Quantitative Lexical Comparison

Michael Cysouw

http://www.eva.mpg.de/~cysouw/teaching/

Language Comparison

What is the reason for a particular similarity between two languages?

- Historical reasons
  - Shared descent
  - Contact
- A-historical reasons
  - Inherent (“universal”) preference
  - Chance

Quantitative Approach

Why should linguists use quantitative methods ?

- Upscaling
  - dealing with more data
- Methodological innovation
  - Using ‘cumbersome’ methods
- Greater precision
  - Beyond discrete ‘yes/no’ results

Lexical Data

What kind of data is used for the language comparison ?

- Traditional trinity of language description
  - Grammar
  - Texts
  - Dictionary
- Generalized dictionary (“constructicon”)
  - Language particular constructions
  - Combination of phonemic structure and meaning/context

Using Lexical Data

What can we do with lexical data ?

- Using lexical variation
  - Meaning
  - Form
- To investigate
  - a-historical general tendencies of lexical structure and sound structure
  - historical processes of meaning change, sound change and language relationship

History of Quantification

- Early quantitative historical analysis
  - Kroeber, & Chrétien (1937, 1939), Chrétien (1943), Kroeber (1960)
- Distribution of reflexes of reconstructions
  - Ross (1950), Davies & Ross (1975)
  - Holm (2000, 2003, 2007)

History of Quantification

- Swadesh-list approach
  - Swadesh (1952), Lees (1953), Merwe (1966)
  - Chretien (1962)
- (silent) further development
  - Dyen et al. (1967), Dyen (1992)
  - Sankoff (1970, 1972)
  - Dobson et al. (1972), Dobson (1978)
  - Black (1976)
  - Embleton (1986, 1991)

History of Quantification

- Chance of cognation
  - Justeson & Stephens (1979)
  - Ringe (1992)
- Russian developments
  - Dolgoposky (1986 [1964])
  - Yakhontov, Starostin

History of Quantification

- Dialectometry
  - Séguy (1973)
  - Goebl (1984, 2006), Embleton (1985)
  - Nerbonne et al (1999), Heeringa (2004), Heeringa et al. (2005, 2006)
- String Alignment
  - Covington (1996, 1998, 2004)
  - Kondrak (2002, 2003), Kondrak & Sherif (2006), Wesley & Kondrak (2005), Bergsma & Kondrak (2007a, 2007b)
  - Rødseth & Sellars (2006)

History of Quantification

- "New synthesis" with bioinformatics
  - Renfrew *et al.* (2000)
  - Warnow *et al. (*1996), Warnow (1997), Ringe *et al.* (1998, 2002), Nakhleh *et al.* (2005a, 2005b)
  - Pagel (2000), Gray & Jordan (2000), Gray & Atkinson (2003), Atkinson & Gray (2005), Atkinson *et al.* (2005, 2008), Pagel *et al.* (2007)
  - Lohr (1999), Haggerty (2000a, 2000b), McMahon & McMahon (2003, 2005), McMahon et al. (2005)
  - Rexova (2002, 2006)
  - Kessler (2001, 2005, 2006, 2007)
  - Brown *et al.* (2008), Holman *et al.* (2008)

Comparison Biology-Linguistics

- Platnick & Cameron (1977)
- Atkinson & Gray (2005)
- Holm (2007)

Data used

- Lexicon approach
  - all available words are included
- Wordlist approach
  - clearly delimited workload
  - control of amount of knowledge
  - selection possibly guides interpretation
  - but: limited information included

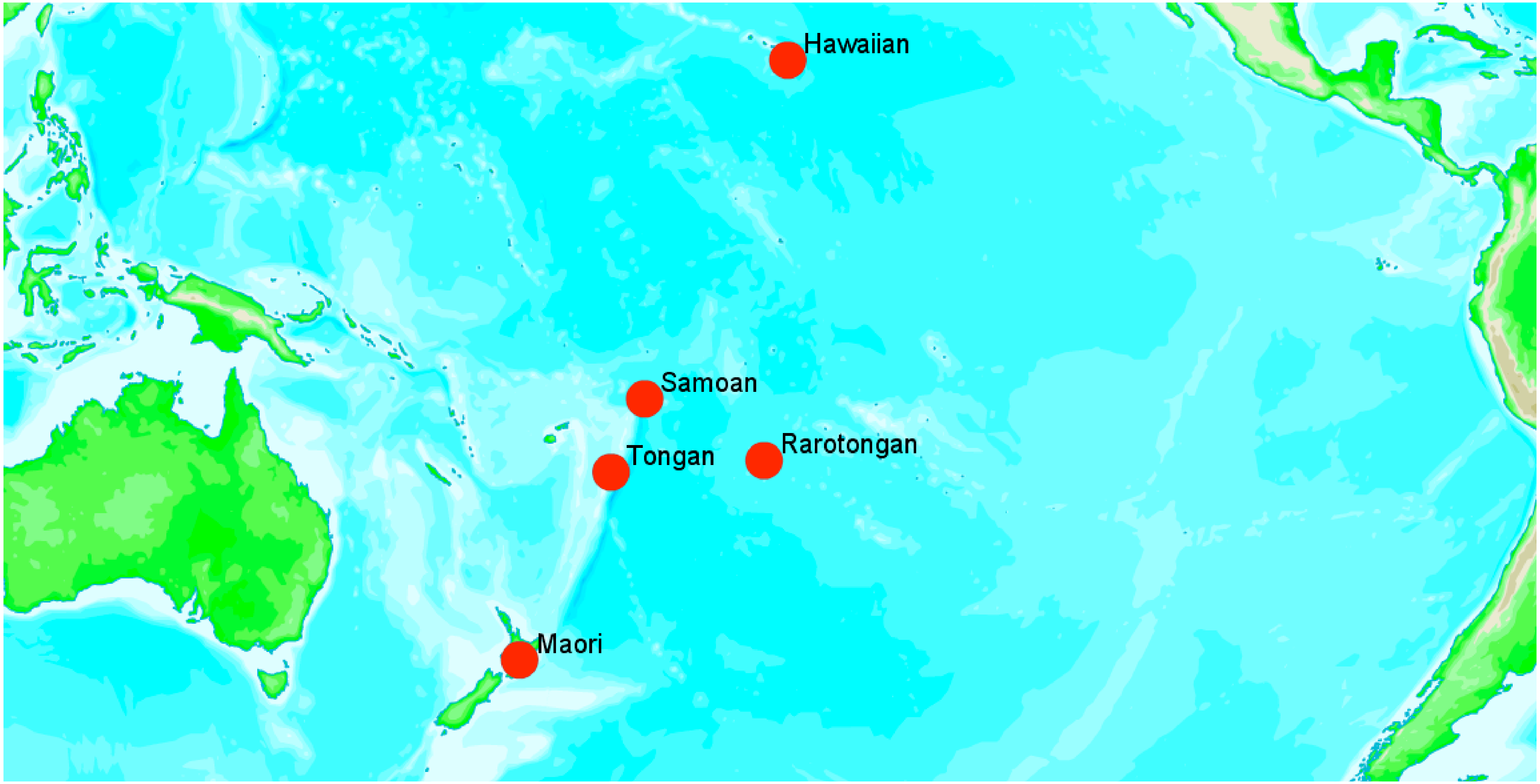
Comparison

- Whole word approach
  - cognate judgements needed
- sub-word approaches
  - sound similarities
  - string similarities
  - character alignment

Comparative Method

- Cognate identification
- Regular sound correspondences
- Correspondence sets
- Reconstruction
- Subgrouping

Polynesian Languages



attention only on the nasals. What will you reconstruct for these? How many nasals do you postulate for Proto-Tulu? State your evidence.
NOTE: *j* = [ǰ], IPA [dʒ]; *ṇ* = IPA [ɳ].

	Shivalli	*Sapaliga*	*gloss*
1.	a:ṇɨ	a:nɨ	‘male’
2.	uṇɨ	a:nɨ	‘dine’
3.	maṇṇɨ	mannɨ	‘soil’
4.	ko:ṇɛ	ko:nɛ	‘room’
5.	e:ṇɨ	ya:nɨ	‘I’
6.	ninɛ	ninɛ	‘wick’
7.	ja:nɛ	da:nɛ	‘what’
8.	sanɛ	tanɛ	‘conceiving’

(Bhat 2001: 11)

Exercise 5.3 Polynesian

The Polynesian languages of the Pacific form a subgroup of the Oceanic branch of the Austronesian family of languages. (1) What are the sound correspondences found in these data? What sound do you reconstruct for the proto-language to represent each sound correspondence set? (2) What sound change or changes have taken place in each of these languages? (3) What is the best reconstruction (proto-form) for 6, 16, 20 and 32? Show how your postulated sound changes apply to each of these to produce the modern forms.
NOTE: <’> = [ʔ].

	Māori	*Tongan*	*Samoan*	*Rarotongan*	*Hawai’ian*	*gloss*
1.	tapu	tapu	tapu	tapu	kapu	‘forbidden’, ‘taboo’
2.	pito	pito	pito	pito	piko	‘navel’
3.	puhi	puhi	—	pu’i	puhi	‘blow’
4.	taha	tafa ‘edge’	tafa	ta’a	kaha	‘side’
5.	tae ‘trash’	ta’e	tae	tae	kae	‘excrement’
6.	taŋata	taŋata	taŋata	taŋata	kanaka	‘man, person’
7.	tai	tahi	tai	tai	kai	‘sea’
8a.	kaha	kafa	’afa	ka’a	’aha	‘strong’
8b.	ma:rohi-	malohi	malosi	ma:ro’i	—	‘strong’
9.	karo	kalo	’alo	karo	’alo	‘dodge’
10.	aka-	aka	a’a	aka	a’a	‘root’
11.	au	’ahu	au	au	au	‘gall’
12.	uru ‘tip of weapon’	’ulu	ulu	uru	ulu ‘centre’	‘head’
13.	uhi	ufi	ufi	u’i	uhi	‘yam’
14.	ahi	afi	afi	a’i	ahi	‘fire’
15.	ɸa:	fa:	fa:	’a:	ha:	‘four’
16.	ɸeke	feke	fe’e	’eke	he’e	‘octopus’
17.	ika	ika	i’a	ika	i’a	‘fish’
18.	ihu	ihu	isu	puta-i’u ‘nostril’ (puta ‘hole’)	ihu	‘nose’
19.	hau ‘wind’ (hauku: ‘dew’ [-ku: ‘showery weather’])	hau	sau	’au	hau	‘dew’
20.	hika	—	si’a	’ika	hi’a	‘firemaking’
21.	hiku ‘fishtail’	hiku	si’u	’iku	hi’u	‘tail’
22.	ake	hake	a’e	ake	a’e	‘up’
23.	uru	—	ulu	uru	ulu	‘enter’
24.	maŋa	maŋa	maŋa	maŋa	mana	‘branch’
25.	mau ‘fixed’	ma’u	mau	mau	mau	‘constant’
26.	mara ‘marinated’	—	mala	mara	mala	‘fermented food’
27.	noho	nofo	nofo	no’o	noho	‘sit’
28.	ŋaru	ŋaru	ŋalu	ŋaru	nalu	‘wave’
29.	ŋutu	ŋutu	ŋutu	ŋutu	nuku	‘mouth’
30.	waka	vaka	va’a	vaka	wa’a	‘canoe’
31.	wae	va’e	vae	vae	wae	‘leg’
32.	raho ‘testicle’	laho	laso	ra’o	laho	‘scrotum’
33.	rou ‘long forked stick’	lohu	lou	rou	lou	‘fruit-picking pole’
34.	roŋo	(loŋo-) (loŋo-a:’a ‘noise’, loŋo-noa ‘silence’)	loŋo	roŋo	lono	‘hear’
35.	rua	-lua (in compounds)	lua	rua	lua	‘two’

	‘taboo’	‘side’	‘trash’
Maori	*tapu*	*taha*	*tae*
Tongan	*tapu*	*tafa*	*ta*'e
Samoan	*tapu*	*tafa*	*tae*
Rarotongan	*tapu*	*ta'a*	*tae*
Hawai’ian	*kapu*	*kaha*	*kae*

	Maori	Tongan	Samoan	Rarot.	Hawai’i	ProtoP.
C1	t	t	t	t	k	*t
C2	p	p	p	p	p	*p
C3	h	h	s	ʔ	h	*s
C4	h	f	f	ʔ	h	*f
C5	ø	ʔ	ø	ø	ø	*ʔ
C6	ŋ	ŋ	ŋ	ŋ	n	*ŋ
C7	ø	h	ø	ø	ø	*h
C8	k	k	ʔ	k	ʔ	*k
C9	m	m	m	m	m	*m
C10	r	l	l	r	l	*L
C11	ɸ	f	f	ʔ	h	*f
C12	n	n	n	n	n	*n
C13	w	v	v	v	w	*V

	Maori	Tongan	Samoan	Rarot.	Hawai'i	ProtoP.
C1	t	t	t	t	k	*t
C2	p	p	p	p	p	*p
C3	h	h	s	ʔ	h	*s
C4	h	f	f	ʔ	h	*f
C5	ø	ʔ	ø	ø	ø	*ʔ
C6	ŋ	ŋ	ŋ	ŋ	n	*ŋ
C7	ø	h	ø	ø	ø	*h
C8	k	k	ʔ	k	ʔ	*k
C9	m	m	m	m	m	*m
C10	r	l	l	r	l	*L
C11	ɸ	f	f	ʔ	h	*f
C12	n	n	n	n	n	*n
C13	w	v	v	v	w	*V

Shared Innovation !

	Maori	Tongan	Samoan	Rarot.	Hawai’i	ProtoP.
C1	t	t	t	t	k	*t
C2	p	p	p	p	p	*p
C3	h	h	s	ʔ	h	*s
C4	h	f	f	ʔ	h	*f
C5	ø	ʔ	ø	ø	ø	*ʔ
C6	ŋ	ŋ	ŋ	ŋ	n	*ŋ
C7	ø	h	ø	ø	ø	*h
C8	k	k	ʔ	k	ʔ	*k
C9	m	m	m	m	m	*m
C10	r	l	l	r	l	*L
C11	ɸ	f	f	ʔ	h	*f
C12	n	n	n	n	n	*n
C13	w	v	v	v	w	*V

	Maori	Tongan	Samoan	Rarot.	Hawai’i	ProtoP.
C1	t	t	t	t	k	*t
C2	p	p	p	p	p	*p
C3	h	h	s	ʔ	h	*s
C4	h	f	f	ʔ	h	*f
C5	ø	ʔ	ø	ø	ø	*ʔ
C6	ŋ	ŋ	ŋ	ŋ	n	*ŋ
C7	ø	h	ø	ø	ø	*h
C8	k	k	ʔ	k	ʔ	*k
C9	m	m	m	m	m	*m
C10	r	l	l	r	l	*L
C11	ɸ	f	f	ʔ	h	*f
C12	n	n	n	n	n	*n
C13	w	v	v	v	w	*V

	Maori	Tongan	Samoan	Rarot.	Hawai’i	ProtoP.
C5	ʔ	ø	ø	ø	ø	*ʔ
C7	h	ø	ø	ø	ø	*h
C3	h	s	h	ʔ	h	*s
C4	f	f	h	ʔ	h	*f
C11	f	f	h	ʔ	ɸ	*f
C8	k	ʔ	ʔ	k	k	*k
C10	l	l	l	r	r	*L
C13	v	v	w	v	w	*V
C9	m	m	m	m	m	*m
C2	p	p	p	p	p	*p
C1	t	t	k	t	t	*t
C12	n	n	n	n	n	*n
C6	ŋ	ŋ	n	ŋ	ŋ	*ŋ

Different approaches

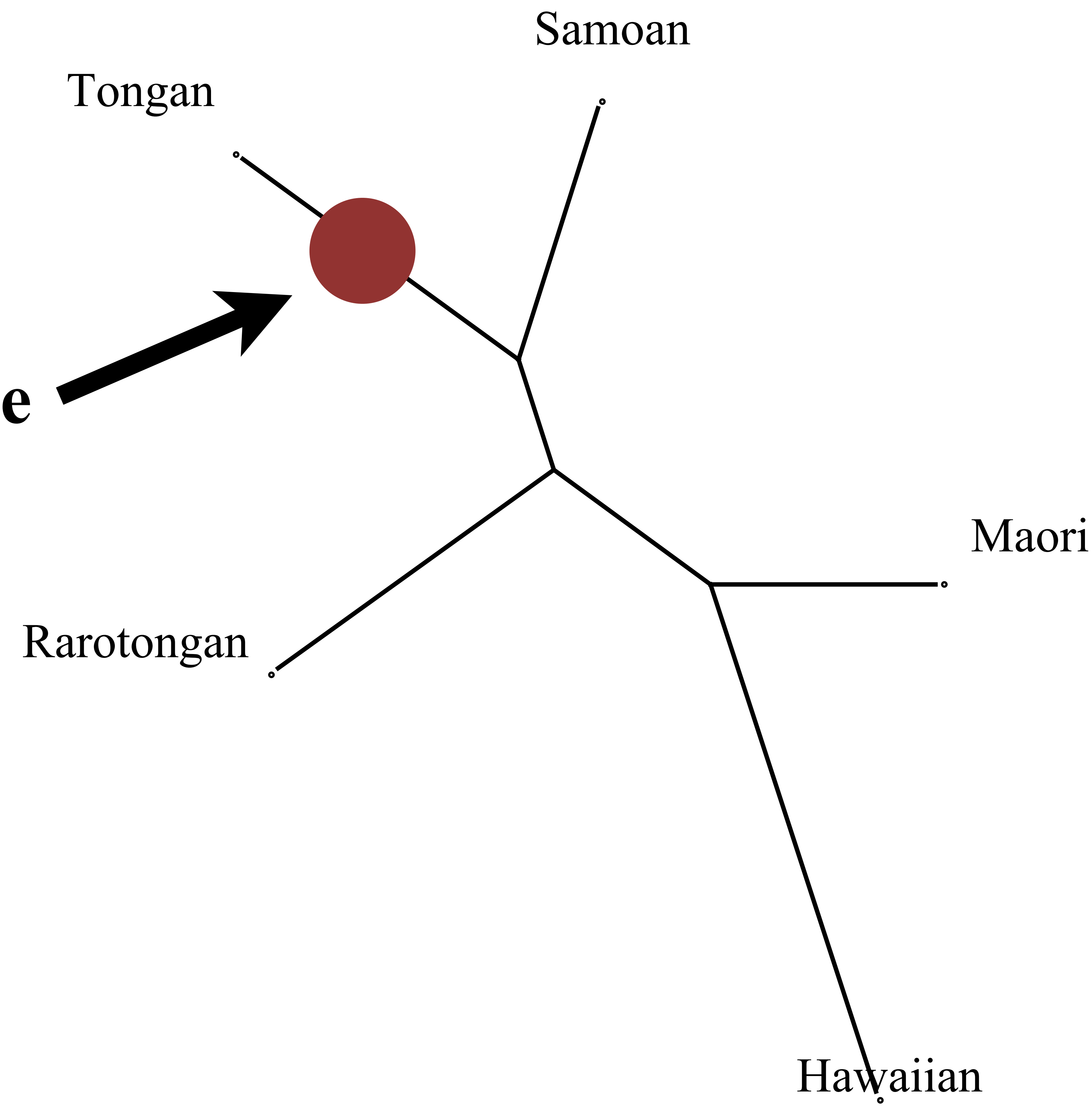
- Historical-comparative approach
  - First reconstruct proto-language
  - Then establish subgrouping of languages
- ‘Bioinformatics’ approach
  - First establish subgrouping (unrooted tree)
  - Then locate proto-language (by outgroup)

Necessary analyses

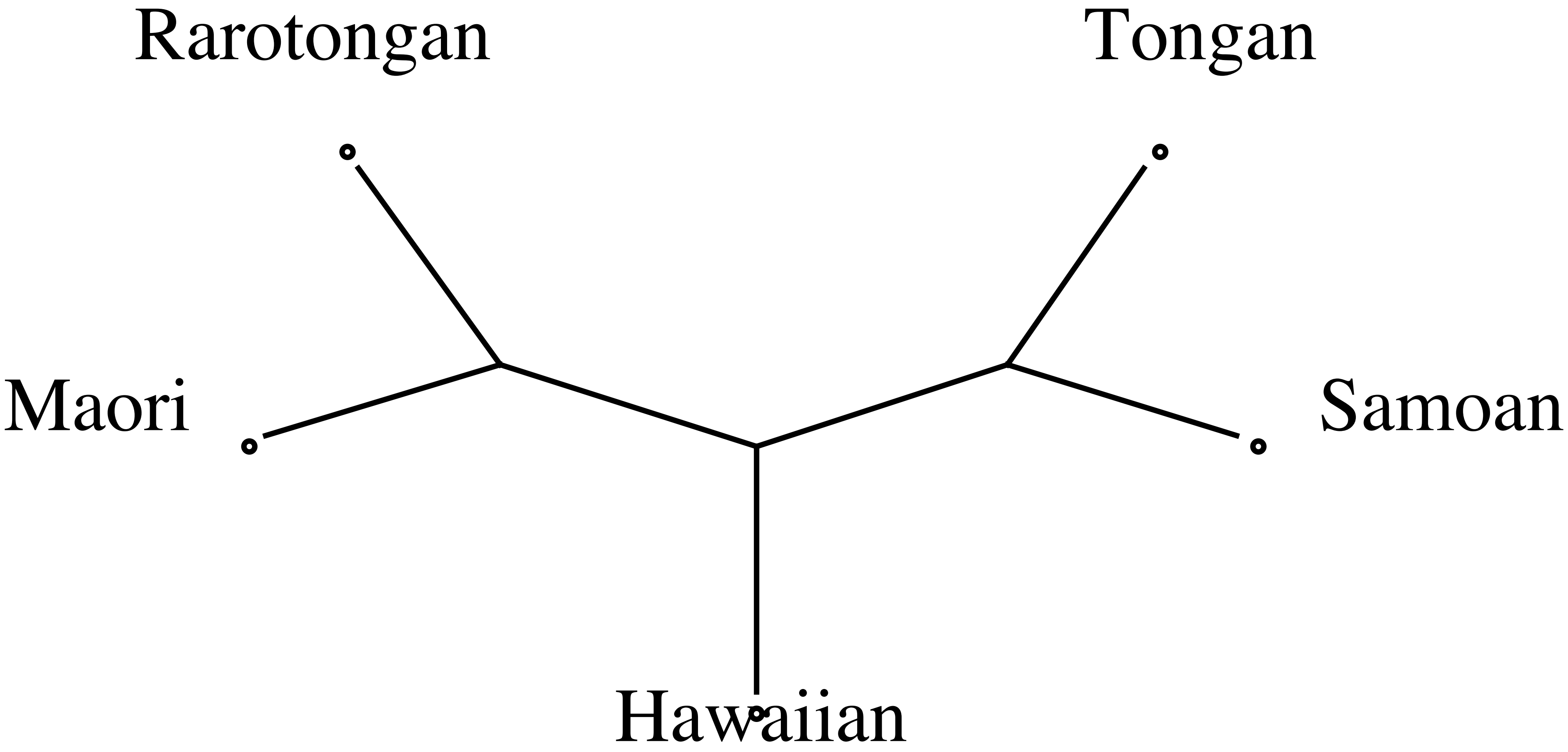
- Subgrouping
  - Similarity-based approaches
  - Parsimony-based approaches

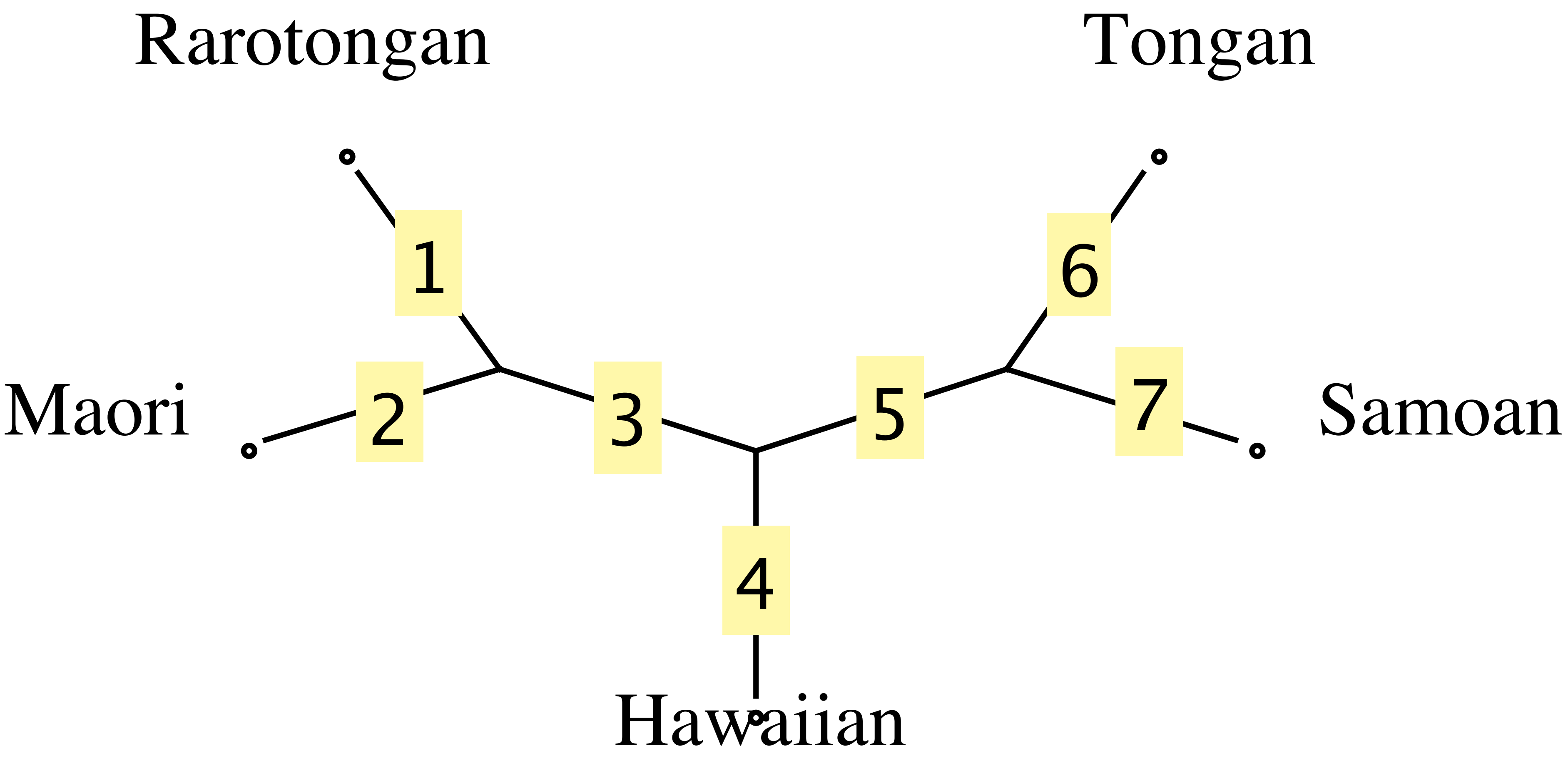
Necessary analyses

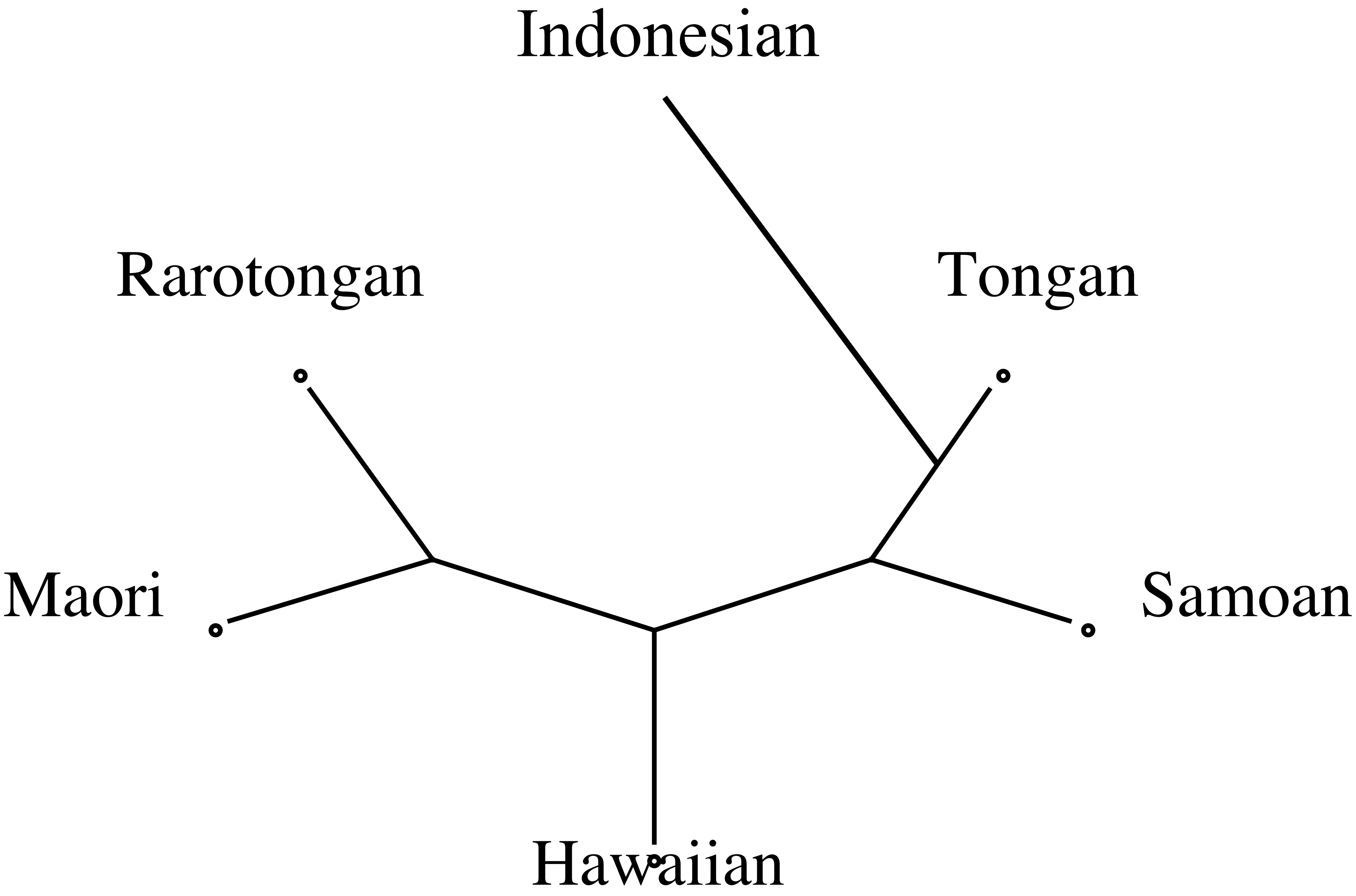
- Subgrouping
  - Similarity-based approaches
  - **Parsimony-based approaches**



Start here


Rarotongan
Tongan
Maori
Samoan
Hawaiian


Rarotongan
Tongan
Maori
Samoan
Hawaiian
1
2
3
4
5
6
7


Indonesian
Rarotongan
Tongan
Maori
Samoan
Hawaiian

Number of Taxa	Number of unrooted trees	Number of rooted trees
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10395	135135
9	135135	2027025
10	2027025	34459425

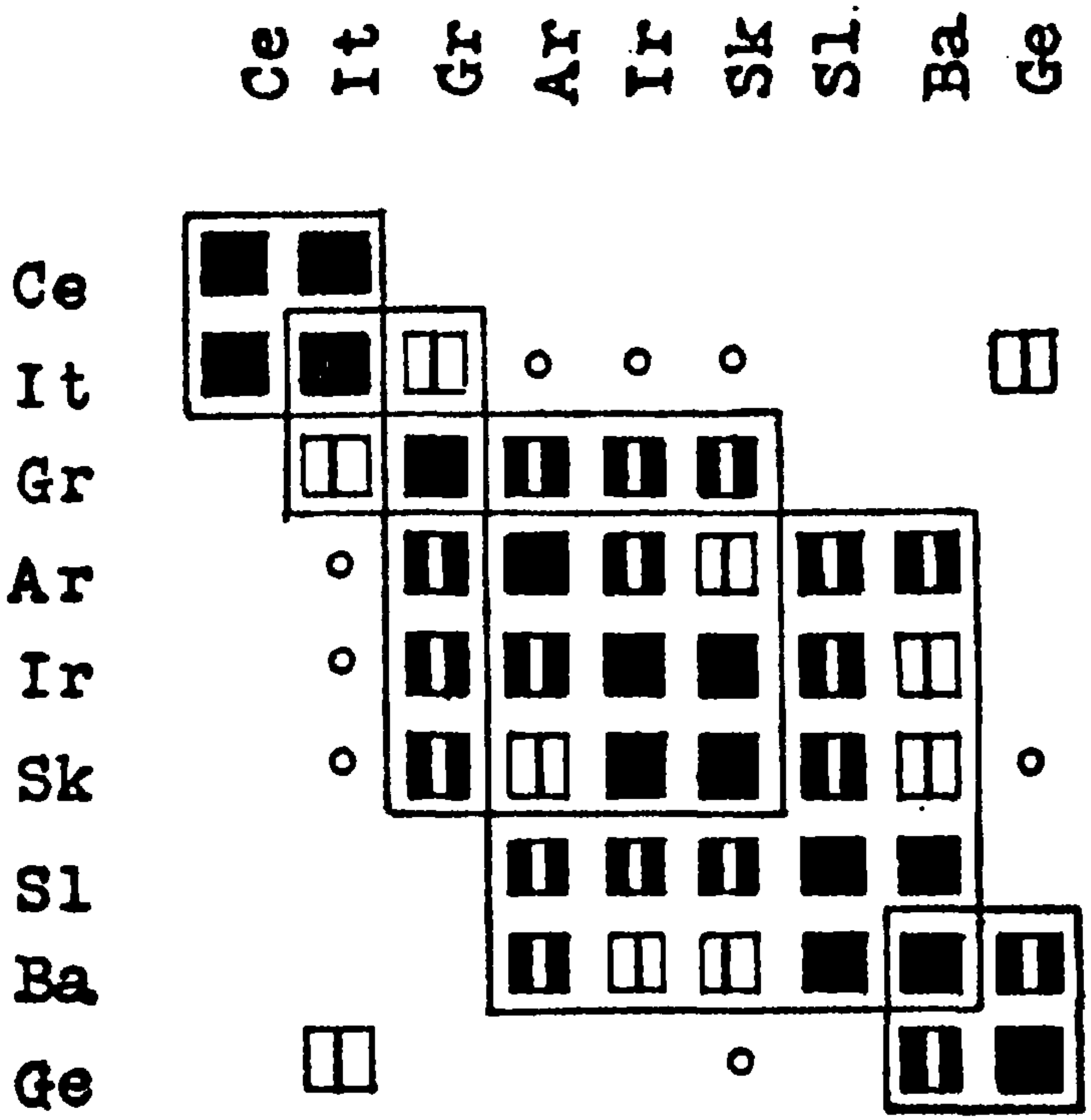
Necessary analyses

- Subgrouping
  - **Similarity-based approaches**
  - Parsimony-based approaches

TABLE V

Kroeber-Chrétien, 74 elements, formula W

	Ce	It	Gr	Ar	Ir	Sk	Sl	Ba	Ge
Ce	1.	.85	.49	.42	.44	.45	.49	.45	.52
It	.85	1.	.57	.39	.30	.32	.41	.42	.57
Gr	.49	.57	1.	.64	.59	.59	.49	.50	.51
Ar	.42	.39	.64	1.	.60	.57	.64	.63	.50
Ir	.44	.30	.59	.60	1.	.87	.62	.55	.46
Sk	.45	.32	.59	.57	.87	1.	.60	.53	.38
Sl	.49	.41	.49	.64	.62	.60	1.	.88	.51
Ba	.45	.42	.50	.63	.55	.53	.88	1.	.64
Ge	.52	.57	.51	.50	.46	.38	.51	.64	1.



	Ce	It	Gr	Ar	Ir	Sk	Sl	Ba	Ge
Ce	■	■							
It	■	■	◫	○	○	○			◫
Gr		◫	■	▮	▮	▮			
Ar		○	▮	■	▮	◫	▮	▮	
Ir		○	▮	▮	■	■	▮	◫	
Sk		○	▮	◫	■	■	▮	◫	○
Sl				▮	▮	▮	■	■	
Ba				▮	◫	◫	■	■	▮
Ge		◫				○		▮	■

KEY

■ 1. to .85 ◫ .57 to .53

▮ .64 to .59 .52 to .41

○ .39 to .30

Fig. 5