Understanding Language Evolution Using an Event-Based Model

John Huelsenbeck
Department of Integrative Biology
johnh@berkeley.edu

Introduction

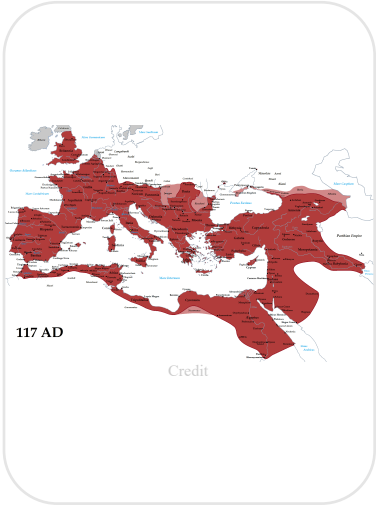
Modern languages are related to one another through a complicated history of divergence and word borrowing. The divergence of languages is caused by the slow change in spoken language as it is passed from parents to offspring. Over time, divergence causes languages to become increasingly different from 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	/mɛ̃/	1
Spanish	/mano/	1
Italian	/maːno/	1
Russian	/rʊ̟ka/	2
Polish	/rɛŋka/	2

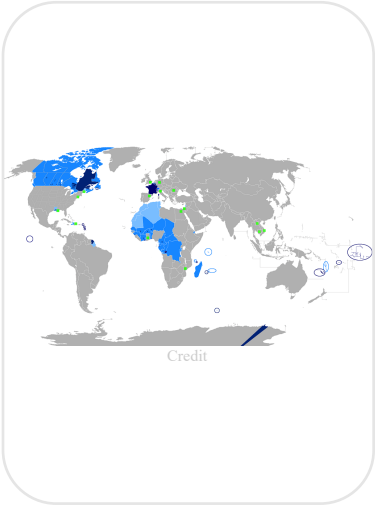
Table 1. Coding of lexical cognates for the word *hand*.

Languages

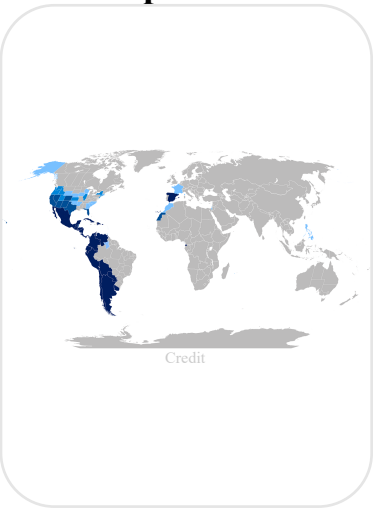
Latin



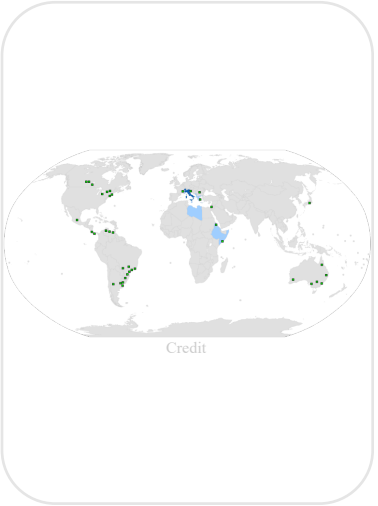
French



Spanish



Italian

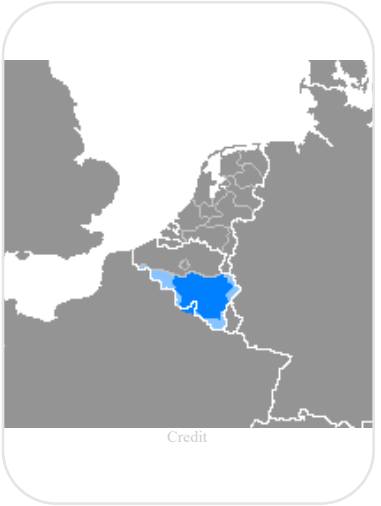


Brazilian Portuguese

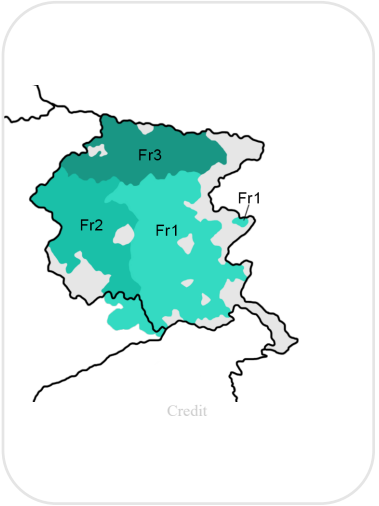
Portuguese

Catalan

Walloon



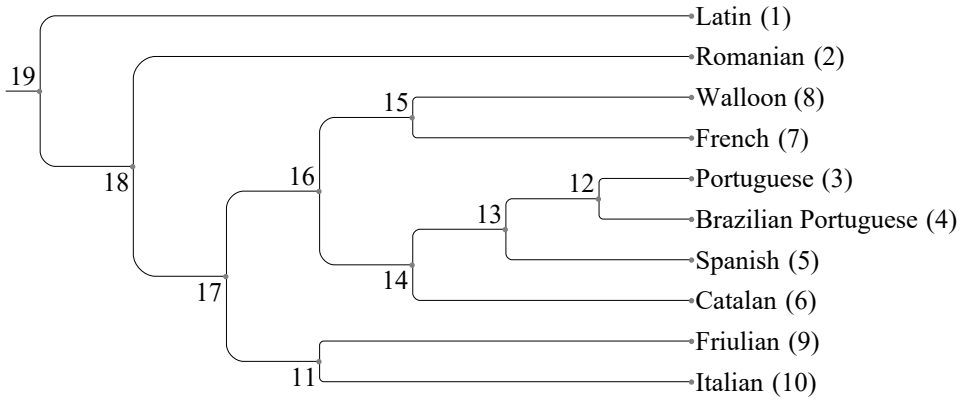
Friulian



Romanian

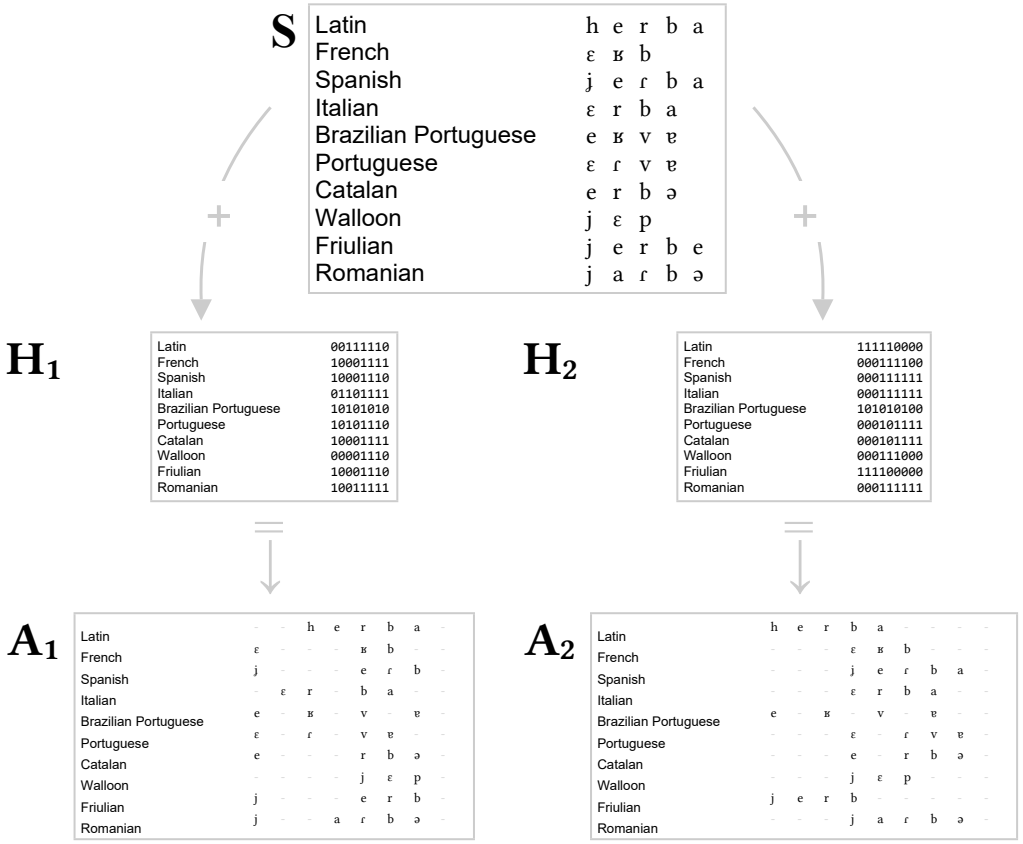
Example Tree

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 f 2 eː 3 m 4 i 5 n 6 a 7 e 8 b 9 ɾ 10 ɐ 11 j 12 œ 13 ɛ 14 u 15 l 16 r 17 x 18 o 19 ʎ 20 ʁ

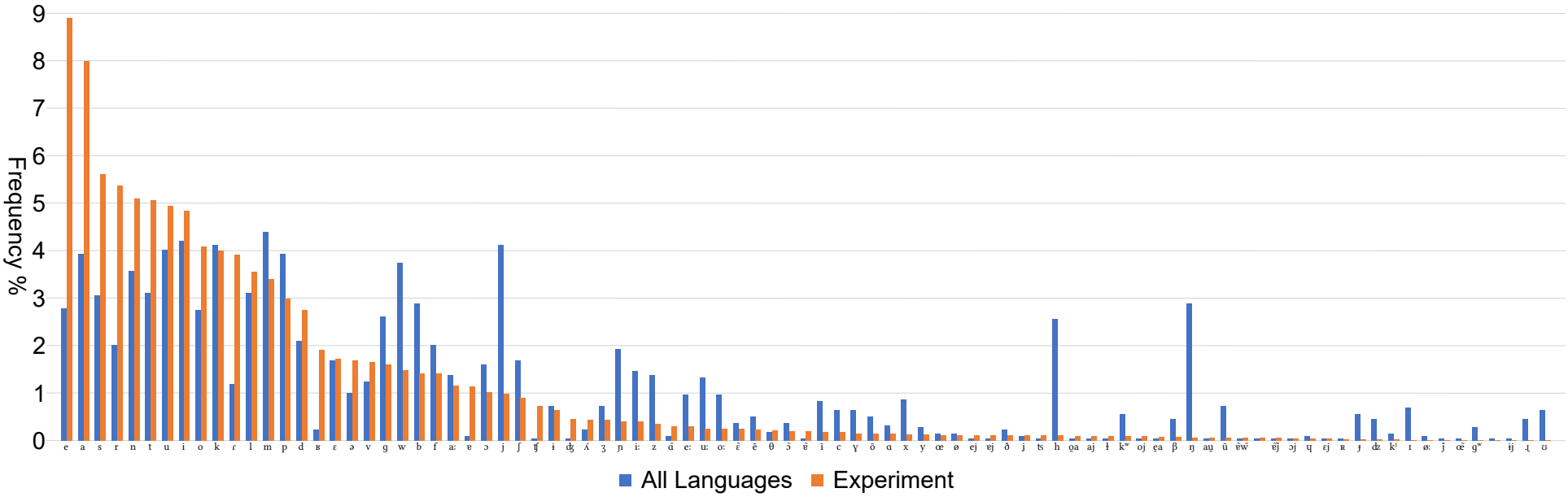
21 iː 22 d 23 ɲ 24 ɔ 25 ə 26 o̯a 27 t 28 ɑ̃ 29 ĩ 30 ɐ̃ 31 ʧ 32 ɨ 33 ɛ̃j̃ 34 s 35 p 36 z 37 ʃ 38 ð 39 oː

40 k 41 ʦ 42 aː 43 ɐ̃j̃ 44 aj 45 θ 46 ej 47 ɐj 48 w 49 ɔ̃ 50 h 51 õ 52 v 53 au̯ 54 ɛ̃ 55 ɐ̃w̃ 56 c

57 ɨj 58 ẽ 59 ɬ 60 ɑ 61 uː 62 ɥ 63 y 64 g 65 ɨw 66 ɔj 67 ʣ 68 ʝ 69 ʤ 70 oj 71 gʷ 72 ŋ 73 ø 74 β

75 e̯a 76 ʁ 77 ɛj 78 kʲ 79 ɪ 80 ʒ 81 ɟ 82 ɣ 83 kʷ 84 ũ 85 ɻ 86 j̃ 87 œ̃ 88 øː 89 ʊ

Prior Segment Frequencies



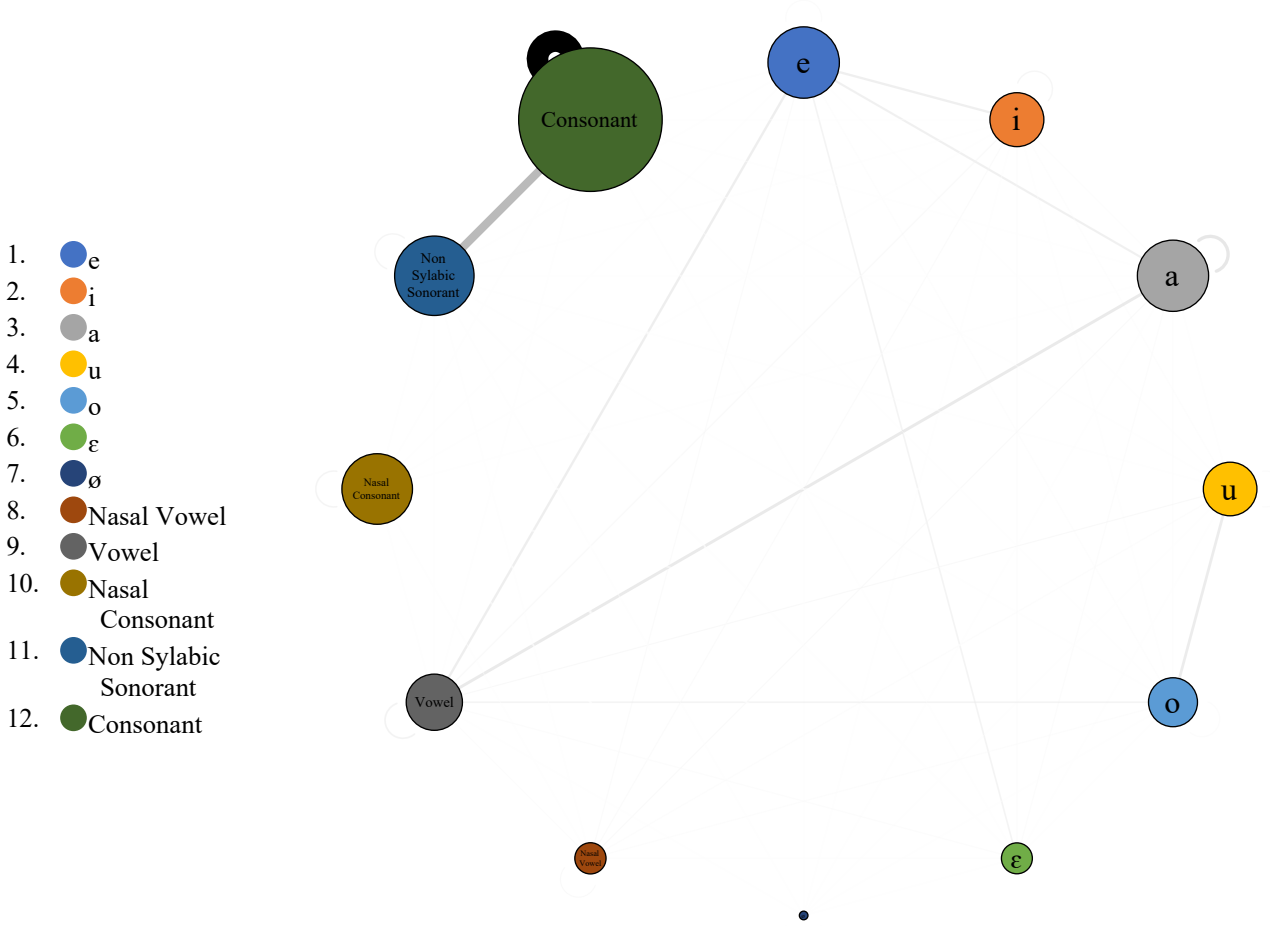
Partition Assignments

Model: “Linguistically Informed”

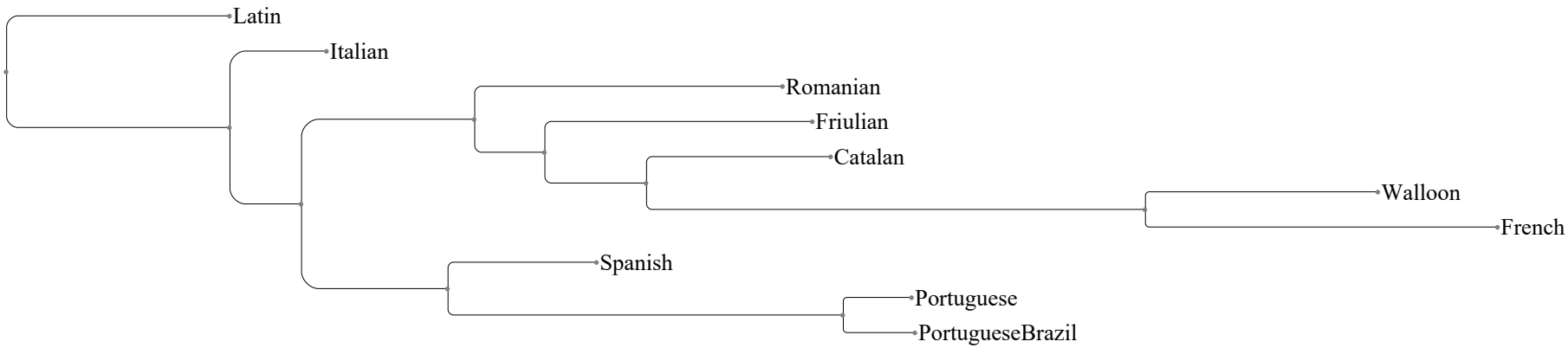
1)	e	eː e
2)	i	i iː
3)	a	a aː
4)	u	u uː
5)	o	o oː
6)	ɛ	ɛ
7)	ø	ø øː
8)	Nasal Vowel	ɑ̃ ĩ ɐ̃ ẽj ɐ̃j ɔ̃ õ ɛ̃ ɐ̃w̃ ẽ ũ œ̃
9)	Vowel	ɐ œ ɔ ə o̯a ɨ aj ej ɐj au̯ ɨj ɑ y ɨw ɔj oj e̯a ɛj ɪ ʊ
10)	Nasal Consonant	m n ɲ ŋ ɲ̃
11)	Non Sylabic Sonorant	j l r w
12)	Consonant	f b ɾ x ʎ ʁ d t ʧ s p z ʃ ð k ʦ θ h v c ɬ ɥ g ʣ ɟ ʤ gʷ β ʁ kʲ ʒ ɟ ɣ kʷ ɻ

Prior Transition Rates

For the ‘Linguistically Informed’ model, states were grouped into 12 sets. Here, the area of the circles is proportional to the occurance frequencies for each group. The width of the lines is proportional to the rates of transition between each partition.



The majority rule consensus tree



Questions