Typologies in equilibrium

Gerhard Jäger

Tübingen University

SLE 2018, Tallinn University

WORDS BONES GENES TOOLS
Tracking Linguistic, Cultural, and Biological Trajectories of the Human Past

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

DFG

Distributional universals

Distributional universals

- common practice since Greenberg (1963):
  - collect a sample of languages
  - classify them according to some typological feature
  - ⇒ skewed distribution indicates something interesting going on

- Problem: languages are not independent samples
- skewed distribution may reflect
  - skewed diversification rate across families
  - properties of an ancestral bottleneck
- balanced sampling mitigates the first, but not the second problem

Distributional universals

Maslova (2000):

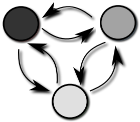
"If the A-distribution for a given typology cannot be assumed to be stationary, a distributional universal cannot be discovered on the basis of purely synchronic statistical data."

"In this case, the only way to discover a distributional universal is to ***estimate transition probabilities*** *and as it were to 'predict' the stationary distribution on the basis of the equations in (1)."*

The phylogenetic comparative method

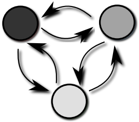
Modeling language change

Markov process

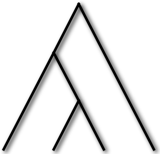


Modeling language change

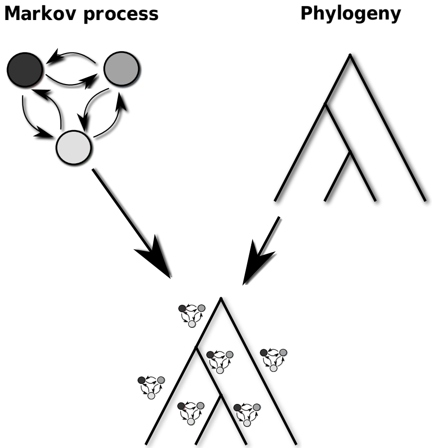
Markov process



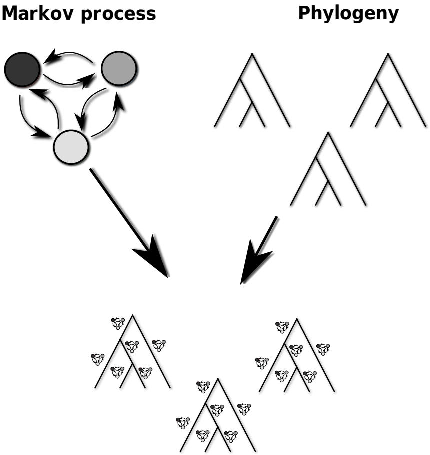
Phylogeny



Modeling language change

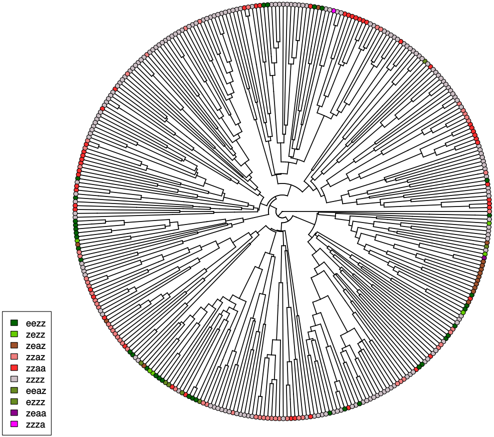


Modeling language change



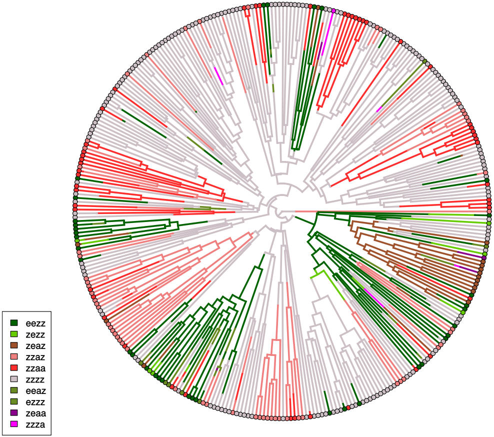
Estimating rates of change

- if phylogeny and states of extant languages are known...



Estimating rates of change

- if phylogeny and states of extant languages are known...
- ... transition rates and ancestral states can be estimated based on Markov model



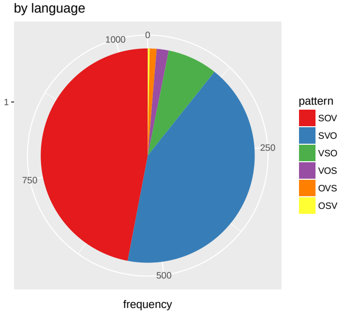
Major word orders

Statistics of major word order distribution

- data: WALS intersected with ASJP
- 1,055 languages, 201 lineages, 71 families with at least 3 languages

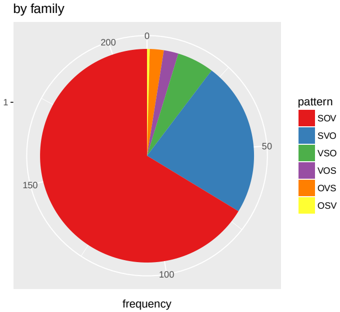
Raw numbers

SOV	SVO	VSO	VOS	OVS	OSV
497	447	78	20	10	3
47.1%	42.4%	7.4%	1.9%	0.9%	0.3%



Weighted by lineages

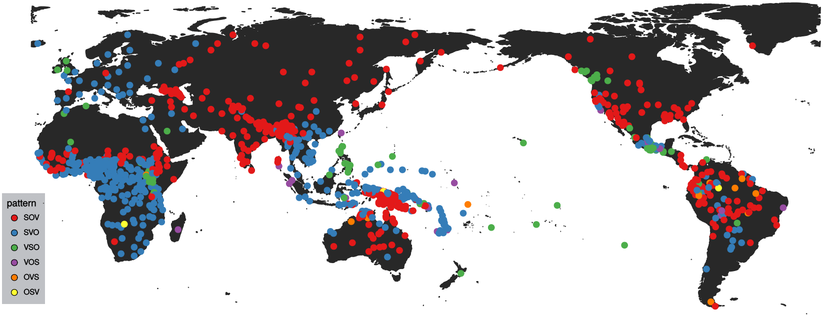
SOV	SVO	VSO	VOS	OVS	OSV
135.1	46.9	10.5	4.0	3.7	0.8
67.2%	23.3%	5.2%	2.0%	1.8%	0.4%



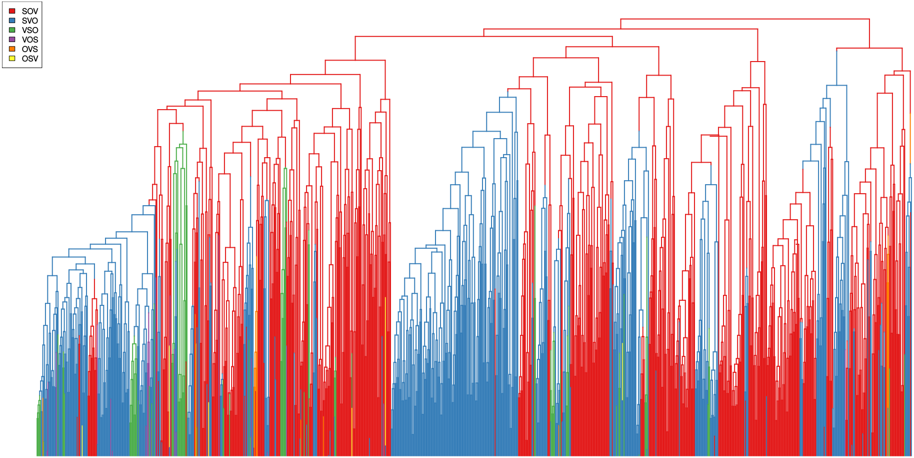
Phylogenetic non-independence

- languages are phylogenetically structured
- if two closely related languages display the same pattern, these are not two independent data points

⇒ we need to control for phylogenetic dependencies

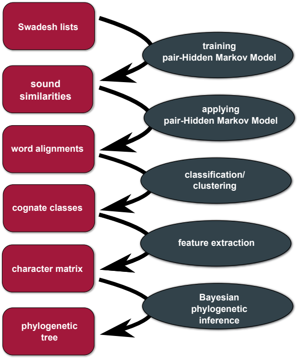


Phylogenetic non-independence

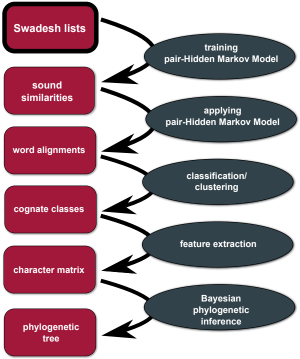


Inferring trees across many families

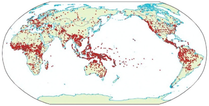
From words to trees



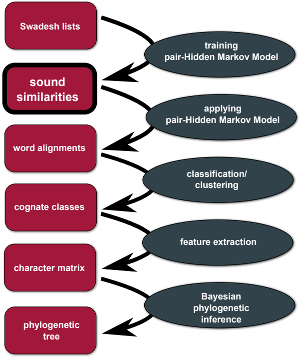
From words to trees

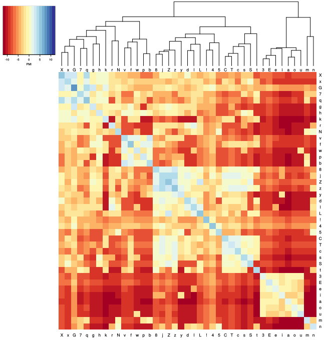


concept	Latin	English
I	ego	Ei
you	tu	yu
we	nos	wi
one	unus	w3n
two	duo	tu
person	persona, homo	pers3n
fish	piskis	fiS
dog	kanis	dag
louse	pedikulus	laus
tree	arbor	tri
leaf	foly∼u*	lif
skin	kutis	skin
blood	saNgw∼is	bl3d
bone	os	bon
horn	kornu	horn
ear	auris	ir
eye	okulus	Ei

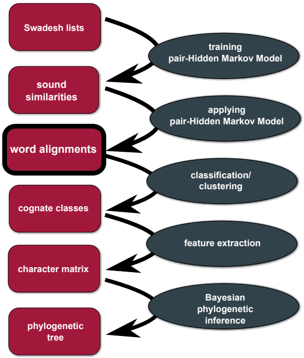


From words to trees



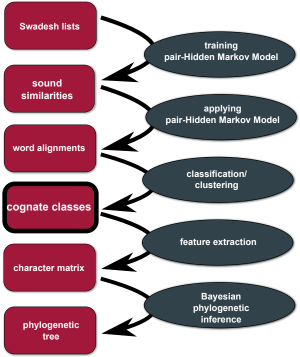


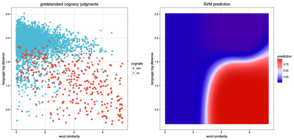
From words to trees



Language	*fish:z*	*tongue:1*	*smoke:1*
Abui-Atangmelang	-af-u		
Abui-Fuimelang	-af-u	tal-i-fi--	
Adang	aab--	tal-E-b---	awai--b-a-n-o-7o-
Blagar-Bakalang	-ab--	--j-e-bur-	--ad--b-a-n-aNka-
Blagar-Bama	aab--	teg-e-bur-	------b-e-n-a-xa-
Blagar-Kulijahi	-ab--	tej-e-bur-	------b-e-n-aNka-
Blagar-Nule	aab--	tej-e-bur-	--ad--b-e-n-aNka-
Blagar-Tuntuli	aab--	tej-e-bur-	a-adgeb-a-n-a-q--
Blagar-Warsalelang	-ab--	tel-e-bur-	a-ad--b-a-n-a-x--
Bunaq			------b-o-t-o-h--
Deing	haf--		------buu-n------
Hamap	7ab--	nar-ø-buN-	------b-a-n-o-7--
Kabola	hab--	tal-e-b---	awal--b-e-n-e-7o-
Kaera-Padangsul	-ab--	talee-b---	a-ad--b-e-naa-x--
Kafoa	-afUi	tal-i-p---	------f-o-n-a----
Kamang	-ap-i	nal---pu--	------p-u-n----a-
Kiraman	-Eb--	nal-i-bar-	--ar--b-a-n-o-kan
Klon	-eb-i	gel-E-b---	--ed-ab-o-n------
Kui	-eb--	tal-i-ber-	--ar--b-o-n-o-k--
Kula	-ap-i	-il-I-p---	------p---n-ekka-
Nedebang	aaf-i	gel-e-fu--	--ar-ab-u-n------
Reta	aab--	nal-e-bul-	a-ad--b-o-n-a----
Sar-Adiabang	haf--	--p-e-fal-	--ar--buu-n------
Sar-Nule	haf--	nal-e-faj-	
Sawila	-ap-i	gal-impuru	------p-u-n-a-ka-
Teiwa-Madar	xaf--	gel-i-vi--	------buu-n------
Wersing	-ap-i	nej-e-bur-	--ad-ap-u-n-a-k--
Wpantar	hap--	nal-e-bu--	------b-unn-a----

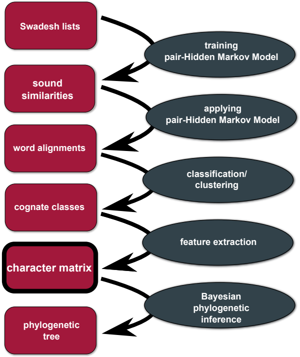
From words to trees





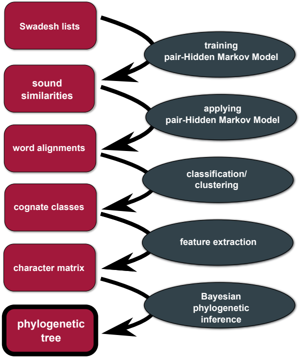
	English	Spanish	Modern Greek	Standard German
I	Ei:A	yo:B	exo:C	iX:D
you	yu:A	ustet:B, tu:C	esi:D	du:E
we	wi:A	nosotros:B	emis:C	vir:A
one	w3n:A	uno:B	enas:C, ena:C	ains:D
two	tu:A	dos:B	8y~o:C, 8io:D	cvai:E
person	pers3n:A	persona:A	an8~ropos:B	mEnS:C
fish	fiS:A	peskado:A, pes:A	psari:B	fiS:A
dog	dag:A	pero:B	sTili:C, sTilos:C	hunt:D
come	k3m:A	veni:B	erx~o:C	kh~om3n:A
sun	s3n:A	sol:B	ily~os:C, iLos:C	zon3:A
star	star:A	estreya:A	asteri:A, astro:A	StErn:A
water	wat3r:A	agw~a:B	nero:C	vas3r:A
stone	ston:A	piedra:B	petra:B	Stain:A
fire	fEir:A	fuego:B	foty~a:C	foia:D
path	pE8:A	senda:B	8romos:C	pf~at:A, vek:D
mountain	maunt3n:A	sero:B, monta5a:A	vuno:C, oros:D	bErk:E
full	ful:A	yeno:B	yematos:C, pliris:D	fol:A
new	nu:A	nuevo:A	neos:A, Tenury~os:B	noi:A
name	nem:A	nombre:A	onoma:A	nam3:A

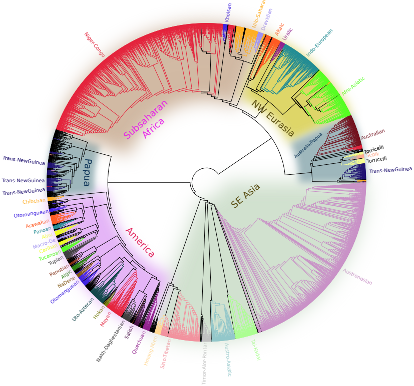
From words to trees




```
TNG.ENGAN.MAIBI                    10000000000000000000000000000000000000000000
TNG.ENGAN.POLE                     00000000000000000000000000000000000010000000
TNG.ENGAN.SAU                      00000000000000000000000000000000000000000000
TNG.ENGAN.YARIBA                   10000000000000000000000000000000000000000000
TNG.FASU.FASU                      00000000000000000000000000000000000010000000
TNG.FASU.NAMUMI                    00000000000000000000000000000000000000010000
TNG.FINISTERRE-HUON.AWARA          00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.BORONG         00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.BURUM          00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.BURUM_MIND     00000000000000000000000000000000001010000000
TNG.FINISTERRE-HUON.DEDUA          00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.HUBE           00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.KATE           00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.KOMBA          00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.KOSORONG       00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.MAPE           00000000000000000000000000000000001000000000
TNG.FINISTERRE-HUON.MAPE_2         00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.MIGABAC        00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.MINDIK         00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.MOMOLILI       00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.NABAK          00000000000000000000000000000000001000000000
TNG.FINISTERRE-HUON.NANKINA        00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.NEK            00000000000000000000000000000000000001000000
TNG.FINISTERRE-HUON.NUKNA          00000000000000000000000000000000000001000000
TNG.FINISTERRE-HUON.ONO            00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.SELEPET        00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.TIMBE          00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.TOBO           00000000000000000000000000000000000010000000
TNG.FINISTERRE-HUON.WANTOAT        00000000000000000000000000000000000110000000
TNG.FINISTERRE-HUON.YOPNO          00000000000000000000000000000000000001000000
TNG.GOILALAN.AFOA                  00000000000000000000000000000000000110000000
TNG.GOILALAN.KUNIMAIPA             00000000000000000000000000000000000010000000
TNG.GOILALAN.MAFULU                00000000000000000000000000000000000010000000
```

From words to trees





Estimating word-order transition patterns

Workflow

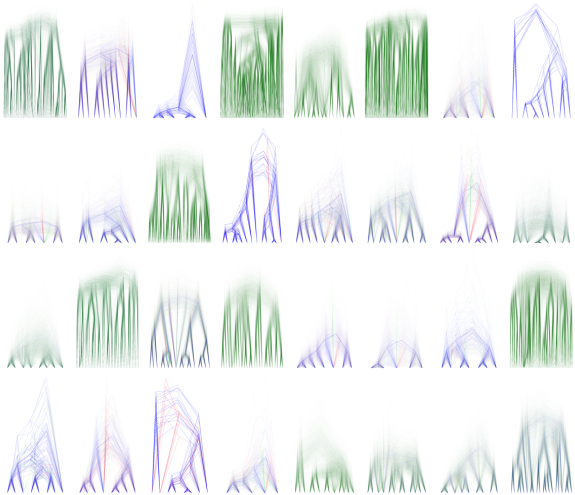
(data from all 77 families with $\geq$ 3 languages in data base; 924 languages in total)

- estimate posterior tree distributions with MrBayes for each family, using Glottolog as constraint tree
- estimate transition rates
- estimate stationary distribution of major word order categories
- apply *stochastic character mapping* (SIMMAP; Bollback 2006)
- estimate expected number of mutations for each transition type

Estimating posterior tree distributions

- using characters extracted from ASJP data (Jäger 2018)
- Glottolog as constraint tree
- Γ-distributed rates
- ascertainment bias correction
- relaxed molecular clock (IGR)
- uniform tree prior
- stop rule: 0.01, samplefreq=1000
- if convergence later than after 1,000,000 steps, sample 1,000 trees from posterior

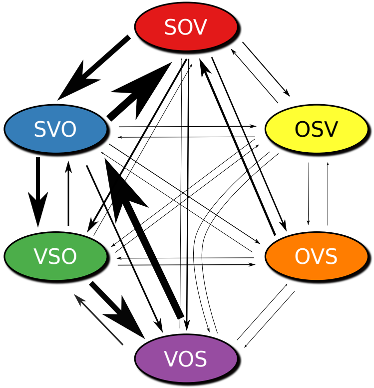
Phylogenetic tree sample



Estimating transition rates

- totally unrestricted model, all 30 transition rates are estimed independently
- implementation using RevBayes (Höhna et al., 2016)

expected strength of flow



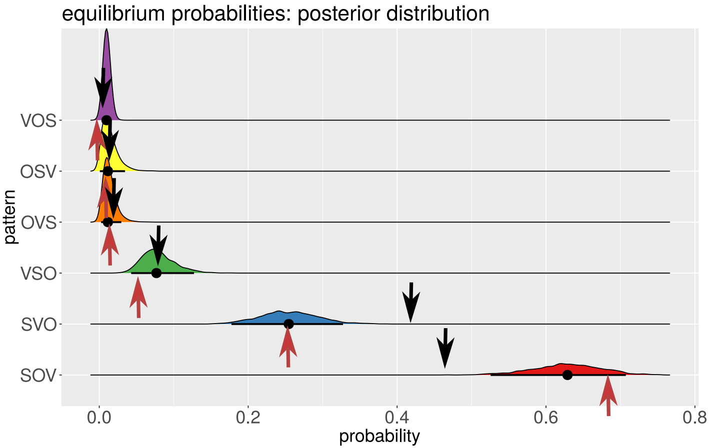
Reconstruction history with SIMMAP

- estimated frequency of mutations within the 77 families under consideration (posterior mean and 95% HPD, 100 simulations

	SOV		SVO		VSO		VOS		OVS		OSV	
SOV	–		51.5	[19; 82]	10.2	[1; 19]	7.5	[0; 29]	5.8	[0; 14]	4.2	[0; 13]
SVO	83.8	[31; 131]	–		22.3	[2; 42]	10.4	[0; 30]	2.8	[0; 8]	3.9	[0; 12]
VSO	1.4	[0; 5]	8.3	[0; 24]	–		29.0	[5; 45]	3.0	[0; 9]	1.1	[0; 5]
VOS	4.3	[0; 15]	141.9	[115; 188]	30.9	[17; 47]	–		2.1	[0; 9]	1.0	[0; 3]
OVS	11.1	[0; 28]	0.8	[0; 4]	1.8	[0; 8]	0.4	[0; 3]	–		0.8	[0; 5]
OSV	4.2	[0; 15]	0.4	[0; 3]	1.9	[0; 11]	1.1	[0; 7]	1.1	[0; 9]	–	

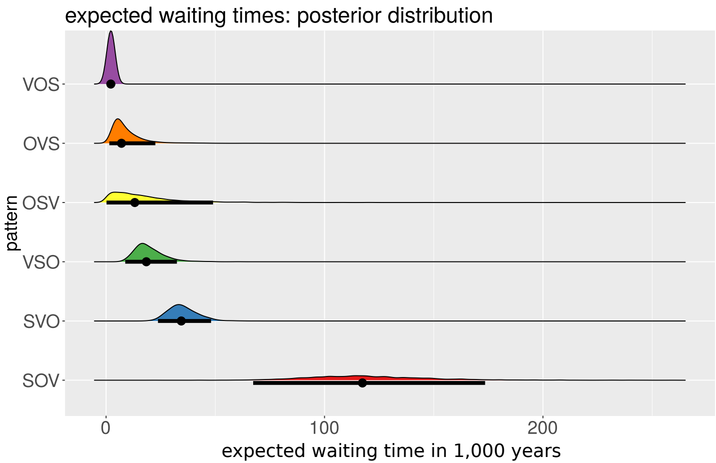
Posterior distributions

Empirical vs. estimated distribution



Posterior distributions

Waiting times



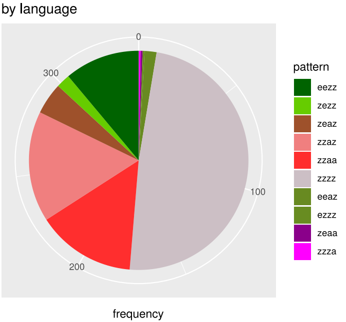
Differential case marking

Statistics of differential case marking distribution

- data: Autotyp intersected with ASJP
- 343 languages, 115 lineages, 29 families with at least 3 languages

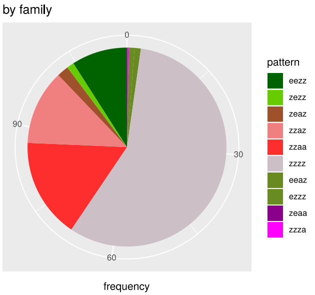
Raw numbers

zzzz	zzaa	zzaz	eezz	zeaz	eeaz	ezzz	zezz	zeaa	zzza
167	50	56	38	16	6	1	7	1	1
48.7%	14.6%	16.3%	11.1%	4.7%	1.7%	0.3%	2.0%	0.3%	0.3%

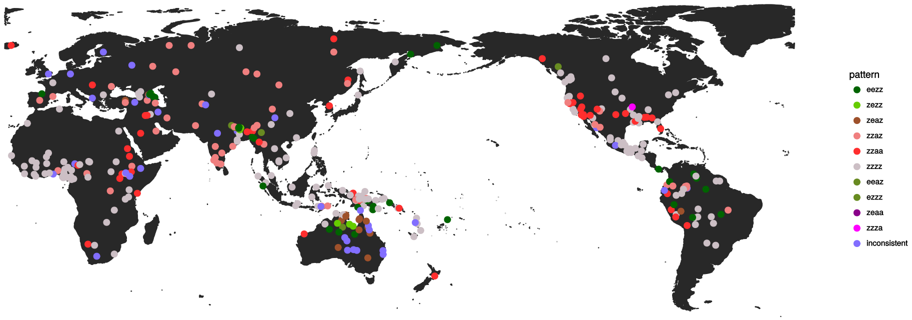


Weighted by lineages

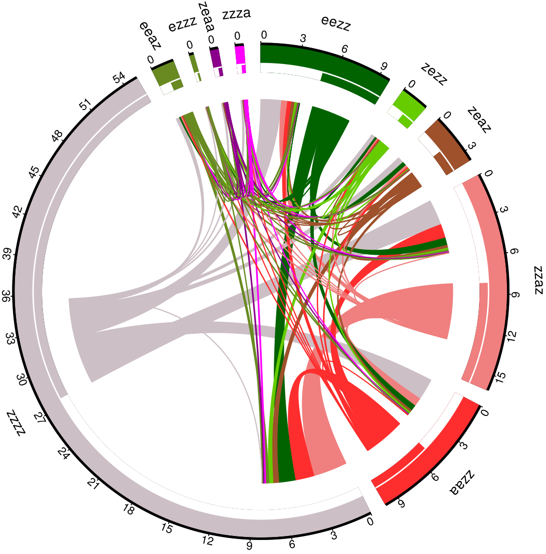
zzzz	zzaa	zzaz	eezz	zeaz	eeaz	ezzz	zezz	zeaa	zzza
65.8	18.6	14.0	10.4	2.2	1.2	1.0	1.3	0.1	0.3
57.2%	16.2%	12.2%	9.1%	1.9%	1.0%	0.9%	1.2%	0.0%	0.3



Geographic distribution



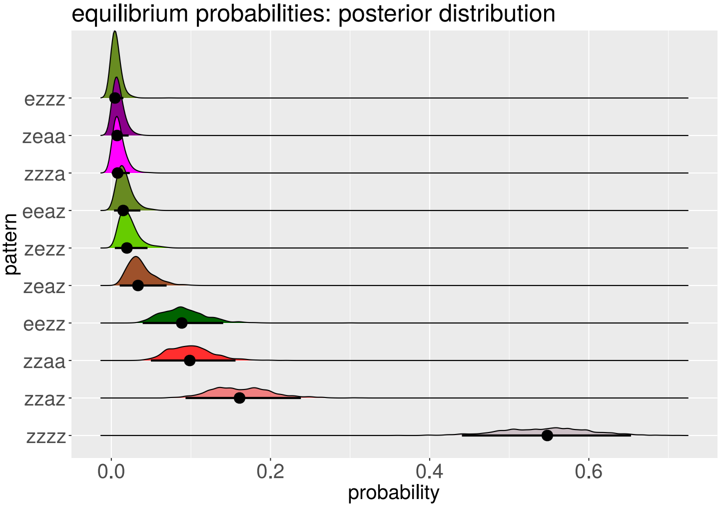
Estimated rates



- green: nominative/accusative
- red: ergative/absolutive
- grey: no case marking
- brown/purple: mixture of accusative and ergative patterns

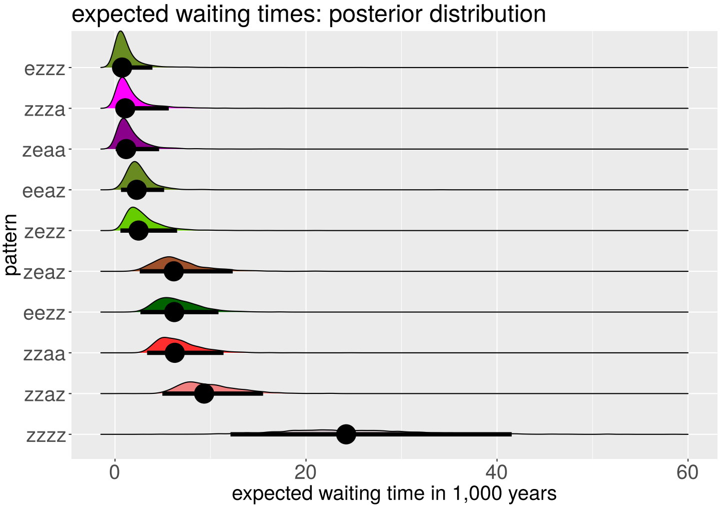
Posterior distributions

Estimated distribution



Posterior distributions

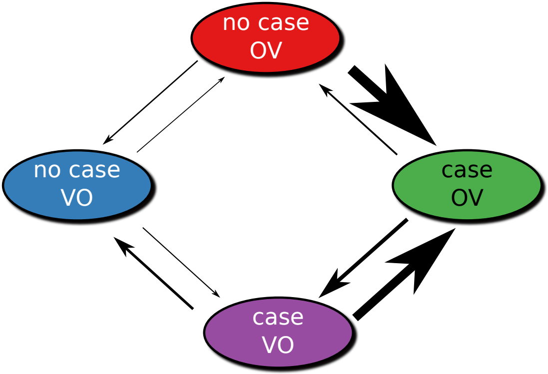
Waiting times

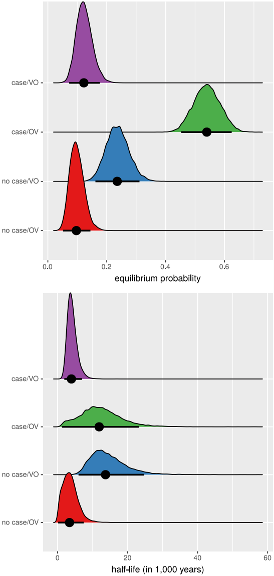


Posterior distribution

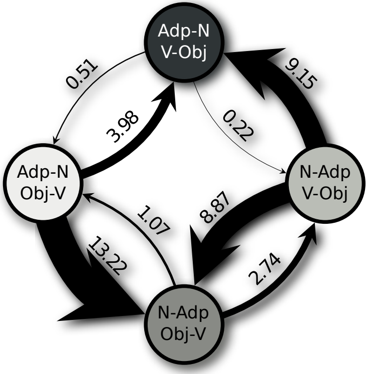
- with 99.7% confidence, accusative languages are more likely than ergative languages
- with 59.6% confidence, consistent accusative is more likely than consistent ergative
- with 100% confidence, differential object marking is more likely than anti-DOM
- with 95.2% confidence, differential subject marking is more likely than anti-DSM

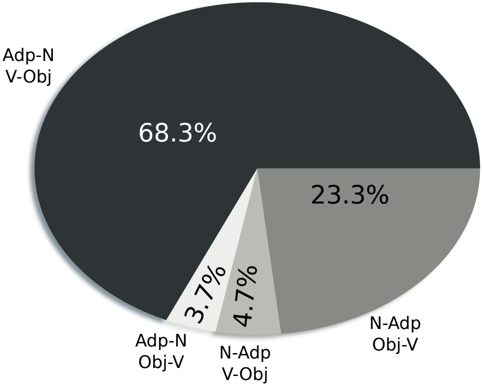
Further variables





Further variables





Conclusion

- Maslova's program can be carried out with phylogenetic comparative method
- future research:
  - equilibrium distributions generally resemble family-wise weighted distributions — bug or feature?
  - hierarchical models instead of one Markov process for all lineages?
  - more data!!! (but there are never enough of them)
  - better methods for feature selection? (Bayes factor test rejected RJ-MCMC)

Jonathan P. Bollback. SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics*, 7(1):88, 2006.

Joseph Greenberg. Some universals of grammar with special reference to the order of meaningful elements. In *Universals of Language*, pages 73–113. MIT Press, Cambridge, MA, 1963.

Sebastian Höhna, Michael J. Landis, Tracy A. Heath, Bastien Boussau, Nicolas Lartillot, Brian R. Moore, John P. Huelsenbeck, and Frederik Ronquist. Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic biology*, 65(4):726–736, 2016.

Gerhard Jäger. Global-scale phylogenetic linguistic inference from lexical resources. arXiv:1802.06079, 2018.

Elena Maslova. A dynamic approach to the verification of distributional universals. *Linguistic Typology*, 4(3):307–333, 2000.

Søren Wichmann, Eric W. Holman, and Cecil H. Brown. The ASJP database (version 18). http://asjp.clld.org/, 2018.