

Treebanks and evolutionary simulation for explaining typological patterns

Treebanks and Linguistic Theory 7
January 24, 2009
Groningen

Rob Malouf
Department of Linguistics and
Asian / Middle Eastern Languages



SAN DIEGO STATE
UNIVERSITY

1

Based on joint work with:

Farrell Ackerman
University of California
at San Diego

James Blevins
Cambridge University

Andrew Wedel
University of Arizona

2

Linguistic typology

- Linguistic typology studies linguistic variation across time and space
- How do human languages differ? How are they the same? Are there true language universals? Are there broad tendencies?
- Even more importantly, why? Where do the observed patterns come from?
- Relation to human cognition and the 'language faculty'

3

Linguistic typology

- Typological databases provide the raw material for typological theorizing
- Phonetics and phonology
 - Phoneme inventories
 - Syllable types and phonotactics
- Morphology
 - Surrey Syncretism Database
- Syntax
 - World Atlas of Language Structures
 - Utrecht Typological Database System

4

Linguistic typology

- Standardized metalanguage ('basic linguistic theory')
- Typological databases list properties at the level of languages or language families
 - WALS entry for [English](#)
 - 'English has SVO word order'
 - 'Dutch has no dominant word order'
- Language databases record facts about linguistic competence
- They necessarily abstract away from a lot of messy details

5

Evolutionary phonology

- Evolutionary phonology (Juliette Blevins) accounts for phonological patterns via sound changes
- Multiple sources for sound pattern similarity (Blevins 2007)
 - Cognitive Factors
 - Direct inheritance
 - Indirect inheritance
 - Phonetic factors
 - Language-specific factors
 - Chance

6

Evolutionary phonology

- Devoicing of word final obstruents ($b, d, g \rightarrow p, t, k$)
- Final devoicing is phonetically a very natural sound change
- Many evolutionary pathways will lead to final devoicing
- The need to maintain lexical contrast will sometimes work against the tendency towards final devoicing
- No natural sound changes lead directly to final voicing, but under the right circumstances it could arise anyway (Breton, Lezgian)

7

Explanation

- Does EP really *explain* the observed variation?
- Generative social science (Epstein 1999) "If you didn't grow it, you didn't explain it."
- "The Generativist's Question: How could the decentralized local interactions of heterogeneous autonomous agents generate the given regularity?"
- "To explain a macroscopic regularity x is to furnish a suitable microspecification that suffices to generate it. The core request is hardly outlandish: To explain a macro- x , please show how it could arise in a plausible society. Demonstrate how a set of recognizable – heterogeneous, autonomous, boundedly rational, locally interacting – agents could actually get there in reasonable time." Epstein (2006:51)

8

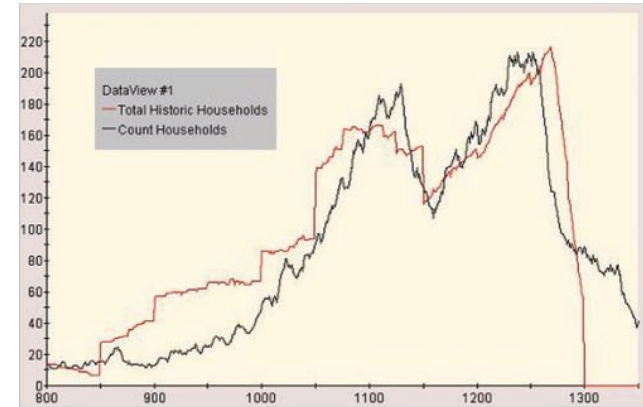
Explanation

- The Anasazi inhabited the American Southwest from c. 1800 BC, experienced a 'Golden Age' beginning c. 900 AD, and disappeared c. 1300 AD
- Various explanations (climate change, internal social changes, external pressure from other groups), all supported by the archeological and geological record
- Axtell, *et al.* (2002) ran detailed agent-based simulations of Kayenta Anasazi (Long House Valley, AZ) agriculture and household creation using reconstructed climate data
- The simulations include many variables, but do not take into account institutional factors (social structure, property rights)

9

Explanation

- The simulation results are a close (but not perfect) fit to the actual population distribution



10

Paradigm Economy Principle

- Carstairs-McCarthy: There are as many inflection classes as there are realizations of the paradigm cell with the most realizations (Paradigm Economy Principle)
- Hungarian (factoring out phonologically conditioned allomorphy):

Sing	1	-ok	-om
	2	-sz	-ol
	3	-Ø	-ik
Pl	1	-unk	-unk
	2	-tok	-tok
	3	-nak	-nak

11

Paradigm Economy Principle

- It's not clear that this relation between number of markers and number of classes is 'surface true' in all languages:
 - Lots of footnotes and qualifications and riders and codicils
 - But, inflectional class systems are reliably less complicated than they could be.
- Tundra Nenets has 14 NOM.SG. forms and 29 ACC.PL. forms, and 94 pairs (more than 29, but much fewer than $14 \times 29 = 8,120$)
- Why?

12

Paradigm Economy Principle

- What is the Paradigm Economy Principle a constraint on? (Carstairs-McCarthy's "Two questions with one answer")
- Representations (Müller 2006)
 - Paradigm economy reveals something about the mental representation of morphology (genetic)
 - But, what about exceptions?
- Usage (Plank 1991)
 - Paradigm economy is a side effect of how language is learned or used (epigenetic)
 - But, why have inflectional classes at all?

13

Paradigm Economy Principle

- A usage-based theory of paradigm economy has to spell out:
 - Exactly what notion of 'simplicity' is relevant, and
 - How that then leads to the observed cross-linguistic patterns
- A paradigm can satisfy the paradigm economy without being particularly simple
- It's not a limitation on the number of classes
- It's not a memory restriction
- Why should language be simple, anyway?

14

The Paradigm Cell Filling Problem

- It is implausible that speakers of languages with complex morphology and multiple inflection classes encounter every inflected form of every word
- **Paradigm Cell Filling Problem:** Given exposure to a novel inflected word form, what licenses reliable inferences about the other word forms in its inflectional family?
- As languages depart from what Lounsbury's (1953) "fictive agglutinative ideal", predicting forms by analogy becomes increasingly important
- Complex paradigms are organized in a way to make the PCFP tractable (Ackerman, Blevins & Malouf 2009; Stump & Finkal 2009)

15

Finnish

(following the classification in Piñel & Pikamäe 1999:758-771)

Nom Sg	Gen Sg	Part Sg	Part Pl	Iness Pl	
ovi	oven	ovea	ovia	ovissa	'door' (8)
kieli	kielen	kieltä	kieliä	kielissä	'language' (32)
vesi	veden	vettä	vesiä	vesissä	'water' (10)
lasi	lasin	lasia	laseja	laseissa	'glass' (4)
nalle	nallen	nallea	nalleja	nalleissa	'teddy' (9)
kirje	kirjeen	kirjettä	kirjeitä	kirjeissä	'letter' (78)

- To deduce the Finnish nominative for *pantti* 'pants', it is enough to know the partitive singular *panttia* (on analogy with *lasi*~*lasia* 'cup')
- Knowing the partitive plural *pantteja* restricts class membership to either 4 or 9
- Knowing the inessive plural *pantteissa* restricts membership to 4, 9, or 78

16

Paradigm entropy

- The **conditional entropy** is the uncertainty in one random variable on average, given that we know the value of another random variable

$$\begin{aligned} H(Y|X) &= - \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2 p(y|x) \\ &= H(X, Y) - H(X) \end{aligned}$$

- The conditional entropy of one cell given another is a measure of inter-predictability with a paradigm (Ackerman, Blevins, and Malouf 2009)
- In our Finnish partial paradigm, $H(\text{GEN.SG.}) = 1.792$ bits but $H(\text{GEN.SG.} | \text{NOM.SG.}) = 1.333$ bits

17

Paradigm entropy

- To extend this to the whole paradigm, we calculate the **expected** conditional entropy

$$E[H(c|c)] = \sum_{c_1, c_2} p(c_1, c_2) H(c_2 | c_1)$$

- The higher the expected conditional entropy, the more difficult it is on average to predict an unknown wordform, given a known wordform
- For the Finnish partial paradigm, the expected conditional entropy is 0.663 bits
- This is an **upper bound** on the actual entropy

18

Paradigm Economy Principle

- A **hypothesis** (Plank 1991, Ackerman & Malouf 2008)

Paradigms which satisfy the Paradigm Economy Principle are simpler w.r.t. the PCFP – and have lower expected conditional entropy – than paradigms that don't

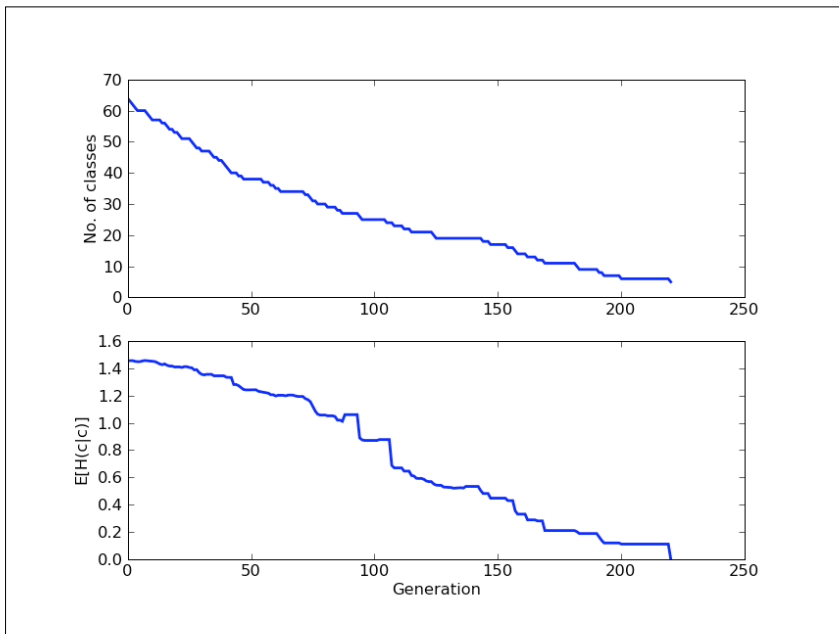
- To test whether this is a possible explanation for the paradigm economy principle, we set up a simple evolution simulation
- Start with a maximally complex paradigm, with c cells, r realizations for each cell, and r^c inflection classes

19

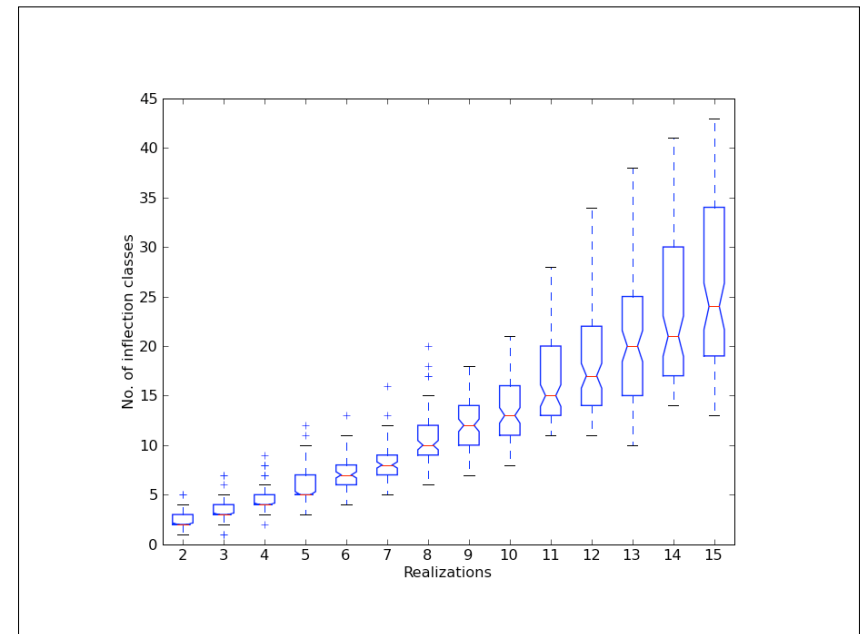
Simulation

- On each generation, we randomly modify the language according to the PCFP
 - Select one cell of one inflection class as given
 - Guess realization of another cell based on the given cell
 - Reconfigure inflection classes as necessary
- Occasionally, a new form is introduced to fill an unknown cell
- Continue until we go $2cr$ generations without reducing the expected conditional entropy

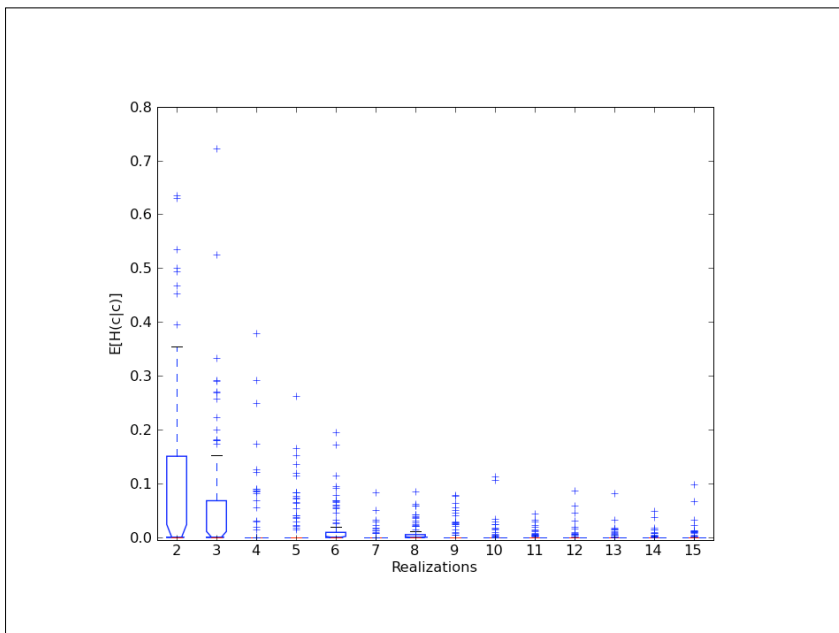
20



21



22



23

Simulation prospects

- We start with a plausible mechanism (small random changes to paradigms under the influence of PCFP)
- In our simulations, this leads to outcomes which are like real languages in some respects
- But, the average number of classes grows a little faster than the number of initial realizations
- And, the final state and entropy depend somewhat on the initial conditions and the distributions of forms
- Unlike the Anasazi case, we don't have enough information to fully evaluate the simulation results

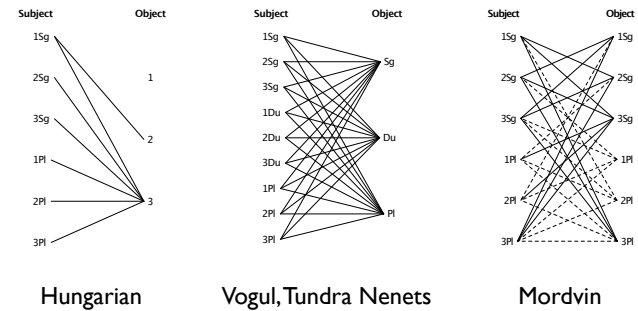
24

Evolutionary typology

- If we want to explain typological patterns via simulation, we need detailed **quantitative** information about lots of languages
- For phonological simulations, segment and syllable inventories may be sufficient (Oudeyer 2002, Wedel & van Volkinburg 2009)
- For simple morphological patterns, lexicons with frequency information can be useful
- To look at the development of other morphological or syntactic constructions, we need linguistically annotated examples of language use in a natural context (i.e., treebanks)

25

Complexification



26

Locality

- **Morphosyntactic locality** (see also Sag 2004, 2008):
 '[L]ocality'. . . refers to the proximity of the agreeing element within the clause structure; a local agreement relation is one which holds between elements of the same simple clause, while a non-local agreement relation is one which may hold between elements of different clauses. (Bresnan and Mchomobo, 1987:52)
- We expect subject-verb agreement to be expressed locally within a clause
- Prenominal relative clauses form a clausal domain

27

Locality

- Quechua (Cole 1987:279; Cinque 2008)
 [nuna tanti-shqa-n] bestya
 man buy-PERF-3 horse.NOM
 'the horse which the man bought'
- Eastern Ostyak (Ackerman & Malouf 2004a):
 [wer-t-äm] kiriw
 make-PART-1SG boat
 'the boat which I will make'
- Eastern Armenian (Ackerman & Malouf 2004b):
 [gn-ac'-əs] hovanoc-ə
 buy-part-1SG umbrella-DEF
 'the umbrella which I bought'

28

Locality

- But, in a few languages, we find a surprising non-local pattern
- Tundra Nenets (Ackerman & Malouf 2004a):

[ta-wi°] te-da
give-PART deer-3SG
'the deer s/he gave'

- Western Armenian (Ackerman & Malouf 2004b):

[im əngeroč-ə nergayatsuts-adz] dəɣa-s
1SG friend-DEF introduce-PPART boy-1SG
'the boy that I introduced to the friend'

Conclusions

- Typologically unusual patterns arise through statistically unlikely evolutionary pathways; typologically usual patterns arise through statistically likely evolutionary pathways
- Simulation allows us to test hypotheses about different evolutionary constraints (language as a complex dynamic system)
- The raw material for (this kind of) typological theorizing needs to be treebanks, from as large and diverse a set of languages as possible
- Reasons to build a treebank, even if the market for QA or MT systems is vanishingly small