A case study in computer-aided typology

Gerhard Jäger

Tübingen University

Symposium *Linguistics Quo Vadis*

MPI Nijmegen, October 2, 2017

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

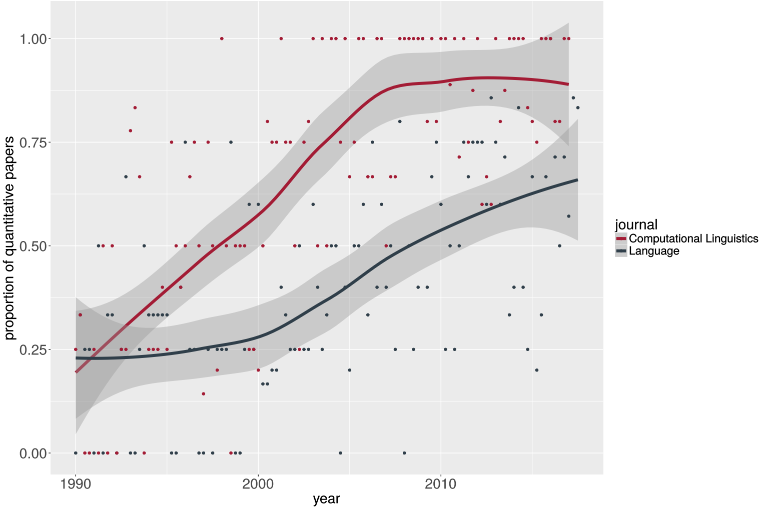
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

DFG

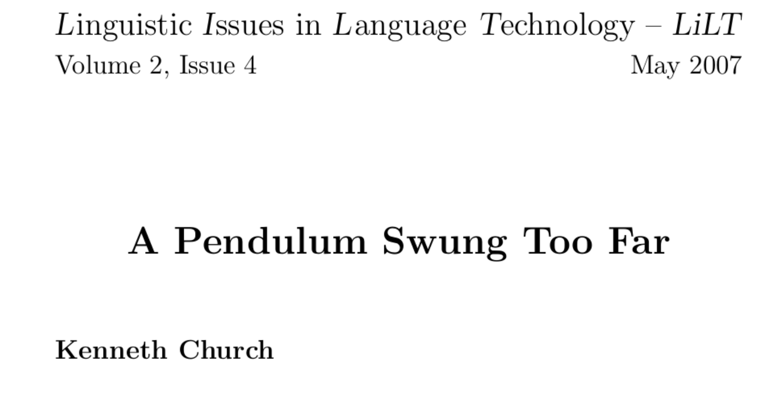
erc
European Research Council
Established by the European Commission

Linguistics Quo Vadis

The ascent of quantitative methods



The ascent of quantitative methods



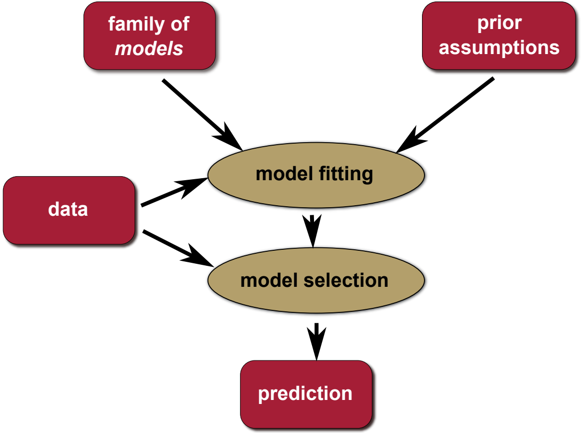
Linguistic Issues in Language Technology – LiLT

Volume 2, Issue 4 May 2007

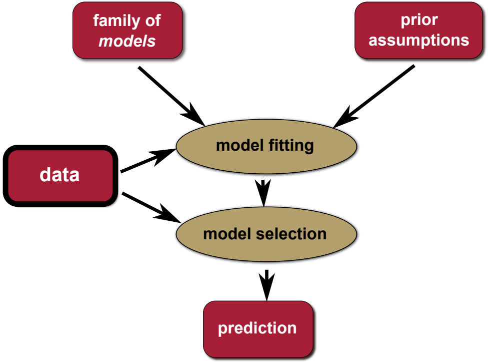
A Pendulum Swung Too Far

Kenneth Church

Statistical modeling of linguistic dynamics

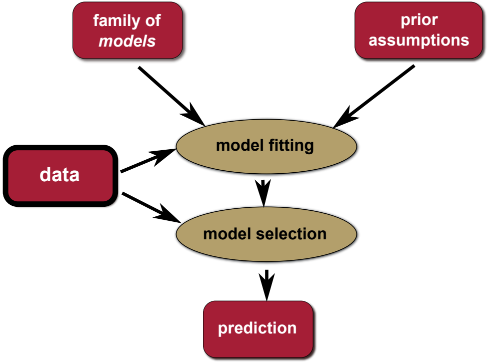


Statistical modeling of linguistic dynamics



small data

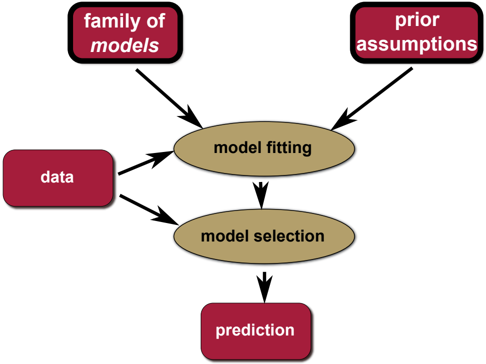
Statistical modeling of linguistic dynamics



Data

- comparative
- multi-modal
- high-quality

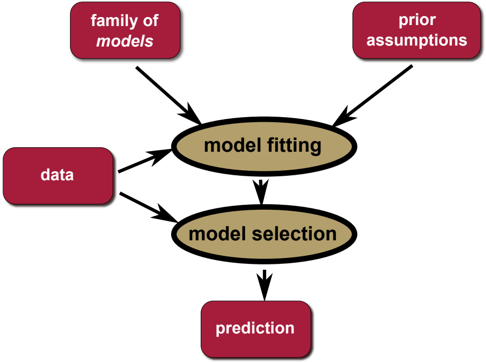
Statistical modeling of linguistic dynamics



Models and priors

- based in **linguistic theory**
- dynamic

Statistical modeling of linguistic dynamics



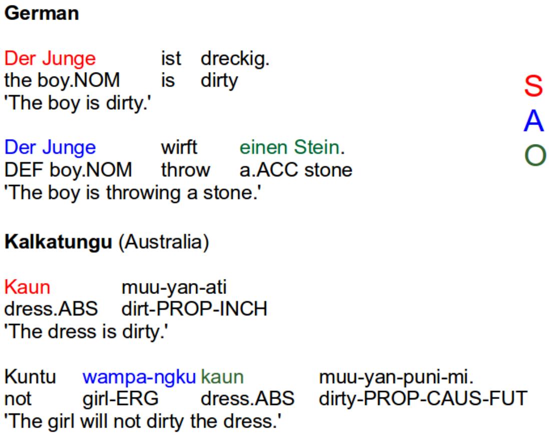
Inference methods

- (approximate) Bayesian computation
- causal inference
- multi-agent simulations
- ...

Case alignment systems

Universal syntactic-semantic primitives

- three universal core roles
  - **S:** intransitive subject
  - **A:** transitive subject
  - **O:** transitive object



German

Der Junge	ist	dreckig.
the boy.NOM	is	dirty

'The boy is dirty.'

Der Junge	wirft	einen Stein.
DEF boy.NOM	throw	a.ACC stone

'The boy is throwing a stone.'

Kalkatungu (Australia)

Kaun	muu-yan-ati
dress.ABS	dirt-PROP-INCH

'The dress is dirty.'

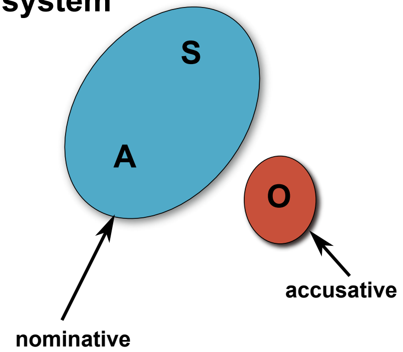
Kuntu	wampa-ngku	kaun	muu-yan-puni-mi.
not	girl-ERG	dress.ABS	dirty-PROP-CAUS-FUT

'The girl will not dirty the dress.'

S
A
O

Alignment systems

Accusative system



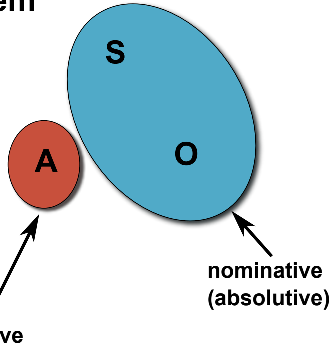
Latin

Puer puellam vidit.
boy.NOM girl.ACC saw *'The boy saw the girl.'*

Puer venit.
boy.NOM came *'The boy came.'*

Alignment systems

Ergative system



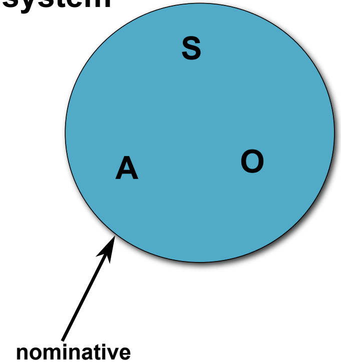
Dyirbal

ŋuma yabu-ŋgu bura-n.
father mother.ERG see-NONFUT
'The mother saw the father.'

ŋuma banaga-nu.
boy.NOM came *'The boy came.'*

Alignment systems

Neutral system



Mandarin

rén lái le.
person come CRS
'The person has come.'

zhāngsān mà lǐsì le ma.
Zhangsan scold Lisi CRS Q
'Did Zhangsan scold Lisi?'

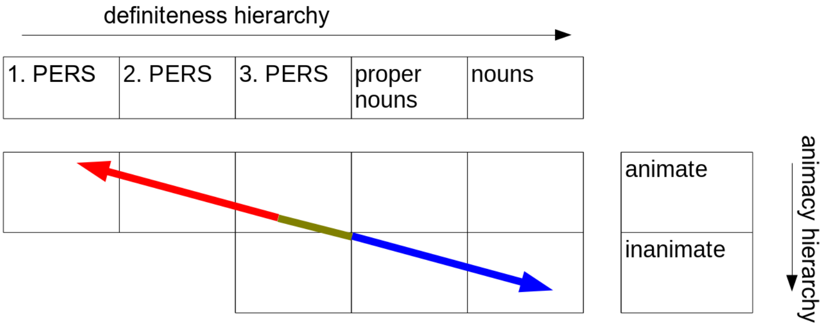
Differential case marking

- many languages have mixed systems
- e.g., some NPs have accusative and some have neutral paradigm, such as Hebrew

(1) Ha-seret her?a ?et-ha-milxama
the-movie showed acc-the-war
'The movie showed the war.'

(2) Ha-seret her?a (*?et-)milxama
the-movie showed (*acc-)war
'The movie showed a war'
(from Aissen, 2003)

Differential case marking



definiteness hierarchy

1. PERS

2. PERS

3. PERS

proper nouns

nouns

animate

inanimate

animacy hierarchy

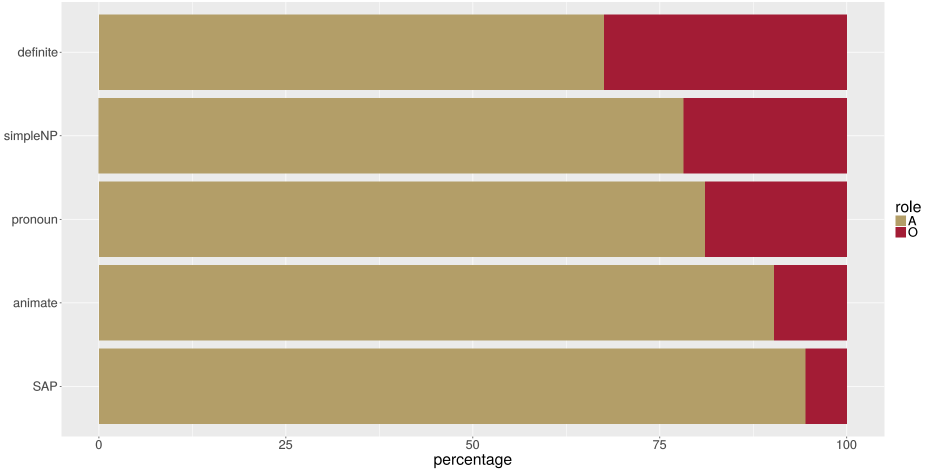
accusative

neutral or tripartite

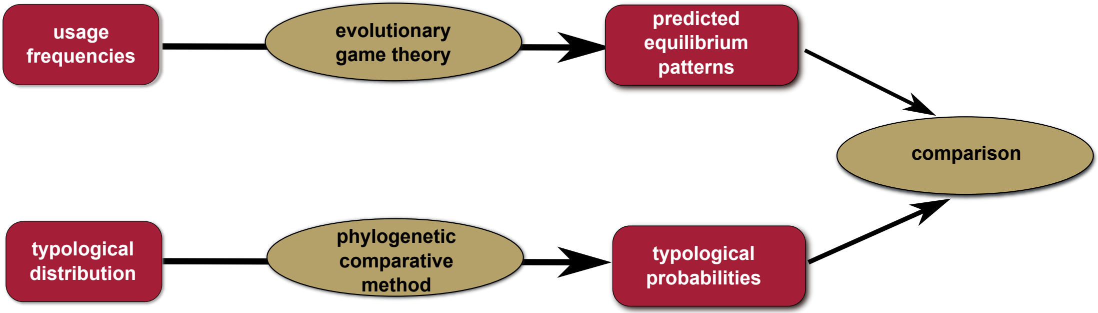
ergative

Functional explanation?

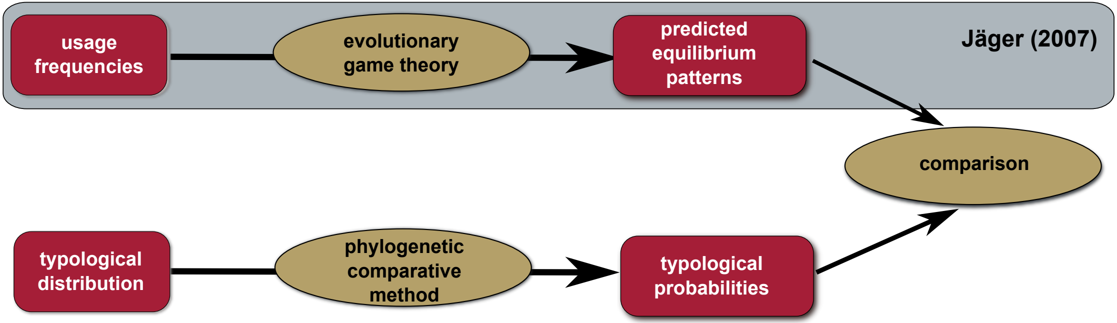
probability P(syntactic role|prominence of NP)



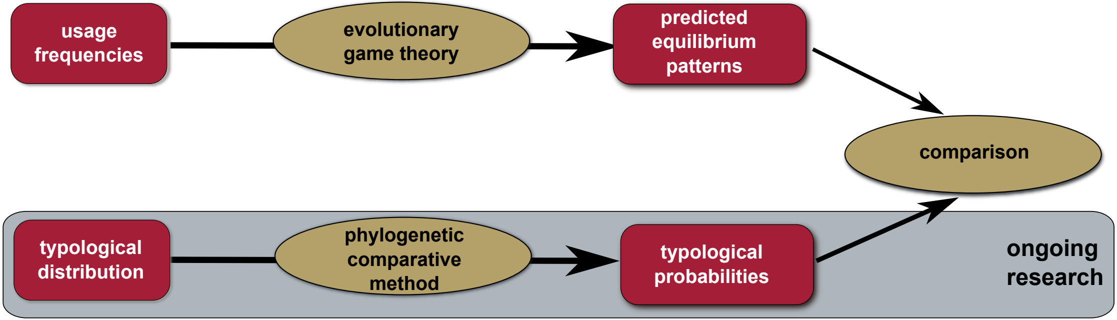
The analysis


usage frequencies
evolutionary game theory
predicted equilibrium patterns
comparison
typological distribution
phylogenetic comparative method
typological probabilities

The analysis

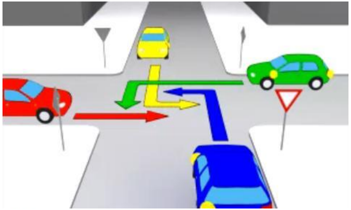


The analysis



Game-theoretic modeling

Game Theory



Rationalistic game theory

- strategic interaction between rational agents
- utility ≈ motivation

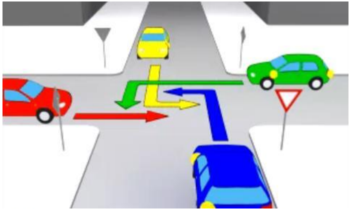
Game Theory

Rationalistic game theory

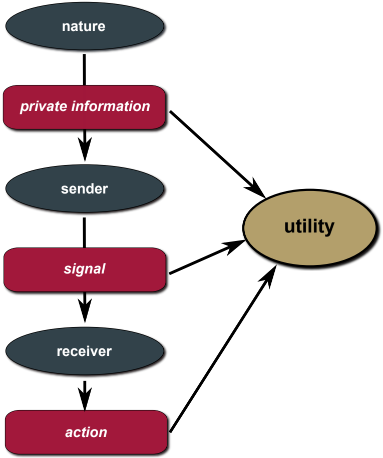
- strategic interaction between rational agents
- utility ≈ motivation

Evolutionary game theory

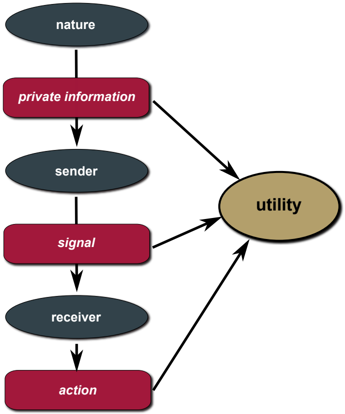
- frequency-dependent Darwinian selection
- utility ≈ fitness



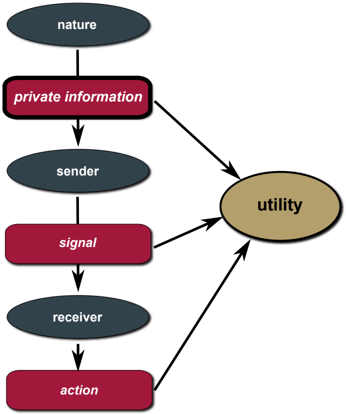
Signaling games



The game of case

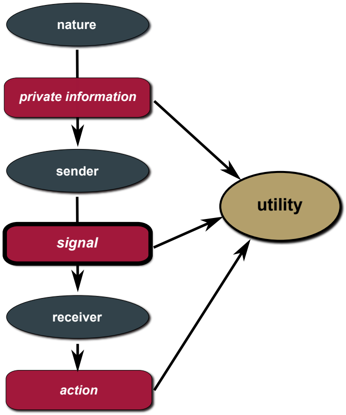


The game of case



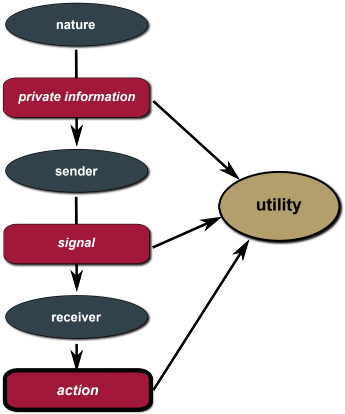
- **private information:** meaning, including linking of NPs to argument slots

The game of case



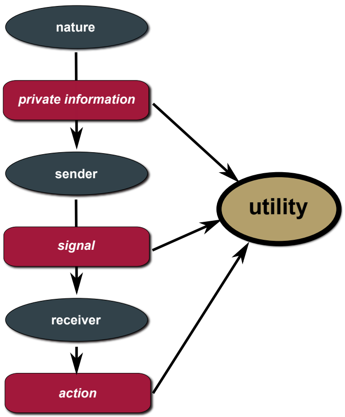
- **private information:** meaning, including linking of NPs to argument slots
- **signal:** case marking of NPs

The game of case



- **private information:** meaning, including linking of NPs to argument slots
- **signal:** case marking of NPs
- **action:** assign NPs to argument slots

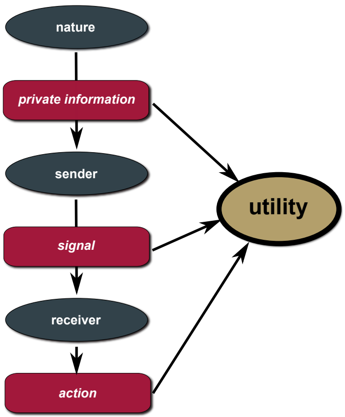
The game of case



- **private information:** meaning, including linking of NPs to argument slots
- **signal:** case marking of NPs
- **action:** assign NPs to argument slots
- **utility:**

$$u(t,m,a) = -k \times c(m) + \begin{cases} 1 & \text{if } a = t \\ 0 & \text{else} \end{cases}$$

The game of case

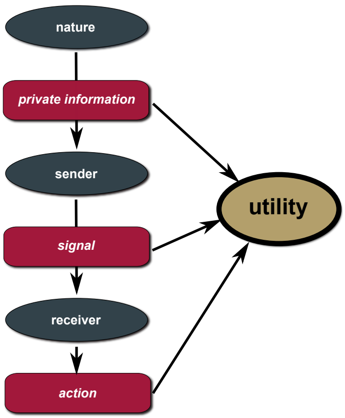


- **private information:** meaning, including linking of NPs to argument slots
- **signal:** case marking of NPs
- **action:** assign NPs to argument slots
- **utility:**

$$u(t, m, a) = -k \times c(m) + \begin{cases} 1 & \textbf{if a} = \textbf{t} \\ 0 & \textbf{else} \end{cases}$$

  - hearer economy

The game of case

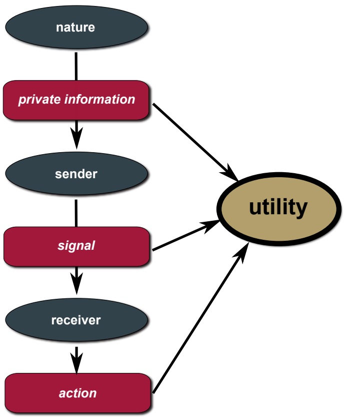


- **private information:** meaning, including linking of NPs to argument slots
- **signal:** case marking of NPs
- **action:** assign NPs to argument slots
- **utility:**

$$u(t, m, a) = -k \times \mathbf{c(m)} + \begin{cases} 1 & \text{if } a = t \\ 0 & \text{else} \end{cases}$$

  - hearer economy
  - speaker economy

The game of case



- **private information:** meaning, including linking of NPs to argument slots
- **signal:** case marking of NPs
- **action:** assign NPs to argument slots
- **utility:**

$$u(t, m, a) = -\mathbf{k} \times c(m) + \begin{cases} 1 & \text{if } a = t \\ 0 & \text{else} \end{cases}$$

  - hearer economy
  - speaker economy
  - relative strength of speaker economy vs. hearer economy

- speaker strategies that will be considered:

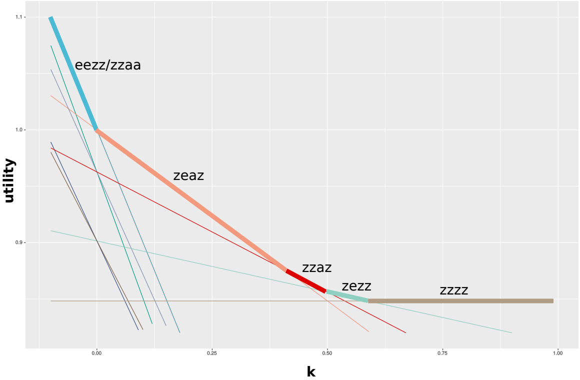
A is prominent	A is non-prominent	O is prominent	O is non-prominent
e(rgative)	e(rgative)	a(ccusative)	a(ccusative)
e	e	a	z(ero)
e	e	z	a
e	e	z	z
e	z	a	a
...	...	...	...
z	e	z	z
z	z	a	a
z	z	a	z
z	z	z	a
z	z	z	z

- hearer strategies:
  - strict rule: ergative means “agent”, and accusative means “object”
  - elsewhere rules:

1. AO: “The first phrase is always the agent.”
2. pA: “Pronouns are agents, and nouns are objects.”
3. pO: “Pronouns are objects, and nouns are agents.”
4. OA: “The first phrase is always the object.”

The game of case

- stochastic evolution always settles for strategy configuration with highest overall utility
- depends on k



Taking stock

Case marking systems participating in stochastically stable equilibria

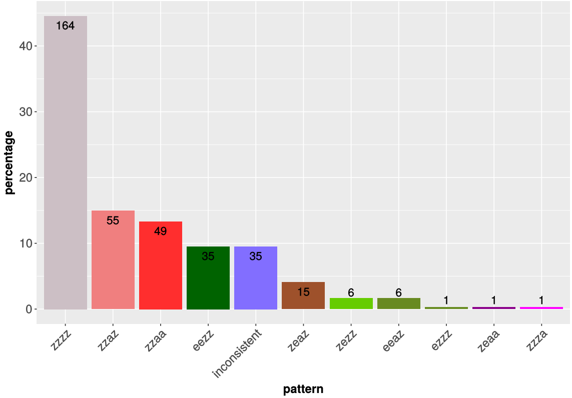
- **eezz:** consistent ergative marking
- **zzaa:** consistent accusative marking
- **zeaz:** split ergative system
- **zezz:** differential subject marking
- **zzaz:** differential object marking
- **zzzz:** no case marking

All stable systems are consistent with prominence hierarchies!

Empirical distribution

Bickel et al.'s (2015) sample

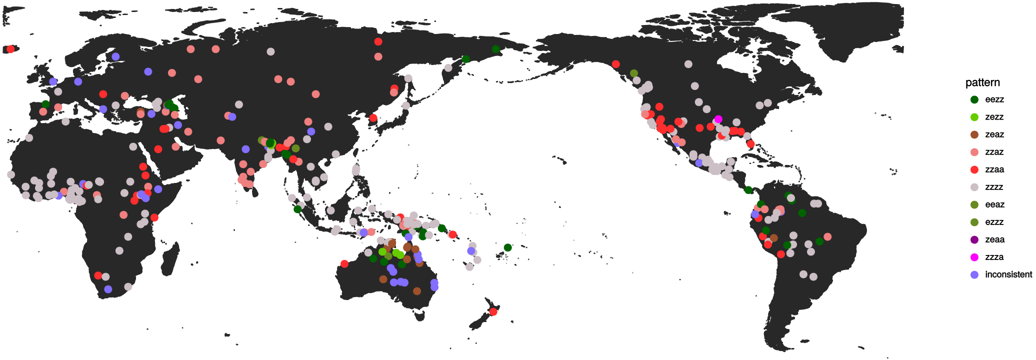
- genetically diverse sample of 460 case marking systems
- used here: 368 systems
  - one system per language
  - only languages with ISO code
  - only languages present in ASJP
- 342 out of 368 systems (88%) are stochastically stable



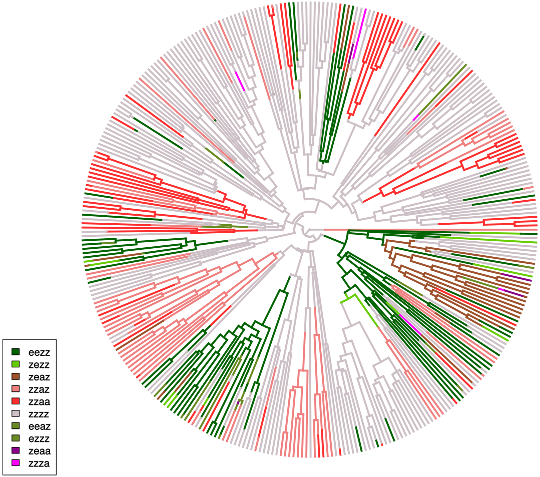
Phylogenetic non-independence

- 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):

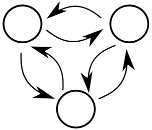
“If the A-distribution for a given typology cannot be assumed to be stationary, a distributional universal cannot be discovered on the basis of purely synchronic statistical data.”

“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

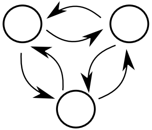
Modeling language change

Markov process

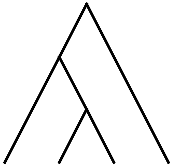


Modeling language change

Markov process

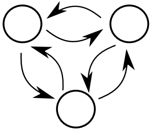


Phylogeny

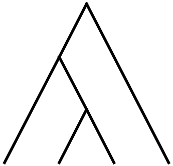


Modeling language change

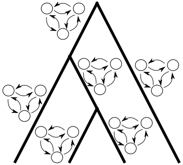
Markov process



Phylogeny

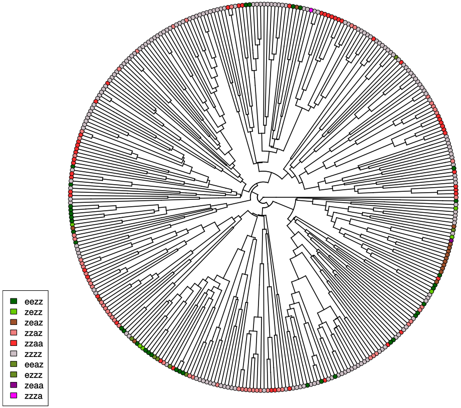


Branching process



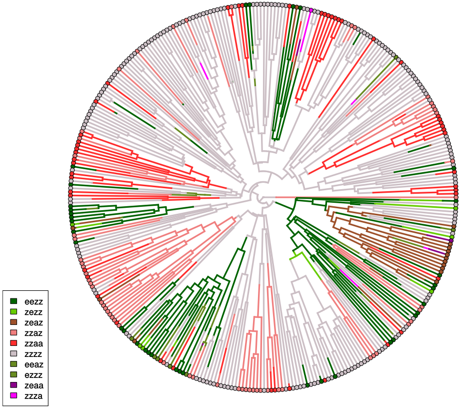
Estimating rates of change

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



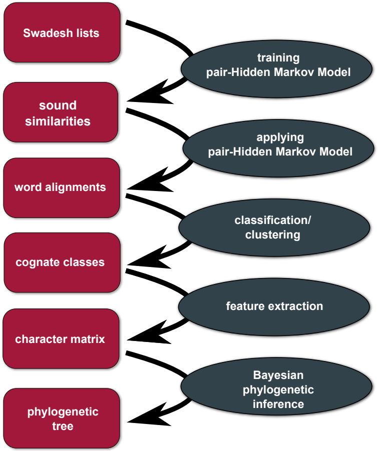
Estimating rates of change

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

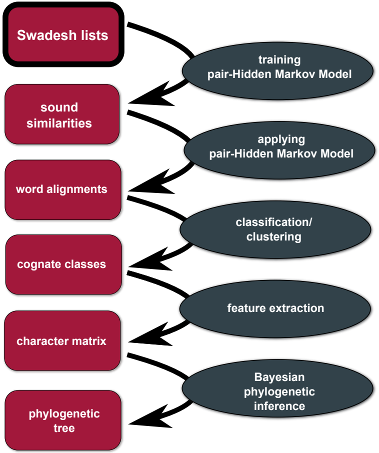


Inferring a world tree of languages

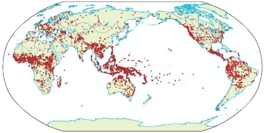
From words to trees



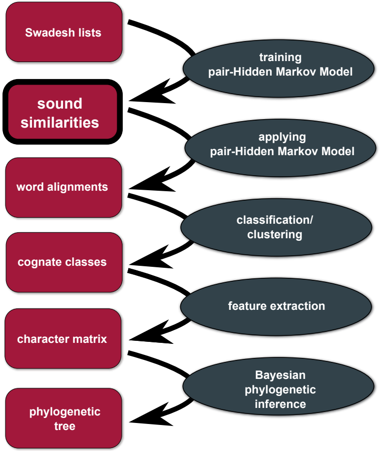
From words to trees



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



Swadesh lists

training
pair-Hidden Markov Model

sound
similarities

applying
pair-Hidden Markov Model

word alignments

classification/
clustering

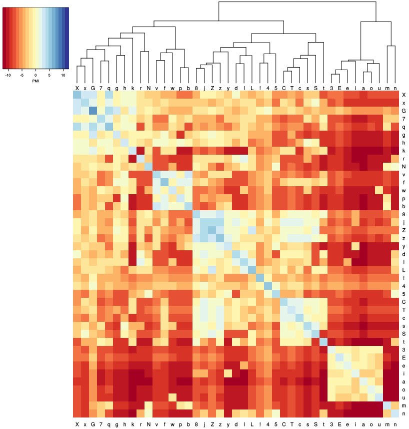
cognate classes

feature extraction

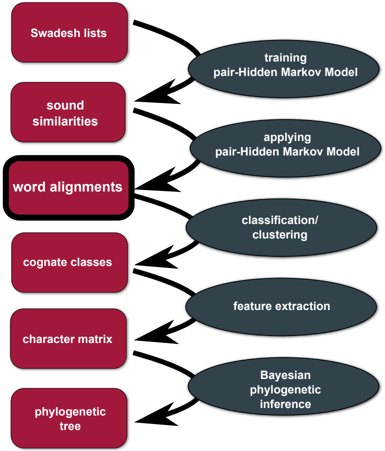
character matrix

Bayesian
phylogenetic
inference

phylogenetic
tree

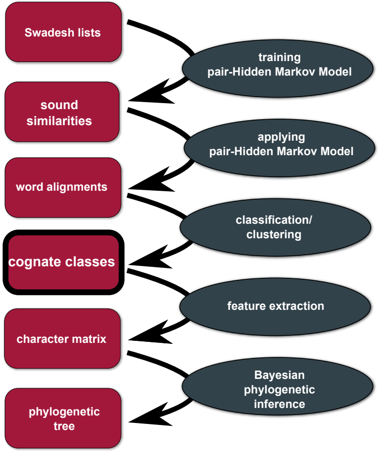


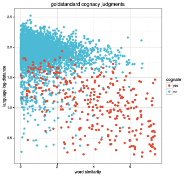
From words to trees

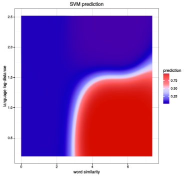


Language	*fish:z*	*tongue:1*	*smoke:1*
Abui-Atangmelang	-af-u		
Abui-Fuimelang	-af-u	tal-i-fi--	
Adang	aab--	tal-E-b---	awai--b-a-n-o-7o-
Blagar-Bakalang	-ab--	--j-e-bur-	--ad--b-a-n-aNka-
Blagar-Bama	aab--	teg-e-bur-	------b-e-n-a-xa-
Blagar-Kulijahi	-ab--	tej-e-bur-	------b-e-n-aNka-
Blagar-Nule	aab--	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

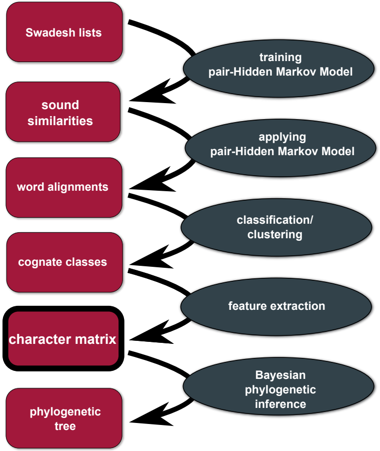



goldstandard cognacy judgments


SVM prediction

	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



Swadesh lists

training
pair-Hidden Markov Model

sound similarities

applying
pair-Hidden Markov Model

word alignments

classification/
clustering

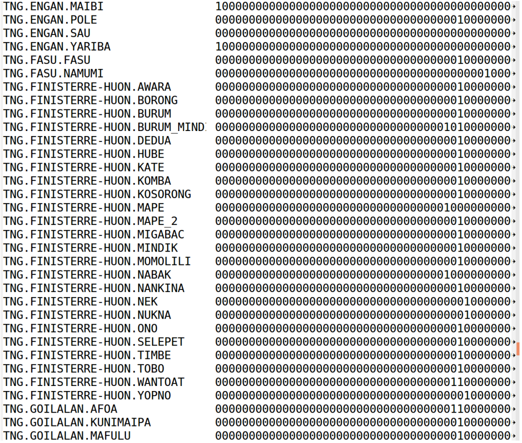
cognate classes

feature extraction

character matrix

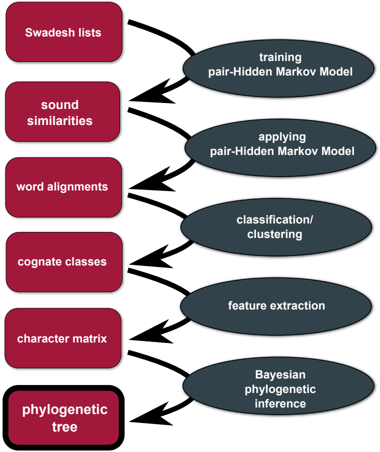
Bayesian
phylogenetic
inference

phylogenetic tree



```
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        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



Swadesh lists

training pair-Hidden Markov Model

sound similarities

applying pair-Hidden Markov Model

word alignments

classification/ clustering

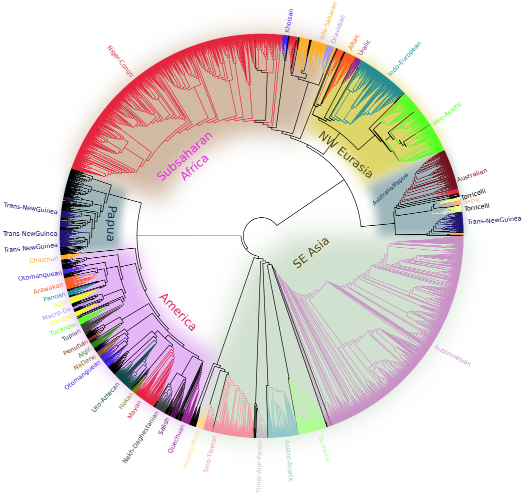
cognate classes

feature extraction

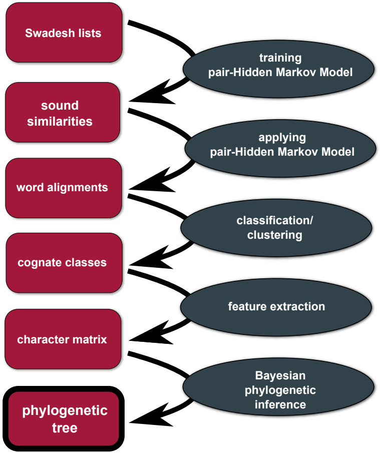
character matrix

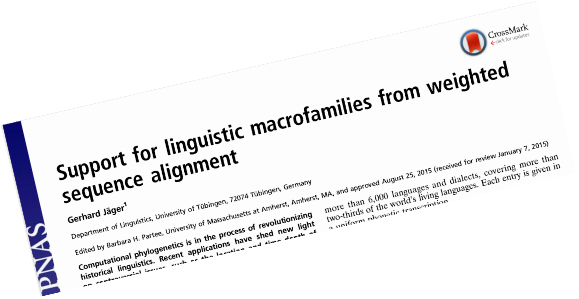
Bayesian phylogenetic inference

phylogenetic tree



From words to trees





PNAS

CrossMark
click for updates

Support for linguistic macrofamilies from weighted sequence alignment

Gerhard Jäger[1]

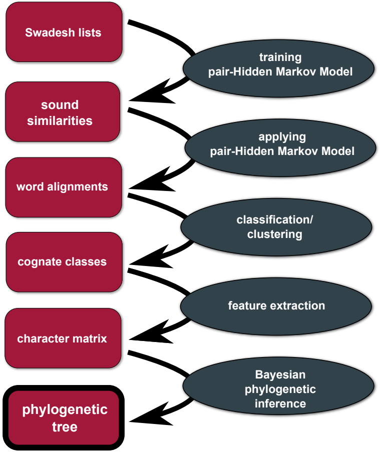
Department of Linguistics, University of Tübingen, 72074 Tübingen, Germany

Edited by Barbara H. Partee, University of Massachusetts at Amherst, Amherst, MA, and approved August 25, 2015 (received for review January 7, 2015)

Computational phylogenetics is in the process of revolutionizing historical linguistics. Recent applications have shed new light on controversial issues, such as the location and time depth of

more than 6,000 languages and dialects, covering more than two-thirds of the world's living languages. Each entry is given in a uniform phonetic transcription

From words to trees



PNAS

CrossMark click for updates

Support for linguistic macrofamilies from weighted sequence alignment

Gerhard Jäger[1]

Department of Linguistics, University of Tübingen, 72074 Tübingen, Germany

Edited by Barbara H. Partee, University of Massachusetts at Amherst, Amherst, MA, and approved August 25, 2015 (received for review January 7, 2015)

Computational phylogenetics ... in the process of revolutionizing historical linguistics. Recent ... tions have shed new light

more than 6,000 languages and dialects, covering more than two-thirds of the world's living languages. Each entry is given in

SCIENTIFIC REPORTS

www.nature.com/scientificreports

OPEN

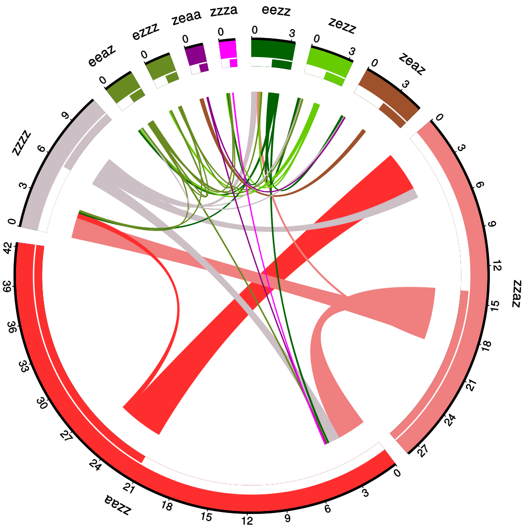
Tracking modern human population history from linguistic and cranial phenotype

Hugo Reyes-Centeno[1,2], Katerina Harvati[1,2] & Gerhard Jäger[2,3]

Received: 21 July 2016

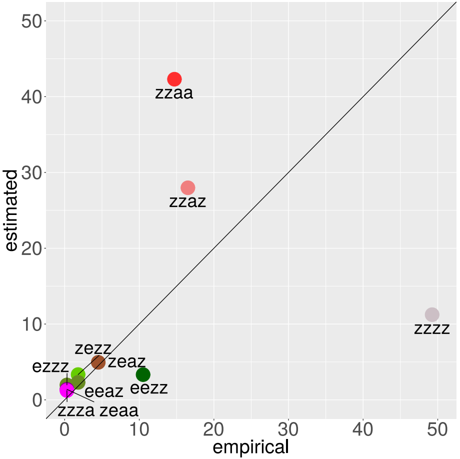
Cases in equilibrium

Phylogenetically estimated Markov chain

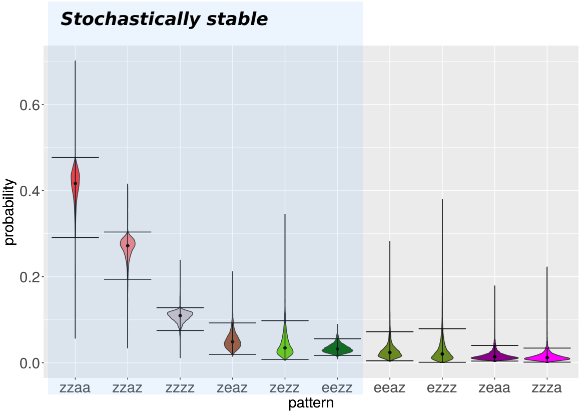


Equilibrium probabilities

Empirical vs. estimated percentages



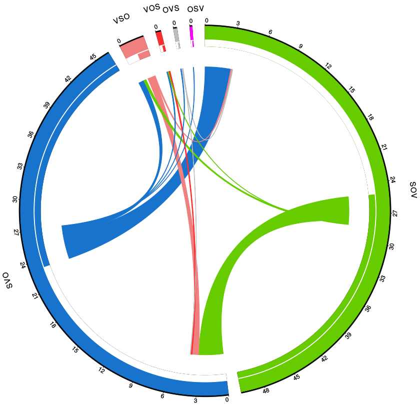
Posterior distribution



Summary

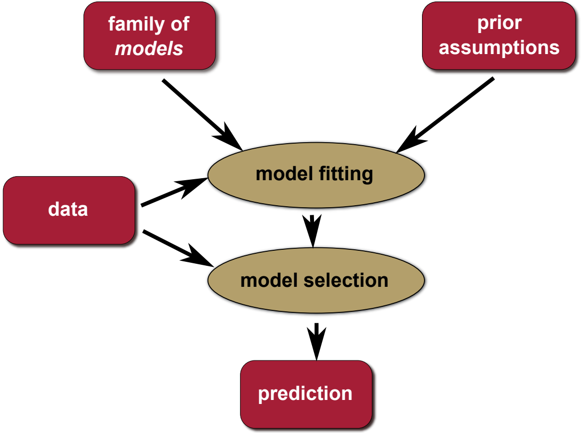
- three patterns occur with probability $> 5\%$ in equilibrium:
  - non-differential accusative marking
  - differential accusative marking
  - no case marking
- all three are predicted to be stochastically stable
- ergative systems are conspicuously underrepresented

- method applicable to many typological issues



Linguistics Quo Vadis (cont.)

Statistical modeling of linguistic dynamics



Topics

Micro-dynamics

- pragmatics
- incremental processing
- language variation

Macro-dynamics

- typology
- historical linguistics
- dialectometry

Data

Micro-dynamics

- corpora
- psycholinguistic experiments
- crowd sourcing

Macro-dynamics

- cross-linguistic databases
- etymological dictionaries
- dialect atlases

Models

Micro-dynamics

- formal semantics and pragmatics
- rationalistic game theory
- classical comparative method

Macro-dynamics

- evolutionary game theory
- phylogenetic inference
- population genetics

Inference methods

Micro- and macro-dynamics

- Bayesian inference
- approximate Bayesian computation
- machine learning
- agent-based simulations
- causal inference

Judith Aissen. Differential object marking: Iconicity vs. economy. *Natural Language and Linguistic Theory*, 21(3):435–483, 2003.

Balthasar Bickel, Alena Witzlack-Makarevich, and Taras Zakharko. Typological evidence against universal effects of referential scales on case alignment. In Ina Bornkessel-Schlesewsky, Andrej L. Malchukov, and Marc D. Richards, editors, *Scales and hierarchies: A cross-disciplinary perspective*, pages 7–43. de Gruyter, Berlin/Munich/Boston, 2015.

Georg Bossong. *Differentielle Objektmarkierung in den neuiranischen Sprachen*. Günther Narr Verlag, Tübingen, 1985.

Kenneth Church. A pendulum swung too far. *Linguistic Issues in Language Technology*, 2(4):1–26, 2007.

Bernard Comrie. *Language Universals and Linguistic Typology*. Basil Blackwell, Oxford, 1981.

Gerhard Jäger. Evolutionary Game Theory and typology: a case study. *Language*, 83(1):74–109, 2007.

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 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.

David Lewis. *Convention*. Harvard University Press, Cambridge, MA, 1969.

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.

Mark Pagel and Andrew Meade. BayesTraits 2.0. software distributed by the authors, November 2014.

Hugo Reyes-Centeno, Katerina Harvati, and Gerhard Jäger. Tracking modern human population history from linguistic and cranial phenotype. *Scientific Reports*, 6, 2016.

Michael Silverstein. Hierarchy of features and ergativity. In R. M. W. Dixon, editor, *Grammatical Categories in Australian Languages*, pages 112–171. Australian Institute of Aboriginal Studies, Canberra, 1976.

Søren Wichmann, Eric W. Holman, and Cecil H. Brown. The ASJP database (version 17). http://asjp.clld.org/, 2016.

H. Peyton Young. The evolution of conventions. *Econometrica*, 61:57–84, 1993.