Frequency of word-use predicts rates of lexical evolution throughout Indo-European history

Mark Pargel[1,2], Quentin D. Atkinson[1] & Andrew Meade[1]

1. School of Biological Sciences, University of Reading, Whiteknights, Reading, Berkshire, RG6 6AS, UK
2. Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA.

http://www.nature.com/nature/journal/v449/n7163/pdf/nature06176.pdf

- Propose that the frequency with which specific words are used in everyday language exerts a general and law-like influence on their rates of evolution

- Frequently used words evolve at slower rates and

  infrequently used words evolve more rapidly.

Eg-

Meaning - **Tail**

- English – tail
- Greek – ovpά
- French – *queue*
- Germans – *schwanz*

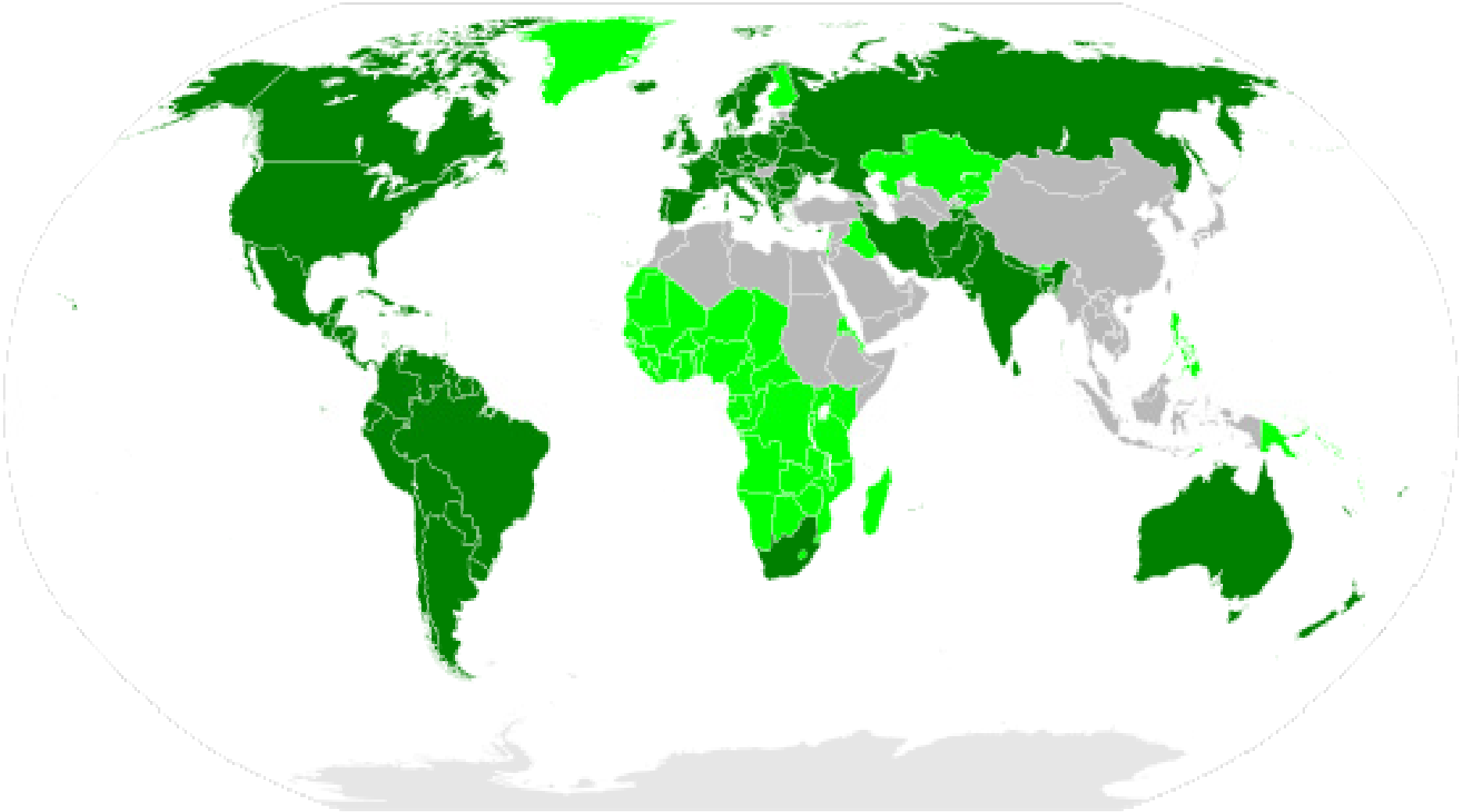
Meaning – **Two (2)**

- English – two
- Greek – σ υ ο (dewo)
- French – deux (du)
- Germans – zwei (zwo)
- Spanish – dos
- Russian – dva

Method

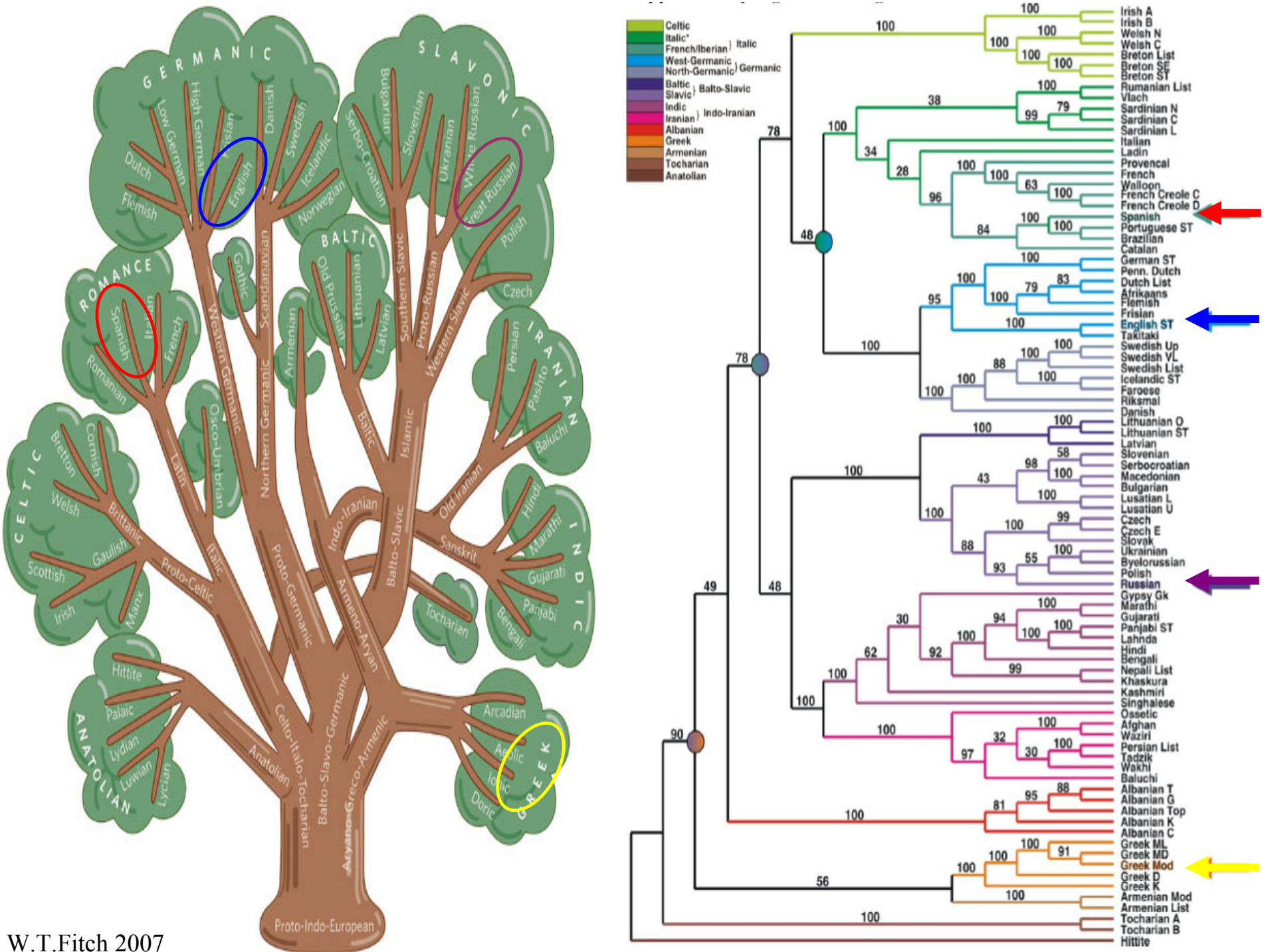
- 87 Indo-European languages
- Used four large and divergent language corpora
  - English,
  - Spanish,
  - Russian
  - Greek
    - (20~100 million words each)
- A comparative database of 200 fundamental vocabulary meanings (Supplementary table S2)

Global distribution of Indo-European Languages



Countries with a majority of speakers of IE languages

Countries with an IE minority language with official status



W.T.Fitch 2007

- Analyzed 200 Meanings by grouping in to cognate sets (1~46 -> for a total 4049) and categorizing in to types

- Examined the relationship between the rates at which Indo- European language speakers adopt new words for a given meaning and the frequency with which those meanings are used in everyday language

Results

- Words such as ‘two’, ‘who’ ‘tongue’, ‘night’, ‘one’ and ‘to die’ predict zero to one cognate replacements per 10,000 years (slow rate of evolution)

- words such as ‘dirty’, ‘to turn’, ‘to stab’ and ‘guts’, predict up to nine cognate replacements in the same time period

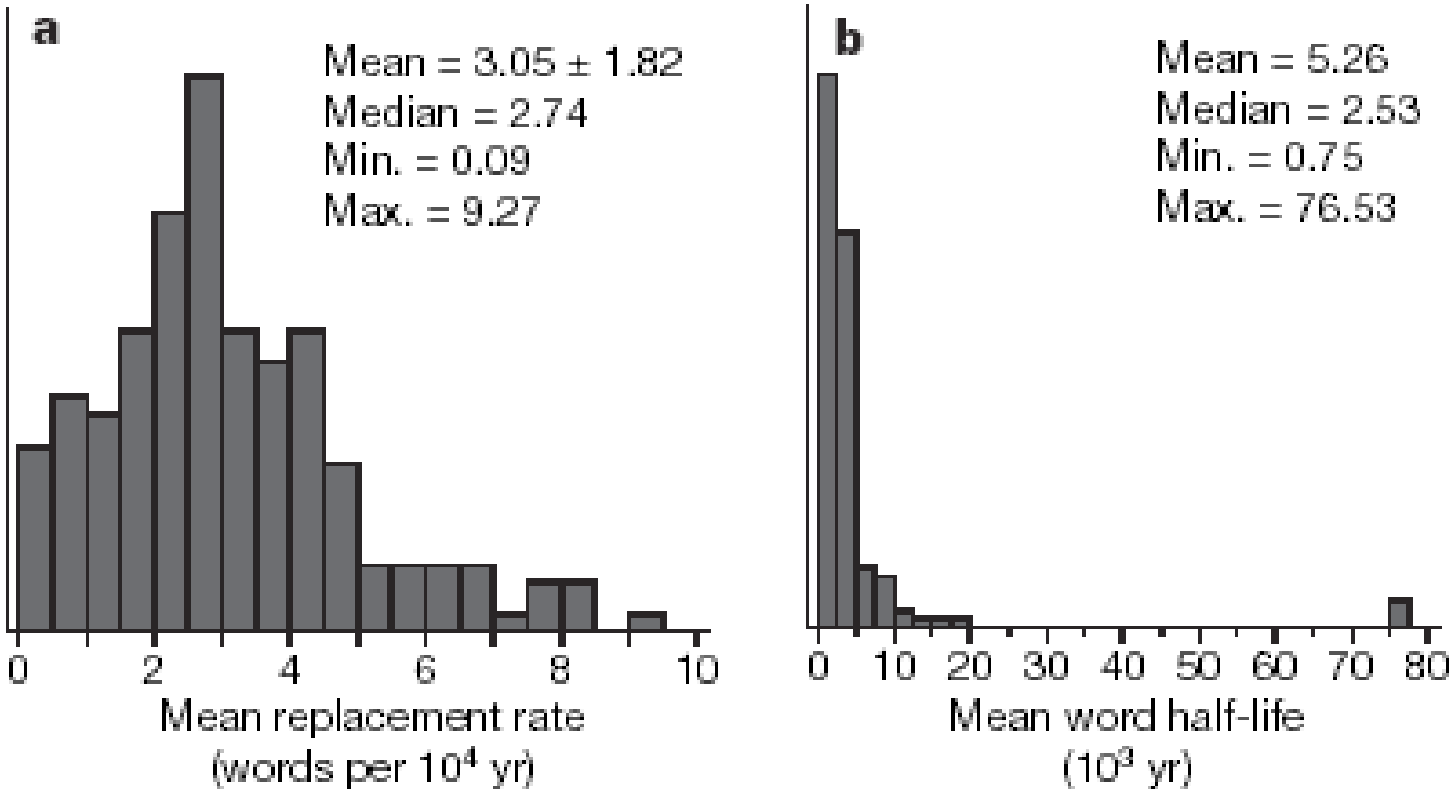


Figure 1 | Frequency plots for rates of lexical evolution in Indo-European across 200 fundamental vocabulary meanings. **a**, The mean estimated rate of cognate replacement for each meaning. **b**, The same rate distribution converted to word half-lives[10], or the time in which there is a 50% chance the word will be replaced by a different non-cognate form. The longest half-lives (76,530 years) are for meanings that show no change across Indo-European (Supplementary Information).

Replacement times vary from 750 years for the fastest evolving words to over 10,000 years for the slowest

- Distribution of word-use frequencies in each language is highly skewed,

  such that most words are used relatively infrequently (fewer than 100 times per million words),

  with a small number of frequently used words (as often as 35,000 times per million words) accounting for most speech.

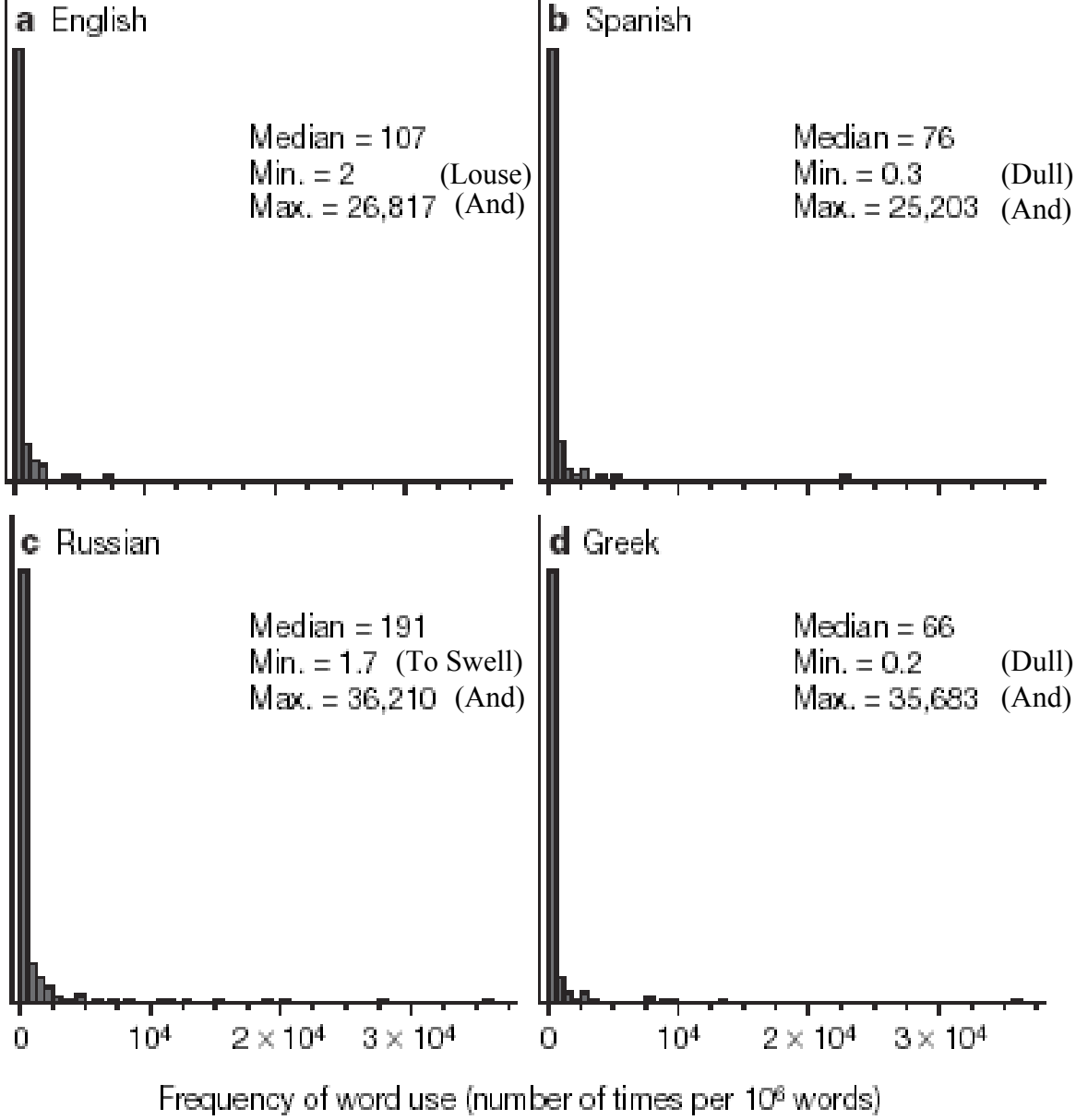
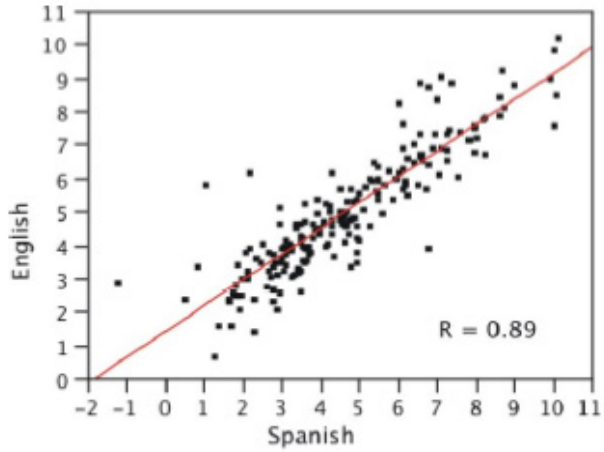
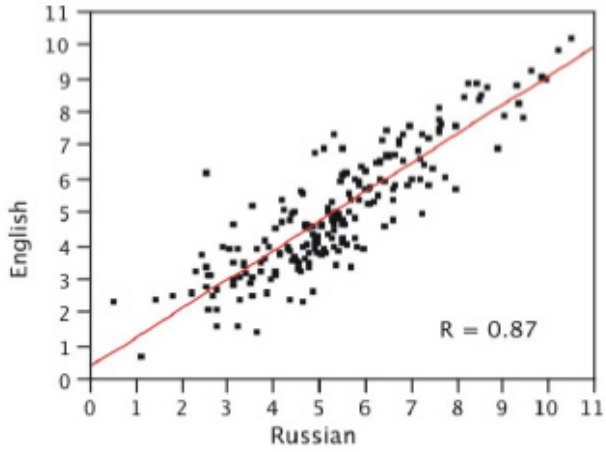
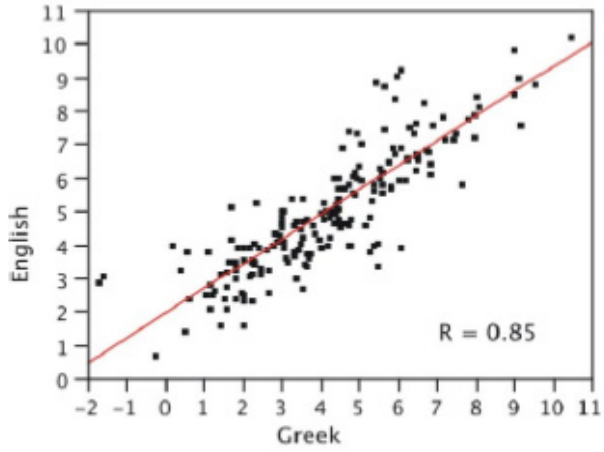


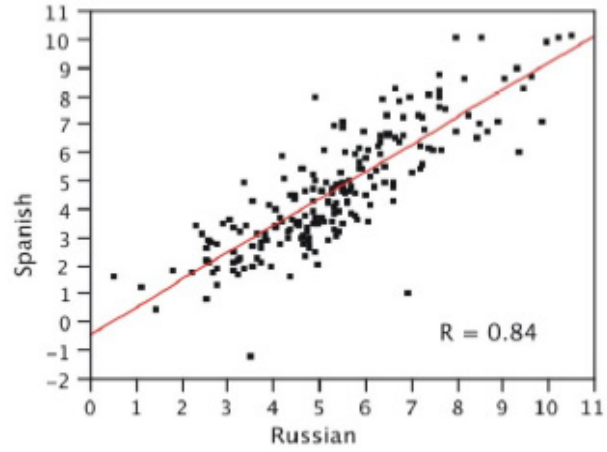
Fig 2. Distribution of frequency of meaning-use for 200 meanings in four Indo European languages.

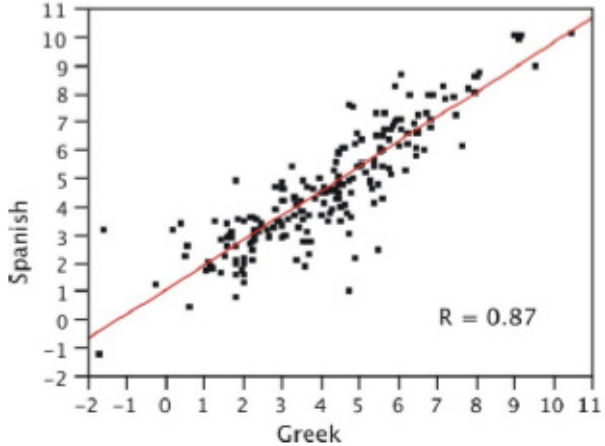
Word-use frequencies are highly correlated among the four languages ($0.78<r<0.89$, mean $r = 0.84$)











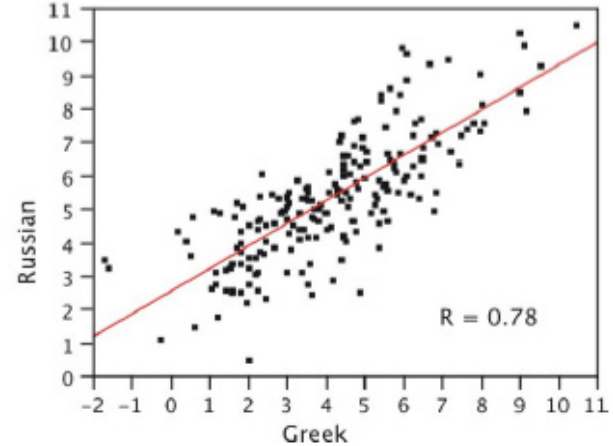


Fig S2

Pair wise correlation between log frequency of words

Words used at high frequency at a one language tend to be used at a high frequency in other languages

categorized meanings as

- nouns,
- adjectives,
- verbs,
- pronouns,
- numbers,
- conjunctions,
- prepositions
- special adverbs

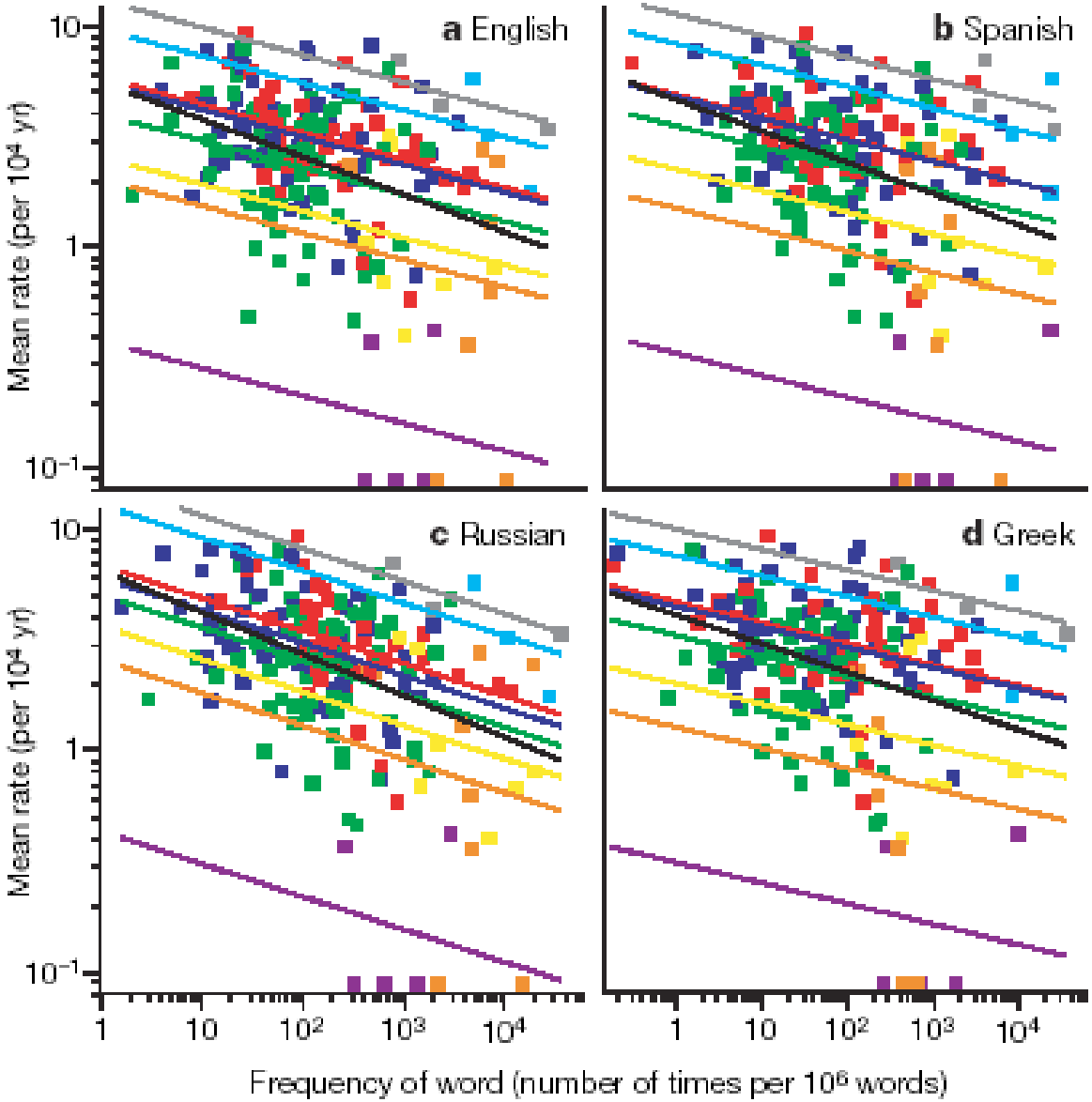


Fig 3. Frequency of meaning-use plotted against estimated rate of lexical evolution for 200 basic meanings in four Indo-European languages.

Conjunctions (grey) evolve fastest, followed by prepositions (turquoise), adjectives (red), verbs (blue), nouns (green), special adverbs (yellow), pronouns (orange) and numbers (purple).

- For a given frequency of meaning-use,
  - Prepositions
  - Conjunctions
  - Adjectives
  - Verbs
  - Nouns
  - Special adverbs
  - Pronouns
  - Numbers

Rate of evolution

- Numbers
- Pronouns
- Special adverbs

- Nouns
- Verbs

- Adjectives
- Conjunctions
- Prepositions

Rate of evolution Low

These parts of speech seem important to the meaning of spoken communication,
So subject to stronger selection.

Rate of evolution High

exact forms may often be less important to conveying meaning

Conclusions

- Frequencies of meaning-use have been remarkably stable throughout the Indo-European history
- The more a meaning is used today, the slower its rate of evolution has been throughout the history of Indo-European.
- The frequency with which different meanings are used in everyday language directly affect the rate at which new words arise and become adopted in populations of speakers
- Some form of the speech is frequency dependant, Purifying selection is responsible for the slow rate of evolution of highly expressed words

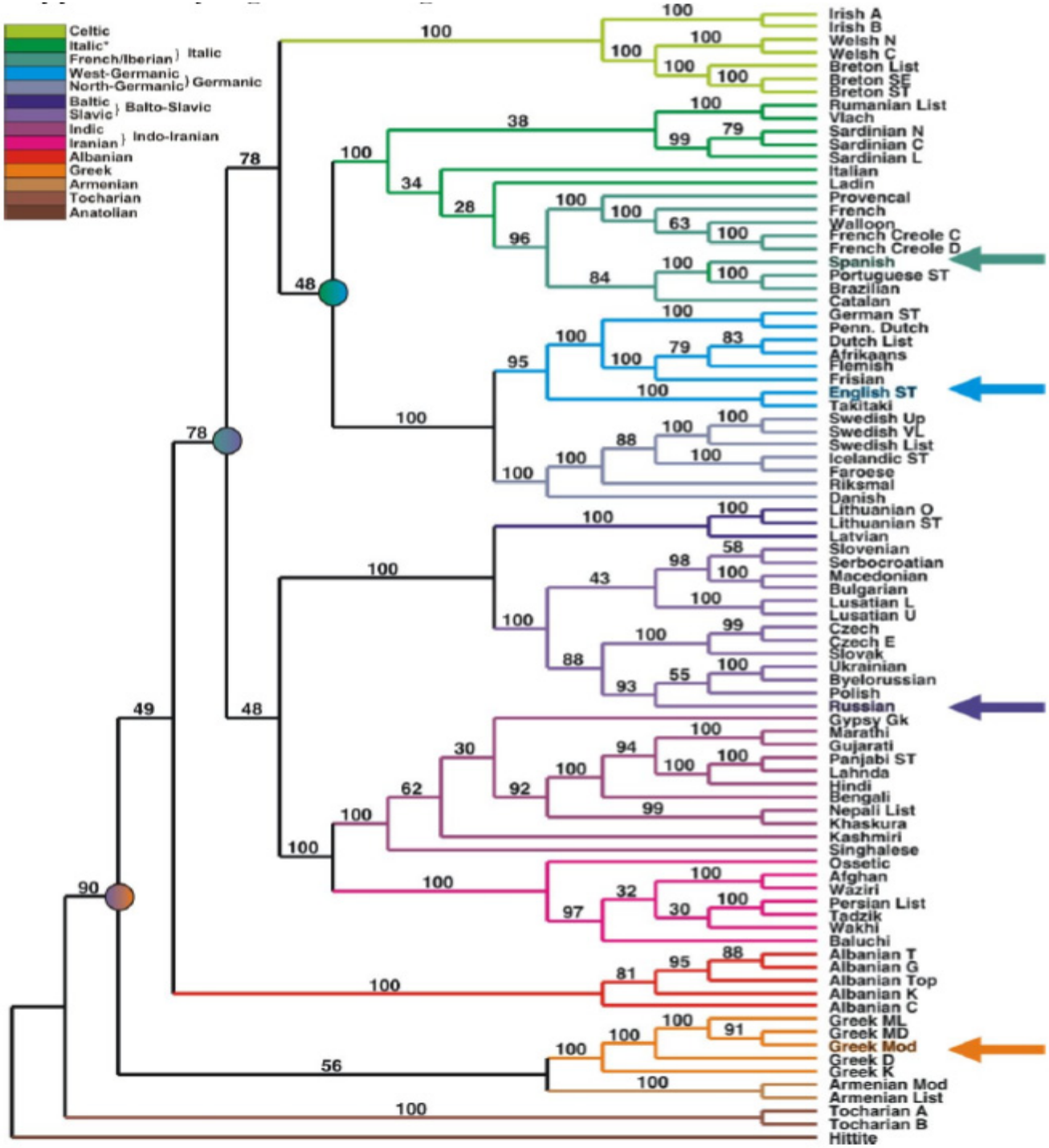
- When more than one word is used to express the same meaning, the relative frequencies of use of the rare words is lower than expected
- Rare complex words are not favored in speech as there's a high chance of misinterpreted
- Owing to the ways that humans use language, some words will evolve slowly and others rapidly across all languages.
- Hmans are capable of producing a culturally transmitted replicator that (perhaps because of the purifying force of spoken word frequency) can have replication accuracy as high as that of some gene

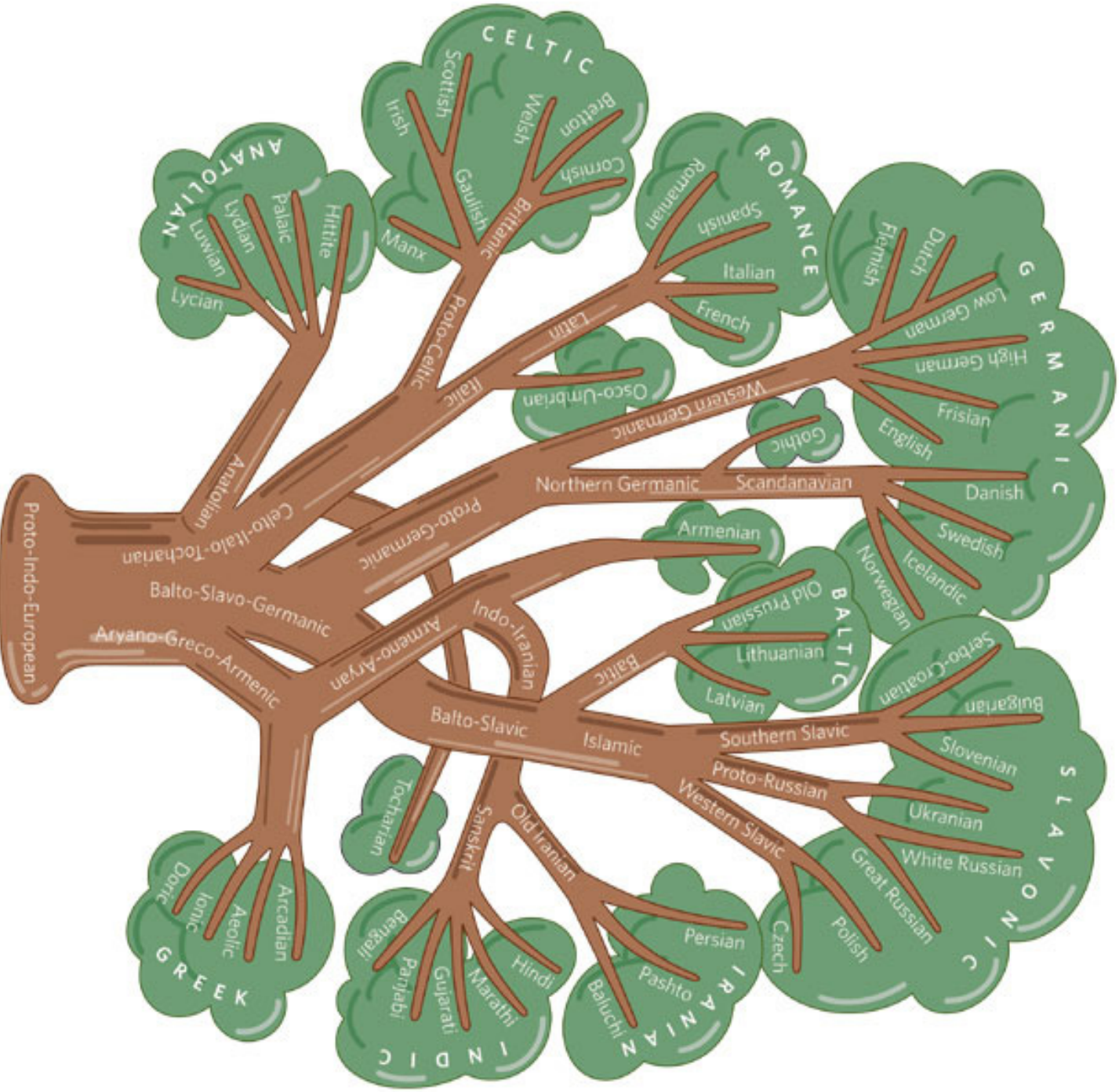
- **Parallels between biological and linguistic evolution**

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

Supplementary Table S2

Thank You


Celtic
Italic*
French/Iberian
Italic
West-Germanic
North-Germanic
Germanic
Baltic
Slavic
Balto-Slavic
Indic
Iranian
Indo-Iranian
Albanian
Greek
Armenian
Tocharian
Anatolian
Irish A
Irish B
Welsh N
Welsh C
Breton List
Breton SE
Breton ST
Rumanian List
Vlach
Sardinian N
Sardinian C
Sardinian L
Italian
Ladin
Provencal
French
Walloon
French Creole C
French Creole D
Spanish
Portuguese ST
Brazilian
Catalan
German ST
Penn. Dutch
Dutch List
Afrikaans
Flemish
Frisian
English ST
Takitaki
Swedish Up
Swedish VL
Swedish List
Icelandic ST
Faroese
Riksmal
Danish
Lithuanian O
Lithuanian ST
Latvian
Slovenian
Serbocroatian
Macedonian
Bulgarian
Lusatian L
Lusatian U
Czech
Czech E
Slovak
Ukrainian
Byelorussian
Polish
Russian
Gypsy Gk
Marathi
Gujarati
Panjabi ST
Lahnda
Hindi
Bengali
Nepali List
Khaskura
Kashmiri
Singhalese
Ossetic
Afghan
Waziri
Persian List
Tadzik
Wakhi
Baluchi
Albanian T
Albanian G
Albanian Top
Albanian K
Albanian C
Greek ML
Greek MD
Greek Mod
Greek D
Greek K
Armenian Mod
Armenian List
Tocharian A
Tocharian B
Hittite


Proto-Indo-European
Anatolian
Celto-Italo-Tocharian
Balto-Slavo-Germanic
Aryano-Greco-Armenic
ANATOLIAN
Hittite
Palaic
Lydian
Luwian
Lycian
CELTIC
Scottish
Irish
Gaulish
Manx
Brittanic
Welsh
Breton
Cornish
Proto-Celtic
Italic
Latin
Osco-Umbrian
ROMANCE
Romanian
Spanish
Italian
French
Proto-Germanic
Western Germanic
Northern Germanic
Gothic
GERMANIC
Flemish
Dutch
Low German
High German
Frisian
English
Scandanavian
Danish
Swedish
Icelandic
Norwegian
Armenian
Armeno-Aryan
Indo-Iranian
Balto-Slavic
Baltic
BALTIC
Old Prussian
Lithuanian
Latvian
Islamic
SLAVONIC
Serbo-Croatian
Bulgarian
Slovenian
Southern Slavic
Proto-Russian
Western Slavic
Ukranian
White Russian
Great Russian
Polish
Czech
Tocharian
Sanskrit
Old Iranian
GREEK
Doric
Ionic
Aeolic
Arcadian
INDIC
Bengali
Panjabi
Gujarati
Marathi
Hindi
IRANIAN
Persian
Pashto
Baluchi

- Social and demographic factors proposed to affect rates of language change within populations of speakers include
  - social status,
  - the strength of social ties
  - the size of the population13 and
  - levels of outside contact.

  These forces may influence rates of evolution on a local and temporally specific scale, but they do not make general predictions across language families about differences in the rate of lexical replacement among meanings.

Cognates

- Cognates are words of similar meaning with systematic sound correspondences indicating they are related by common ancestry.

  For example, cognates meaning ‘water’ exist in
  - English (water)
  - German (wasser)
  - Swedish (vatten) and
  - Gothic (wato)

  reflecting descent from proto-Germanic (*water).

- Markov Chain Monte Carlo (MCMC) methods

  (which include random walk Monte Carlo methods), are a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its equilibrium distribution.

  The state of the chain after a large number of steps is then used as a sample from the desired distribution. The quality of the sample improves as a function of the number of steps.

- A Markov chain

  is a sequence of random property $X1$, $X2$, $X3$, ... with the Markov property, namely that, given the present state, the future and past states are independent.

- Markov property

  Having the Markov property means the next state solely depends on the present state, but not on the previous states.