QUALITATIVE AND QUANTITATIVE APPROACHES TO LANGUAGE DIVERSITY
What they can, can’t, and may be able to tell us about human (pre)history

Paul Heggarty
Linguistics, Max Planck Institute for Evolutionary Anthropology, Leipzig
Paul.Heggarty@gmail.com — http://eva-mpg.academia.edu/PaulHeggarty
LINGUISTIC DIVERSITY: WHAT DO WE MEAN?
→ = Expansions as proposed by the farming/language dispersal hypothesis.
Linguistic record of human past is not only about families and relatedness!
OR: DIVERSITY/PATTERNS IN LANGUAGE STRUCTURES — E.G. TONE

Not random — but not necessarily because of relatedness: e.g. tonogenesis.

OR: DIVERSITY/PATTERNS FROM A GENERATIVIST PERSPECTIVE?

Pino and friends, audience: you tell me...
LINGUISTIC AREAS: CONVERGENCE OUT OF LINEAGE DIVERSITY

Not rare — just *some* examples.
DIVERSITY WITHIN MAJOR FAMILIES: TREELIKE OR CONTINUOUS?

Indic, Arabic, ‘Chinese’, Bantu, Mayan, Quechua, Algonquian, Italy, Scandinavia, Switzerland (formerly much more of Europe).
DIVERSITY LINGUISTICS: RETROSPECT AND PROSPECT

1-3 May 2015 (Leipzig, Germany)

Closing conference of the Department of Linguistics at the Max Planck Institute for Evolutionary Anthropology

This conference is the final event taking place at MPI EVA’s Department of Linguistics, which will close at the end of May 2015 with Bernard Comrie’s retirement. To celebrate eighteen years of research on the diversity of human languages, we invited all present and former members and guests of the department to submit an abstract.

The conference thus reflects current activities within the broad area of “diversity linguistics”: research on little-studied languages, language typology and universals, comparative/historical linguistics, and evolutionary linguistics. It will provide a representative overview of past achievements and future prospects of research within the various subfields of diversity linguistics.
PATTERNS ON DIFFERENT DIMENSIONS: OVERLAPS & CONTRASTS

DIVERGENT LANGUAGE FAMILIES … LANGUAGE CONVERGENCE AREAS

↑ Güldemann (2010):
“Sprachraum” and geography: linguistic macro-areas in Africa
WHEN DIVERSITY IS NOT DIVERSITY

— EXAMPLE PROBLEMS
**PHONEMIC ‘DIVERSITY’?**

**OUT OF AFRICA ...**

STRUCTURE AND RATE OF CHANGE

- Phoneme inventories and how fast they change: e.g. over 2000 years.

<table>
<thead>
<tr>
<th>Language</th>
<th>Vowels</th>
<th>Consonants</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latin</td>
<td>10*</td>
<td>17</td>
<td>27</td>
</tr>
<tr>
<td>Romanian</td>
<td>7</td>
<td>20</td>
<td>27</td>
</tr>
<tr>
<td>Italian</td>
<td>7</td>
<td>23</td>
<td>30</td>
</tr>
<tr>
<td>French</td>
<td>16</td>
<td>21</td>
<td>37</td>
</tr>
<tr>
<td>Catalan</td>
<td>8</td>
<td>25</td>
<td>33</td>
</tr>
<tr>
<td>Spanish</td>
<td>5</td>
<td>19</td>
<td>24</td>
</tr>
<tr>
<td>Portuguese:</td>
<td>14</td>
<td>23</td>
<td>37</td>
</tr>
<tr>
<td>Portugal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portuguese:</td>
<td>16</td>
<td>23</td>
<td>39</td>
</tr>
<tr>
<td>Brazil</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Counts fall or — mostly — rise.

- Founder effects model not applicable. (Not ‘diversity’ but counts.)
**MISAPPLYING FOUNDER EFFECT LOGIC**

- The phonemes of a language are not a ‘phoneme pool’.
- **Counts** are not “diversity” = limb ‘diversity’, ‘limb pool’?

**THE ‘FOUNDING’ OF AMERICAN ENGLISH?**

We hold these truths to be self-evident, that all men are created equal ...

\[ \downarrow \]

founder effect

\[ \downarrow \]

Wi huld fisi drufs tu bi silf-ividind, vat all min ari criadid igual ...

\[ \downarrow \]

founder effect

\[ \downarrow \]

DNA

ACGT

\[ \downarrow \]

founder effect

\[ \downarrow \]

A-G-
• Known histories: language here almost all replaced in last 4000 years.

• Pattern is in reverse:
  – Bantu picks up more phonemes by contact as wave of advance spreads.
OUT OF AFRICA ... OR OUT OF EASTER ISLAND?

• Cysouw et al. (2012) on results in Atkinson (2011):
  an artefact of ... suboptimal data, biased methodology, unjustified assumptions.
ANYONE FOR FISHING?

Everett (2013): Evidence for direct geographic influences on linguistic sounds: the case of ejectives

57 of 92 (62%) languages with ejectives are located in high elevation ‘zones’, which are defined here as major regions greater than 1500 m in altitude, plus land within 200 km of such a region of high altitude.

Creanza et al. (2015): A comparison of worldwide phonemic and genetic variation

Everett et al. (2015): Climate, vocal folds, and tonal languages: connecting the physiological and geographic dots
LANGUAGE DIVERSITY DATABASES: QUALITATIVE VS. QUANTITATIVE?
PUTTING MEANINGFUL NUMBERS ON LANGUAGE?


Maddieson (2013: WALS 2a): Vowel Quality Inventories
QUANTIFICATION, RULE 1: **DO NOT ‘BIN’ CONTINUOUS DATA**

“Due to uncertainty in ascertaining exact inventory counts across languages, the WALS data are binned into ranges for:

- **Vowel:** small [2-4] medium [5-6] large [7-14]
- **Consonant:** small [6-14] mod. small [15-18] average [19-25] mod. large [26-33] large [34+]
- **Tone:** no tone simple tone complex tone ...

... diversity.”

Atkinson (2011: SI 2)
Maddieson (2013: WALS 2a): Vowel Quality Inventories
WHEN A PHONEME INVENTORY IS NOT A PHONEME INVENTORY...

• Vowel quality — not phoneme — inventories.

2. Establishing the values

When vowel qualities are counted in this way in the sample of languages surveyed for this chapter, the average number of vowels in a language is just fractionally below 6. The smallest vowel quality inventory recorded is 2 and the largest 14.

Only one language in the sample, German, uses 14 vowel qualities and only 2 make use of 13, namely the variety of British English included here and Bété.

– Long and short variants of the same vowel are always counted once

– Nasalized vowels do not add to the inventory as long as a non-nasalized counterpart occurs, and so on.

Maddieson (2013: WALS 2a)
**WHEN 7 = 13, BUT NOT 5**

- **Spanish:** \(5 = \text{white}\)
- **Latin:** \(5 = \text{white}\)  
  \((5 \text{ long } + 5 \text{ short})\)
- **Italian:** \(7 = \text{red}\)  
  \((5 \text{ basic, } + 2 /ɛ/ /ɔ/ \text{ if stressed})\)
- **English:** \(13 = \text{red}\)
- **German:** \(14 = \text{red}\)
- **[red] = [red]** \(i.e. 7 = 13 = 14\)
- **[red] ≠ [white]** \(i.e. 7 ≠ 5\)

So in vowel quality inventory, Italian is ...

- Identical to English, German, the most extreme of all languages in sample.
- Completely different to Spanish, Latin, just on either side of mean (6).
QUALITATIVE OR QUANTITATIVE?


Comrie (2013: WALS 98a): Alignment of Case Marking of Full Noun Phrases
But the main recurrent difficulty is that in many languages, different kinds of full noun phrases partake of different case marking patterns. For instance, in Spanish the accusative marker, the preposition a, is found (roughly) only with specific, animate noun phrases, so that strictly speaking a noun phrase like the male proper name Juan has a nominative–accusative case marking system, while the inanimate noun phase el libro ‘the book’ has a neutral case marking system, as illustrated partially in (7).

(7) Spanish

a. María vio a Juan.
   Mary see.AOR.3SG ACC John
   ‘Mary saw John.’

b. María vio el libro.
   Mary see.AOR.3SG the book
   ‘Mary saw the book.’

‘Qualitative’ justification ↑...
... but quantitative chaos:

A. The policy that has been followed in assigning such languages to types has been to maximize the occurrence of overt case marking. Thus, if a language has an optional accusative case marker, or one that occurs only under certain specified circumstances, then this has been given priority and taken as critical. This policy decision needs to be taken into account consistently in interpreting the maps. (For details on how decisions were taken for individual languages, reference should be made to the electronic version of this atlas.) Thus, Spanish and Burmese come out as accusative, Araona and Goonyandi as ergative, and Hindi as tripartite.

• “Maximise ... priority ... critical.” Any = All. 1% = 100% 0.01 = 1.
  – 0.01 is closer to 1 than to 0. 0.01 is 1.

• A meaningful representation of language data in numbers?

• An atlas for display purposes ... not a database for quantitative purposes.
QUALITATIVE OR QUANTITATIVE?

Atlas of Pidgin and Creole Language Structures — APiCS
www.apics-online.info — Haspelmath et al. (2013: APiCS 19)
19 Interrogative pronouns

Here we look at the four interrogative pronouns ‘who’, ‘where’, ‘when’, and ‘how’ and ask whether they are expressed as simple, monomorphemic words as in the European lexifiers, or as compound expressions consisting of a generic noun and an adnominal interrogative word.

Compound interrogatives translate literally as 'which person', 'what place', 'what hour', 'what manner' or similar. From the first through the last value, languages show an increasing amount of compound expressions, and thus an increasing degree of distance from the European lexifiers.

Two languages A and B, entirely opposite values on all four data points:

- Language A: who  where  [what time]  [what way] = ✓ ✓ ✗ ✗
- Language B: [what person] [what place] when  how = ✗ ✗ ✓ ✓ ✓

→ Languages A and B = ‘two compound expressions’ = orange, identical.
OTHER PROBLEMS WITH WALS FOR REFERENCE, QUANTITATIVE PURPOSES...

- **Classification into WALS ‘families’:**
  - e.g. *Khoisan, *Altaic, *Australian, *Nilo-Saharan, etc...
  - Very controversial, non-entities, faces in the fire?

- **Coverage:**
  - Sparse: on average, c. 200 languages per feature = under 3%.
  - Inconsistent: languages covered varies widely from feature to feature.
**Why Linguistic Diversity?**

**What Shaped Linguistic Patterns?**
DADDY, WHERE DO LANGUAGE FAMILIES COME FROM?
### SOUND CHANGES IN VARIOUS REGIONS

(in approximate order)

**INPUT:** The *same original form* in all regions

<table>
<thead>
<tr>
<th>Time</th>
<th>Bucharest</th>
<th>Florence</th>
<th>Lisbon</th>
<th>Madrid</th>
<th>Barcelona</th>
<th>Paris</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
</tr>
<tr>
<td>2</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
</tr>
<tr>
<td>3</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
</tr>
<tr>
<td>4</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
</tr>
<tr>
<td>5</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
</tr>
<tr>
<td>6</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
</tr>
<tr>
<td>7</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
</tr>
<tr>
<td>8</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
</tr>
<tr>
<td>9</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
</tr>
<tr>
<td>10</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
</tr>
<tr>
<td>11</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
<td>[sɛptɛ]</td>
</tr>
</tbody>
</table>

**OUTPUT:** *different descendant forms* from one region to

<table>
<thead>
<tr>
<th></th>
<th>[ʃapte]</th>
<th>[ʃete]</th>
<th>[ʃetə]</th>
<th>[ʃete]</th>
<th>[ʃete]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[ʃapte]</td>
<td>[ʃete]</td>
<td>[ʃetə]</td>
<td>[ʃete]</td>
<td>[ʃetə]</td>
</tr>
<tr>
<td></td>
<td>[ʃapte]</td>
<td>[ʃete]</td>
<td>[ʃetə]</td>
<td>[ʃete]</td>
<td>[ʃetə]</td>
</tr>
<tr>
<td></td>
<td>[ʃapte]</td>
<td>[ʃete]</td>
<td>[ʃetə]</td>
<td>[ʃete]</td>
<td>[ʃetə]</td>
</tr>
<tr>
<td></td>
<td>[ʃapte]</td>
<td>[ʃete]</td>
<td>[ʃetə]</td>
<td>[ʃete]</td>
<td>[ʃetə]</td>
</tr>
<tr>
<td></td>
<td>[ʃapte]</td>
<td>[ʃete]</td>
<td>[ʃetə]</td>
<td>[ʃete]</td>
<td>[ʃetə]</td>
</tr>
</tbody>
</table>
WHAT DO LANGUAGE FAMILIES MEAN?

TODAY

Language families do not happen by chance, for no reason.
**TALKING ‘IN ROMAN’: ROMAN(I)CĒ ...**

- Romansch, Romanian, ‘Romance’ languages ... = ‘Neo-Latin’.

(N.B. All Roman ... but not all Romance: language clues to the past.)
PRINCIPLE: **LINGUISTIC EFFECTS ← REAL-WORLD CAUSES**

Heggarty & Renfrew (2014: 19-21)

**Language families only exist because of powerful expansive processes.**

[= Linguistic evidence of past processes impacting on populations and cultures.]
EXPANSIVE PROCESSES: BUT WHICH?

NOT JUST FAMILIES …

Traditional ‘family preference’, but …

…. linguistics has far more to say on human origins and interactions.

On all other ‘diversity dimensions’ of linguistic panorama.
DIFFERENT EFFECTS ← DIFFERENT CAUSES

General principle, applies to patterns in all dimensions of linguistic panorama.

Heggarty & Renfrew (2014: 21-24); Heggarty (2014)

LANGUAGE FAMILIES
← expansive, divergent processes

LANGUAGE AREAS
← convergent processes.

Clear-cut: Member of family, yes or no?

Diffuse: core vs. peripheral members.
WHICH PROCESSES — WHICH LINGUISTIC EFFECTS?

‘Family preference’: all processes explain only families, divergence?

No prehistory of linguistic convergence areas, of diversity hotspots.

Some processes / contexts invoked for (divergent) families are in fact a more natural explanation for convergent areas instead...
Archaeological Patterns: Core vs. Periphery

• D’Altroy (2014: 9) *The Incas* — on theories of empires:

Over the last few decades, the most widely used approach in anthropology and history divides empires into their core and periphery.

Neolithic Animal Domesticates

Gothic Architecture!
GENETIC PATTERNS: CLINES AND DIFFUSION

Balaresque et al. (2010)
[Paternal lineage — modern DNA] →

Haak et al. (2010)  [Maternal lineage — ancient vs. modern DNA]
HOW THE PATTERNS RELATE...

FAMILIES AND ‘ANTI-FAMILIES’?
PATTERNS ON DIFFERENT DIMENSIONS: OVERLAPS & CONTRASTS

DIVERGENT LANGUAGE FAMILIES ... LANGUAGE CONVERGENCE AREAS

↑ Güldemann (2010):
“Sprachraum” and geography: linguistic macro-areas in Africa
**COMPLEMENTARY DISTRIBUTION: FAMILIES VS. DIVERSITY HOTSPOTS**

- Old World vs. New?
- Background diversity arisen since first settlement.
- ‘Neolithic Revolution’: → Farming/language dispersals?
- Diversity hotspots? = Where the great families just didn’t reach?
- (But many complexities …)
### Families vs. Areas: Opposites?

<table>
<thead>
<tr>
<th></th>
<th>Language Families</th>
<th>Language Areas</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source(s)</strong></td>
<td>single, common</td>
<td>multiple, independent</td>
</tr>
<tr>
<td><strong>= Origins</strong></td>
<td>homogeneity</td>
<td>diversity</td>
</tr>
<tr>
<td><strong>Structures ...</strong></td>
<td>homogeneity → diversity</td>
<td>diversity → homogeneity</td>
</tr>
<tr>
<td><strong>Process</strong></td>
<td>divergence</td>
<td>convergence</td>
</tr>
</tbody>
</table>

**Images:**
- Map of Romance languages in Europe.
- Map of Southeast Asia.
ONE FAMILY, MORE THAN ONE AREA: TIBETO-BURMAN

• One family, straddled over two contrasting convergence areas:
  – Matisoff’s (1990: 113) “Sinosphere” vs. “Indosphere”.

• Tibeto-Burman languages diverge:
  – Some to Sinosphere … others to Indosphere characteristics.
ONE AREA, MORE THAN ONE FAMILY: ‘BALKANISATION’

Change towards the characteristics of the area
= change away from related languages not in the area.
HOW AREAL CONVERGENCE CREATES FAMILY DIVERGENCE

Muysken et al. “Traces of Contact” Project

Structural features: subordination, argument marking, noun phrase, TAME
FAMILY VS. CONVERGENCE? A CASE-STUDY

• Some isolates + several small families, widely (but thinly) spread:
  – Uralic, Turkic, Mongolic, Tungusic.

‘Altaic’: divergent macro-family — or convergence area?
PATTERNS AND CAUSATION

• **CORE** vs. **PERIPHERY**
  Altai vs. Uralic, Korean, Japanese.
  = Typical of convergence areas.

• ‘Mobility’, nomadism, very low population density …
  → Family expansions, divergence, ‘spread zone’?
  Or → Intense long-range contact → convergence (Steppe ‘confederations’).

☒ A diverging ‘Altaic’ family. ☑ A North Eurasian convergence area.
**Language Structures and the Holy Grail**

- ‘Ultra-stable’ structures / parameters → reveal deepest families, prehistory?

↓ Phylogeny of Austronesian

*Gray et al. (2009)*

Structural isoglosses within Austronesian

*Donohue & Denham (2010)*
MASS LANGUAGE SHIFT: STRUCTURES MORE STABLE THAN FAMILIES

• The same deep structural features:
  
  – Resistant to internal change:
    → genealogically most stable — but only if transmission is vertical!
  
  – Resistant even through language shift, carried over into new language:
    → genealogically least stable in mass language shift.

  = ‘Stable’ in speaker population, even when they switch lineage.

• Features so structurally stable ... they are ‘genealogically unstable’!
  → Less diagnostic of deep genealogy than ‘Austronesian’ lexis!
How To Work It All Out?
QUALITATIVE + QUANTITATIVE: ‘GLOTTOBANK’?

New databases: world-scale, and specifically for quantitative applications...

- Grambank                      Harald Hammarström
- Lexibank                      Simon Greenhill & Russell Gray
- Phonobank                     Mattis List
- IElex and URAlex              Michael Dunn
- Paradigms and deep signal     Nick Evans

- New models: constantly refined to get closer to how languages really work.
- New co-operation: with ancient DNA, archaeological science...
WHAT PLACE FOR LINGUISTICS
IN THE ‘SCIENCE OF HUMAN HISTORY’?
Some papers on these themes:

http://eva-mpg.academia.edu/PaulHeggarty

Paul.Heggarty@gmail.com


http://dx.doi.org/10.1126/science.1166858

http://dx.doi.org/10.1371/journal.pbio.1000536


http://apics-online.info/parameters/19

www.routledge.com/books/details/9780415527897.

www.cambridge.org/ec/academic/subjects/archaeology/prehistory/cambridge-world-prehistory.