Statistical estimation of diachronic stability from synchronic data

Gerhard Jäger

Tübingen University

Cape Town, July 6, 2018

Introduction

From the workshop description

"The workshop starts from the null hypothesis that diachronically stable properties are those that appear as the typologically most frequent ones, and that cross-linguistic rarity correlates with diachronic instability."

Inferring diachronic stability of a feature from its typological frequency is potentially fallacious for three reasons:

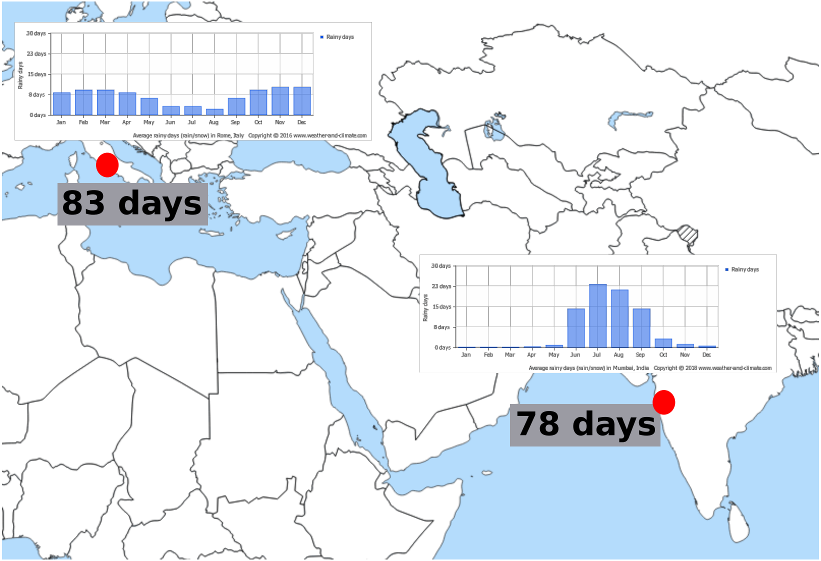
1. Processes of different rates may lead to identical equilibrium distributions.
2. Individual languages are not independent random samples, since genetically related languages are likely to have similar typological profiles.
3. The stability of a feature value might depend on the value of other, correlated features.

Frequency, stability, and Markov chains

Rainy days per year in Mumbay and Rome

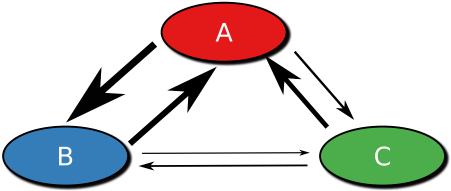
source: `https://weather-and-climate.com`

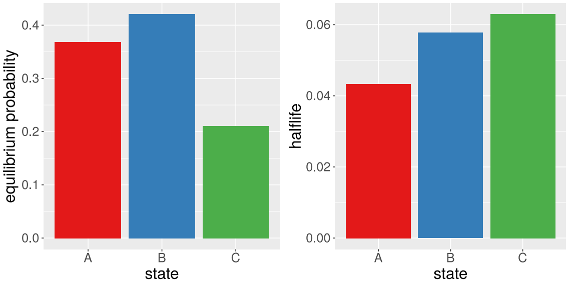
Rainy days per year in Mumbay and Rome



source: `https://weather-and-climate.com`

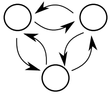
Markov chains



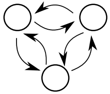


Phylogenetic structure

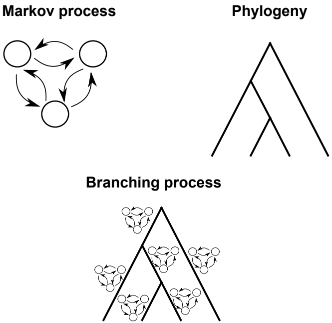
Markov process



Phylogenetic structure



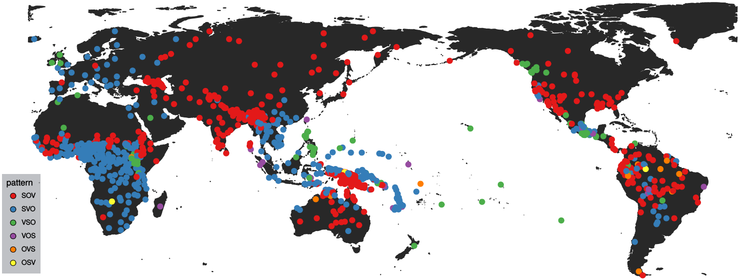
Phylogenetic structure



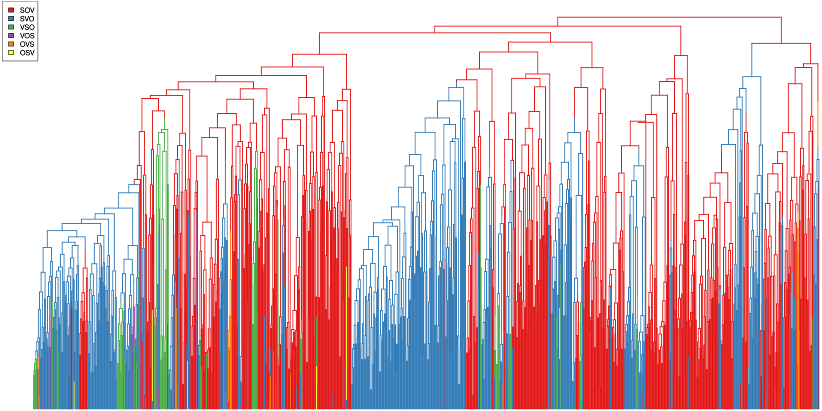
Phylogenetic non-independence

- 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



Phylogenetic non-independence

Maslova (2000):

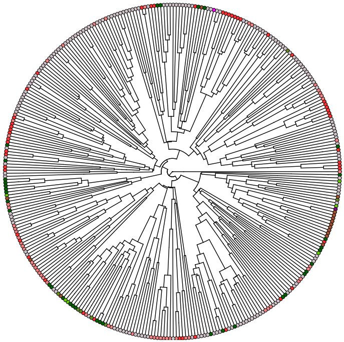
> *"If the A-distribution for a given typology cannot be assumed to be stationary, a distributional universal cannot be discovered on the basis of purely synchronic statistical data."*

> *"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

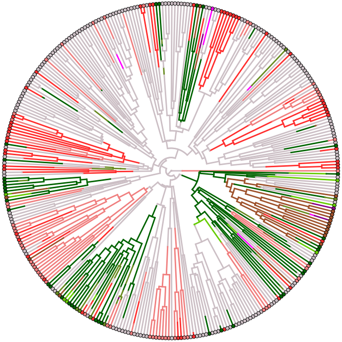
Estimating rates of change

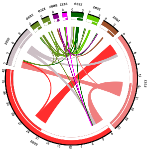
- if phylogeny and states of extant languages are known...



Estimating rates of change

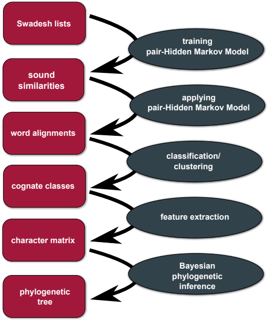
- if phylogeny and states of extant languages are known...
- ... transition rates and ancestral states can be estimated based on Markov model



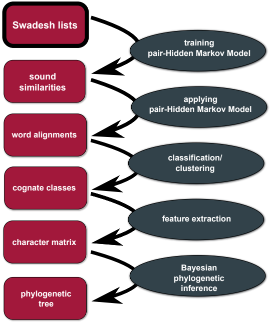


Inferring a world tree of languages

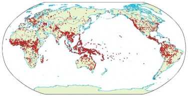
From words to trees



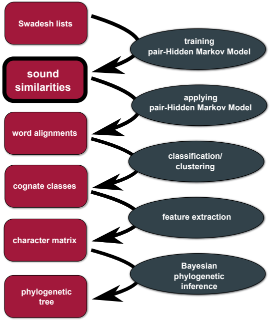
From words to trees

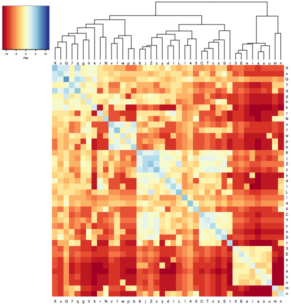


concept	Latin	English
I	ego	Ei
you	tu	yu
we	nos	wi
one	unus	w3n
two	duo	tu
person	persona, homo	pers3n
fish	piskis	fiS
dog	kanis	dag
louse	pedikulus	laus
tree	arbor	tri
leaf	foly~u*	lif
skin	kutis	skin
blood	saNgw~is	bl3d
bone	os	bon
horn	kornu	horn
ear	auris	ir
eye	okulus	Ei

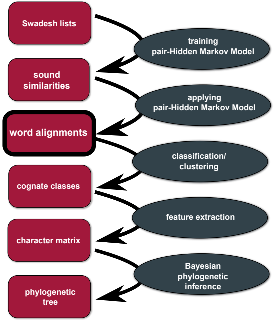


From words to trees



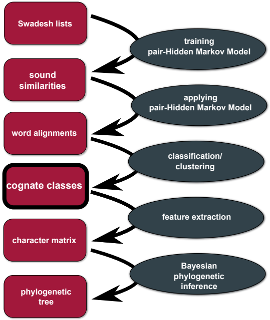


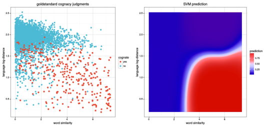
From words to trees



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----

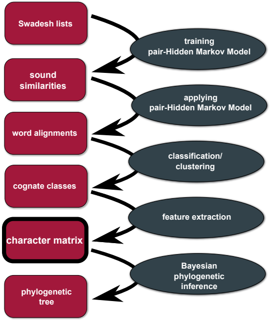
From words to trees





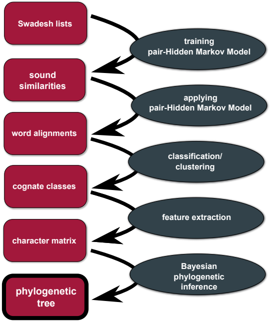
	English	Spanish	Modern Greek	Standard German
I	Ei:A	yo:B	exo:C	iX:D
you	yu:A	ustet:B, tu:C	esi:D	du:E
we	wi:A	nosotros:B	emis:C	vir:A
one	w3n:A	uno:B	enas:C, ena:C	ains:D
two	tu:A	dos:B	8y~o:C, 8io:D	cvai:E
person	pers3n:A	persona:A	an8~ropos:B	mEnS:C
fish	fiS:A	peskado:A, pes:A	psari:B	fiS:A
dog	dag:A	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

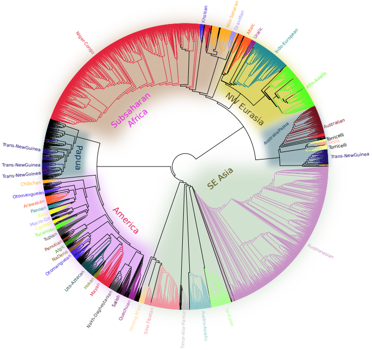
From words to trees



TNG.ENGAN.MAIBI	10000000000000000000000000000000000000000000
TNG.ENGAN.POLE	00000000000000000000000000000000000010000000
TNG.ENGAN.SAU	00000000000000000000000000000000000000000000
TNG.ENGAN.YARIBA	10000000000000000000000000000000000000000000
TNG.FASU.FASU	00000000000000000000000000000000000010000000
TNG.FASU.NAMUMI	00000000000000000000000000000000000000001000
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	00000000000000000000000000000000000100000000
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	00000000000000000000000000000000000100000000
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

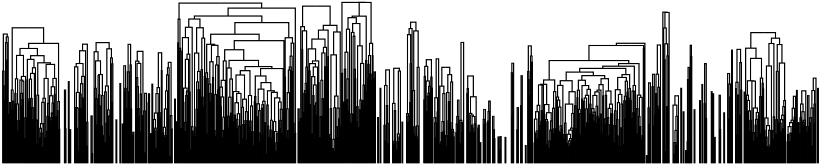
From words to trees





From tree to forest

- branch lengths within Glottolog families estimated from lexical data
- calibration: Proto-Austronesian $\sim$ 5,000 years
- branches above family level effictively set to infinity



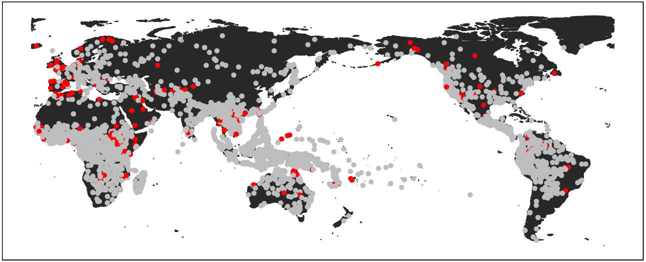
Case study 1: Rare consonants

Synchronic statistics

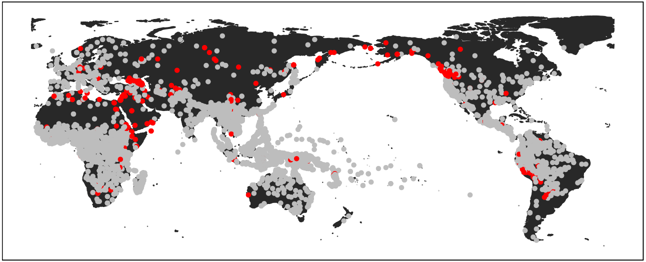
- data: ASJP word lists (word lists from ca. 6,000 living languages and dialects; Wichmann et al. 2016)
- variables:
  - *voiceless and voiced dental fricative* (transcribed as `8`)
  - *voiceless and voiced uvular fricative, voiceless and voiced pharyngeal fricative* (transcribed as `X`)

	8	**X**
raw numbers	334	378
average	5.7%	6.6%
weighted by family	14.6	22.2
average	4.6%	7.0%

Dental fricative



Uvular or pharyngeal fricative



Phylogenetic estimates

	8	**X**
equilibrium probability	5.5%	7.4%
half-life present (kyrs)	1.8	4.6
half-life absent (kyrs)	30.1	58.4

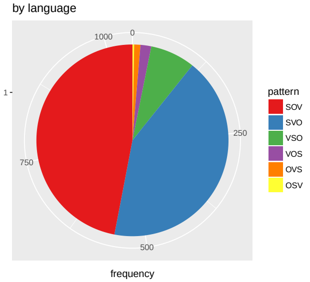
Case study 2: Major word orders

Statistics of major word order distribution

- data: WALS intersected with ASJP
- 1,045 languages, 211 lineages

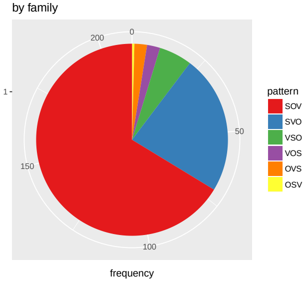
Raw numbers

SOV	SVO	VSO	VOS	OVS	OSV
491	442	79	19	11	3
47.0%	42.3%	7.6%	1.8%	1.1%	0.3%

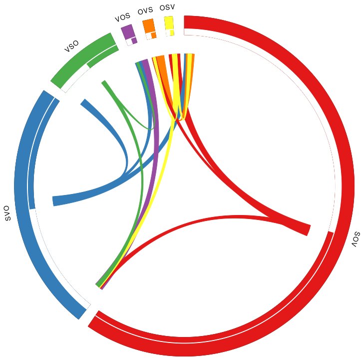


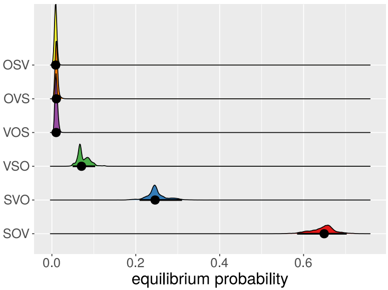
Weighted by lineages

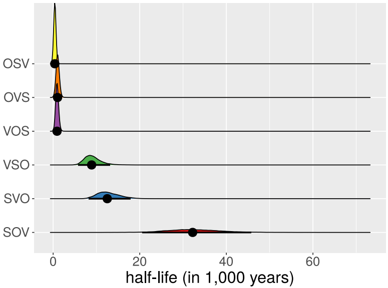
SOV	SVO	VSO	VOS	OVS	OSV
139.1	49.3	11.8	4.7	4.5	0.8
66.3%	23.4%	5.6%	2.2%	2.1%	0.4%



Phylogenetically estimated Markov process







Case study 3: Word order and case

Statistics

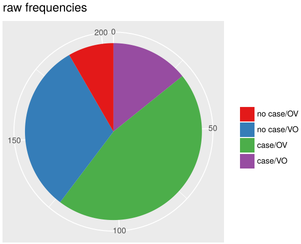
- data: WALS intersected with ASJP
- 204 languages, 103 lineages

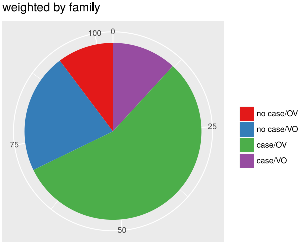
Raw numbers

no case/OV	no case/VO	case/OV	case/VO
17	64	94	29
8.3%	31.4%	46.1%	14.2%

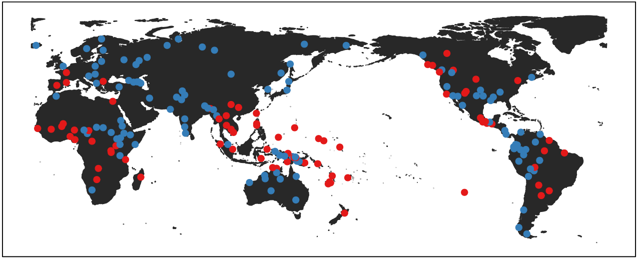
Weighted by lineages

no case/OV	no case/VO	case/OV	case/VO
10.6	22.6	57.7	12.2
10.3%	21.9%	56.0%	11.8%

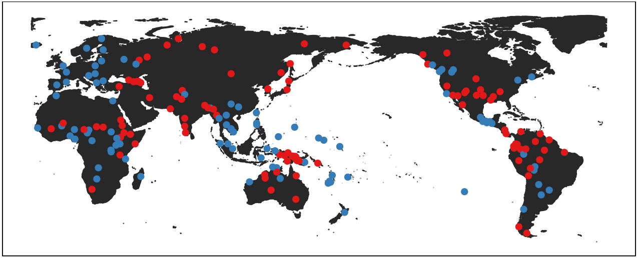




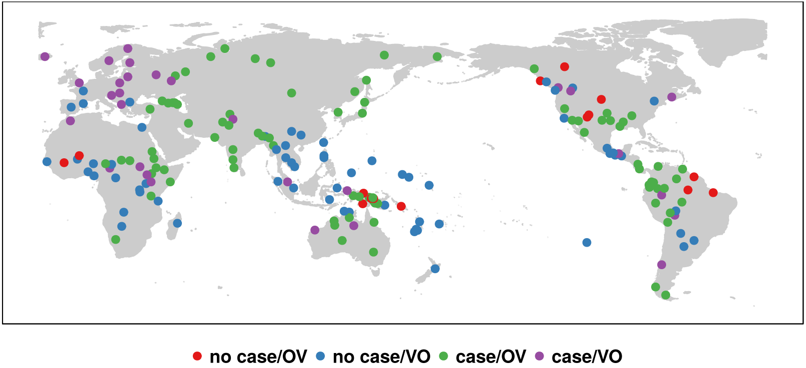
Case



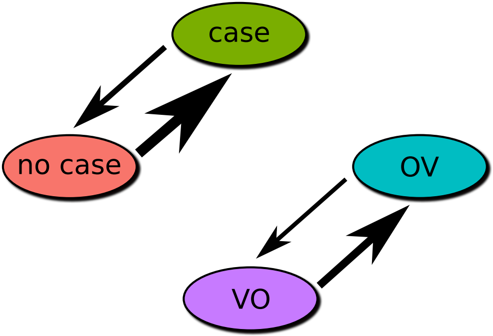
VO vs. OV

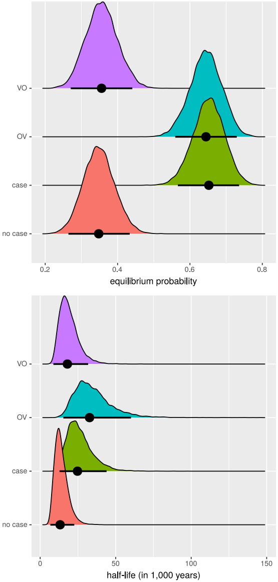


case + word order

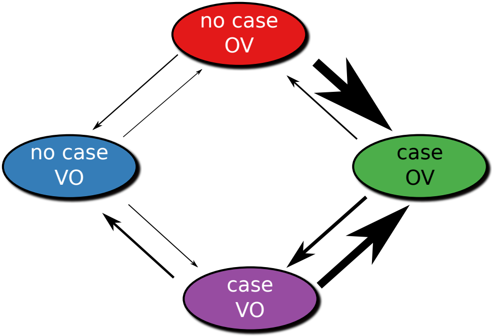


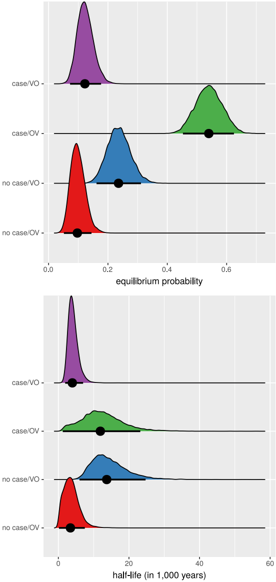
Phylogenetically estimated Markov process: features individually





Phylogenetically estimated Markov process: dependent features





Conclusion

Conclusion

- connection between cross-linguistic frequency and diachronic stability is loose at best
- to assess diachronic stability, we need information on
  - phylogenetic structure
  - branch lengths
- stability of feature values may depend on other features → potentially complex causal network between typological variables, waiting to be explored
- todo:
  - comparison to related but different approaches, such as Bickel's Family Bias Method (Bickel, 2013) or Greenhill et al.'s (2017) approach
  - factoring in language contact
  - non-homogeneous Markov chains?

Balthasar Bickel. Distributional biases in language families. In *Language Typology and Historical Contingency: In honor of Johanna Nichols*, pages 415–444. John Benjamins, Amsterdam, 2013.

Simon J Greenhill, Chieh-Hsi Wu, Xia Hua, Michael Dunn, Stephen C Levinson, and Russell D Gray. Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences*, 114(42):E8822–E8829, 2017.

Gerhard Jäger. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change*, 3(2):245–291, 2013.

Gerhard Jäger. Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences*, 112(41):12752–12757, 2015. doi: 10.1073/pnas.1500331112.

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

Gerhard Jäger and Søren Wichmann. Inferring the world tree of languages from word lists. In S. G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Feher, and T. Verhoef, editors, *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11)*, 2016. Available online: http://evolang.org/neworleans/papers/147.html.

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 17). http://asjp.clld.org/, 2016.