A case study in computer-aided typology

Gerhard Jäger

Tübingen University

Symposium Linguistics Quo Vadis

MPI Nijmegen, October 2, 2017
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

- family of models
- prior assumptions
- model fitting
- model selection
- prediction

Gerhard Jäger (Tübingen)
Statistical modeling of linguistic dynamics

- Data
- Family of models
- Prior assumptions
- Model fitting
- Model selection
- Prediction

Small data
Statistical modeling of linguistic dynamics

- family of models
- prior assumptions
- model fitting
- model selection
- prediction

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
  - ...

Diagram:
-amily of models → model fitting → model selection → prediction
- data → model fitting
- prior assumptions
Case alignment systems
Universal syntactic-semantic primitives

- three universal core roles
  - \textbf{S}: intransitive subject
  - \textbf{A}: transitive subject
  - \textbf{O}: transitive object

\textbf{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.'

\textbf{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.'
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

\[
\begin{array}{c}
S \\
A \\
O
\end{array}
\]

nominative (absolutive)

ergative

Dyirbal

\textit{ŋuma yabu-ŋu bura-n.}
father mother.ERG see-NONFUT
'The mother saw the father.'

\textit{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

accusative  neutral or tripartite  ergative
Functional explanation?

probability $P(\text{syntactic role} | \text{prominence of NP})$
The analysis

usage frequencies → evolutionary game theory → predicted equilibrium patterns

typological distribution → phylogenetic comparative method → typological probabilities

comparison
The analysis

- Usage frequencies
- Evolutionary game theory
- Predicted equilibrium patterns
- Phylogenetic comparative method
- Typological distribution
- Typological probabilities
- Comparison

Jäger (2007)
The analysis

usage frequencies → evolutionary game theory → predicted equilibrium patterns → comparison

typological distribution → phylogenetic comparative method → typological probabilities → ongoing research
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 $\approx$ motivation

Evolutionary game theory
- frequency-dependent Darwinian selection
- utility $\approx$ fitness
Signaling games

nature

private information

sender

signal

receiver

action

utility
The game of case

nature

private information

sender

signal

receiver

action

utility

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

signal: case marking of NPs

action: assign NPs to argument slots

utility: hearer economy, speaker economy, relative strength of speaker economy vs. hearer economy

Gerhard Jäger (Tübingen)
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**:
  \[
  u(t, m, a) = -k \times c(m) + \begin{cases} 
  1 & \text{if } a = t \\
  0 & \text{else}
  \end{cases}
  \]
  - 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**:

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

- 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**:

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

- hearer economy
- speaker economy
- relative strength of speaker economy vs. hearer economy
Game-theoretic modeling

- speaker strategies that will be considered:

<table>
<thead>
<tr>
<th>A is prominent</th>
<th>A is non-prominent</th>
<th>O is prominent</th>
<th>O is non-prominent</th>
</tr>
</thead>
<tbody>
<tr>
<td>e(rgative)</td>
<td>e(rgative)</td>
<td>a(ccusative)</td>
<td>a(ccusative)</td>
</tr>
<tr>
<td>e</td>
<td>e</td>
<td>a</td>
<td>z(ero)</td>
</tr>
<tr>
<td>e</td>
<td>e</td>
<td>z</td>
<td>a</td>
</tr>
<tr>
<td>e</td>
<td>e</td>
<td>z</td>
<td>z</td>
</tr>
<tr>
<td>e</td>
<td>z</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>z</td>
<td>e</td>
<td>z</td>
<td>z</td>
</tr>
<tr>
<td>z</td>
<td>z</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>z</td>
<td>z</td>
<td>a</td>
<td>z</td>
</tr>
<tr>
<td>z</td>
<td>z</td>
<td>z</td>
<td>a</td>
</tr>
<tr>
<td>z</td>
<td>z</td>
<td>z</td>
<td>z</td>
</tr>
</tbody>
</table>
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
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
  \[ \Rightarrow \] 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

- Diagram of a Markov process in the context of language change.
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
Inferring a world tree of languages

From words to trees

Swadesh lists
sound similarities
word alignments
cognate classes
character matrix
phylogenetic tree

training pair-Hidden Markov Model
applying pair-Hidden Markov Model
classification/clustering
feature extraction
Bayesian phylogenetic inference

Gerhard Jäger (Tübingen)

Computer-aided typology
Linguistics Quo Vadis
From words to trees

- Swadesh lists
- sound similarities
- word alignments
- cognate classes
- character matrix
- phylogenetic tree

- training pair-Hidden Markov Model
- applying pair-Hidden Markov Model
- classification/clustering
- feature extraction
- Bayesian phylogenetic inference

<table>
<thead>
<tr>
<th>concept</th>
<th>Latin</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>ego</td>
<td>Ei</td>
</tr>
<tr>
<td>you</td>
<td>tu</td>
<td>yu</td>
</tr>
<tr>
<td>we</td>
<td>nos</td>
<td>wi</td>
</tr>
<tr>
<td>one</td>
<td>unus</td>
<td>w3n</td>
</tr>
<tr>
<td>two</td>
<td>duo</td>
<td>tu</td>
</tr>
<tr>
<td>person</td>
<td>persona, homo</td>
<td>pers3n</td>
</tr>
<tr>
<td>fish</td>
<td>piscis</td>
<td>fiS</td>
</tr>
<tr>
<td>dog</td>
<td>kanis</td>
<td>dag</td>
</tr>
<tr>
<td>louse</td>
<td>pedikulus</td>
<td>laus</td>
</tr>
<tr>
<td>tree</td>
<td>arbor</td>
<td>tri</td>
</tr>
<tr>
<td>leaf</td>
<td>foly-ú*</td>
<td>lif</td>
</tr>
<tr>
<td>skin</td>
<td>kutis</td>
<td>skin</td>
</tr>
<tr>
<td>blood</td>
<td>saNgw-ís</td>
<td>bl3d</td>
</tr>
<tr>
<td>bone</td>
<td>os</td>
<td>bon</td>
</tr>
<tr>
<td>horn</td>
<td>kornu-ís</td>
<td>horn</td>
</tr>
<tr>
<td>ear</td>
<td>auris</td>
<td>ir</td>
</tr>
<tr>
<td>eye</td>
<td>okulus</td>
<td>Ei</td>
</tr>
</tbody>
</table>
From words to trees

- Swadesh lists
- Sound similarities
- Word alignments
- Cognate classes
- Character matrix
- Phylogenetic tree

Training pair-Hidden Markov Model
Applying pair-Hidden Markov Model
Classification/clustering
Feature extraction
Bayesian phylogenetic inference
Inferring a world tree of languages

From words to trees

- Swadesh lists
- Sound similarities
- Word alignments
- Cognate classes
- Character matrix
- Phylogenetic tree

Training pair-Hidden Markov Model

Applying pair-Hidden Markov Model

Classification/clustering

Feature extraction

Bayesian phylogenetic inference

Language | fish:ə | tongue:ɨ | smoke:ɨ
---|---|---|---
Abui-Atangmelang | -aː/u | -u | -u
Abui-Fuimelang | -aː/u | -u | -u
Adang | aː/ə | -u | -u
Blagar-Bakalang | -aː/ə | -u | -u
Blagar-Bama | aː/ə | -u | -u
Blagar-Kulijahi | aː/ə | -u | -u
Blagar-Nule | aː/ə | -u | -u
Blagar-Tuntuli | aː/ə | -u | -u
Blagar-Warsalelang | aː/ə | -u | -u
Bunaq | aː/ə | -u | -u
Deing | -aː/ə | -u | -u
Hamap | 7aː/ə | -u | -u
Kabola | aː/ə | -u | -u
Kala-Padangusul | aː/ə | -u | -u
Kafoa | aː/ə | -u | -u
Kamang | aː/ə | -u | -u
Kiraman | aː/ə | -u | -u
Klon | aː/ə | -u | -u
Kuli | aː/ə | -u | -u
Kula | aː/ə | -u | -u
Nedebang | aː/ə | -u | -u
Reta | aː/ə | -u | -u
Sar-Adiabang | aː/ə | -u | -u
Sar-Nule | aː/ə | -u | -u
Sawila | aː/ə | -u | -u
Tswa-Madar | aː/ə | -u | -u
Wersing | aː/ə | -u | -u
Wpantar | aː/ə | -u | -u
Inferring a world tree of languages

From words to trees

- Swadesh lists
- sound similarities
- word alignments
- cognate classes
- character matrix
- phylogenetic tree

- training pair-Hidden Markov Model
- applying pair-Hidden Markov Model
- classification/clustering
- feature extraction
- Bayesian phylogenetic inference

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
<th>Modern Greek</th>
<th>Standard German</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>yo:B</td>
<td>eso:C</td>
<td>iX:D</td>
</tr>
<tr>
<td>yes</td>
<td>ustet:B</td>
<td>esi:D</td>
<td>dsl:E</td>
</tr>
<tr>
<td>we</td>
<td>nosotros:B</td>
<td>emas:C</td>
<td>vir:A</td>
</tr>
<tr>
<td>one</td>
<td>uno:B</td>
<td>emas:C, ena:C</td>
<td>ams:D</td>
</tr>
<tr>
<td>two</td>
<td>dos:B</td>
<td>By-&quot;o:C, Bia:D</td>
<td>cval:E</td>
</tr>
<tr>
<td>person</td>
<td>personas:A</td>
<td>amB-ropos:B</td>
<td>mENs:C</td>
</tr>
<tr>
<td>flash</td>
<td>peskado:A, pos:A</td>
<td>psar1:C</td>
<td>fiS:A</td>
</tr>
<tr>
<td>dog</td>
<td>pero:B</td>
<td>stTili:C, stTilos:C</td>
<td>hunt:B</td>
</tr>
<tr>
<td>come</td>
<td>veni:B</td>
<td>erx-o:C</td>
<td>khr-oam3n:A</td>
</tr>
<tr>
<td>sem</td>
<td>sol:B</td>
<td>ily-os:C, 1oos:C</td>
<td>zom3:A</td>
</tr>
<tr>
<td>star</td>
<td>estrey:A</td>
<td>asteri:A, astro:A</td>
<td>Stiri:A</td>
</tr>
<tr>
<td>water</td>
<td>agua-a:B</td>
<td>nero:C</td>
<td>vauC:A</td>
</tr>
<tr>
<td>stone</td>
<td>piedra:B</td>
<td>petsa:B</td>
<td>Stain:A</td>
</tr>
<tr>
<td>fire</td>
<td>fuego:B</td>
<td>foty-a:C</td>
<td>feisD:</td>
</tr>
<tr>
<td>path</td>
<td>senda:B</td>
<td>cronos:C</td>
<td>pf-vt-A, vek:D</td>
</tr>
<tr>
<td>mountain</td>
<td>monte:3n:A</td>
<td>vaneC:C, enne:D</td>
<td>hErick:E</td>
</tr>
<tr>
<td>ful</td>
<td>yeno:B</td>
<td>yenalC:C, pliris:D</td>
<td>folic:A</td>
</tr>
<tr>
<td>new</td>
<td>muevo:A</td>
<td>neos:A, Tenury-os:B</td>
<td>nos:A</td>
</tr>
<tr>
<td>name</td>
<td>nombre:A</td>
<td>onomes:A</td>
<td>nam5:A</td>
</tr>
</tbody>
</table>
Inferring a world tree of languages

From words to trees

- Swadesh lists
- Sound similarities
- Word alignments
- Cognate classes
- Character matrix
- Phylogenetic tree

- Training pair-Hidden Markov Model
- Applying pair-Hidden Markov Model
- Classification/clustering
- Feature extraction
- Bayesian phylogenetic inference

- TNG. ENGAN. MAIDI
- TNG. ENGAN. POLE
- TNG. ENGAN. SAU
- TNG. ENGAN. YARIBA
- TNG. FASU. MAIDI
- TNG. FASU. NAMUMI
- TNG. FINISTERRE-HUON. AWARA
- TNG. FINISTERRE-HUON. BORONG
- TNG. FINISTERRE-HUON. BURUM
- TNG. FINISTERRE-HUON. BURUM MIND
- TNG. FINISTERRE-HUON. DEDUA
- TNG. FINISTERRE-HUON. HUBE
- TNG. FINISTERRE-HUON. KATE
- TNG. FINISTERRE-HUON. Komba
- TNG. FINISTERRE-HUON. KOSORONG
- TNG. FINISTERRE-HUON. MAPE
- TNG. FINISTERRE-HUON. MAPE_2
- TNG. FINISTERRE-HUON. MIGABAC
- TNG. FINISTERRE-HUON. MINDIK
- TNG. FINISTERRE-HUON. MOMOLILI
- TNG. FINISTERRE-HUON. NABAK
- TNG. FINISTERRE-HUON. NANKINA
- TNG. FINISTERRE-HUON. NEK
- TNG. FINISTERRE-HUON. NUKNA
- TNG. FINISTERRE-HUON. ONO
- TNG. FINISTERRE-HUON. SELEPET
- TNG. FINISTERRE-HUON. TIMBE
- TNG. FINISTERRE-HUON. TOBO
- TNG. FINISTERRE-HUON. MANTOAT
- TNG. FINISTERRE-HUON. YOPNO
- TNG. GOILALAN. AFDA
- TNG. GOILALAN. KUNIMAIPA
- TNG. GOILALAN. MAFULU

Gerhard Jäger (Tübingen)
Inferring a world tree of languages

From words to trees

- Swadesh lists
- sound similarities
- word alignments
- cognate classes
- character matrix
- phylogenetic tree

- training pair-Hidden Markov Model
- applying pair-Hidden Markov Model
- classification/clustering
- feature extraction
- Bayesian phylogenetic inference

- Khoisan
- Niger-Congo
- Nilo-Saharan
- Afro-Asiatic
- Indo-European
- Uralic
- Altaic
- Ainu
- Nakh-Daghestanian
- Dravidian
- Sino-Tibetan
- Hmong-Mien
- Tai-Kadai
- Austro-Asiatic
- Austronesian
- Sepik
- Torricelli
- Trans-New Guinea
- Australian
- NaDene
- Algic
- Uto-Aztecan
- Salish
- Penutian
- Hokan
- Otomanguean
- Mayan
- Chibchan
- Tucanoan
- Panoan
- Quechuan
- Arawakan
- Cariban
- Tupian
- Macro-Ge
- Sub-Saharan Africa
- SE Asia
- America

Gerhard Jäger (Tübingen)

Computer-aided typology

Linguistics Quo Vadis
Inferring a world tree of languages

From words to trees

- Swadesh lists
- Sound similarities
- Word alignments
- Cognate classes
- Character matrix
- Phylogenetic tree

- Training pair-Hidden Markov Model
- Applying pair-Hidden Markov Model
- Classification/clustering
- Feature extraction
- Bayesian phylogenetic inference

Support for linguistic macrofamilies from weighted sequence alignment

Gerhard Jäger
Department of Linguistics, University of Tübingen, Oskar-von-Miller-Ring 15, 72076 Tübingen, Germany

Computational phylogenetics is in the process of revolutionizing historical linguistics. Recent applications have shown new insights into the evolutionary history of languages. This work focuses on developing computational methods for inferring phylogenetic relationships between languages from various sources of data, such as Swadesh lists, and applying these methods to real-world linguistic problems. The goal is to provide a robust and accurate framework for understanding the relationships between languages and their historical development. 
Inferring a world tree of languages

From words to trees

- Swadesh lists
- Sound similarities
- Word alignments
- Cognate classes
- Character matrix
- Phylogenetic tree

Training pair-Hidden Markov Model

Applying pair-Hidden Markov Model

Classification/clustering

Feature extraction

Bayesian phylogenetic inference

Support for linguistic macrofamilies from weighted sequence alignment

Gerhard Jäger (Tübingen)
Cases in equilibrium
Phylogenetically estimated Markov chain
Equilibrium probabilities

Empirical vs. estimated percentages

Posterior distribution

Stochastically stable

Gerhard Jäger (Tübingen)

Computer-aided typology

Linguistics Quo Vadis
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
Statistical modeling of linguistic dynamics

- family of models
- prior assumptions
- model fitting
- model selection
- prediction

Gerhard Jäger (Tübingen)
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


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


