A Bayesian test of the lineage-specificity of word-order correlations

Gerhard Jäger, Gwendolyn Berger, Isabella Boga, Thora Daneyko & Luana Vaduva

Tübingen University

DIP Colloquium Amsterdam

January 25, 2018

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

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

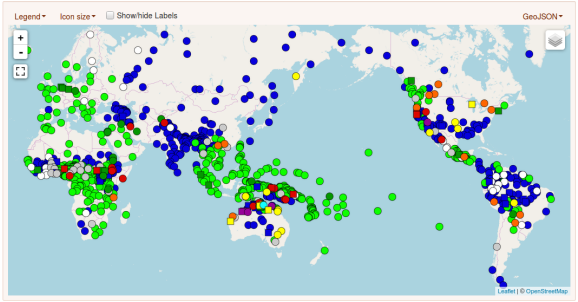
DFG

erc
European Research Council
Established by the European Commission

Introduction

Word order correlations

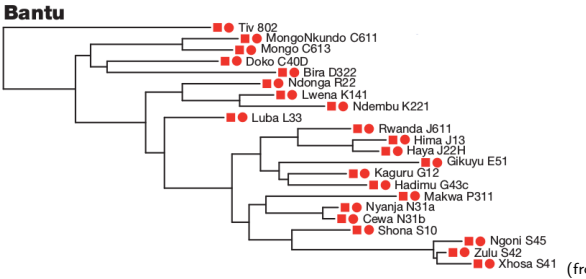
- Greenberg, Keenan, Lehmann etc.: general tendency for languages to be either consistently head-initial or consistently head-final
- alternative account (Dryer, Hawkins): phrases are consistently left- or consistently right-branching
- can be formalized as collection of implicative universals, such as
  With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional. (Greenberg's Universal 4)
- both generativist and functional/historical explanations in the literature



Phylogenetic non-independence

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

$\Rightarrow$ we need to control for phylogenetic dependencies



Dunn et al., 2011)

Phylogenetic non-independence

Maslova (2000):

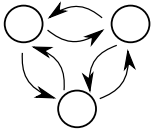
"If the A-distribution for a given typology cannot be assumed to be stationary, a distributional universal cannot be discovered on the basis of purely synchronic statistical data."

"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

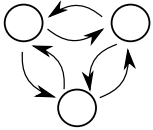
Modeling language change

Markov process



Modeling language change

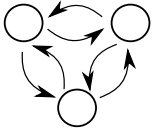
Markov process



Phylogeny

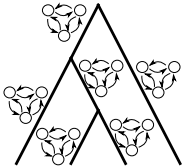
Modeling language change

Markov process



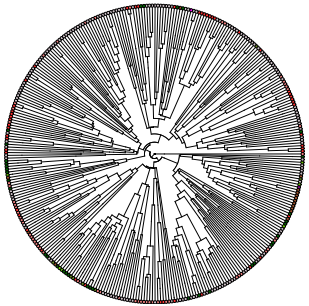
Phylogeny

Branching process



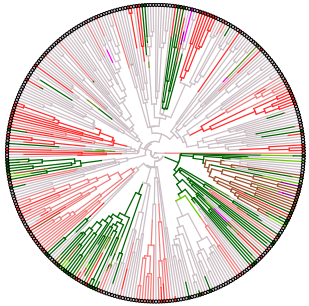
Estimating rates of change

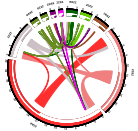
- if phylogeny and states of extant languages are known...



Estimating rates of change

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



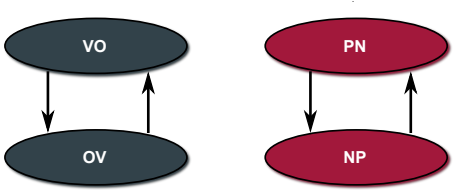


Correlation between features

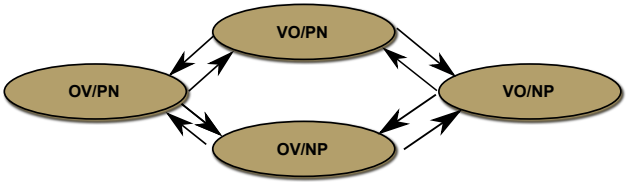
Pagel and Meade (2006)

- construct two types of Markov processes:
  - **independent:** the two features evolve according to independend Markov processes
  - **dependent:** rates of change in one feature depends on state of the other feature
- fit both models to the data
- apply statistical model comparison

Independent model



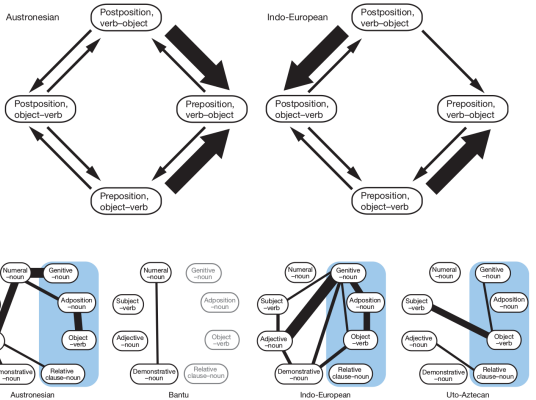
Dependent model



Dunn et al. (2011)

Dunn et al. (2011)

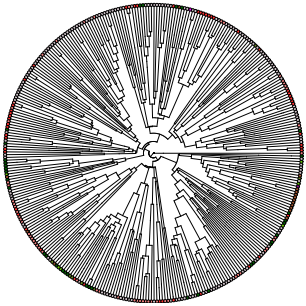
- all 28 pairs of 8 word-order features considered
- 4 language families: Austronesian, Bantu, Indo-European, and Uto-Aztecan
- main finding: wildly different results between families
- conclusion:
  word-order correlations are lineage-specific



Universal and lineage-specific models

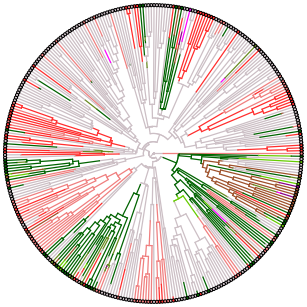
Estimating rates of change

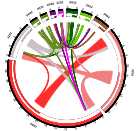
- if phylogeny and states of extant languages are known...



Estimating rates of change

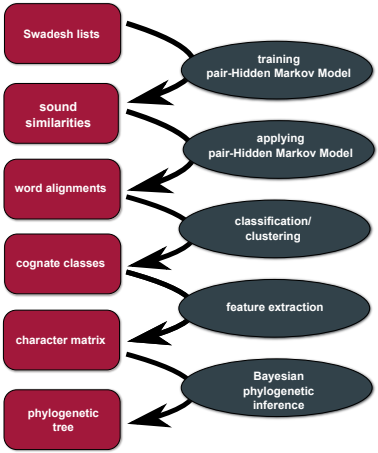
- if phylogeny and states of extant languages are known...
- ... transition rates and ancestral states can be estimated based on Markov model



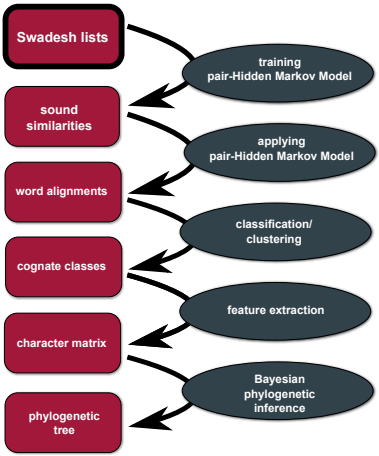


Inferring a world tree of languages

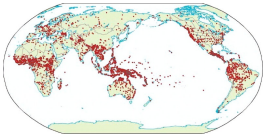
From words to trees



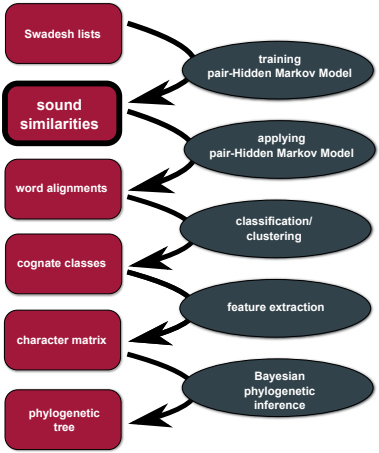
From words to trees

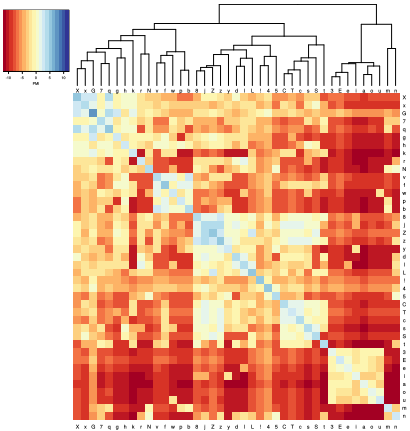


concept	Latin	English
I	ego	Ei
you	tu	yu
we	nos	wi
one	unus	w3n
two	duo	tu
person	persona, homo	pers3n
fish	piskis	fiS
dog	kanis	dag
louse	pedikulus	laus
tree	arbor	tri
leaf	foly∼u*	lif
skin	kutis	skin
blood	saNgw∼is	bl3d
bone	os	bon
horn	kornu	horn
ear	auris	ir
eye	okulus	Ei

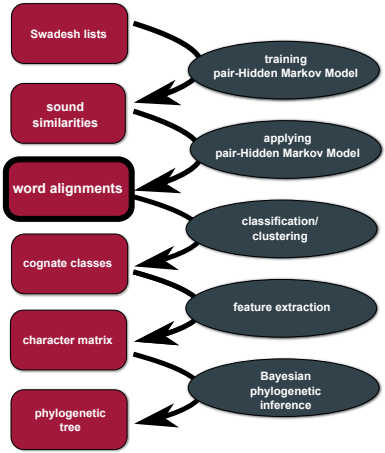


From words to trees



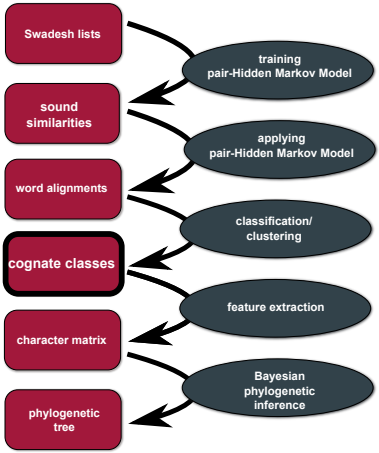


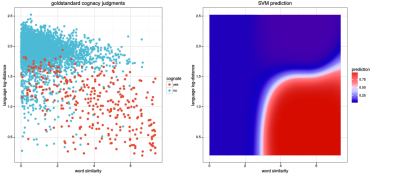
From words to trees



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----

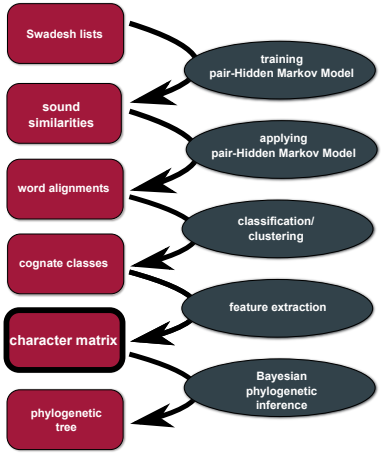
From words to trees





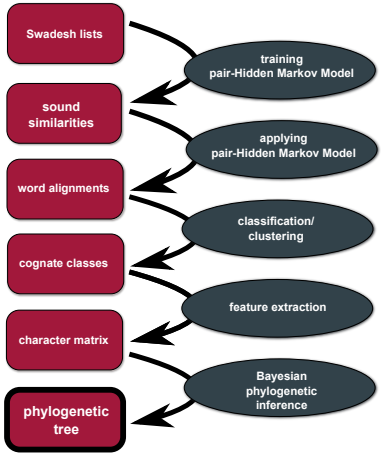
	English	Spanish	Modern Greek	Standard German
I	Ei:A	yo:B	exo:C	iX:D
you	yu:A	ustet:B, tu:C	esi:D	du:E
we	wi:A	nosotros:B	emis:C	vir:A
one	wXn:A	uno:B	enas:C, ena:C	ains:D
two	tu:A	dos:B	8y~o:C, 8io:D	cvai:E
person	pers3n:A	persona:A	an8~ropos:B	mEnS:C
fish	fiS:A	peskado:A, pes:A	psari:B	fiS:A
dog	dag:A	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	pE3:A	senda:B	8romos:C	pf~at:A, vek:D
mountain	maunt3n:A	sero:B, montaña: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

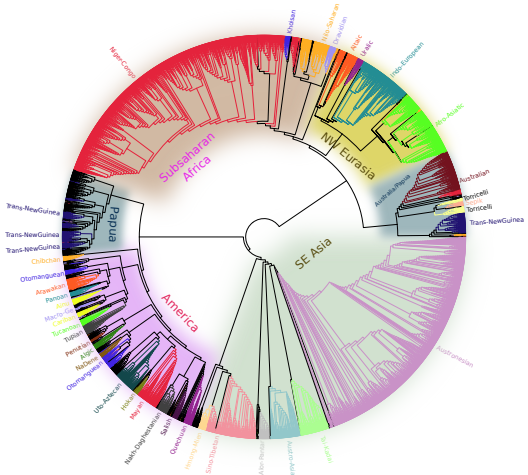
From words to trees



TNG.ENGAN.MAIBI	1000000000000000000000000000000000000000000
TNG.ENGAN.POLE	0000000000000000000000000000000000000010000
TNG.ENGAN.SAU	0000000000000000000000000000000000000000000
TNG.ENGAN.YARIBA	1000000000000000000000000000000000000000000
TNG.FASU.FASU	0000000000000000000000000000000000000100000
TNG.FASU.NAMUMI	0000000000000000000000000000000000000001000
TNG.FINISTERRE-HUON.AWARA	0000000000000000000000000000000000000100000
TNG.FINISTERRE-HUON.BORONG	0000000000000000000000000000000000000010000
TNG.FINISTERRE-HUON.BURUM	0000000000000000000000000000000000000010000
TNG.FINISTERRE-HUON.BURUM_MIND	0000000000000000000000000000000000001010000
TNG.FINISTERRE-HUON.DEDUA	0000000000000000000000000000000000000100000
TNG.FINISTERRE-HUON.HUBE	0000000000000000000000000000000000000010000
TNG.FINISTERRE-HUON.KATE	0000000000000000000000000000000000000010000
TNG.FINISTERRE-HUON.KOMBA	0000000000000000000000000000000000000010000
TNG.FINISTERRE-HUON.KOSORONG	0000000000000000000000000000000000000010000
TNG.FINISTERRE-HUON.MAPE	0000000000000000000000000000000000000100000
TNG.FINISTERRE-HUON.MAPE_2	0000000000000000000000000000000000000010000
TNG.FINISTERRE-HUON.MIGABAC	0000000000000000000000000000000000000010000
TNG.FINISTERRE-HUON.MINDIK	0000000000000000000000000000000000000010000
TNG.FINISTERRE-HUON.MOMOLILI	0000000000000000000000000000000000000100000
TNG.FINISTERRE-HUON.NABAK	0000000000000000000000000000000000001000000
TNG.FINISTERRE-HUON.NANKINA	0000000000000000000000000000000000000100000
TNG.FINISTERRE-HUON.NEK	0000000000000000000000000000000000000010000
TNG.FINISTERRE-HUON.NUKNA	0000000000000000000000000000000000000100000
TNG.FINISTERRE-HUON.ONO	0000000000000000000000000000000000000010000
TNG.FINISTERRE-HUON.SELEPET	0000000000000000000000000000000000000100000
TNG.FINISTERRE-HUON.TIMBE	0000000000000000000000000000000000000010000
TNG.FINISTERRE-HUON.TOBO	0000000000000000000000000000000000000100000
TNG.FINISTERRE-HUON.WANTOAT	0000000000000000000000000000000000000110000
TNG.FINISTERRE-HUON.YOPNO	0000000000000000000000000000000000000100000
TNG.GOILALAN.AFOA	0000000000000000000000000000000000000110000
TNG.GOILALAN.KUNIMAIPA	0000000000000000000000000000000000000100000
TNG.GOILALAN.MAFULU	0000000000000000000000000000000000000010000

From words to trees





Universal and lineage-specific models

This study

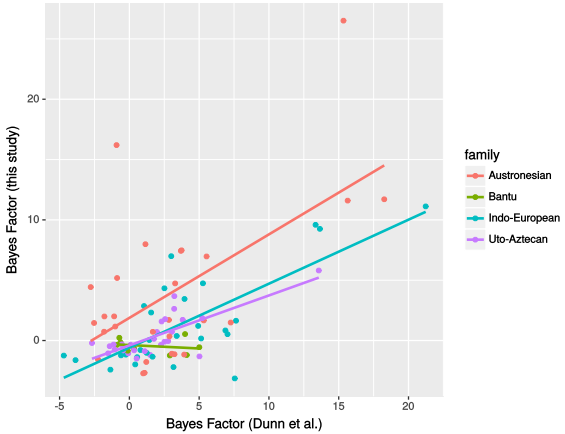
Experiments

1. replication of Dunn et al. (2011) with different data
2. model comparison: universal vs. lineage-specific correlations
3. word-order correlations across a world-tree of languages
4. automatically identifying lineage-specificity

Data

- **word-order data:** WALS
- **phylogeny:**
  - ASJP word lists (Wichmann et al., 2016)
  - feature extraction (automatic cognate detection, *inter alia*) $\rightsquigarrow$ character matrix
  - Maximum-Likelihood phylogenetic inference with Glottolog (Hammarström et al., 2016) tree as backbone
  - advantages over hand-coded Swadesh lists
    - applicable across language familes
    - covers more languages than those for which expert cognate judgments are available
  - 1004 languages in total
  - Austronesian: 123; Bantu: 41; Indo-European: 53; Uto-Aztecan: 13

Replication of Dunn et al.

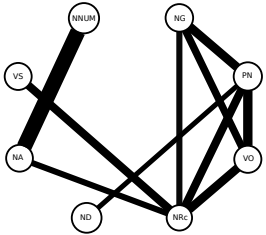


Comparing universal and lineage-specific models

- so far: fitting a separate model for each language family
  - **advantage:** good fit of the lineage-specific data
  - **disadvantage:** many parameters (8 per family for a dependent model)
- statistical model comparison: quantifying to what degree the data support the excess parameters of lineage-specific models
- models to be compared:
  - **universal:** one set of rates (8 parameters), applying to all 4 families
  - **lineage specific:** a separate set of rates for each family
- comparison via **Bayes Factor**
  (implementation with RevBayes; Höhna et al. 2016)

Results

- very strong evidence for universality:
  - noun-adjective ↔ noun-numeral
  - adposition-noun ↔ verb-object
- strong evidence for universality:
  - adposition-noun ↔ verb-object ↔ noun-genitive ↔ noun-relative clause
- strong or very strong evidence for lineage specificity:
  - behavior of noun-adjective and noun-numeral



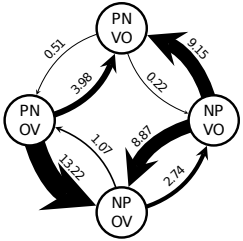
feature pair	Bayes Factor
NA-NNum	16.24
PN-VO	15.22
PN-NG	9.45
VO-NRc	9.21
PN-NRc	8.69
NRc-VS	8.18
NG-VO	7.92
NG-NRc	6.55
NA-NRc	6.49
PN-ND	5.42
ND-NRc	4.32
VO-VS	3.15
PN-VS	1.71
NA-ND	0.54
ND-VO	0.37
NA-VO	-2.07
ND-NG	-3.17
NA-PN	-3.40
NNum-VS	-8.13
NNum-NRc	-8.40
NA-VS	-9.66
NG-VS	-9.84
NA-NG	-10.94
ND-NNum	-12.12
ND-VS	-15.01
PN-NNum	-16.37
NNum-VO	-17.57
NG-NNum	-28.63

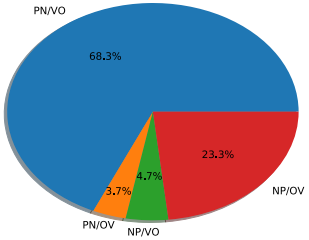
universal

lineage-specific

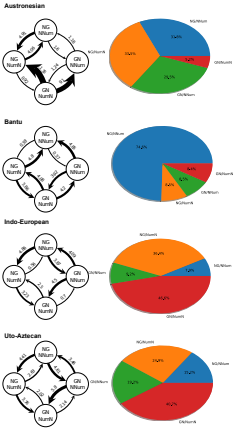
Results

universal (PN/VO)

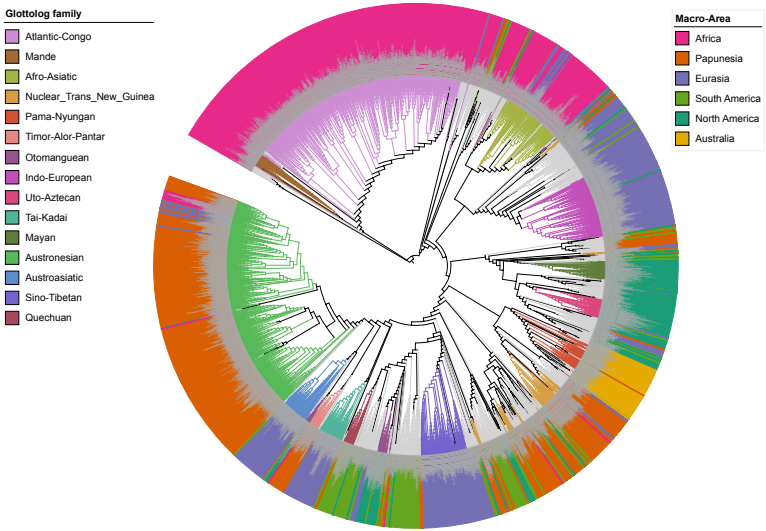




lineage-specific (NG/NNum)

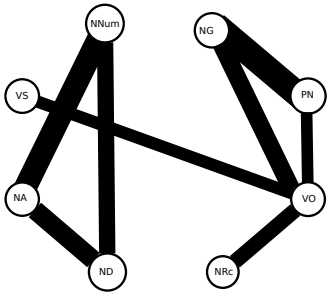


Using the world tree

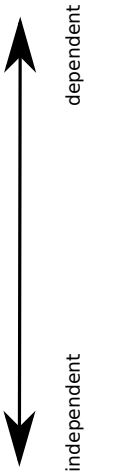


Results

- strong evidence for dependent model for 21 out of 28 feature pairs
- no evidence for independent model
- strongest evidence (BF>100) supports Dryer (1992)

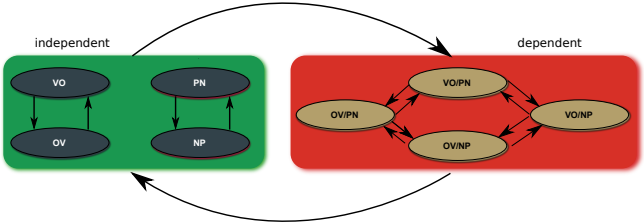


feature pair	Bayes Factor
PN-VO	267.83
PN-NG	220.74
NA-NNum	192.78
NA-ND	163.62
NG-VO	152.64
ND-NNum	140.17
VO-NRc	129.74
VO-VS	105.73
NG-NRc	99.82
PN-NRc	99.28
NA-NRc	84.36
NG-VS	83.68
ND-NRc	71.32
PN-VS	57.51
NNum-VS	37.25
NNum-NRc	36.54
NRc-VS	17.28
ND-NG	16.75
NA-NG	16.55
ND-VO	14.00
NNum-VO	12.43
PN-ND	6.99
NA-VS	5.91
NA-PN	3.84
NA-VO	3.24
ND-VS	1.25
PN-NNum	-0.75
NG-NNum	-2.38

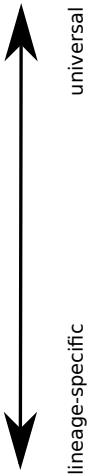


Automatically identifying lineage-specificity

- lineages with different dynamics can be inferred automatically on the world tree
- latest version of *BayesTraits* (v. 3) implements a model ("discrete covarion model") where languages can be either in a dependent or an independent state
- statistical model comparison between universal and lineage-dependent model (in this sense)

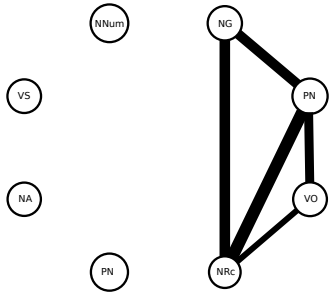


feature pair	Bayes Factor
PN-NRc	0.42
NG-NRc	-0.90
PN-NG	-1.37
PN-VO	-2.29
VO-NRc	-4.86
NA-ND	-11.66
NA-NRc	-21.42
ND-NNum	-22.86
ND-NRc	-23.16
NG-VO	-25.20
PN-VS	-25.70
ND-VS	-28.63
NG-VS	-29.05
VO-VS	-29.74
PN-ND	-30.35
ND-VO	-30.90
NA-NNum	-31.42
ND-NG	-37.75
NA-VS	-40.18
NRc-VS	-44.06
NA-PN	-44.25
NNum-VS	-45.30
NA-VO	-49.34
NNum-NRc	-53.38
PN-NNum	-55.88
NA-NG	-58.86
NNum-VO	-64.76
NG-NNum	-66.61



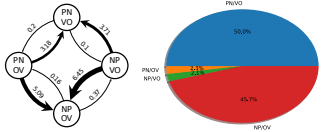
Automatically identifying lineage-specificity

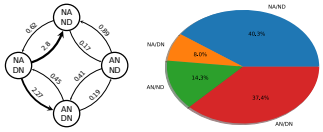
- no evidence for truely universal dependent model
- equivocal evidence for 5 feature pairs
- define a cluster for which there was strong evidence for universality in experiment 2

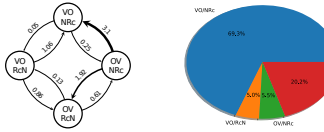


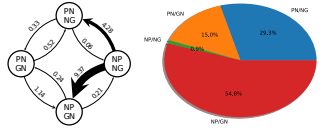
feature pair	Bayes Factor
PN-NRc	0.42
NG-NRc	-0.90
PN-NG	-1.37
PN-VO	-2.29
VO-NRc	-4.86
NA-ND	-11.66
NA-NRc	-21.42
ND-NNum	-22.86
ND-NRc	-23.16
NG-VO	-25.20
PN-VS	-25.70
ND-VS	-28.63
NG-VS	-29.05
VO-VS	-29.74
PN-ND	-30.35
ND-VO	-30.90
NA-NNum	-31.42
ND-NG	-37.75
NA-VS	-40.18
NRc-VS	-44.06
NA-PN	-44.25
NNum-VS	-45.30
NA-VO	-49.34
NNum-NRc	-53.38
PN-NNum	-55.88
NA-NG	-58.86
NNum-VO	-64.76
NG-NNum	-66.61

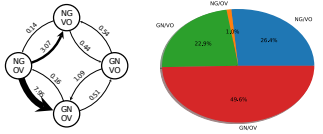
What the dependencies look like

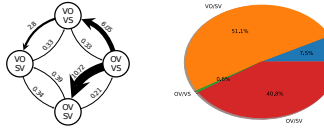


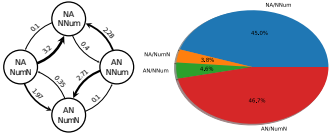


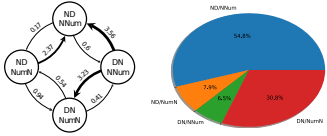


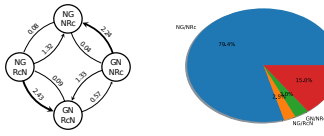












Conclusion

Conclusion

- empirical
  - *universal* vs. *lineage-specific* is not an absolute distinction, but a matter of degree
  - some "classical" word-order correlation fall very close to the universal end
- methodological
  - important to fit statistical model across language-families

Our co-authors

Thora Daneyko

Luana Vaduva

Isabella Boga

Gwendolyn Berger

Matthew S. Dryer. The Greenbergian word order correlations. *Language*, 68(1):81–138, 1992.
Michael Dunn, Simon J. Greenhill, Stephen Levinson, and Russell D. Gray. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82, 2011.
Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. *Glottolog 2.7*. Max Planck Institute for the Science of Human History, Jena, 2016. Available online at http://glottolog.org, Accessed on 2017-01-29.
Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie. The World Atlas of Language Structures online. Max Planck Digital Library, Munich, 2008. http://wals.info/.
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.
Elena Maslova. A dynamic approach to the verification of distributional universals. *Linguistic Typology*, 4(3):307–333, 2000.
Mark Pagel and Andrew Meade. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *The American Naturalist*, 167(6):808–825, 2006.
Søren Wichmann, Eric W. Holman, and Cecil H. Brown. The ASJP database (version 17). http://asjp.clld.org/, 2016.