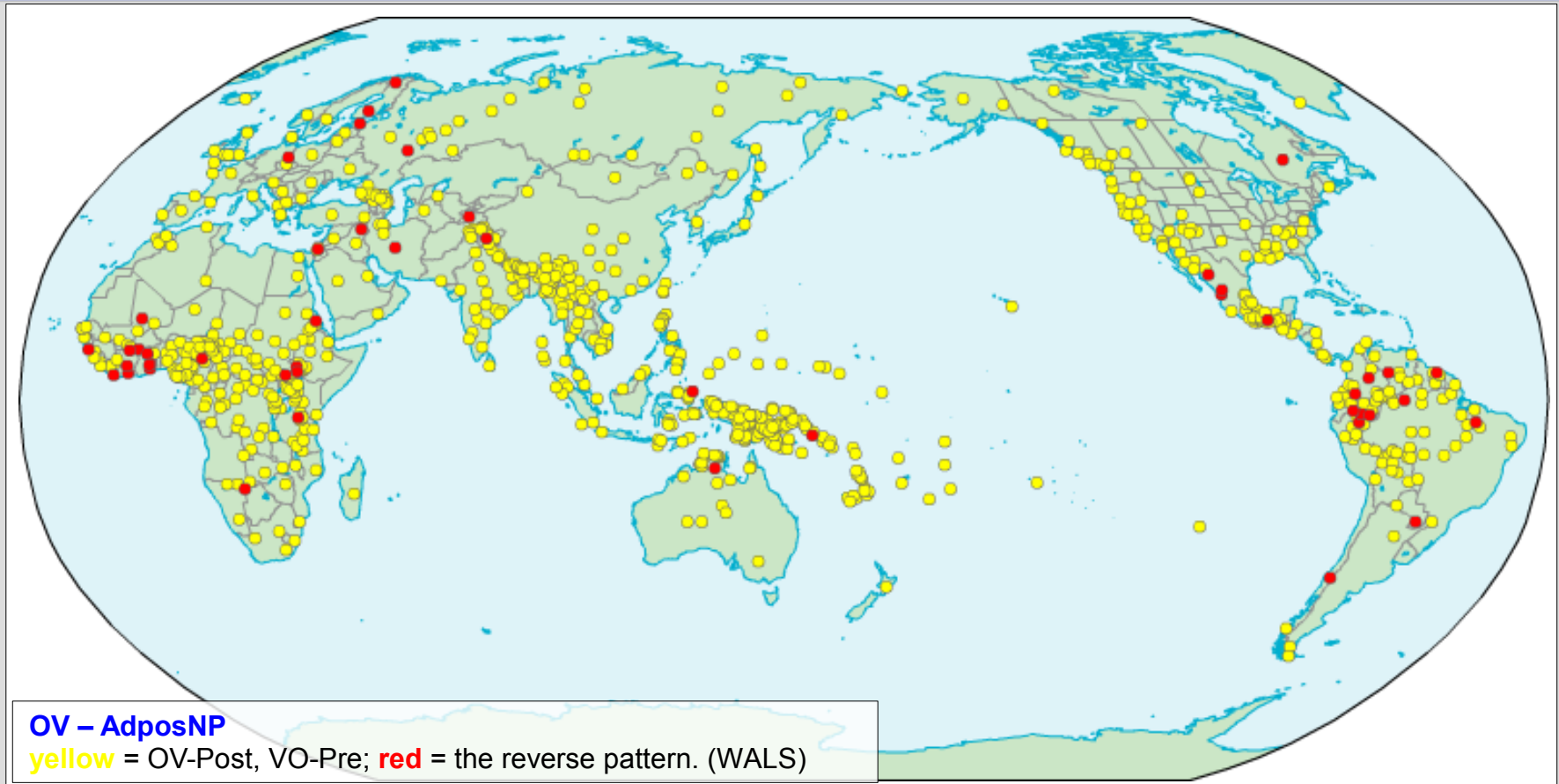


Statistical approaches to linguistic typology

delineating the roles of geography and common descent



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

Dan Dediu
dan.dediu@ed.ac.uk

Typological correlations

- **association** between two linguistic features in a large sample of languages;
- features that can be treated as **binary**:
 - syllable codas**: allowed / not allowed
 - retroflex consonants**: present / absent
 - definite article**: present / absent
 - case marking**: present / absent
 - order of verb and object**: VO / OV

Typological correlations: explanations

Historical relatedness:

- sharing through common descent

Geographical proximity:

- linguistic areas

Cognitive constraints:

- history & geography do **not** explain the correlation
- Greenberg (e.g. VO/OV and Nadj/AdjN)
- overall structural congruity and/or cognitive factors
- Hawkins; cultural evolution of language (e.g. Kirby, Hurford)
- **controversy** in the field concerning this type of explanation.

Cognitive constraints

Data are **insufficient** to settle the questions scientifically:

1. insufficient information on **historical relatedness**;
2. disagreement about **whether certain correlations really exist**:
 - e.g. headedness direction: Greenberg, Hawkins vs. Dryer (WALS) → correlation found only in the Old World?

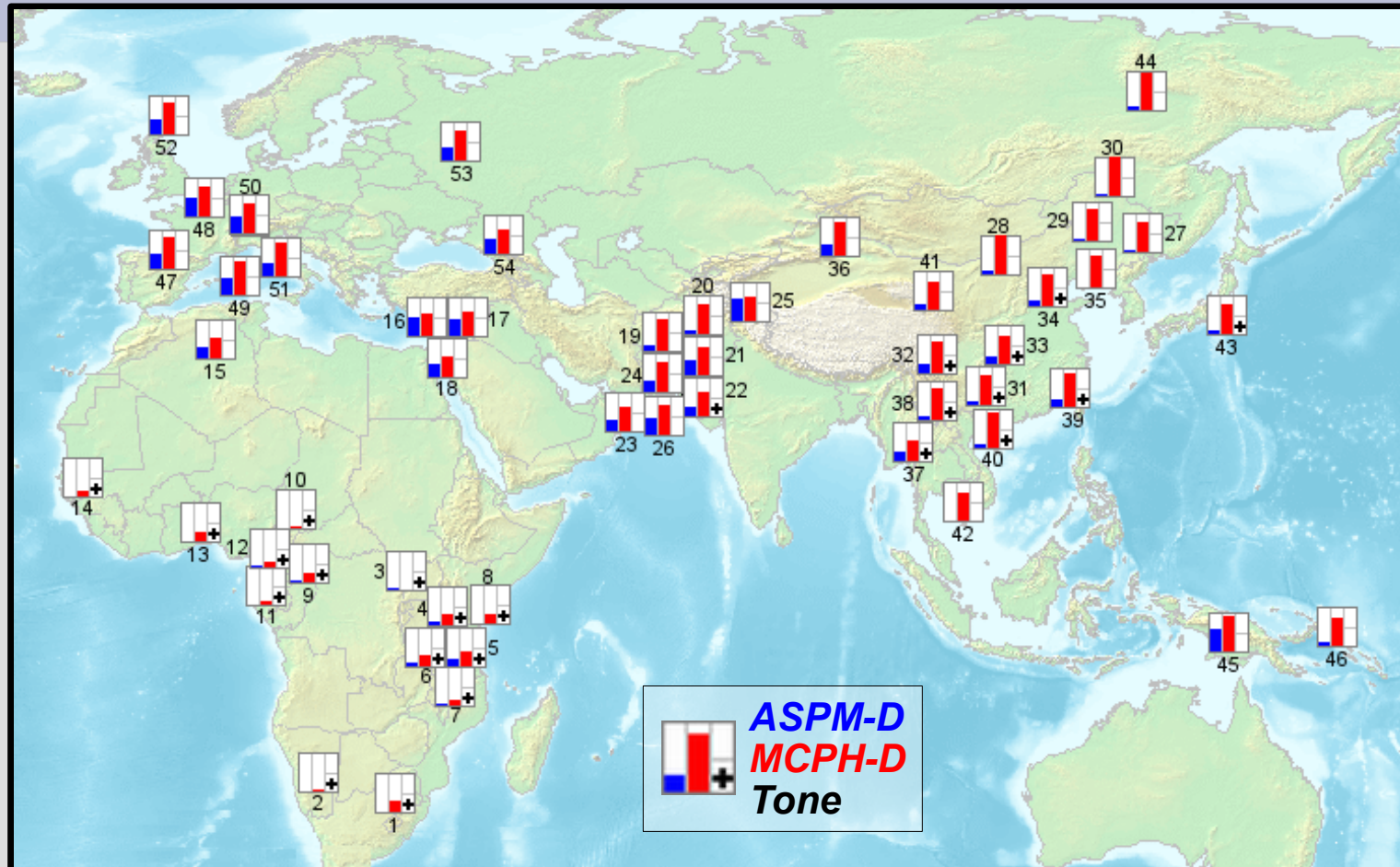
WALS: maps → **human pattern-recognition** abilities

Quantitative statistical methods:

- improvement on pattern recognition?

Background study: the relationship between *tone* and two brain-related *human genes*

(Dediu & Ladd, 2007)



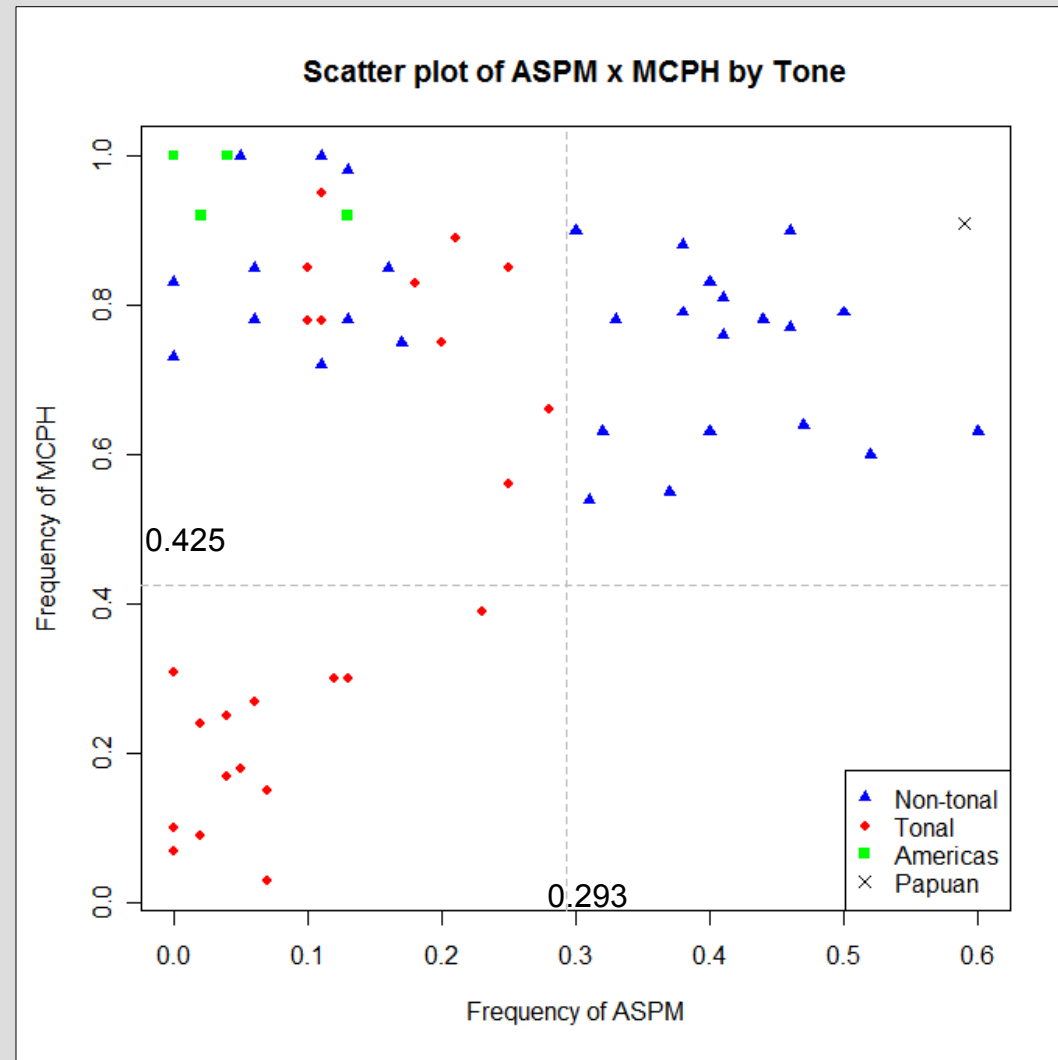
1. SE-SW Bantu, 2. San, 3. Mbuti, 4. Masai, 5. Sandawe, 6. Burunge, 7. Turu, 8. NE Bantu, 9. Biaka, 10. Zime], 11. Bakola, 12. Bamoun, 13. Yoruba, 14. Mandenka, 15. Mozabite, 16. Druze, 17. Palestinian, 18. Bedouin, 19. Hazara, 20. Balochi, 21. Pathan, 22. Burusho, 23. Makrani, 24. Brahui, 25. Kalash, 26. Sindhi, 27. Hezhen, 28. Mongola, 29. Daur, 30. Orogen, 31. Miaozi, 32. Yizu, 33. Tujia, 34. Han, 35. Xibo, 36. Uygur, 37. Dai, 38. Lahu, 39. She, 40. Naxi, 41. Tu, 42. Cambodian, 43. Japanese, 44. Yakut, 45. Papuan, 46. NAN Melanesian, 47. French Basque, 48. French, 49. Sardinian, 50. N Italian, 51. Tuscan, 52. Orcadian, 53. Russian, 54. Adygei.

Background study: database and results

- **49** Old World *populations*
- **981** *genetic markers* from public databases
- **26** *binary linguistic features*

Results:

- $r_{ASPM, Tone} = -0.53$,
- $r_{MCPH, Tone} = -0.54$, $p < 0.05$,
- **top 1.4%**
- logistic regression: $p < 0.05$,
- Nagelkerke's $R^2 = 0.53$,
- **73%** correct classification,
- **top 2.7%**
- Mantel $(ASPM, MCPH)$ vs Tone:
 $r_{geo} = 0.291$, $p = 0.003$
- $r_{geo\&hist} = 0.283$, $p = 0.000$



The proposed method

Sampling:

- our sample was predetermined by the availability of genetic data;
- future work might use (controlled) genealogical sampling (e.g., Bickel *in press*).

Coding:

- binary coding:
 - ensures uniformity and comparability;
 - powerful statistical methods.

General method:

- compute the correlations between features (Pearson's $r \equiv$ phi correlation coefficient; inferential or randomization significance/confidence intervals);
- compare the correlation(s) of interest with the entire database of correlations;
- control for geographic and linguistic effects.

The linguistic features (26)

ConsCat (0 = small, moderately small & average, 1 = moderately large or large)

VowelsCat (0 = small & average, 1 = moderately large or large)

UvularC (0 = none, 1 = uvular stops, uvular continuants or both)

GlottC (0 = no glottalized consonants, 1 = any category of glottalized consonants)

VelarNasal (0 = no velar nasal, 1 = initial velar nasal or not initial velar nasal)

FrontRdV (0 = none, 1 = high, mid or both)

Codas (0 = no codas allowed, 1 = otherwise)

OnsetClust (0 = no onset clusters allowed, 1 = otherwise)

WALSSylStr (0 = simple or moderately complex, 1 = complex)

Tone (0 = no tones, 1 = simple or complex)

RareC (0 = none, 1 = clicks, labial-velar, pharyngeals or 'th' sounds)

Affixation (0 = little affixation, 1 = strong & weak suffixing, equal suffixing and prefixing, weak & strong prefixing)

CaseAffixes (0 = yes, 1 = no case affixes or adpositional clitics)

NumClassifiers (0 = no, 1 = optional or obligatory)

TenseAspect (0 = no tense-aspect inflection, 1 = tense-aspect prefixes, suffixes, tone or mixed type)

MorphImpv (0 = no second person imperatives, 1 = second singular, plural or number-neutral)

SVWO (0 = SV, 1 = VS)

OVWO (0 = OV, 1 = VO)

AdposNP (0 = postpositions, 1 = prepositions)

GenNoun (0 = genitive-noun, 1 = noun-genitive)

AdjNoun (0 = adjective-noun, 1 = noun-adjective)

NumNoun (0 = numeral-noun, 1 = noun-numeral)

InterrPhr (0 = not initial interrogative phrase, 1 = initial interrogative phrase)

Passive (0 = absent, 1 = present)

NomLoc (0 = different (split-language), 1 = identical (share-language))

ZeroCopula (0 = impossible, 1 = possible)

Correlations between linguistic features

Correlations between pairs of linguistic features:

- **325** such pairs
- **Pearson** correlations between values
- **Mantel** correlations between linguistic distances (0 order), also controlling for:
 - **geographic proximity** (1st order partial Mantel controlling for land distances)
 - **historic relatedness** (1st order partial Mantel controlling for historical distances)
 - **geography and history** (2nd order partial Mantel)
- Holm's (1979) **multiple comparisons correction**

In general:

- Pearson's r is **much larger** (in absolute value) than Mantel's r ;
- controlling for geography, history and both **slightly decreases** Mantel's r (RareC-AdjNoun: 1st order $r = 0.014$, 2nd order (geo) $r = 0.0079$, 2nd order (hist) $r = -0.0035$, 3rd order $r = -0.0003$, all n.s.);
- they tend to **agree**¹ (high correlations and concordances);
- Pearson's p-values: **inferential and randomization agree**¹;

Correlations between linguistic features (2)

32 pairs² have at least one correlation signif. ($\alpha = 0.05$) and **23** have all signif.:

- some are “**definitional**”/”**logical**”:

Codas-WALSSylStr, OnsetClust-WALSSylStr ← **syllable structure**

Affixation-TenseAspect, Affixation-MorphImpv, TenseAspect-MorphImpv ← **morphology**

- **head position**:

OVWO-AdposNP, OVWO-GenNoun, AdposNP-GenNoun, AdposNP-AdjNoun, GenNoun-AdjNoun, OVWO-AdjNoun

- **word order & affixation**:

CaseAffixes-AdjNoun, CaseAffixes-OVWO, CaseAffixes-AdposNP ← **negative corr. (head final)**

- **tone & syllable structure**:

Tone-Codas, Tone-WALSSylStr

- some “**odd**” ones:

NumNoun-VowelsCat, **NumNoun**-Codas, **NumNoun**-WALSSylStr, **NumNoun**-Tone

NumClassifiers-Affixation, **NumClassifiers**-MorphImpv, **NumClassifiers**-TenseAspect

The relationship of linguistic features with history and geography

- **Mantel** correlation with historical linguistic distance;
- The **intra- vs. inter-linguistic family** linguistic distances.

13 features have a significant ($\alpha=0.05$) relationship with history³:

Tone, NumNoun, Affixation, WALSSylStr, NumClassifiers, TenseAspect, OVWO, AdposNP, **NomLoc**, AdjNoun, Codas, CaseAffixes, RareC.

- **Mantel** correlation with geographical (land) distance;
- **Spatial autocorrelation** (*Moran's I* and *Geary's c*);
- (Semi)**variograms**.

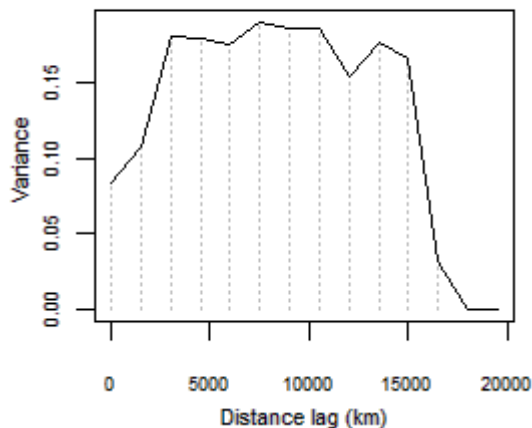
14 features have significant ($\alpha=0.05$) relationship with geography⁴:

NumClassifiers, Codas, AdjNoun, Affixation, WALSSylStr, NumNoun, RareC, TenseAspect, **MorphImpv**, OVWO, CaseAffixes, Tone, VelarNasal, AdposNP.

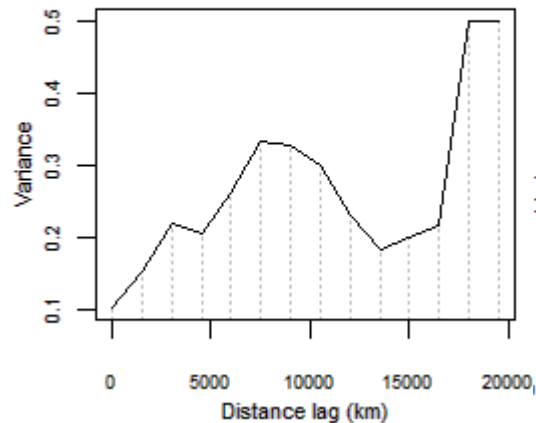
The relationship of linguistic features with geography: (Semi)variograms

1. Abrupt increase in variance, followed by a plateau (e.g., MorphImpv, GenNoun, FrontRdV, ConsCat, Affixation, AdjNoun).
2. Gradual increase in variance until a (local) maximum is reached, followed by a decrease in variance at medium scales and again followed by an increase in variance for large scales (ZeroCopula, WALSSylStr, VowelsCat, Tone, OVWO, OnsetClust, Codas, CaseAffixes, AdposNP).
3. Monotonic increase in variance with the spatial lag (Passive, NumNoun).
4. Very rugged pattern (GlottC, InterrPhr, SVWO, TenseAspect, VelarNasal).

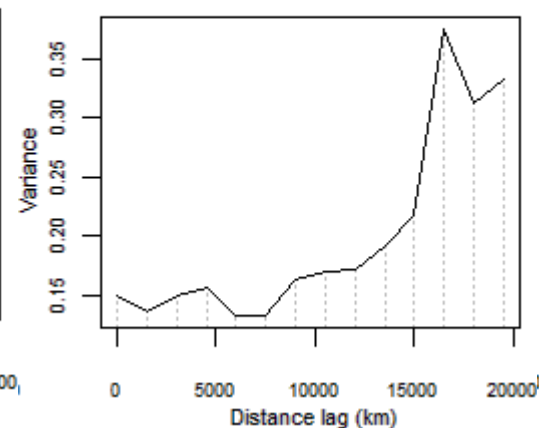
Variogram of MorphImpv
for lag increment 1500km



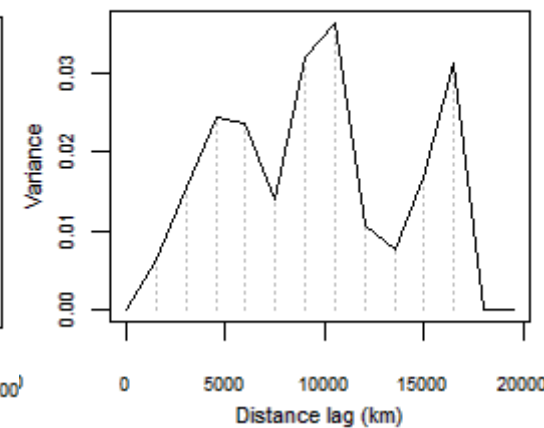
Variogram of Tone
for lag increment 1500km



Variogram of Passive
for lag increment 1500km



Variogram of SVWO
for lag increment 1500km



Conclusions

- quantitative approach to looking for relationships between linguistic features;
- techniques for studying the effects of shared history and spatial proximity;
- still in a preliminary stage;
- adaptation of techniques from spatial statistics, ecology, geostatistics, etc.

The need for larger and standardized databases, designed not only for map generation but also for quantitative approaches.

Probably better to have fewer linguistic features in more populations and with standardized coding (preferably binary or interval/scale).

Further Info and Acknowledgements

Dediu, D. & Ladd, D.R. (2007). Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, *ASPM* and *Microcephalin*, *PNAS* 104:10944-10949.

Summary & further information:

<http://www.ling.ed.ac.uk/~s0340638/tonegenes/tonegenessummary.html>

We thank:

B. Connell, C. Kutsch Lojenga, H. Eaton, J. A. Edmondson, J. Hurford, K. Bostoen, L. Ziwo, M. Blackings, N. Fabb, O. Stegen, R. Asher, R. Ridouane, M. Endl, and J. Roberts for [primary language data](#);

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**