Relationships between Genetic and Linguistic Diversities: The Case of Tone

MRI Scans of a microcephalic (left, 13 yo) and a normal (right, 11 yo) brains (Bond et al., Nature Genetics 32: 316-320, 2002)

World map of tonality: no tone, simple and complex tones. Generated with WALS (Haspelmath et al., 2005)

International Society for Human Ethology (ISHE)
Human Ethology Summer School
Max-Planck-Institute, Andechs

Dan Dediu
dan.dediu@ed.ac.uk

D. Robert Ladd
bob@ling.ed.ac.uk

LINGUISTICS + ENGLISH LANGUAGE

Linguistics and English Language
University of Edinburgh
The Genetics of Interindividual Differences for Language

**Behaviour genetics** applied to speech and language:

**Heritability** → important caveats in interpretation
- various aspects of language (abilities and disabilities): important heriabilities → genes play an important role
Genetic correlations → both generalist and specialist genes
Pathologies – quantitatively different (the lower and of the distribution)
Most probable model: many genes with small effects as opposed to few genes with big effects.

**FOXP2**: developmental verbal dyspraxia (OMIM 602081), included SLI.
- three mutations: KE family, CS and another proband (MacDermot et al., 2005);
- quantitative deficit hypothesis;
- complex phenotype: articulatory problems, cognitive impairments, language impairments;
- evolution: very conserved, 2 “human specific” mutations (one also in *Carnivora*);
- age estimates: very problematic (i.e., Enard et al., 2002: 0ya, 95% CI 0-120kya);
→ function not known
Correlations between the genetic structure of populations and the languages they speak → *genetically related populations tend to speak related languages/dialects*

The language/farming co-dispersal hypothesis (Renfrew, Diamond, Cavalli-Sforza, etc.).

Such correlations are **spurious** in the (statistical) sense that they are due to a **deeper causal factor**, namely **demographic phenomena**: migrations, splits, admixture, separation by distance/topography/ecology, etc.

→ the correlations are accidental (from the genes and languages' points of view).
Non-spurious Correlations Between Genetic and Linguistic Population Diversities

The arguments *pro*:
- **evolutionary theory**:
  - language evolution involved *genetic changes*
  - evolution requires *variation*
- **inter-individual variation**:
  - *genetic differences* for language
- **population differences**:
  - *inter-population genetic diversity*, also for coding genes (brain ?)

The arguments *contra*:
- **linguistic communicative power uniformitarianism**:
  - all languages are *equally expressive*
- **linguistic capacity uniformitarianism**:
  - all (*normal*) humans can *learn any* $L_i$ they are exposed to

**Conclusion**:
- there are *no a priori grounds* for not having inter-population linguistic differences influenced by inter-population genetic differences
→ non-spurious correlations between genetic and linguistic diversities
A possible example: 
**ASPM, Microcephalin and Tone**

Two human genes involved in brain growth and development:
- **ASPM** (Abnormal SPindle-like, Microcephaly-associated/MCPH5, OMIM 605481, 1q31)
- **Microcephalin** (MCPH1, OMIM 607117, 8p23)

Homozygous deleterious mutations → primary microcephaly

Exact function(s): unknown:
- **ASPM**: involved in cell cycle regulation; mitotic spindle poles orientation in neuroepithelial cells → number of asymmetric divisions (?)
- **Microcephalin**: DNA damage response; regulator of chromosome condensation → neural progenitor death (?)

Accelerated evolution in the lineage leading to humans.

Recently (Mekel-Bobrov et al., 2005; Evans et al., 2005), derived haplogroups (ASPM-D, MCPH-D):
- signs of positive natural selection
- geographic structure
- phenotypic effect(s): unknown, but seemingly do not concern:
  - incidence of schizophrenia (Rivero et al., 2006)
  - variation in intelligence (IQ) (Mekel-Bobrov et al., 2007; Rushton et al., 2007)
  - normal brain/head size variation (Woods et al., 2006; Rushton et al., 2007)
  - prosocial attitudes/altruism (Rushton et al., 2007)
Tone languages: Structure and Geography

Pitch is organized into tone phonemes, and is an integral part of the phonological form of morphemes.

Tone phonemes may be levels (High, Low, etc.) or contours (Rise, Fall, etc.); in many tone languages contours are clearly phonologically composite (e.g. Rise = Low High).

Morphemes may consist only of a toneme; Grammatical functions: just by tone changes.

Canonical tone languages: every syllable has a tone phoneme in addition to segmental phonemes.

Tone languages:
- sub-Saharan Africa;
- East and Southeast Asia;
- Central America/Caribbean/Amazonia.

Yoruba tone contrasts
- ìgbá (M-H) ‘calabash’
- ègbá (L-H) ‘[type of tree]’
- ìgbà (M-M) ‘200’
- ègbà (L-L) ‘time, period’

Dinka grammatical tone
- mòc a lé:y man INFL challenge-2sg “you are challenging the man”
- mòc a lè:y man INFL challenge-PASSIVE “the man is being challenged”
The Geographic Distribution of ASPM, Microcephalin and Tone

ASPM, Microcephalin and Tone: Database and Results

- **49** Old World populations
- **981** genetic markers from public databases: ALFRED (Rajeevan et al., 2003; Osier et al., 2002) and Human Diversity Panel Genotypes (Rosenberg et al., 2002)
- **26** binary linguistic features covering many aspects of linguistics (Hapemath et al., 2005 + many, many others)

**Results:**
- \( r_{ASPM,\text{Tone}} = -0.53, r_{MCPH,\text{Tone}} = -0.54, p < 0.05 \) and in the top 1.4% of the sample
- Logistic regression: Tone on ASPM & Microcephalin: \( p < 0.05 \), Nagelkerke's \( R^2 = 0.53 \), 73% correct classification, in the top 2.7% of the sample.

**Mantel correlations** between distances: geographic, genetic (Nei, 1972), feature linguistic (Euclidean) & historical linguistic (Nettle & Harriss, 2003): (ASPM, MCPH) & Tone:
- controlling geography:
  \( r = 0.291, p = 0.003 \)
- and also controlling hist. ling.:
  \( r = 0.283, p = 0.000 \)
**ASPM, Microcephalin and Tone: Conclusions**

The correlation between Tone, *ASPM* and *MCPH* holds, even after controlling geographical and historical linguistic factors → supports the claim that it is non-spurious.

*Americas* as expected.  
*Australia* – very important.  
*India* – also potentially interesting.  
*PNG* – probably contaminated.

**Future work:**  
- more data/better stats  
- computer modeling (data analysis)  
- experimental work (soon)
The **Mechanism**: From Individual Genetic Biases to Language Change

1. **Genetics**
   - Ontogeny/environmental modulation
   - **Individual linguistic bias**
   - Population
   - Language transmission
   - Language change

1: influences on:
- production (/r/ vs. /ɪ/)
- perception (F0/VOT disambiguation)
- temporal resolution
- phonological working memory, etc.

2: acquisition:
- \( L_1 \) (ILM, etc.)
- \( L_2 \) (language shift; Ostler, 2005, etc.)
So, *What does this *Mean*?*

Inter-population linguistic/cultural differences are influenced by genetic differences.

There are linguistic/cultural aspects very **stable** in time.

Supports a **gradual, accretionary** model for language evolution.

Could **non-tonality** be the derived state?

**Language** = a *mosaic* of very old, old, recent and still evolving aspects.

What this does **not** mean:
- that there are genes “for Chinese”;
- that “races” are “not equal”;
- that there are genes which “allow”/”forbid” you to learn Chinese;
- that there is “genetic determinism” in language acquisition/learning;
- etc.
Further Info and Acknowledgements


Summary & further information: [http://www.ling.ed.ac.uk/~s0340638/tonegenes/tonegenessummary.html](http://www.ling.ed.ac.uk/~s0340638/tonegenes/tonegenessummary.html)

**We thank:**
A. Dima for help with statistics;
J. Hurford, S. Kirby, R. McMahon, D. Nettle, S. Della Sala, T. Bates, and P. Wong for discussions and comments.

ORS,
University of Edinburgh,
Leverhulme Trust,
ESRC
Example:
Some Diary Data
From a child acquiring English, around the “50-word stage” (handout).

Minimal pairs/sets involving pitch and/or voice quality, ~16 months (persisted 3-4 months):

<table>
<thead>
<tr>
<th>[ka] (high-pitched, breathy)</th>
<th>‘cat’</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ka] (modal)</td>
<td>‘bottle/cup/milk’</td>
</tr>
<tr>
<td>[ka] (low-pitched, pressed)</td>
<td>‘duck’</td>
</tr>
<tr>
<td>[a] (modal)</td>
<td>‘rabbit’ or ‘hummus’</td>
</tr>
<tr>
<td>[a] (low-pitched, pressed)</td>
<td>‘dog’, later also ‘tiger’</td>
</tr>
<tr>
<td>[a] (voiceless)</td>
<td>‘hot’</td>
</tr>
</tbody>
</table>

At ~17 months a few other such contrasts were added (lasted for 1-2 months):

<table>
<thead>
<tr>
<th>[ti] (high-pitched, breathy)</th>
<th>‘teddy’</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ti] (modal)</td>
<td>‘teeth’ or ‘chair’ or ‘juice’</td>
</tr>
<tr>
<td>[kika] (modal)</td>
<td>‘scaffolding’</td>
</tr>
<tr>
<td>[kika] (voiceless)</td>
<td>‘leopard’</td>
</tr>
</tbody>
</table>

Note also isolated words:

<table>
<thead>
<tr>
<th>[nənənə] (high-pitched, breathy, throughout)</th>
<th>‘penguin’ (ca. 16 months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[kuka] (modal 1st syll., low pressed 2nd syll.)</td>
<td>‘duck book’ (ca. 16 months)</td>
</tr>
</tbody>
</table>

There is a similar anecdotal report in Crystal (1975: 148f).
Could it be that children who use non-segmental features in their early lexicon this way are biased to expect simultaneous phonemes?
Population bias for/against tone should manifest itself in patterns of tonogenesis and tone loss. Both processes are known to have happened in widely separated (and genetically different) areas of the globe: bad news for our claim?

We propose to distinguish two types of tonogenesis:

- **secondary split**
- **spontaneous reanalysis**.

Tonogenesis is usually linked to the loss of some other phonological distinction:

- loss of final syllables (e.g. Scandinavia)
- loss of voicing contrasts (e.g. SE Asia)
- loss of laryngeal contrasts (e.g. Athabaskan)

In **secondary split**, tonogenesis is a compensatory mechanism: the loss of another distinction “forces” speakers to rely on a previously allophonic difference to preserve the cue to some important grammatical or lexical distinction. Genetic biases are at least partially irrelevant.

In **spontaneous reanalysis**, tonogenesis involves a reinterpretation of the relation between acoustic cues and phonological categories. Genetic biases could be active.
Voicing-based tonogenesis as spontaneous reanalysis

The cues to the distinction between “[ba]” and “[pa]” are distributed throughout most of the syllable.

A sequential bias will favor treating the burst+VOT as one phonological event and the voiced vowel as the next. A simultaneous bias will favor treating the combination of VOT and higher/lower F0 as a phonological event simultaneous with the vowel event.

The same acoustic signal can be interpreted phonologically in two different ways; there is no need to invoke secondary split as an explanation for this kind of tonogenesis.
Pearce’s work in progress on Kera (Chadic)

Kera shows three types of stop-vowel sequences, conventionally notated as e.g. [pá], [pā] and [bà] (conventional notation devised by European linguists based on impressionistic transcription.)

But VOT and F0 covary (and substantially overlap). Tone is primary basis of 3-way distinction for most native speakers and “voicing” not relevant (or not even perceived?) except that speakers living in town and exposed to French seem to be developing a voicing distinction and reducing the tonal distinction to [±High], with subtle but measurable effects on their production.
Kera: Diagrams

Kera village men: normalized F0 and VOT

Town Kera men: normalized F0 and VOT, 5 speakers
The data: populations

- The two original papers (Mekel-Bobrov et al., 2005; Evans et al., 2005) use 59 poorly characterized world-wide populations, which we managed to identify (see handout).
- We excluded 10 populations from the analysis:
  - 5 American populations (Karitiana, Surui, Colombian, Pima & Maya) due to poor sampling of the Americas – used as a test case;
  - 4 African populations (Masai, Sandawe, Burunge & Zime) due to unavailability of genetic data;
  - Papuan due to its vagueness and very high suspicions of European contamination of the sample.

Our sample: 49 Old World populations.

Australia: unfortunately missing but very important for our hypothesis – Australian languages are strongly non-tonal.
Caveats: difficulty of obtaining genetic samples without European and East Asian admixture.
The data: genetics & linguistics

For each of our 49 populations frequency information for 981 new genetic variants was gathered from public databases:

- ALFRED (ALlele FREquency Database; Rajeevan et al., 2003; Osier et al., 2002);
- Human Diversity Panel Genotypes (Rosenberg et al., 2002).

Selection criterion: to provide information for at least 44 of the 49 populations.

For each of our 49 populations we determined the language(s) spoken (Gordon, 2005; Campbell, 2000; Haspelmath et al., 2005).

The 141 linguistic features in Haspelmath et al. (2005) were carefully inspected and 24 met the selection criteria:

1. best coverage of the 49 populations;
2. meaningful coding as a binary variable: due to statistical considerations and uniformity of coding.

2 new features added by us
→ a total of 26 binary linguistic features.

Analyzing the linguistic data

General obs.: Holm's (1979) multiple comparisons correction was systematically applied in all appropriate cases

Missing data analysis: OK. Linguistic features' distribution: mostly in the 20%-80% bounds, except NumClassifiers, TenseAspect, FrontRdV & SVWO.

Correlations between all pairs of linguistic features (Pearson's $r \equiv \Phi$ correlation coefficient):
$N = 325$, mean $= 0.012$ & sd $= 0.274$.

Only 23 correlations are significant at the 0.05 level (see handout).

Alternative codings for Papuan (0, 1 or missing data) suggest that it can be safely neglected.
Analyzing the genetic data

The correlations (Pearson's $r$) between all pairs of genetic variants fit a normal distribution:
$N = 482,653$, mean $= 0.024$ & sd $= 0.225$. 
Correlations between genetic variants & linguistic features

Pearson's $r$ (≡ Point-biserial correlation coefficient) for all pairs of genetic variants and linguistic features fit a normal distribution:

$N = 25,558$, mean $= -0.006$ & sd $= 0.218$.

The correlations:
- **Tone & ASPM**: $r = -0.53$;
- **Tone & MCPH**: $r = -0.54$,
  both significant at $p < 0.05$ (two-tailed) and in the top 1.4% of the empirical distribution.
Correlations between linguistic features & **pairs** of genetic variants

Logistic regressions of all linguistic features on all **pairs** of genetic variants \((N = 11,582,690\) converged\): heavily skewed towards poor fits.

- **The logistic regression of **tone** on \((ASPM \& MCPH)\) is very good:**
  - **Nagelkerke's \(R^2 = 0.528\),**
  - **correct classification: 73%**
  - and in the **top 2.7%** of the empirical distribution.

- **Estimates:**
  - **Intercept = 4.478, \(p = 0.015\)**
  - \(B_{ASPM} = -7.170, p = 0.010\)
  - \(B_{MCPH} = -4.952, p = 0.026\)**

**Due to NumClassifiers & SVWO (skewed & missing data)**

**Histogram of Nagelkerke \(R^2\) for all linguistic features and genetic markers**

---

\(5\%\) tail

**Tone on ASPM & MCPH**

---
Controlling for geography and history: general considerations

**General method:** compute the correlation between distance/similarity matrices – Mantel (1967) corrects the $p$-level for matrices (still Pearson's $r$, though).

$1^\text{st}$ & $2^\text{nd}$ order partial correlations – controls for one (or more) variables.

- **Geographic distances** – estimate distances on land (handout):
  - great circle distances between locations on the same continent;
  - intercontinental paths pass through connection points: Damascus, Bangkok and Fairbanks.

- **Genetic distances** – how genetically different are two populations (handout):
  - Nei's $D$ (1972), depends on genetic variants' frequencies in the two populations.

- **Linguistic (feature) distances** – how different are two languages (handout):
  - weighted Euclidean distances on the linguistic features' space;
  - three methods: (1) equal weighting of features, (2) weighting proportional to the feature's information entropy and (3) inversely proportional $\rightarrow$ equivalent.

- **Linguistic (historical) distances** – how related (historically) are two languages (handout):
  - the method of Nettle & Harriss (2003), compatible also with Poloni et al. (1997)
  - distance is: 1, same language; 2, languages in the same branch of a family; 3, languages in different branches of same family; or 4, languages not demonstrably related (using Gordon, 2005).
Controlling for geography and history: results

<table>
<thead>
<tr>
<th>Correlation Type</th>
<th>Correlation Coefficient</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geography &amp; genetics</td>
<td>$r = 0.509$</td>
<td>$0.000$</td>
</tr>
<tr>
<td>Geography &amp; linguistics</td>
<td>$r = 0.283$</td>
<td>$0.000$</td>
</tr>
<tr>
<td>Genetics &amp; linguistics</td>
<td>$r = 0.162$</td>
<td>$0.011$</td>
</tr>
<tr>
<td>Genetics &amp; linguistics, controlling geography</td>
<td>$r = 0.021$</td>
<td>$0.407$</td>
</tr>
</tbody>
</table>

→ **in general**, the correlation between genes & languages is explained by **geography**.

<table>
<thead>
<tr>
<th>Correlation Type</th>
<th>Correlation Coefficient</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geography &amp; tone</td>
<td>$r = 0.169$</td>
<td>$0.015$</td>
</tr>
<tr>
<td>Geography &amp; ASPM</td>
<td>$r = 0.074$</td>
<td>$1.000$</td>
</tr>
<tr>
<td>Geography &amp; MCPH</td>
<td>$r = 0.543$</td>
<td>$0.000$</td>
</tr>
<tr>
<td>(ASPM, MCPH) &amp; tone</td>
<td>$r = 0.333$</td>
<td>$0.000$</td>
</tr>
<tr>
<td>(ASPM, MCPH) &amp; tone, controlling geography</td>
<td>$r = 0.291$</td>
<td>$0.003$</td>
</tr>
</tbody>
</table>

→ **ASPM, MCPH and tone** correlate even after controlling for geography.

<table>
<thead>
<tr>
<th>Correlation Type</th>
<th>Correlation Coefficient</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>controlling <em>historical linguistics</em></td>
<td>$r = 0.271$</td>
<td>$0.000$</td>
</tr>
<tr>
<td>controlling <em>historical linguistics &amp; geography</em></td>
<td>$r = 0.283$</td>
<td>$0.000$</td>
</tr>
</tbody>
</table>

**Tone, ASPM & MCPH** – more similar inside than across linguistic families (independent samples t-test). The partial Mantel correlations between (ASPM, MCPH) and **tone**:

• controlling *historical linguistics*                  | $r = 0.271$              | $0.000$   |
• controlling *historical linguistics & geography*      | $r = 0.283$              | $0.000$   |

• The correlation between **ASPM, MCPH** and **tone** is **significant and important** even after controlling for geographical and historical linguistic factors.