Understanding Language Evolution Using an Event-Based Model

John Huelsenbeck
Department of Integrative Biology
University of California, Berkeley
johnh@berkeley.edu

Introduction

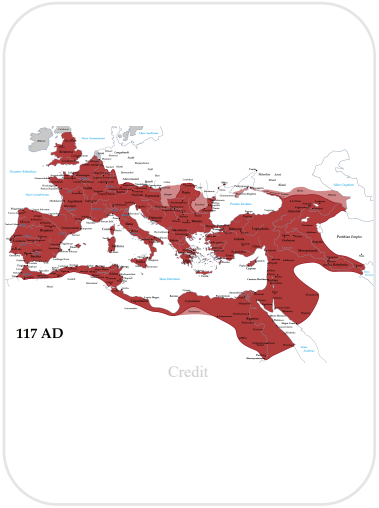
Modern languages are related to one another through a complicated history of divergence and word borrowing. The divergence of languages is caused by the slow change in spoken language as it is passed from parents to offspring. Over time, divergence causes languages to become increasingly different from one another, ultimately to the point where they are mutually unintelligible. Languages that were spoken by the same human group more recently in time are considered to be more closely related to each other than they are to groups that spoke the language more distantly in time; this relatedness information can be depicted by a tree-like diagram called a 'phylogeny.' Linguistic borrowing, by contrast, causes languages to become more similar to one another.

Language	IPA	Coding
English	/hænd/	0
German	/hant/	0
French	/mẽ---/	1
Spanish	/mano-/	1
Italian	/maːno-/	1
Russian	/rʊka/	2
Polish	/rɛŋka/	2

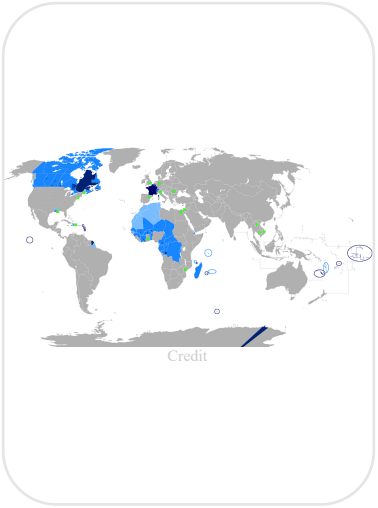
Table 1. Coding of lexical cognates for the word *hand*.

Languages

Latin

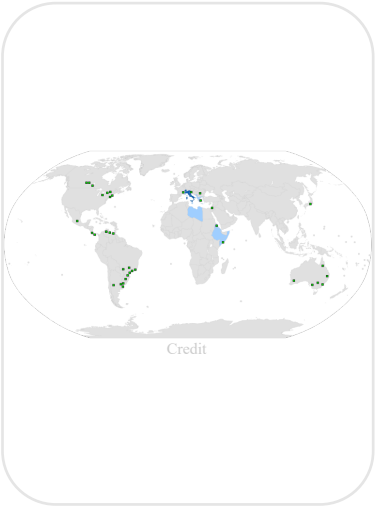


French



Spanish

Italian



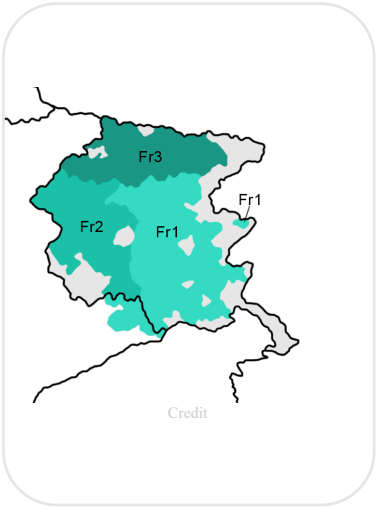
Brazilian Portuguese

Portuguese

Catalan

Walloon

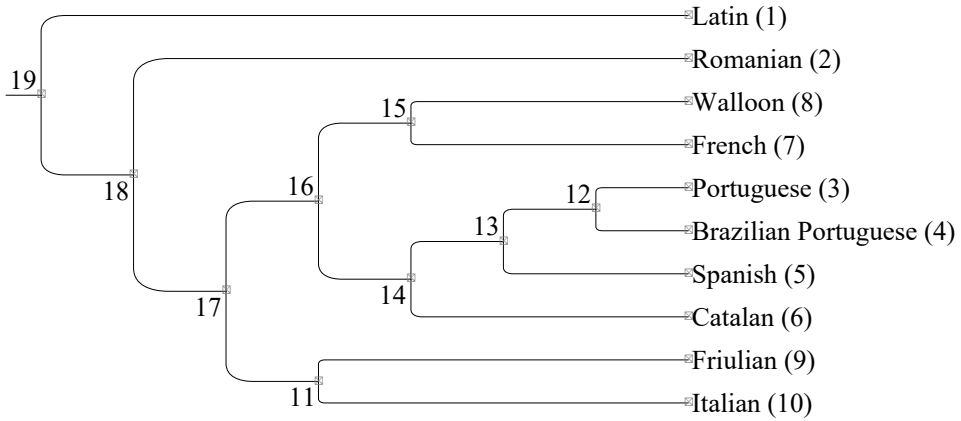
Friulian



Romanian

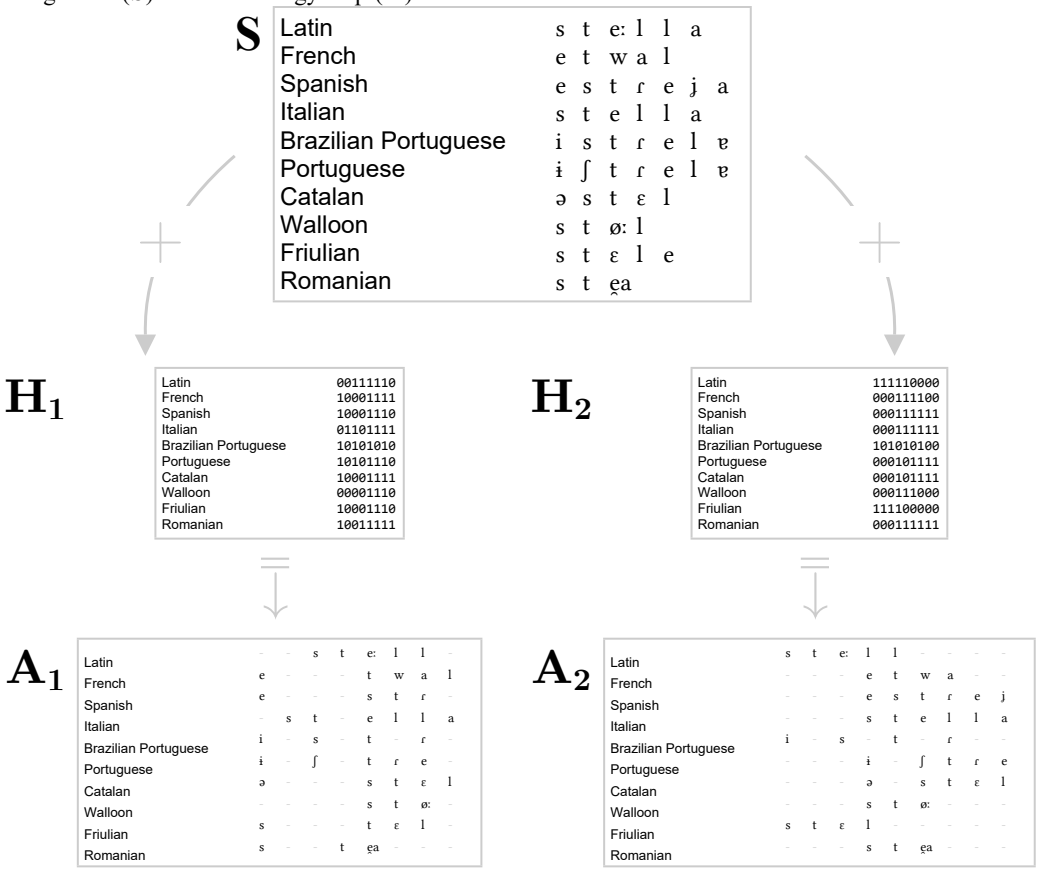
Example Tree

An example tree showing the relationships of $N = 10$ languages.



Alignment

Alignments (**A**) are formed from the observed segments (**S**) and a homology map (**H**).



Character Assignments

Each segment gets a different number

1 e 2 g 3 oː 4 ʒ 5 ə 6 j̟ 7 o 8 i 9 w 10 ɔ 11 d̻ʒ̻ 12 ɪ 13 j 14 n 15 s 16 u 17 ʃ 18 oj 19 v

20 b 21 iː 22 ʎ 23 l 24 r 25 ɨ 26 h 27 k 28 t 29 a 30 ɐ 31 uː 32 y 33 d 34 f 35 eː 36 m

37 ɾ

38 œ 39 ɛ 40 x 41 ʁ 42 ɲ 43 o̟a 44 ẽj̃ 45 ɔ̃ 46 ɑ̃ 47 ĩ 48 ɐ̃ 49 ʧ 50 p 51 z 52 ð 53 ʦ 54 aː

55 ɐ̃j̃

56 ʁ 57 aj 58 ɛː 59 θ 60 ej 61 ɐj 62 õ 63 ɛ̃ 64 ɐ̃w̃ 65 ẽ 66 c 67 ɨj 68 β 69 ɬ 70 ɥ 71 g

72 ɨw

73 ʣ 74 øː 75 ɔː 76 gʷ 77 ŋ 78 õː 79 ç 80 au̯ 81 ø 82 e̟a 83 kʲ 84 ã 85 ɟ 86 ɑ 87 ɣ 88 ɛj

89 kʷ

90 ẽː 91 ẽu̯ 92 o(w) 93 n̪ 94 ʧ 95 ʦ 96 ũ 97 ɻ 98 ɑː 99 j̃ 100 œ̃ 101 dː 102 kː 103 u̯ 104 e̟

105 ŋʷ

106 ʊ

Partition Assignments

Model: “Linguistically Informed”

1 **Nasal Vowel**

ẽj̃ ɔ̃ ɑ̃ ĩ ɐ̃ ɐ̃j̃ õ ɛ̃ ɐ̃w̃ ẽ õː ã ẽː ẽu̯ ũ œ̃

2 **Vowel**

e oː ə o i ɔ ɪ u oj iː ɨ a ɐ uː y eː œ ɛ o̯a aː aj ɛː ej ɐj ɨj ɨw øː ɔː au̯ ø e̯a ɑ ɛj o(w) ɑː u̯ e̯ ʊ

3 **Nasal Consonant**

n m ɲ ŋ n̪ j̃ ŋʷ

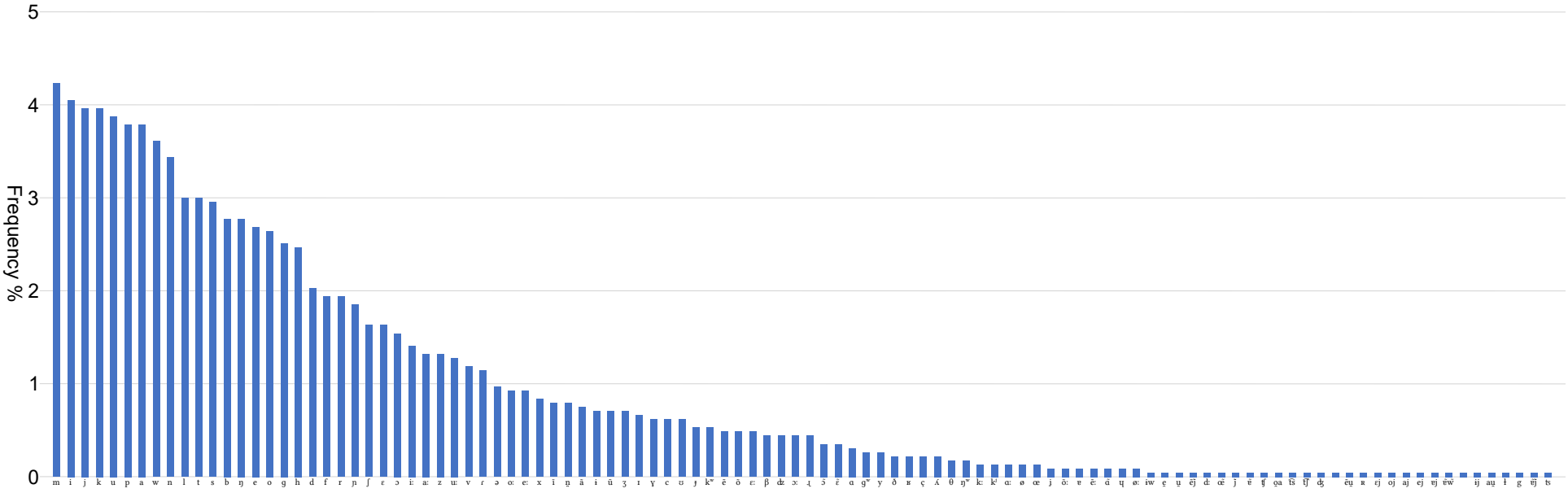
4 **Non Sylabic Sonorant**

w j l r

5 **Consonant**

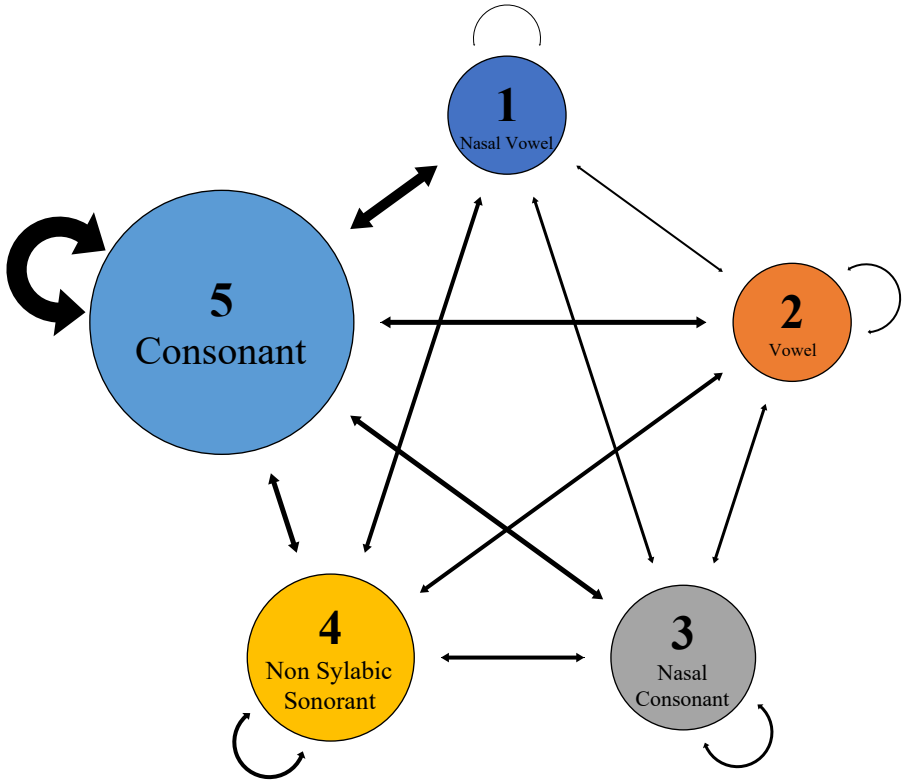
g ʒ ɟ ʤ s ʃ v b ʎ h k t d f ɾ x ʁ ʧ p z ð ʦ ʀ θ c β ɬ ɥ g ʣ gʷ ç kʲ ɟ ɣ kʷ t͡ʃ t͡s ɻ dː kː

Prior Segment Frequencies

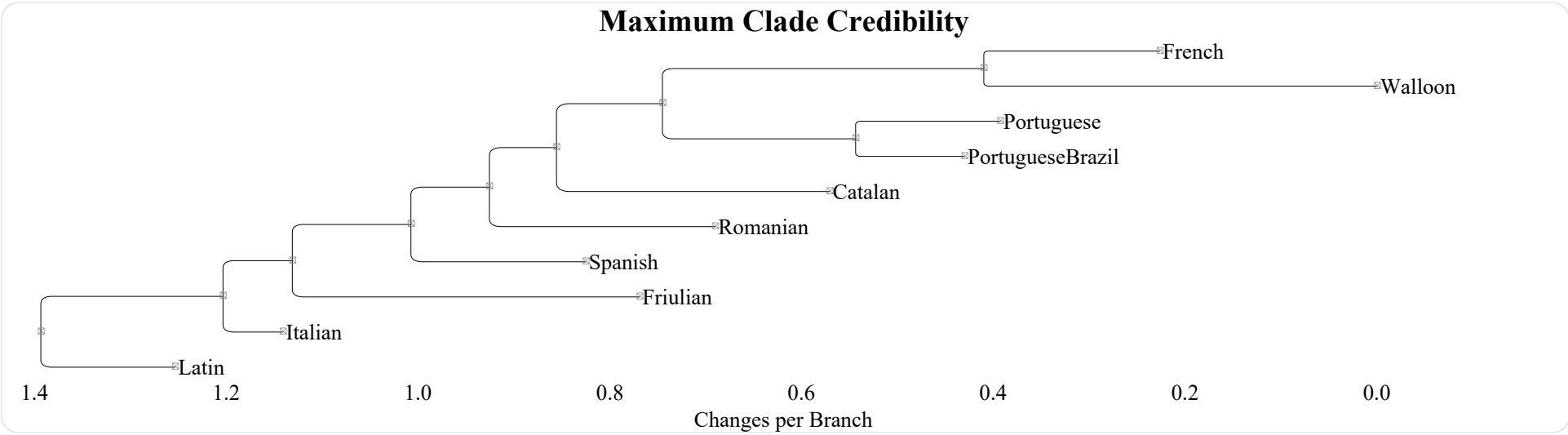


Transition Rates Between Partitions

For the ‘Linguistically Informed’ model, states were grouped into five sets: Nasal Vowel (1), Vowel (2), Nasal Consonant (3), Non Sylabic Sonorant (4) and Consonant (5). Here, the area of the circles is proportional to the estimated equilibrium frequencies for each group. The width of the arrows is proportional to the estimated rates. Note that rates of change are much greater from one word segment to another when the change is within the word segment group than it is when the change is between word segments in different groups.



Results


Maximum Clade Credibility
French
Walloon
Portuguese
PortugueseBrazil
Catalan
Romanian
Spanish
Friulian
Italian
Latin
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Changes per Branch

Questions