise sich später wieder differenziirte, bis endlich die Mannigfaltigkeit unserer Epoche erstand." (Schleicher 1853: 786)

What does it do and what does it assume?

- The Tree Model assumes that language diversify through abrupt splits from an ancestor language and isolated development of the new varieties.
- Related languages are sorted into smaller genetic groups ('subgrouping') based on *shared innovations* established through the *Historical Comparative Method*. The languages in these subgroups are more closely related.
- Contact phenomena (e.g. borrowing) are treated as a kind of anomaly since the Tree Model only considers inherited material ('vertical descent').

How are the findings depicted and how to evaluate them?

- Traditional Trees usually proceed from top to bottom or left to right, with the proto-language (German *Ursprache*) at the top or left side.
- Every node in the tree represents a split. The language that heads the node is the proto-language from which the languages at the end of each line below the node developed. All languages below a node form a *subgroup*.

Example 1 – Semitic

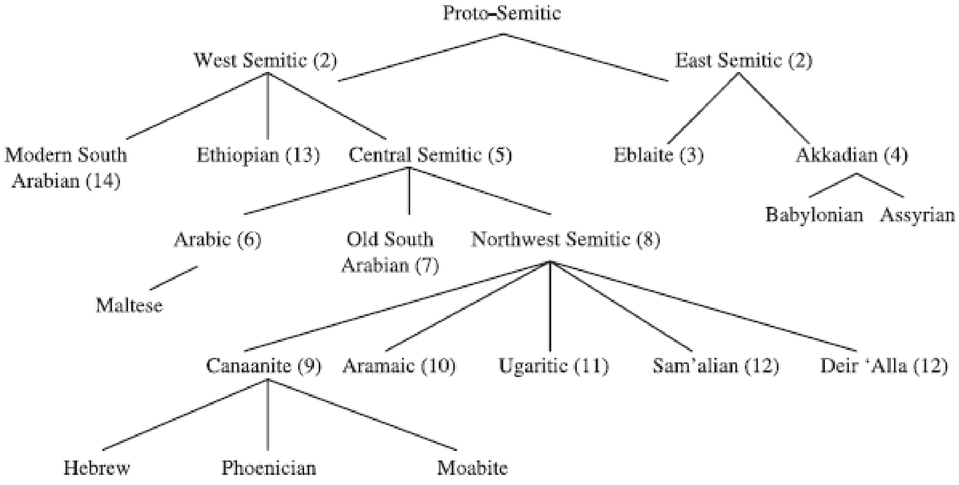


Fig. 1. The subgrouping of the Semitic language family.

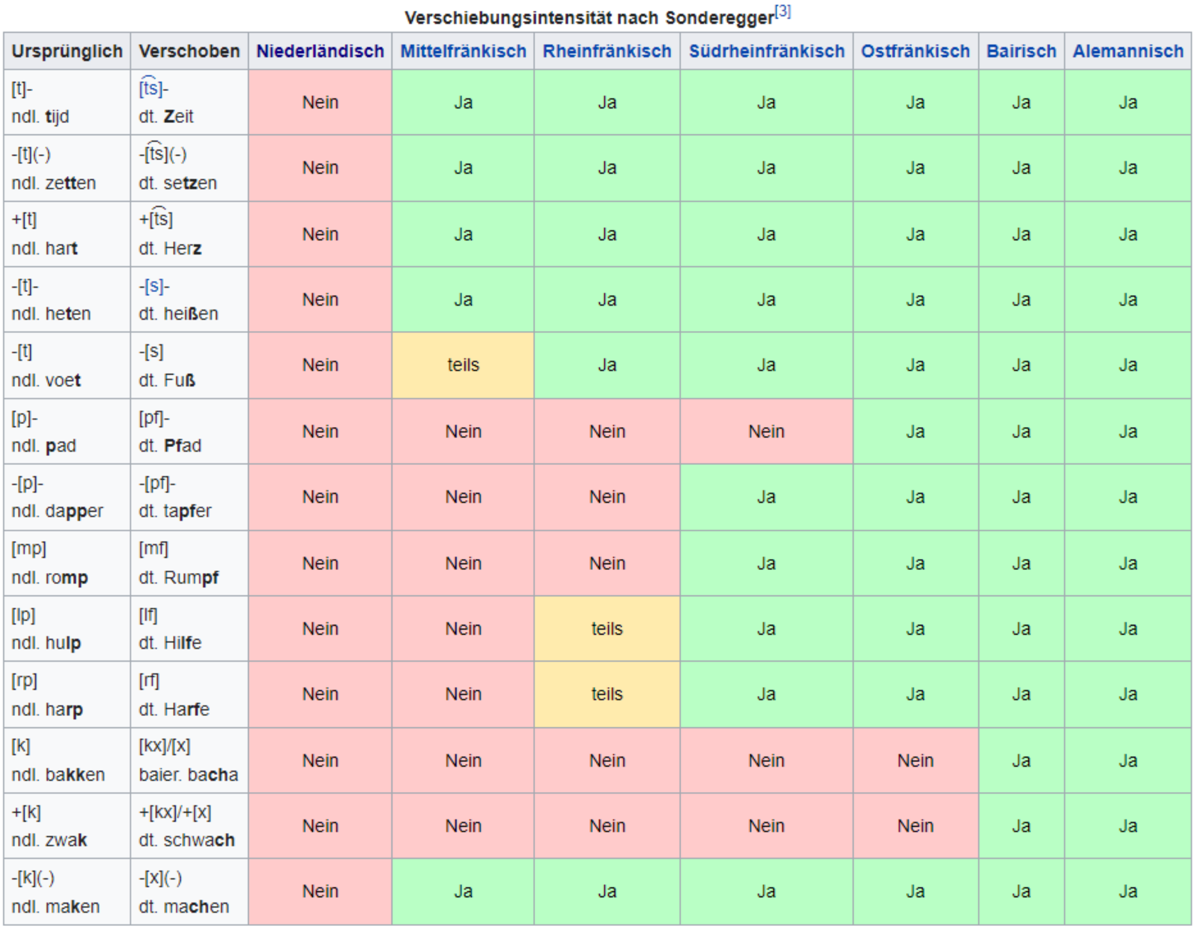
This family tree for the Semitic languages was proposed by Rubin (2008). You can see that there is one primary split from Proto-Semitic into West Semitic and East Semitic. This split is based on the following shared innovations:

- West Semitic innovated a suffixing past tense from participles + enclitic pronouns that replaced the older prefixing past
- East Semitic innovated a new perfect tense with an infix *-t-*
- East Semitic innovated a set of possessive adjectives

These three innovations are uniquely shared between the West or East Semitic languages.

Example 2 – West Germanic dialect continuum

The following table shows phonetic features of several West Germanic dialects in the Netherlands and Germany.



Verschiebungsintensität nach Sonderegger[3]

Ursprünglich	Verschoben	Niederländisch	Mittelfränkisch	Rheinfränkisch	Südrheinfränkisch	Ostfränkisch	Bairisch	Alemannisch
[t]- ndl. **t**ijd	[t͡s]- dt. **Z**eit	Nein	Ja	Ja	Ja	Ja	Ja	Ja
-[t](-) ndl. ze**tt**en	-[t͡s](-) dt. se**tz**en	Nein	Ja	Ja	Ja	Ja	Ja	Ja
+[t] ndl. har**t**	+[t͡s] dt. Her**z**	Nein	Ja	Ja	Ja	Ja	Ja	Ja
-[t]- ndl. he**t**en	-[s]- dt. hei**ß**en	Nein	Ja	Ja	Ja	Ja	Ja	Ja
-[t] ndl. voe**t**	-[s] dt. Fu**ß**	Nein	teils	Ja	Ja	Ja	Ja	Ja
[p]- ndl. **p**ad	[pf]- dt. **Pf**ad	Nein	Nein	Nein	Nein	Ja	Ja	Ja
-[p]- ndl. da**pp**er	-[pf]- dt. ta**pf**er	Nein	Nein	Nein	Ja	Ja	Ja	Ja
[mp] ndl. ro**mp**	[mf] dt. Rum**pf**	Nein	Nein	Nein	Ja	Ja	Ja	Ja
[lp] ndl. hu**lp**	[lf] dt. Hi**lf**e	Nein	Nein	teils	Ja	Ja	Ja	Ja
[rp] ndl. ha**rp**	[rf] dt. Ha**rf**e	Nein	Nein	teils	Ja	Ja	Ja	Ja
[k] ndl. ba**kk**en	[kx]/[x] baier. ba**ch**a	Nein	Nein	Nein	Nein	Nein	Ja	Ja
+[k] ndl. zwa**k**	+[kx]/+[x] dt. schwa**ch**	Nein	Nein	Nein	Nein	Nein	Ja	Ja
-[k](-) ndl. ma**k**en	-[x](-) dt. ma**ch**en	Nein	Ja	Ja	Ja	Ja	Ja	Ja

Traditional Lexicostatistics

Traditional Lexicostatistics is an early quantitative approach to genetic classification based on shared lexicon. It was developed in the 1950s by Morris Swadesh (1909-1967) with the aim to introduce a means for dating branching events.

> “[T]he fundamental everyday vocabulary of any language – as against the specialized or "cultural" vocabulary – changes at a relatively constant rate. The percentage of retained elements in a suitable test vocabulary therefore indicates the elapsed time.” (Swadesh 1952: 452)

What does it do and what does it assume?

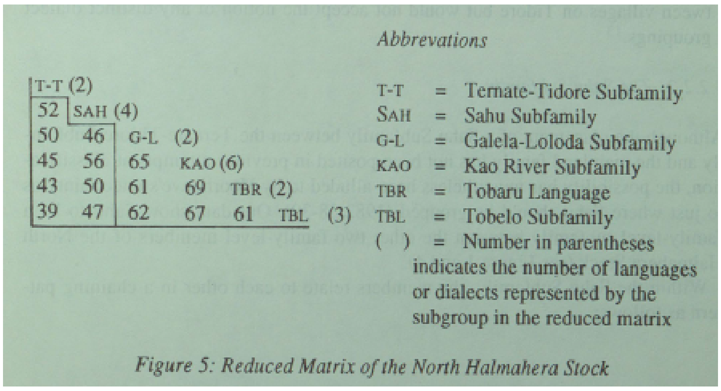
- Lexicostatistics measures relatedness between two languages based on lexical similarity.
- Standardized word lists (‘Swadesh lists’) are use which cover is allegedly ‘stable’ vocabulary which is unlikely to be borrowed (‘basic vocabulary’).
- Lexicostatistics assumes that lexical similarity of basic vocabulary is lost at a consonant rate, this is used to date branching events = glottochronology (in analogy to carbon dating)

How are the findings depicted and how to evaluate them?

- Sometimes the results of a lexicostatistic survey are depicted in a tree, these can be read like traditional trees (but be careful: they are based on a completely different methods).
- Another means of depicting lexicostatistic results are matrices that show the similarity between each compared language.

Example 3 – The North Halmahera Languages

The North Halmahera languages are a small Papuan (non-Austronesian) family of about 10 languages, spoken in the North Moluccas (East Indonesia). In 1994, Grimes & Grimes published a lexicostatistic survey of the languages and proposed a classification which is now used in Glottolog (cf fig. 3). The classification is based on similarity, according to the percentages scheme shown in fig. 2.
The sound correspondences in fig. 4 were proposed by Voorhoeve (1994).



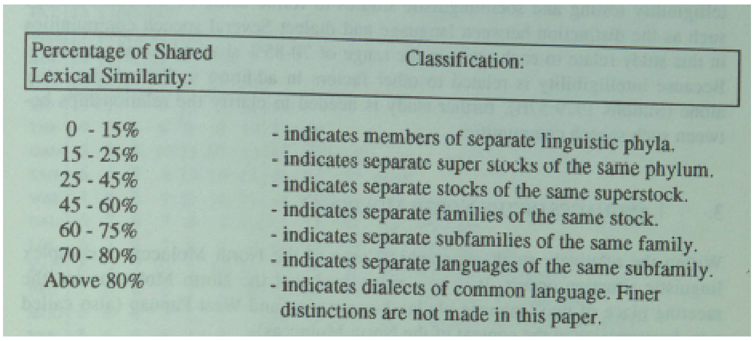
T-T (2)					
52	SAH (4)				
50	46	G-L (2)			
45	56	65	KAO (6)		
43	50	61	69	TBR (2)	
39	47	62	67	61	TBL (3)

Abbrevations

T-T	=	Ternate-Tidore Subfamily
SAH	=	Sahu Subfamily
G-L	=	Galela-Loloda Subfamily
KAO	=	Kao River Subfamily
TBR	=	Tobaru Language
TBL	=	Tobelo Subfamily
( )	=	Number in parentheses indicates the number of languages or dialects represented by the subgroup in the reduced matrix

Figure 5: Reduced Matrix of the North Halmahera Stock

Figure 1: Lexicostatistic matrix for North Halmahera



Percentage of Shared Lexical Similarity:	Classification:
0 - 15%	- indicates members of separate linguistic phyla.
15 - 25%	- indicates separate super stocks of the same phylum.
25 - 45%	- indicates separate stocks of the same superstock.
45 - 60%	- indicates separate families of the same stock.
60 - 75%	- indicates separate subfamilies of the same family.
70 - 80%	- indicates separate languages of the same subfamily.
Above 80%	- indicates dialects of common language. Finer distinctions are not made in this paper.

Figure 2: Percentages scheme

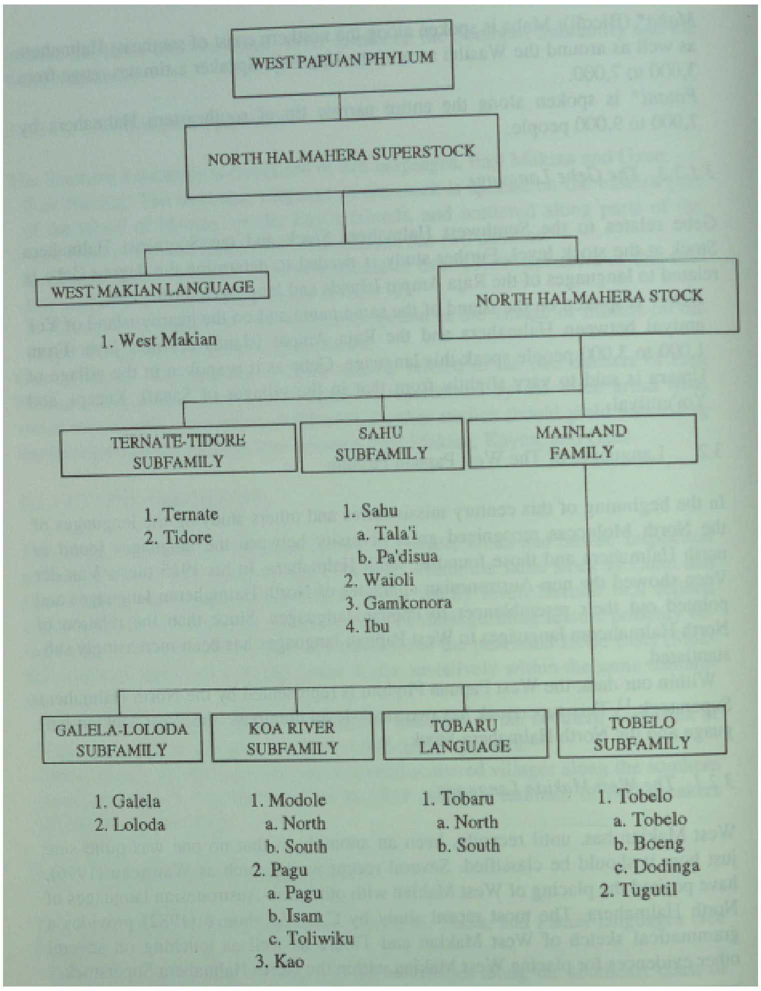


Figure 3: Classification of NH according to Grimes & Grimes

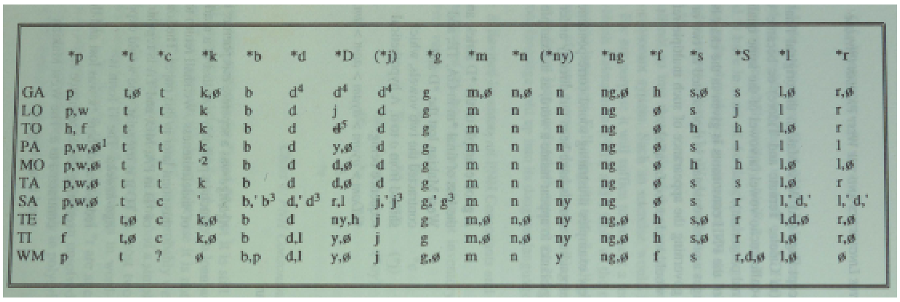
Bayesian Phylogenetics

Bayesian Phylogenetics is a new method for doing lexicostatistics which is based on methods from biology and Bayesian statistics. It is claimed to balance out the shortcomings of Traditional Lexicostatistics and Traditional Trees. It is very important to understand that it is not just a 'modern' method to infer a genetic Tree but that it relies on different theoretical and methodological assumptions (cf Carling et al. (2022)).

> "The approach infers a sample of trees that have a high probability of explaining the patterns in the data under a specified model of character change." (Greenhill et al. 2021: 229)

What does it do and what does it assume?

- Bayesian Phylogenetics can process large data sets (better than humans).



	*p	*t	*c	*k	*b	*d	*D	(*j)	*g	*m	*n	(*ny)	*ng	*f	*s	*S	*l	*r
GA	p	t,ø	t	k,ø	b	d^{4}	d^{4}	d^{4}	g	m,ø	n,ø	n	ng,ø	h	s,ø	s	l,ø	r,ø
LO	p,w	t	t	k	b	d	j	d	g	m	n	n	ng	ø	s	j	l	r
TO	h, f	t	t	k	b	d	ɗ^{5}	d	g	m	n	n	ng	ø	h	h	l,ø	r
PA	p,w,ø^{1}	t	t	k	b	d	y,ø	d	g	m	n	n	ng	ø	s	l	l	l
MO	p,w,ø	t	t	'^{2}	b	d	d,ø	d	g	m	n	n	ng	ø	h	h	l,ø	l,ø
TA	p,w,ø	t	t	k	b	d	d,ø	d	g	m	n	n	ng	ø	s	s	l,ø	r
SA	p,w,ø	t	c	'	b,' b^{3}	d,' d^{3}	r,l	j,' j^{3}	g,' g^{3}	m	n	ny	ng	ø	s	r	l,' d,'	l,' d,'
TE	f	t,ø	c	k,ø	b	d	ny,h	j	g	m,ø	n,ø	ny	ng,ø	h	s,ø	r	l,d,ø	r,ø
TI	f	t,ø	c	k,ø	b	d,l	y,ø	j	g	m,ø	n,ø	ny	ng,ø	h	s,ø	r	l,ø	r,ø
WM	p	t	?	ø	b,p	d,l	y,ø	j	g,ø	m	n	y	ng,ø	f	s	r,d,ø	l,ø	ø

Figure 4: Sound correspondences according to Voorhoeve (1994)

- It can: “evaluate subgrouping hypothesis”, “date language divergences”, “estimate ancestral states”, “infer rates of change in lexical and grammatical traits”, “test hypotheses of functional dependencies in linguistic features”, “infer geographic homelands and migration routes” (Greenhill et al. 2021: 228-229)
- Instead of one tree, a a bunch of trees (forest) is generated. The trees in the forest are rated due to their probability. There are different ways to generate forests.
- Like lexicostatistics, Bayesian phylogenetics can be misused, especially when non-specialists take data from random languages without establishing relatedness or cognacy first.
- The data comes in the form of word lists (e.g. Swadesh’s 200-meaning list), coded for cognacy beforehand.
- Bayesian phylogenetics does not directly use sound changes as diagnostics for subgrouping, instead cognates are assumed to entail these innovations.
- Splits are dated based on historical information which is then extrapolated to parts of the tree for which no historical records exist, instead of using set rates for all language families (‘relaxed clock’).

How are the findings depicted and how to evaluate them?

- *Consensus trees* show the common denominator of the forest, the little numbers next to the branches show the probability of each subgroup (‘clade’) aka the percentage of generated trees in the forest that include this subgroup
- *Densitrees* show all probably generated trees on top of each other and hence the uncertainties

Example 4 – Austronesian

The following classifications of Austronesian languages were published in Greenhill & Gray (2009) and Greenhill (2021). They show different languages.

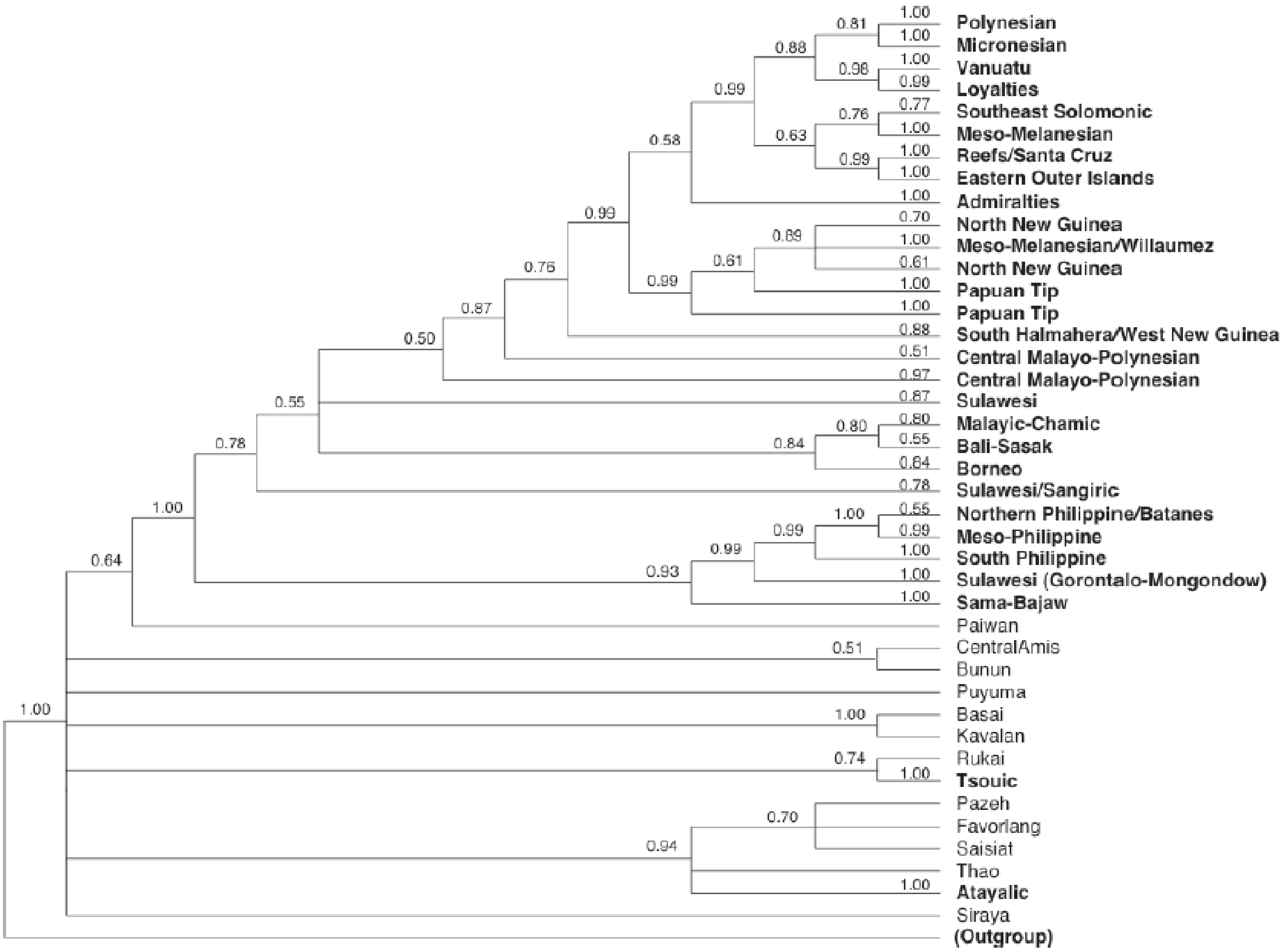


Figure 2: The majority-rule consensus tree of all post burn-in Austronesian trees. Labels in bold represent subgroups of languages, normally-weighted labels denote languages. Where subgroups appear twice in the tree this indicates that they are not monophyletic (e.g. Central Malayo-Polynesian). The numbers on the branches denote the posterior probability of each node. For example, the split between the Northern- and Meso-Philippine languages is strongly supported (1.00). Posterior probability values below 0.50 are considered weak and are not included.

Figure 5: Consensus Tree for Austronesian

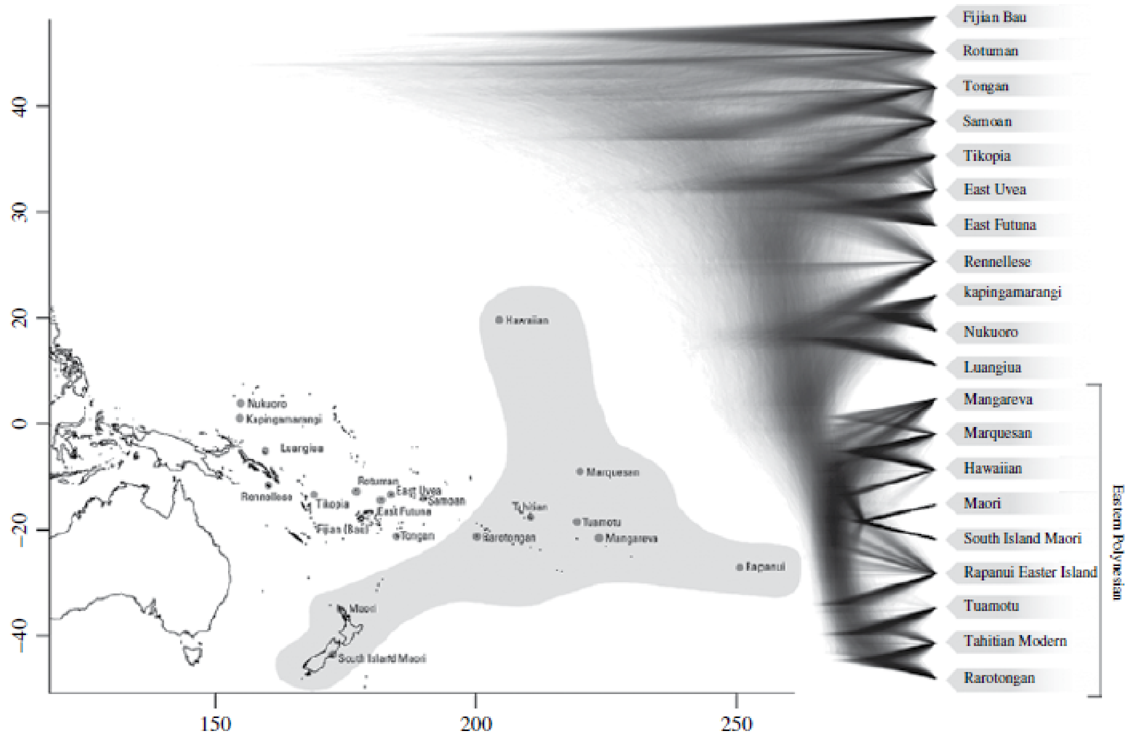


Figure 11.3 A densitree and map of the Central Pacific languages constructed from the posterior distribution of trees from a Bayesian analysis of basic vocabulary. Note that the densitree reveals considerable conflicting signal that cannot be captured in single tree.

Figure 6: Densitree for Austronesian

	how do languages diversify?	what is a subgroup?	which method?	what kind of data is used?	how is it depicted?
Trad. Tree Model					
Trad. Lexicostatistics					
Bayesian Phylogenetics					
Trad. Wave Model					
Hist. Glottometry					

Wave Tasks

1. Read through the short summary of the traditional Wave Model. Have a look at the examples.
2. The figure in example 5 shows the ‘original’ isogloss map by Bloomfield. In which ways does it illustrate the basic assumptions of the Wave Model, and which assumptions aren’t depicted?
3. Let’s have a look at the West Germanic dialects in example 2 again. Draw a rudimentary isogloss map for the features in the table.
4. Read through the short summary of Historical Glottometry. Have a look at example 7.
5. How does the glottometric diagram in example 7 differ from Bloomfield’s isogloss map? Which subgroups have a high cohesiveness or subgroupiness?
6. Trick question: Can you infer from the glottometric diagram in which order the innovations took place?
7. Try to fill out the remaining rows of the table on p. 9.
8. What do you like or don’t like about the Wave Model and Historical Glottometry? Can you think of cases for which it is not applicable? Which problems did you encounter in the trick question?

The Traditional Wave Model

The Wave Model (German *Wellentheorie*) was developed as an alternative to the Tree Model. It is usually attributed to Johannes Schmidt (1843-1901, Schmidt (1872)) but was actually proposed a little earlier by Hugo Schuchardt (1842-1927, Schuchardt (1868), cf also Schuchardt (1885)).

> "Ich möchte an [stelle des baumes] das bild der welle setzen, welche sich in concentrischen mit der entfernung vom mittelpunkte immer schwächer werdenden ringen ausbreitet." (Schmidt 1872: 27)

What does it do and what does it assume?

- The Wave Model assumes that dialect continuums are the most common state of languages (languages are never uniform, huge intra-language variation is the norm), therefore the Wave Model claims to have greater historical accuracy than the Tree Model.
- Innovations are assumed to start in one variety and then 'wave' over adjacent varieties which adopt the innovations, likewise innovations within a variety (e.g. sound changes) start in one word and then 'wave' over to other words. Diversification evolves through slow divergence of varieties first into dialects and then into languages.
- "Since later changes may not cover the same area, there may be no sharp boundaries between neighbouring dialects or languages; rather, the greater the distance between them, the fewer linguistic traits dialects or languages may share." (Campbell 2013: 188)

How are the findings depicted and how to evaluate them?

- 'Wave diagrams' usually show a number of languages/varieties, displayed according to their geographic location. This depiction was popularized by Bloomfield (1933)316.
- Lines are used to show the overlapping shared features, all varieties within a circle share a specific feature.

Example 5 – Bloomfield's isoglosses for Indo-European

This is the isogloss map published by Bloomfield (1933) which served as role model for many depictions of the Wave Model.

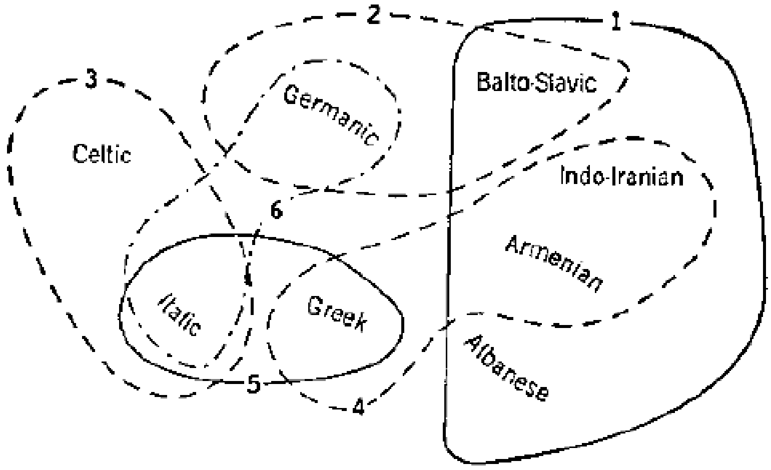
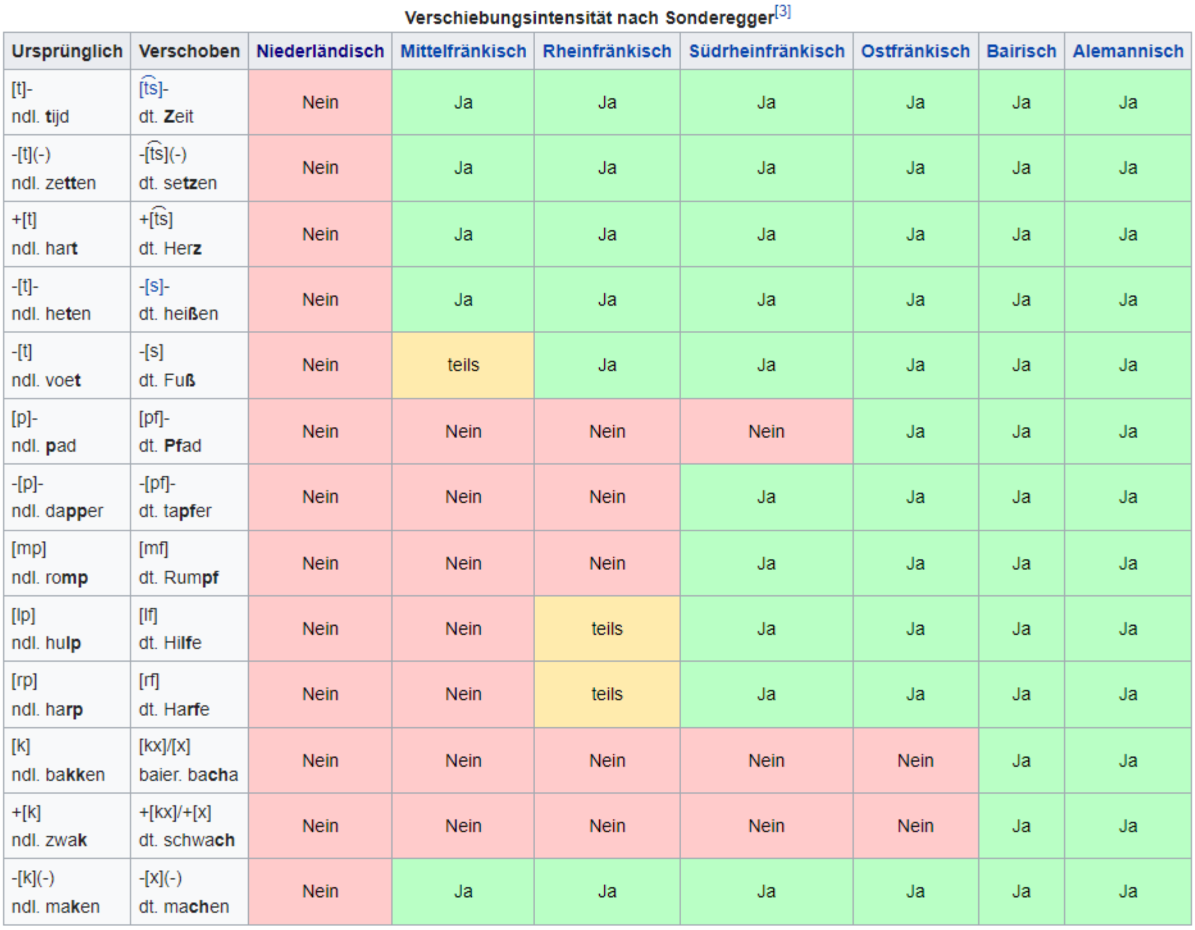


FIGURE 3. Some overlapping features of special resemblance among the Indo-European languages, conflicting with the family-tree diagram. — Adapted from Schrader.

1. Sibilants for velars in certain forms.
2. Case-endings with [m] for [bh].
3. Passive-voice endings with [r].
4. Prefix ['e-] in past tenses.
5. Feminine nouns with masculine suffixes.
6. Perfect tense used as general past tense.

Example 6 – West Germanic dialect continuum

The following table shows phonetic features of several West Germanic dialects in the Netherlands and Germany.



Verschiebungsintensität nach Sonderegger[3]

Ursprünglich	Verschoben	Niederländisch	Mittelfränkisch	Rheinfränkisch	Südrheinfränkisch	Ostfränkisch	Bairisch	Alemannisch
[t]- ndl. tijd	[t͡s]- dt. Zeit	Nein	Ja	Ja	Ja	Ja	Ja	Ja
-[t](-) ndl. zetten	-[t͡s](-) dt. setzen	Nein	Ja	Ja	Ja	Ja	Ja	Ja
+[t] ndl. hart	+[t͡s] dt. Herz	Nein	Ja	Ja	Ja	Ja	Ja	Ja
-[t]- ndl. heten	-[s]- dt. heißen	Nein	Ja	Ja	Ja	Ja	Ja	Ja
-[t] ndl. voet	-[s] dt. Fuß	Nein	teils	Ja	Ja	Ja	Ja	Ja
[p]- ndl. pad	[pf]- dt. Pfad	Nein	Nein	Nein	Nein	Ja	Ja	Ja
-[p]- ndl. dapper	-[pf]- dt. tapfer	Nein	Nein	Nein	Ja	Ja	Ja	Ja
[mp] ndl. romp	[mf] dt. Rumpf	Nein	Nein	Nein	Ja	Ja	Ja	Ja
[lp] ndl. hulp	[lf] dt. Hilfe	Nein	Nein	teils	Ja	Ja	Ja	Ja
[rp] ndl. harp	[rf] dt. Harfe	Nein	Nein	teils	Ja	Ja	Ja	Ja
[k] ndl. bakken	[kx]/[x] baier. bacha	Nein	Nein	Nein	Nein	Nein	Ja	Ja
+[k] ndl. zwak	+[kx]/+[x] dt. schwach	Nein	Nein	Nein	Nein	Nein	Ja	Ja
-[k](-) ndl. maken	-[x](-) dt. machen	Nein	Ja	Ja	Ja	Ja	Ja	Ja

Historical Glottometry

Glottometry was devised by Alexandre François and Siva Kalyan as a modern, Wave Model-based alternative to the Tree Model.

> "The objective of Historical Glottometry is to identify genealogical subgroups in a language family, and measure their relative strengths so as to assess their historical patterns of distribution across social networks." (François 2014: 173)

What does it do and what does it assume?

- Historical Glottometry adopts the basic assumptions of the Wave Model that innovations 'wave' over varieties and languages develop from dialect continuums. On the other hand, it also assumes that 'related' varieties are derived from the same common ancestor. It therefore mostly deals with 'waves' within related (and mutual intelligible) varieties.
- The Historical Comparative Method is used to find shared features between languages.
- Varieties can belong to several diffusion areas (unlike in Trees where every variety or subgroups is headed by one node). "[I]t is often the case [...] that an innovation only spreads partway through a population before that population splits. In this situation, an innovation need not be passed on to all of the descendants of the language it occurs in, but only to some of them." (Kalyan & François 2019: 168)
- Historical Glottometry accepts that languages may diversify due to splits, but such cases are regarded as rare and a special case of 'waves'.

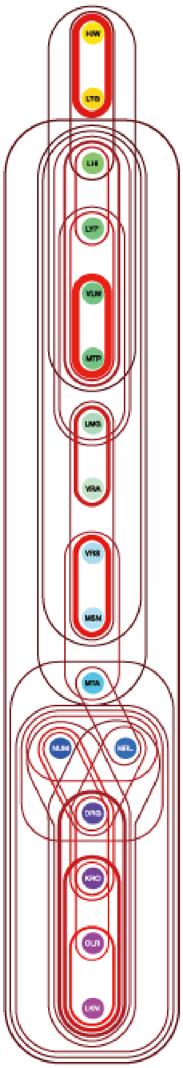
How are the findings depicted and how to evaluate them?

- The diagrams are similar to isogloss maps but the thickness and darkness of the lines is used to show degrees of relationship.

- The darkness of the lines shows the ‘cohesiveness’ of a subgroup: “the proportion of supporting evidence with respect to the entire set of relevant evidence” (how many innovations does a subgroup share compared to the total number of scrutinized innovations?)
- The thickness of the lines shows the ‘subgroupiness’: “the product of the cohesiveness rate (k) with the number of exclusively shared innovations” (how many innovations does a subgroup share compared to the total number of scrutinized innovations and how many does it share exclusively?)

Example 7 – The languages auf Vanuatu

This diagram shows subgroups of languages of Vanuatu, more precisely of the Torres and Banks islands (Kalyan & François 2018).



Figures 5-11 A glottometric diagram of the Torres–Banks languages

Literature

Bloomfield, Leonard. 1933. *Language*. London: Allen & Unwin.

Campbell, Lyle. 2013. *Historical linguistics: an introduction*. 3. ed. Edinburgh: Edinburgh Univ. Press. 538 pp.

Carling, Gerd, Chundra Cathcart & Erich Round. 2022. Reconstructing the origins of language families and variation. In Nathalie Gontier, Andy Lock & Chris Sinha (eds.), *The Oxford Handbook of Human Symbolic Evolution*, 1st edn. Oxford University Press. `https://doi.org/10.1093/oxfordhb/9780198813781.013.34`.

François, Alexandre. 2014. Trees, waves and linkages: Models of Language Diversification. In Claire Bowern & Bethany Evans (eds.), *The Routledge Handbook of Historical Linguistics* (Routledge Handbooks in Linguistics), 161–189. London; New York: Routledge. `https://doi.org/10.4324/9781315794013.ch6`.

Greenhill, Simon J & Russell D Gray. 2009. Austronesian language phylogenies: myths and misconceptions about Bayesian computational methods. In *Austronesian historical linguistics and culture history: a festschrift for Robert Blust*, 375–397. Canberra: Pacific Linguistics.

Greenhill, Simon J. 2021. Bayesian Phylogenetic Methods.

Greenhill, Simon J., Paul Heggarty & Russell D. Gray. 2021. Bayesian Phylolinguistics. In Richard D. Janda, Brian D. Joseph & Barbara S. Vance (eds.), *Handbook of historical linguistics* (Blackwell Handbooks in Linguistics volume 2), 226–253. Hoboken, NJ, USA: Wiley.

Grimes, Charles E. & Barbara D. Grimes. 1984. Languages of the North Moluccas. In *Maluku dan Irian Jaya*, vol. III:1 (Buletin LEKNAS: Terbitan Khusus), 35–63. Jakarta: LEKNAS-LIPI.

Kalyan, Siva & Alexandre François. 2018. Freeing the Comparative Method from the Tree Model: A Framework for Historical Glottometry. *Senri Ethnological Studies* 98. 59–89.

Kalyan, Siva & Alexandre François. 2019. When the waves meet the trees: A response to Jacques and List. *Journal of Historical Linguistics* 9(1). 168–177. `https://doi.org/10.1075/jhl.18019.kal`.

List, Johann-Mattis, Jananan Sylvestre Pathmanathan, Philippe Lopez & Eric Bapteste. 2016. Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics. *Biology Direct* 11(1). 39. `https://doi.org/10.1186/s13062-016-0145-2`.

Rubin, Aaron D. 2008. The Subgrouping of the Semitic Languages: The Subgrouping of the Semitic Languages. *Language and Linguistics Compass* 2(1). 79–102. `https://doi.org/10.1111/j.1749-818X.2007.00044.x`.

Schleicher, A. 1853. Die ersten Spaltungen des indogermanischen Urvolkes. *Allgemeine Monatsschrift für Wissenschaft und Literatur* 3. 786–787. `https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_2381174` (15 November, 2022).

Schmidt, Johannes. 1872. *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*. Weimar: H. Böhlau.

Schuchardt, Hugo Ernestus Mario. 1868. *Der Vokalismus des Vulgärlateins*. Leipzig: Teubner.

Schuchardt, Hugo Ernestus Mario. 1885. *Ueber die Lautgesetzte. gegen die Junggrammatiker*. Berlin: Robert Oppenheim.

Swadesh, Morris. 1952. Lexico-Statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* 96(4). 452–463.

Voorhoeve, C. L. 1994. Comparative linguistics and the West Papuan phylum. In E. K. M. Masinambow (ed.), *Maluku dan Irian Jaya*, vol. 3 (Buletin LEKNAS: Terbitan Khusus), 65–90. Jakarta: Lembaga Ekonomi dan Kemasyarakatan Nasional, LIPI.