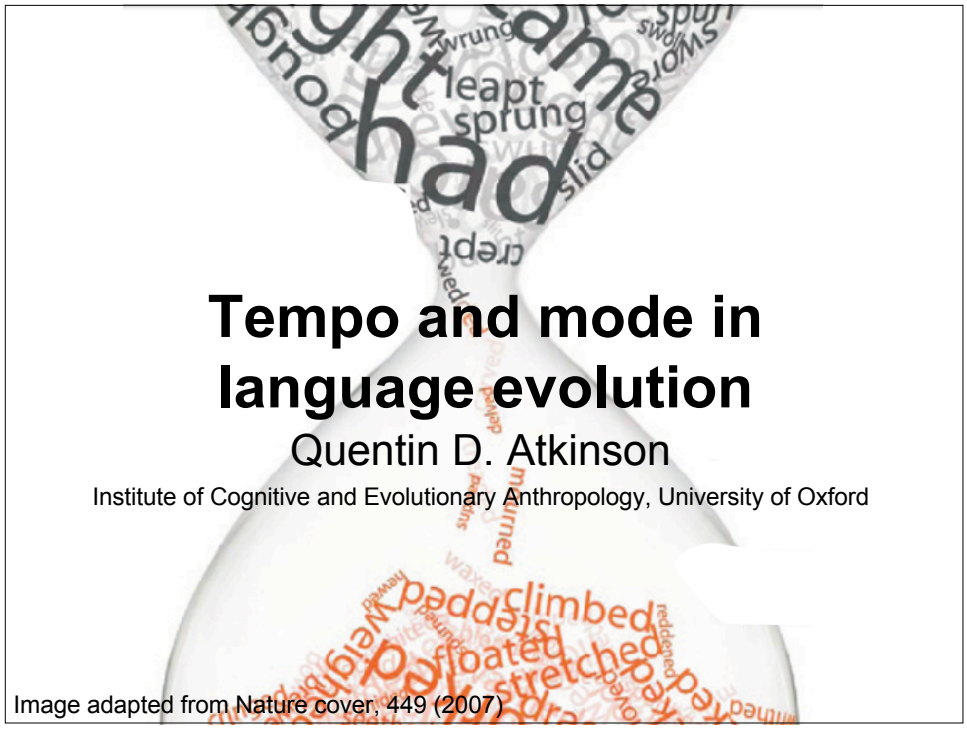


Tempo and mode in language evolution

Quentin D. Atkinson

Institute of Cognitive and Evolutionary Anthropology, University of Oxford

Image adapted from Nature cover, 449 (2007)

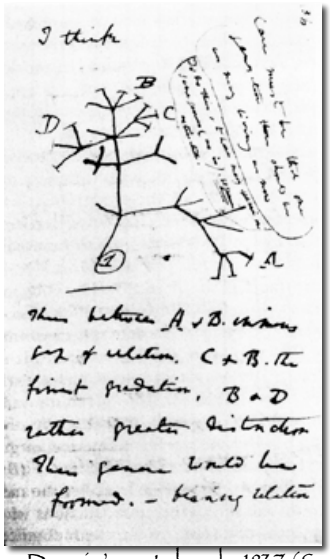
"The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel. … We find in distinct languages striking homologies due to community of descent, and analogies due to a similar process of formation"

-Charles Darwin (The Descent of Man, 1871)

“Curious Parallels”

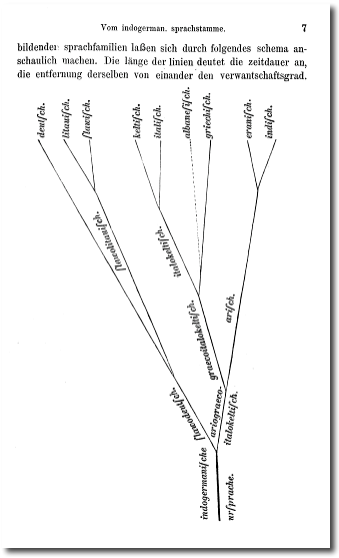
Biological Evolution	Language Evolution
Discrete heritable units – e.g. genetic code, morphology, behaviour	Discrete heritable units – e.g. lexicon, syntax, and phonology
Homology	Cognates
Mutation – e.g. Base-pair substitutions	Innovation – e.g. Sound changes
Drift	Drift
Natural selection	Social selection
Cladogenesis – e.g. allopatric speciation (geographic separation) and sympatric speciation (ecological/reproductive separation)	Lineage splits – e.g. geographical separation and social separation
Anagenesis	Change without split
Horizontal gene transfer – e.g. hybridisation	Borrowing
Plant Hybrids – e.g. wheat, strawberry	Language Creoles – e.g. Surinamese
Correlated genotypes/phenotypes – e.g. allometry, pleiotropy.	Correlated cultural terms – e.g. ‘five’ and ‘hand’.
Geographic clines	Dialects/Dialect chains
Fossils	Ancient Texts
Extinction	Language death

Tree of life



Darwin’s notebook, 1837 (Syndics of Cambridge Univ. Lib.)

Tree of languages



Schleicher, 1865

Tempo and Mode in Evolution

George Gaylord Simpson, 1944

Tempo - variation in rates of evolution and factors affecting rates of evolution

Mode - Speciation and major evolutionary transitions

"The basic problems of evolution are so broad that they cannot hopefully be attacked from the point of view of a single scientific discipline. Synthesis has become both more necessary and more difficult as evolutionary studies have become more diffuse and more specialized. Knowing more and more about less and less may mean that relationships are lost and that the grand pattern and great processes of life are overlooked."

Stochastic models of biological evolution…

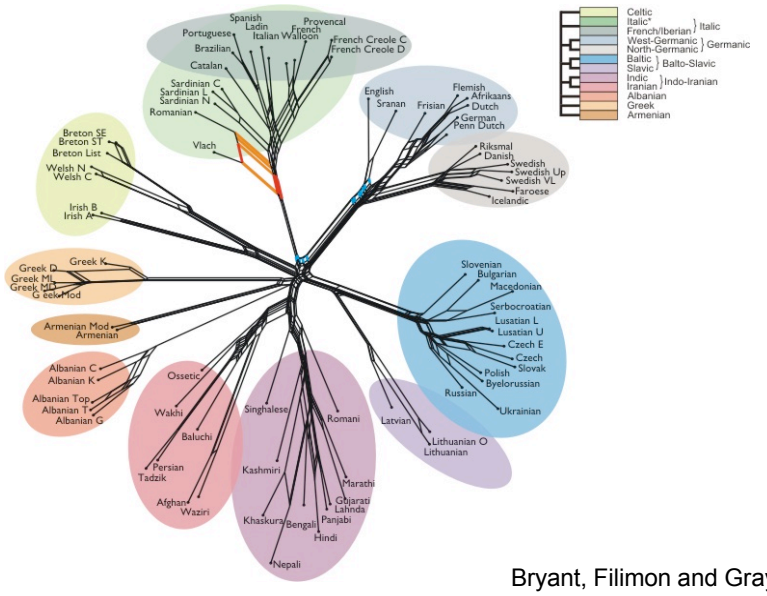
- Nucleotide and amino acid substitution, selection, migration, drift, speciation rates, lineage coalescence, phylogeny, autocorrelation within and between genes, recombination, morphological evolution, correlated evolution, population size, sex ratios, inclusive fitness, multi-level selection, frequency dependent selection, purifying selection, ancestral state reconstruction, haplotype clines, phylogeography…

Language "genes" (cognates)

English	*here*	*sea*	*water*	*when*
German	*hier*	*See, Meer*	*Wasser*	*wann*
French	*ici*	*mer*	*eau*	*quand*
Italian	*qui, qua*	*mare*	*acqua*	*quando*
Greek	*edo*	*thalasa*	*nero*	*pote*
Hittite	*ka*	*aruna-*	*watar*	*kuwapi*

Meaning	here				sea				water			when
English	1	0	0	0	1	0	0	0	1	0	0	1
German	1	0	0	0	1	1	0	0	1	0	0	1
French	0	1	0	0	0	1	0	0	0	1	0	1
Italian	0	1	0	0	0	1	0	0	0	1	0	1
Greek	0	0	1	0	0	0	1	0	0	0	1	1
Hittite	0	0	0	1	0	0	0	1	1	0	0	1

Is an evolutionary tree a good model?

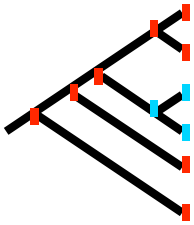


Bryant, Filimon and Gray, 2005

Tree building

- MCMC 40M iterations
  - Burnin 2.5M iterations
  - Posterior distribution of 1000 trees
- 2 state, time-reversible model in BayesPhylogenies

	0	1
0	$-u\pi_1$	$u\pi_1$
1	$u\pi_0$	$-u\pi_0$



- gamma distributed rates across sites

The Indo-European Language Family Tree

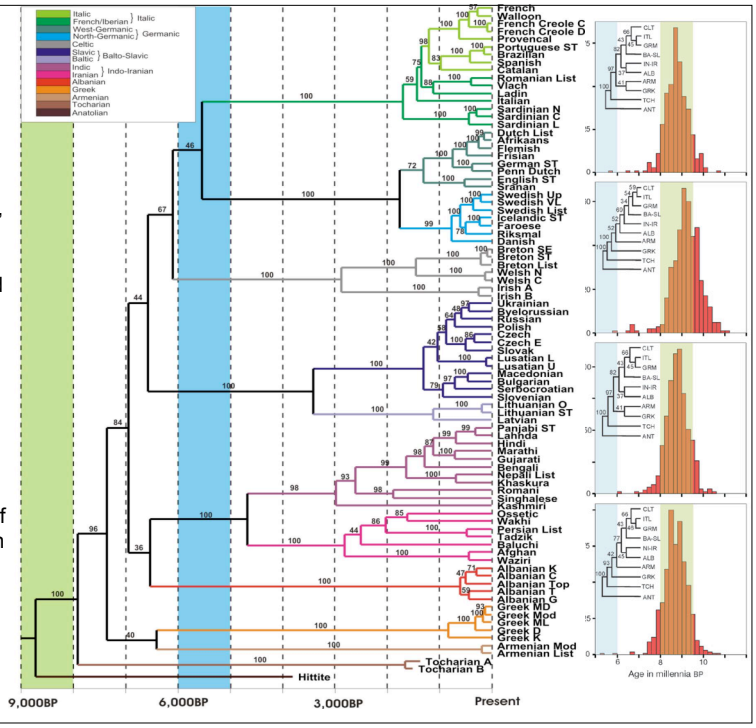
Gray and Atkinson, *Nature*, 2003

Swadesh 200 word list
basic vocabulary terms
E.g. kinship terms, body parts, numbers
2449 cognate sets

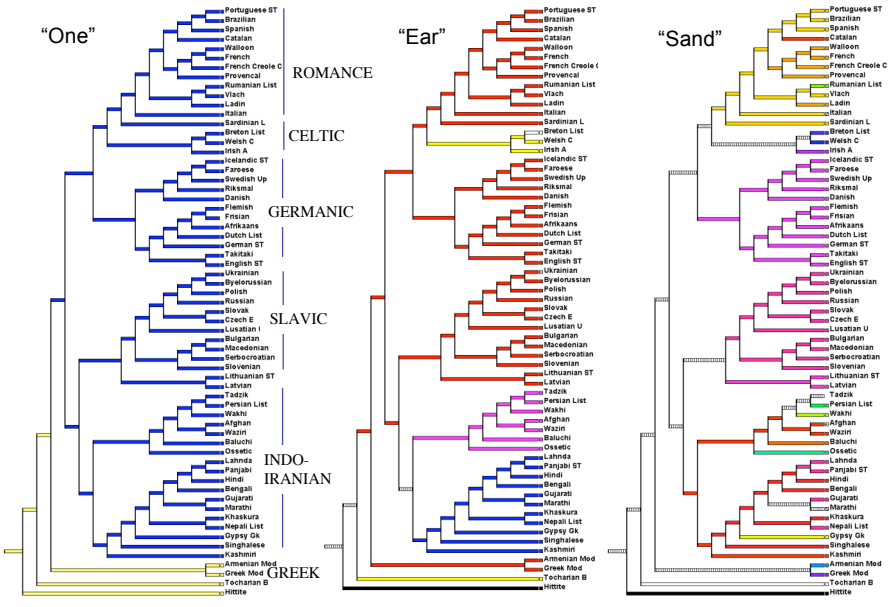
Likelihood model of cognate birth/death

Branch-lengths = time

Phylogenetic uncertainty



I-E tree showing variation in rates of lexical replacement



Some examples of meanings with small and large numbers of cognate sets

Cognate sets	Examples
1	two, three, five, I, who
2	**one**, four, we
3	how
4	name, tongue
6	**ear**, night, thou
10	day, to live, mother, salt, when
27	bark (of a tree), to count, to dig, to float, to flow, if, rub, **sand**, straight, woods
46	dirty (the most variable word)

Coding the cognate data

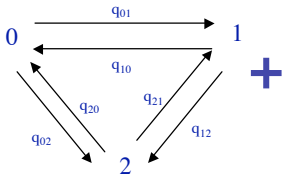
English	*here*	*sea*	*water*	*when*
German	*hier*	*See, Meer*	*Wasser*	*wann*
French	*ici*	*mer*	*eau*	*quand*
Italian	*qui, qua*	*mare*	*acqua*	*quando*
Greek	*edo*	*thalasa*	*nero*	*pote*
Hittite	*ka*	*aruna-*	*watar*	*kuwapi*

English	0	0	0	0
German	0	0, 1	0	0
French	1	1	1	0
Italian	1	1	1	0
Greek	2	2	2	0
Hittite	3	3	0	0

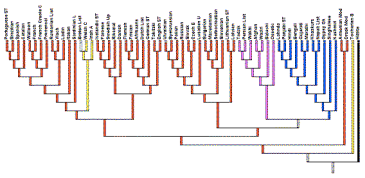
Estimating rates of word evolution on a phylogeny

Languages		meanings		
English	here	sea	water 0	when
German	hier	see, meer	wasser 0	wan
French	ici	mer	eau 1	quand
Italian	qui, qua	mare	acqua 1	quando
Greek	edo	thalasa	nero 2	pote
Hittite	ka	aruna-	watar 0	kuwapi

transition model (e.g., water)

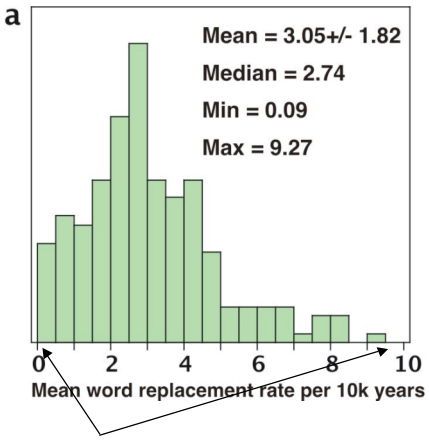


phylogeny

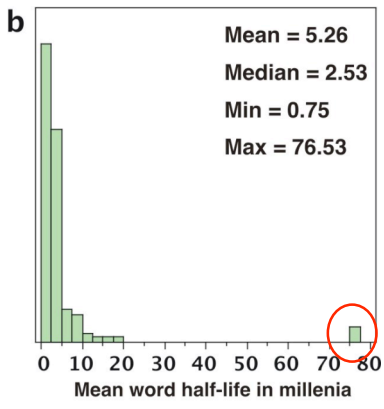


numerical estimates of transition rates, *q* (scaled as expected changes per ten thousand years)

Distribution of word replacement rates
(rates of lexical evolution)



100-fold rate variation



Correlated rates in Bantu
(Pagel & Meade, 2006)

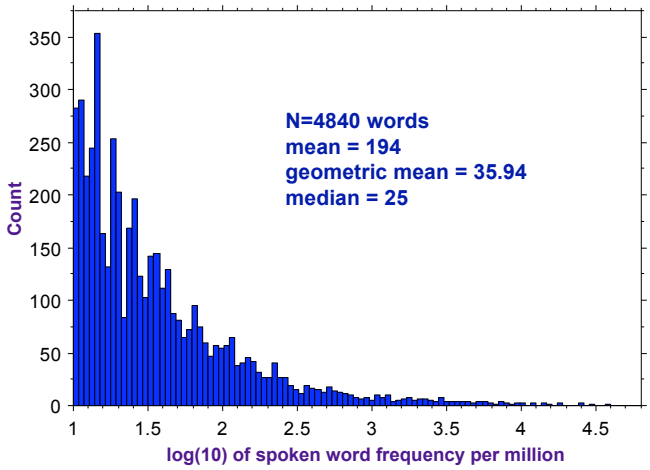
"Among the most important factors that may or do influence both the rate and the pattern of evolution are variability, rate of mutation, character of mutations, length of generations, size of populations, and natural selection."

What predicts variation in rates of evolution?

genes
directional versus purifying selection (conserved and non-conserved elements), expression levels, population size

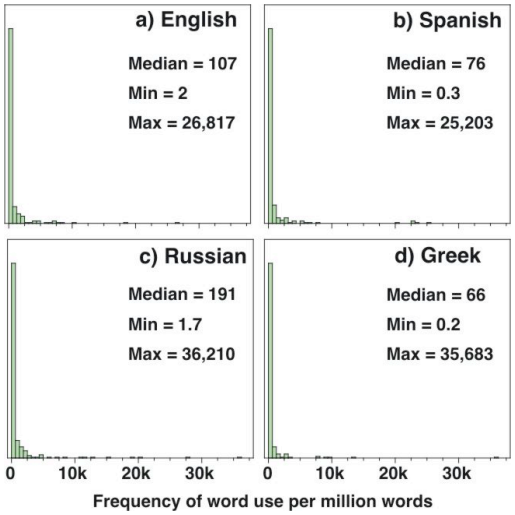
words
word frequency
Paul (1880) and Zipf (1947), but not tested.

Spoken word frequency in the British National Corpus



Distribution of frequency of word use

(20-100 million words)



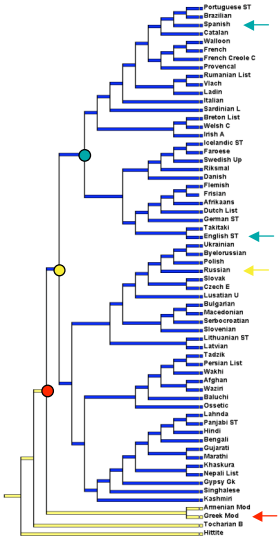
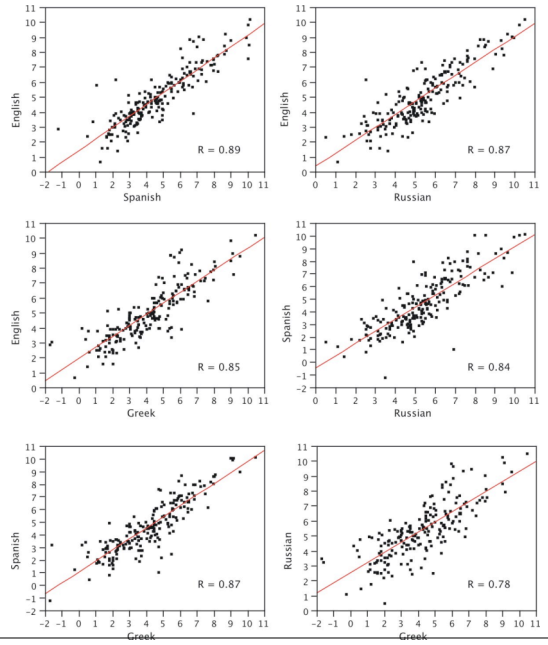


Figure from Pagel et al., *Nature*, 2007.

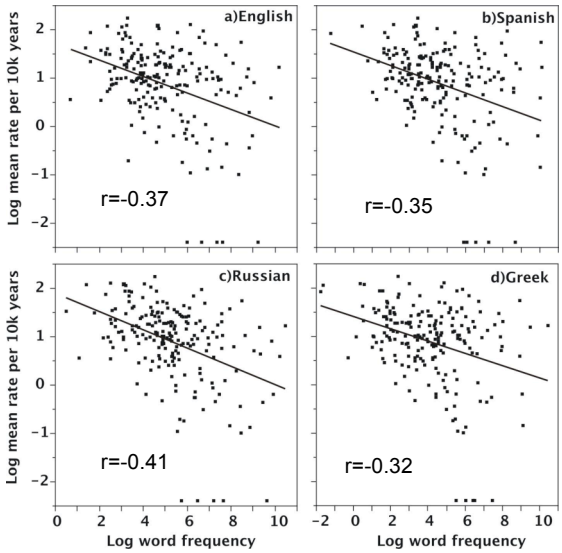
Correlations between frequencies of word use



average of the six pairwise correlations = 0.84

range: 0.78-0.89

Frequency vs rate of lexical evolution



Parts of speech

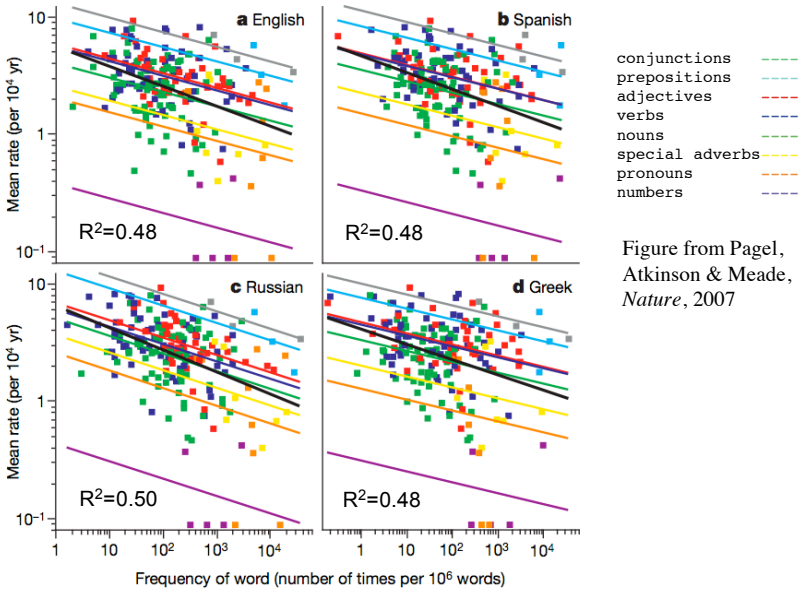
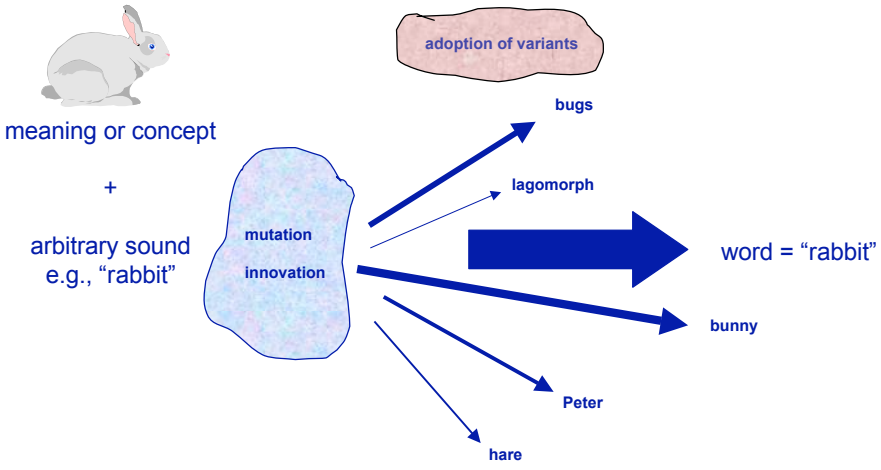


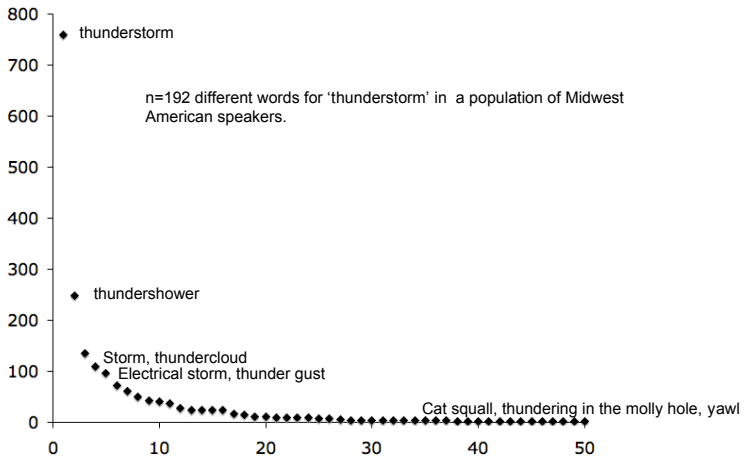
Figure from Pagel, Atkinson & Meade, *Nature*, 2007

Two models of how frequency influences the rate of lexical evolution

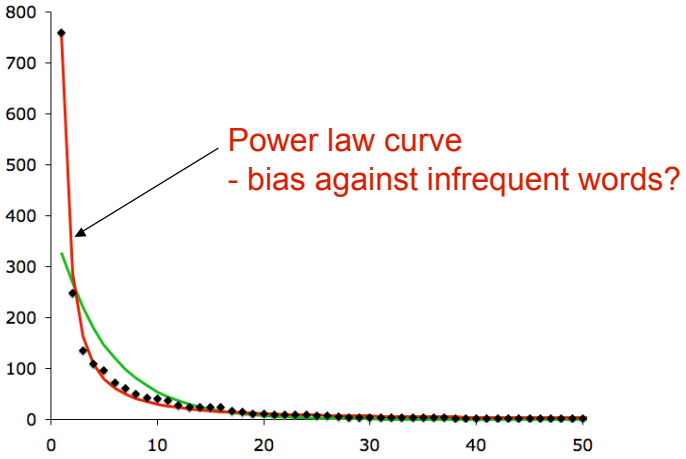
i) reduced mutation **ii) matching-purifying model**



Word frequency distribution for “Thunderstorm”



Word frequency distribution for “Thunderstorm”



What can we say about rates of lexical replacement…

Frequency of word use and POS account for 50% of variation in rates of evolution across 87 languages representing ~130,000 language-years of evolution

Frequency may act to reinforce the status quo or as a linguistic form of 'purifying selection' affecting the choice of words

The mechanism is expected to operate similarly across all languages and time scales, and makes predictions about specific meanings. (e.g. Indo-European and Bantu correlation).

Some insights for cultural evolution

languages evolve initially in less frequently used parts of vocabulary, retaining mutual intelligibility for longer

high frequency words may be less likely to be borrowed

cultural replicators can evolve more slowly than some human genes (e.g., compare "five" with lactase gene) -- some words persisting for tens of thousands of years

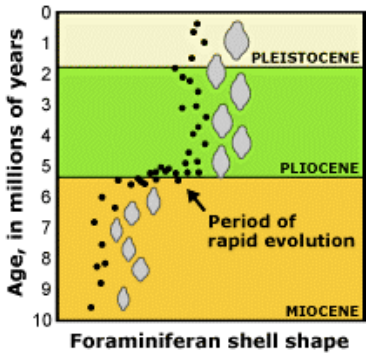
slow evolution raises possibility of deep linguistic reconstructions

Modes

- Speciation
- Phyletic evolution
- Quantum evolution

Punctuated Equilibrium and the fossil record

- Eldredge and Gould 1972
- long periods of stability or stasis followed by short punctuational bursts associated with speciation



Species Formation through Punctuated Gradualism in Planktonic Foraminifera Bjorn A. Malmgren; W. A. Berggren; G. P. Lohmann. *Science*, 225 (4659): 317-319.

Large Punctuational Contribution of Speciation to Evolutionary Divergence at the Molecular Level

Mark Pagel,* Chris Venditti, Andrew Meade

A long-standing debate in evolutionary biology concerns whether species diverge gradually through time or by punctuational episodes at the time of speciation. We found that approximately 22% of substitutional changes at the DNA level can be attributed to punctuational evolution, and the remainder accumulates from background gradual divergence. Punctuational effects occur at more than twice the rate in plants and fungi than in animals, but the proportion of total divergence attributable to punctuational change does not vary among these groups. Punctuational changes cause departures from a clock-like tempo of evolution, suggesting that they should be accounted for in deriving dates from phylogenies. Punctuational episodes of evolution may play a larger role in promoting evolutionary divergence than has previously been appreciated.

Pagel, M. et al. (2006). *Science* **314**: 119-21.

Curious Parallels

Biological Evolution	Language Evolution
Discrete heritable units – e.g. genetic code, morphology, behaviour	Discrete heritable units – e.g. lexicon, syntax, and phonology
Homology	Cognates
Mutation – e.g. Base-pair substitutions	Innovation – e.g. Sound changes
Drift	Drift
Natural selection	Social selection
Cladogenesis – e.g. allopatric speciation (geographic separation) and sympatric speciation (ecological/reproductive separation)	Lineage splits – e.g. geographical separation and social separation
Anagenesis	Change without split
Horizontal gene transfer – e.g. hybridisation	Borrowing
Plant Hybrids – e.g. wheat, strawberry	Language Creoles – e.g. Surinamese
Correlated genotypes/phenotypes – e.g. allometry, pleiotropy.	Correlated cultural terms – e.g. 'five' and 'hand'.
Geographic clines	Dialects/Dialect chains
Fossils	Ancient Texts
Extinction	Language death

Is language evolution also punctuated?

- We might expect that whatever causes punctuational species evolution may have a linguistic analogue.
- Dixon (1997) posited punctuational language evolution, explicitly drawing on analogy with biology (Eldredge and Gould, 1972).
- Goodenough (1992) describes languages slowly accumulating phonological, morphological and lexical changes until a threshold is reached and the system is rapidly restructured.

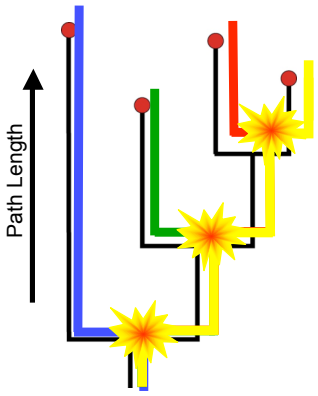
Motivating questions:

- Is language evolution punctuated at splitting events?
- If present, how big is the punctuational effect in languages?
- What could cause a punctuational effect?
- What are the implications of this for understanding language evolution?

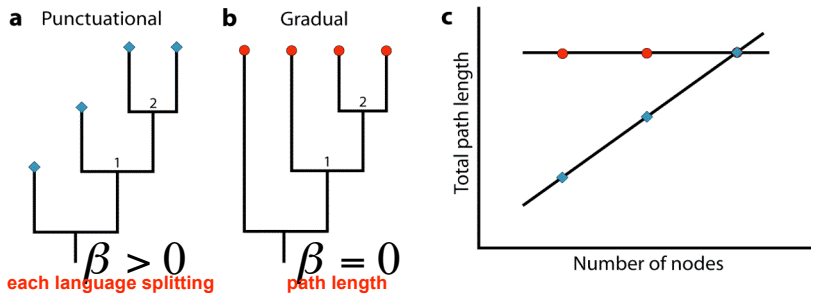
Phylogenies, Nodes and Path Length

Phylogenies Record:

- net-language splits represented by ***nodes*** of the tree
- branches measure evolutionary divergence between splitting events
- the sum of the branch lengths, from root to tip of the tree is called the ***path length***



Path lengths may contain components derived from punctuational and gradual processes



$\beta > 0$ — each language splitting event makes some contribution to path length

$\beta = 0$ — path length accumulates as a function of time

$$path\ length = n\beta + g$$

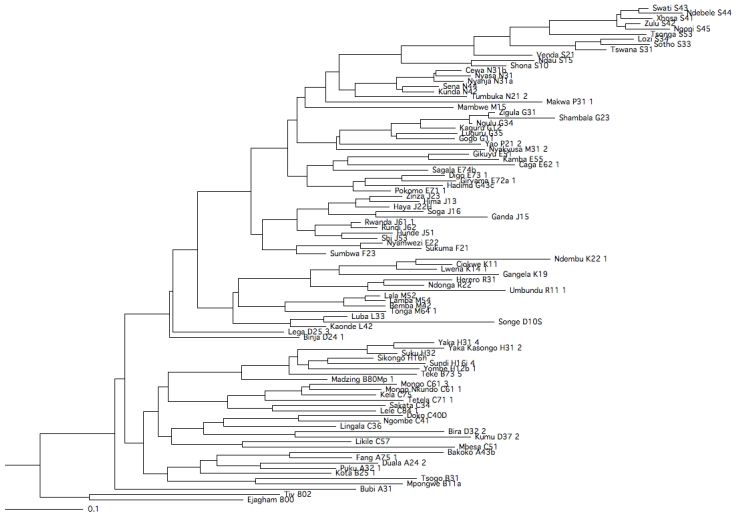
The data

- Requirements:
  - Lexical cognate data
  - Established language families
  - Reliable coding
  - Relatively well-sampled
- Three datasets identified:
  - Austronesian - 200 meanings in 328 languages
  - Bantu - 100 meanings in 95 languages
  - Indo-European - 200 meanings in 63 languages

Tree building

- for each data set, we derived a Bayesian posterior distribution of phylogenetic trees
  - Binary coding of cognate presence and absence
  - Based on a range of models of cognate gain and loss
    - Here report 1 parameter w/ gamma distributed rate variation
  - Sample of 1,000 trees per language family
- calculated the relationship between path lengths and number of nodes in each tree of the sample
- generalised least squares framework in which non-independence among languages that arises from shared ancestry is statistically controlled

A punctuational tree…



Do the data sets show evidence of a punctuational effect ($\beta > 0$)?

- Austronesian
  - **punctuational effect** in ~100% of trees
- Bantu
  - **punctuational effect** in ~98% of trees
- Indo-European
  - **punctuational effect** in ~67% of trees

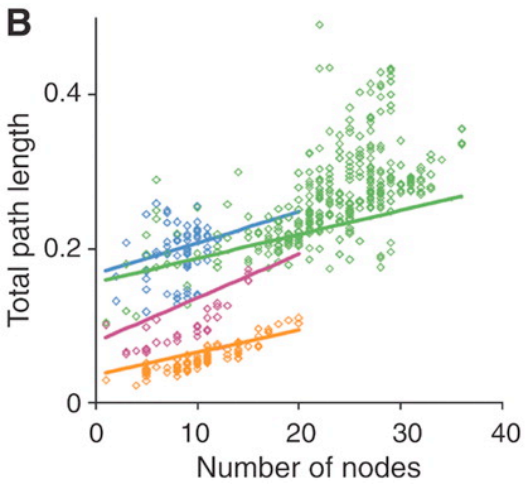


Figure from Atkinson et al., Science, 2008

Estimating the punctuational effect of language divergence on overall lexical evolution

β measures the increase in the number of changes per language divergence event on the tree

- but the absolute value depends on the rate of change used to infer the tree

Number of branches in a bifurcating tree

Tree length – sum of all branch lengths

This ratio measures proportion of tree length attributable to punctuational effects

What proportion of lexical evolution is attributable to punctuational effects?

- Bantu
  - Punctuational effect accounts for 31% of lexical evolution
- Indo-European
  - Punctuational effect accounts for 21% of lexical evolution
- Austronesian
  - Punctuational effect accounts for 10% of lexical evolution
- Polynesian
  - Punctuational effect accounts for 33% of lexical evolution
- Sequence Evolution
  - Punctuational effect accounts for ~22% of sequence evolution (Pagel et al., 2006)

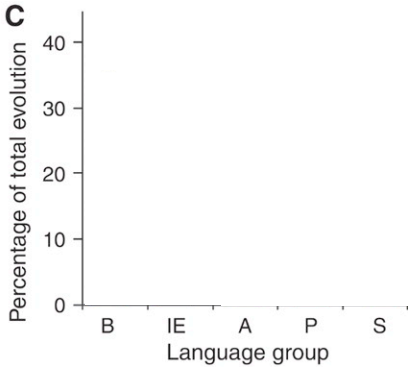


Figure from Atkinson et al., Science, 2008

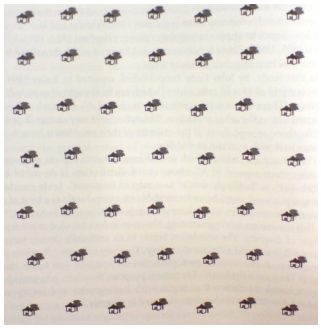
Simulations

- Simulated data using cognate birth/death model in *TraitLab**
  - No evidence of punctuated evolution
- Simulated w/ borrowing
  - No evidence of punctuated evolution
- Simulated w/ local borrowing
  - No evidence of punctuated evolution

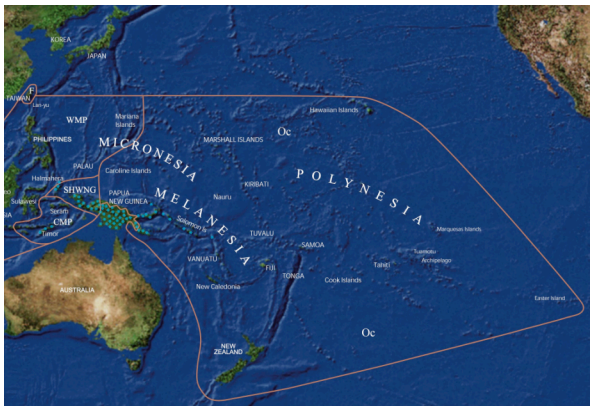
* Q. D. Atkinson, G. K. Nicholls, D. Welch, R. D. Gray, *Transactions of the Philological Society* **103**, 193 (2005).

Possible Mechanisms

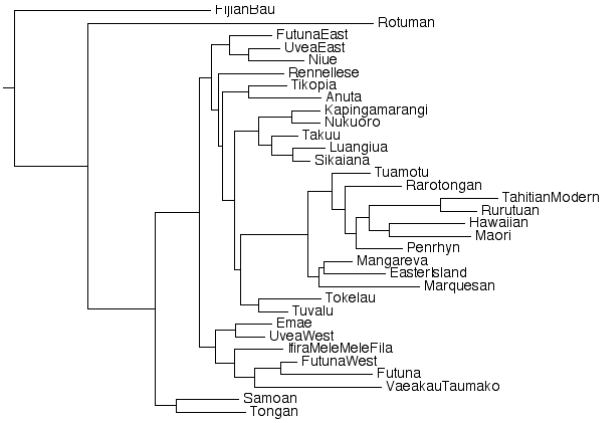
1. small founder population
   - Nettle (1999) - computer simulation
   - Simulated word propagation through populations in a grid
   - Smaller populations evolve at faster rates
   - Biological analogy - “founder effect”
   - May be a similar mechanism that causes increased rates in low frequency words.
   - Kirch and Green, 1987 - founder events in settlement of the Pacific lead to increased rates of change



A Polynesian “founder effect”



A Polynesian "founder effect"



Austronesian - ~10% of evolution due to PE
Polynesian - ~30% of evolution due to PE

Possible Mechanisms

2. Social Identity
   - "The underlying cause of sociolinguistic differences…is the human instinct to establish and maintain social identity"

     - Chambers (1995, p 250)
   - Martha's Vineyard (Labov, 1963)
   - Noah Webster - "as an independent nation, our honor requires us to have a system of our own, in language as well as government"
     - Noah Webster, *Dissertations on the English Language* (1789, p. 20).
   - Social identity drives language diversification
     - Recently separated languages
     - Sympatric language divergence - perhaps due to class/prestige differences

Implications of findings…

- language splitting events have a punctuational effect on lexical evolution
- This effect is substantial, and potentially a ubiquitous property of language evolution
- There may be more than one process causing punctuational language change
  - Founder effect or social identity
  - perhaps an allopatric vs. sympatric distinction?

General Conclusions

Computational phylogenetic methods and comparative data allow us to develop an understanding of factors affecting rates of language change.

This approach holds the promise of identifying nomothetic laws governing the tempo and mode of cultural replicators like language.

Thanks to:

Mark Pagel

Chris Venditti

Andrew Meade

Russell Gray

Simon Greenhill

The Leverhulme Trust