Understanding Language Evolution Using an Event-Based Model

John Huelsenbeck

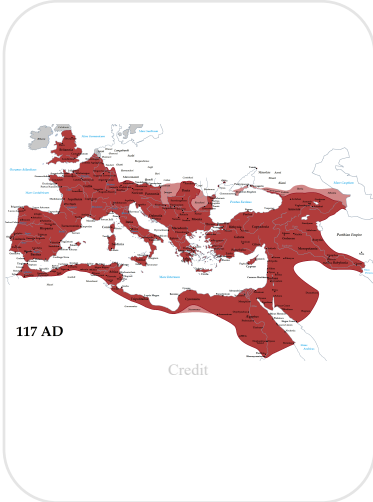Department of Integrative Biology
johnh@berkeley.edu

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.

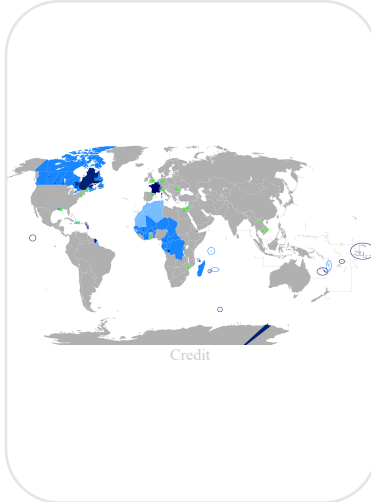| Language	IPA	Coding
English	/hænd/	0
German	/hant/	0
French	/mɛ̃/	1
Spanish	/mano/	1
Italian	/maːno/	1
Russian	/rʊka/	2
Polish	/rɛŋka/	2

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

Latin

French

Spanish
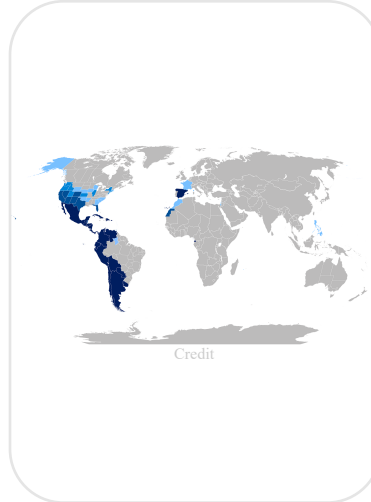
Italian

Brazilian Portuguese

117 AD

Credit

Credit
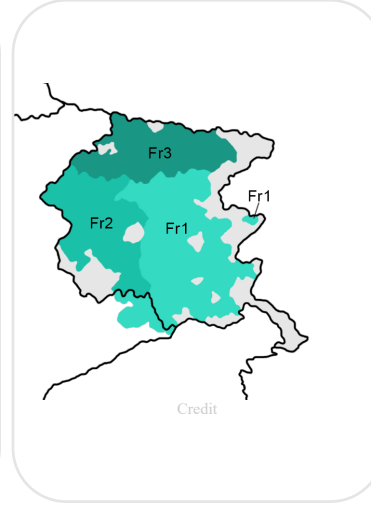
Credit

Credit

Credit
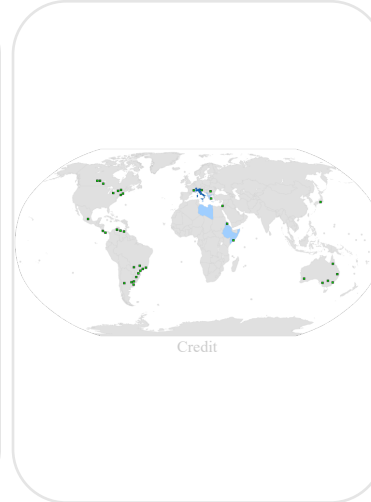
Portuguese

Catalan

Walloon

Friulian

Romanian

Credit

Credit

Credit

Fr3

Fr1

Fr2

Fr1

Credit

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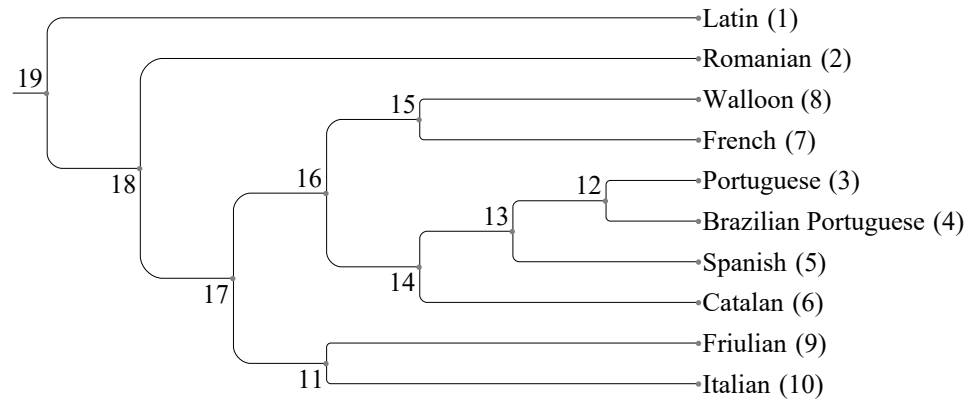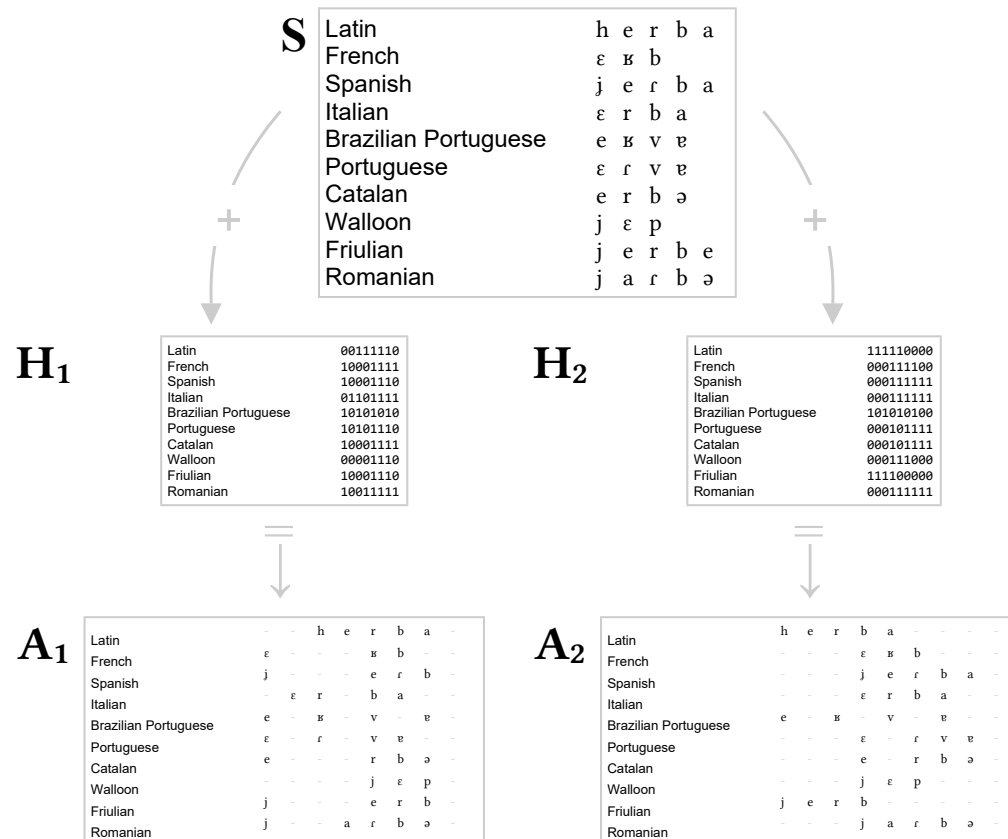
An example tree showing the relationships of $N = 10$ languages.

Alignment

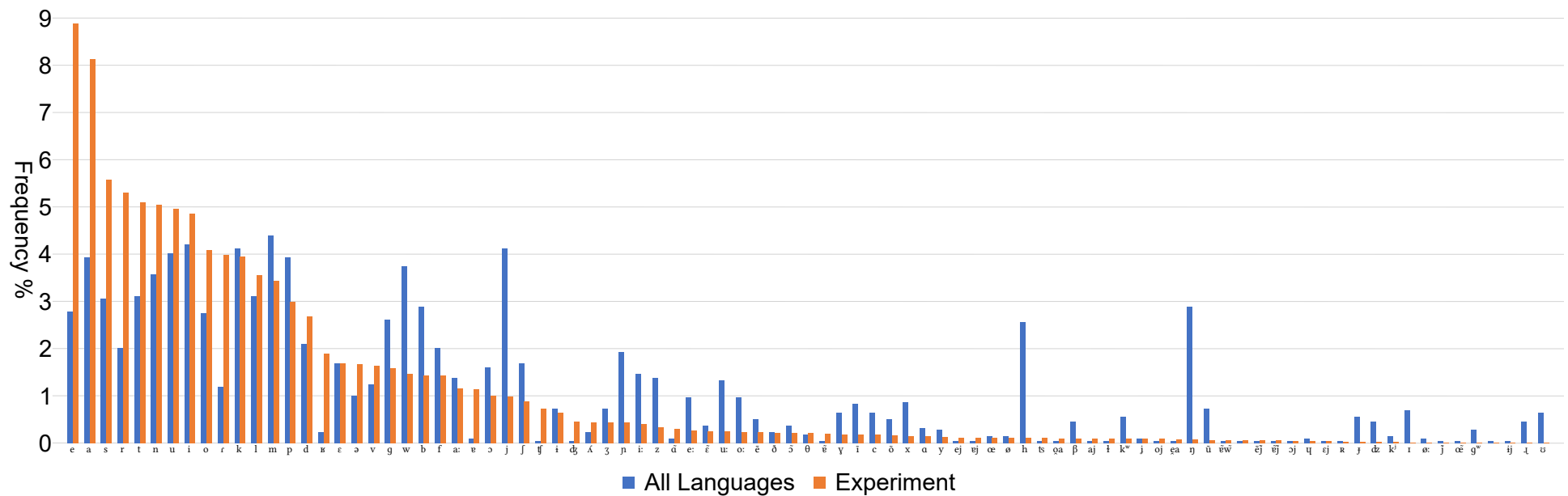Alignments (**A**) are formed from the observed segments (**S**) and a homology map (**H**).

S

Latin	h	e	r	b	a
French	ɛ	ʁ	b		
Spanish	j	e	ɾ	b	a
Italian	ɛ	r	b	a	
Brazilian Portuguese	e	ʁ	v	ɐ	
Portuguese	ɛ	ɾ	v	ɐ	
Catalan	e	r	b	ə	
Walloon	j	ɛ	p		
Friulian	j	e	r	b	e
Romanian	j	a	ɾ	b	ə

H₁

Latin	00111110
French	10001111
Spanish	10001110
Italian	01101111
Brazilian Portuguese	10101010
Portuguese	10101110
Catalan	10001111
Walloon	00001110
Friulian	10001110
Romanian	10011111

H₂

Latin	111110000
French	000111100
Spanish	000111111
Italian	000111111
Brazilian Portuguese	101010100
Portuguese	000101111
Catalan	000101111
Walloon	000111000
Friulian	111100000
Romanian	000111111

A₁

Latin			h	e	r	b	a	
French	ɛ				ʁ	b		
Spanish	j				e	ɾ	b	
Italian		ɛ	r		b	a		
Brazilian Portuguese	e		ʁ		v		ɐ	
Portuguese	ɛ		ɾ		v	ɐ		
Catalan	e				r	b	ə	
Walloon					j	ɛ	p	
Friulian	j				e	r	b	
Romanian	j			a	ɾ	b	ə	

A₂

Latin	h	e	r	b	a			
French				ɛ	ʁ	b		
Spanish				j	e	ɾ	b	a
Italian				ɛ	r	b	a	
Brazilian Portuguese	e		ʁ		v		ɐ	
Portuguese				ɛ		ɾ	v	ɐ
Catalan				e		r	b	ə
Walloon				j	ɛ	p		
Friulian	j	e	r	b				
Romanian				j	a	ɾ	b	ə

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 ɐ̯	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 c	25 ə	26 o̯a	27 t	28 ɑ̃	29 ĩ	30 ẽ̯	31 ʧ	32 ɨ	33 ẽj̃	34 s	35 p	36 z	37 ʃ	38 ð	39 oː
40 k	41 ʦ	42 aː	43 ẽj̃	44 aj	45 θ	46 ej	47 ɐj	48 w	49 ɔ̃	50 h	51 õ	52 v	53 au̯	54 ɛ̃	55 ẽw̃	56 c
57 ɨj	58 ẽ	59 β	60 ł	61 ɑ	62 uːn	63 ɥ	64 y	65 g	66 ɨw	67 ɔj	68 dz	69 ɟ	70 ʤ	71 oj	72 gʷ	73 ŋ	74 ø
75 ẹa	76 ʀ	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

Model: "Linguistically Informed"

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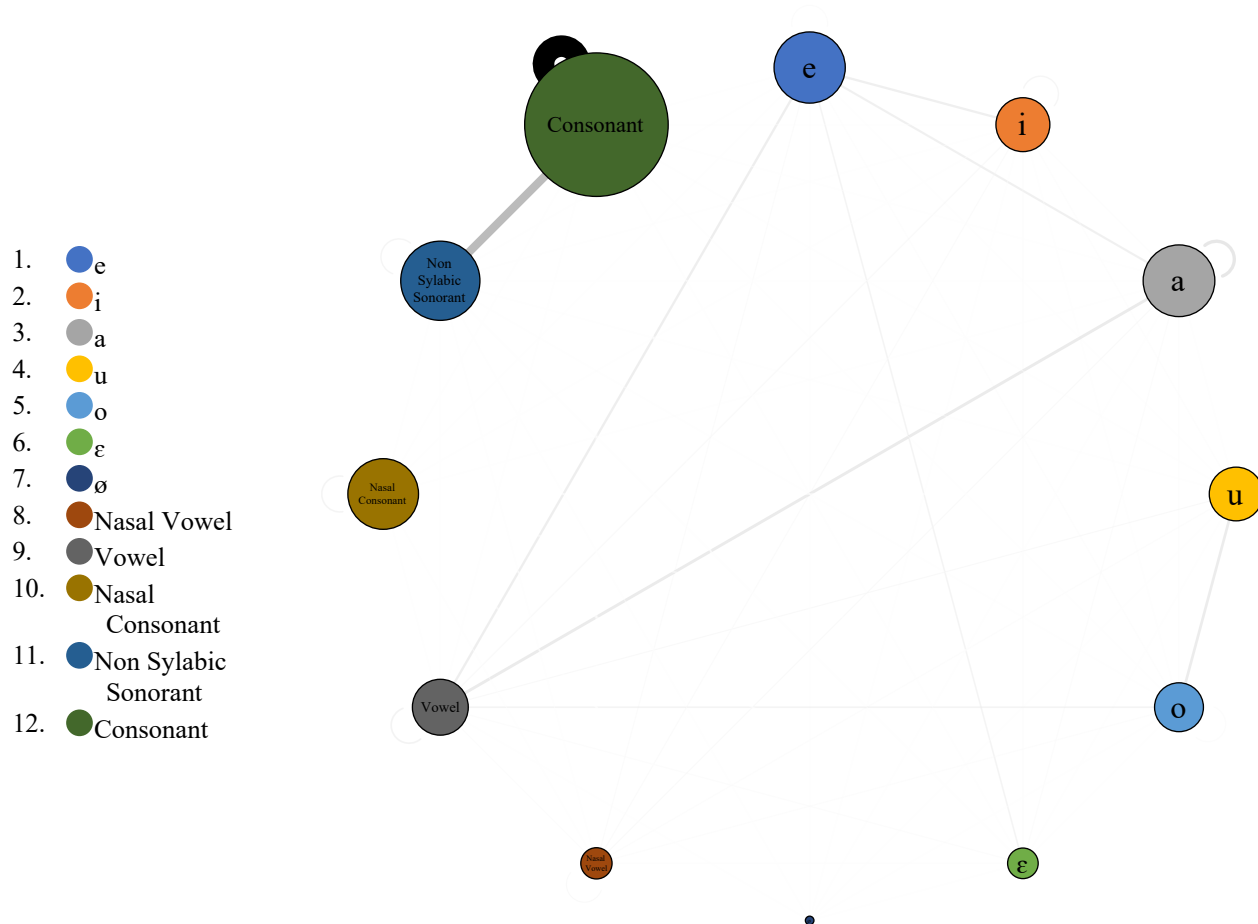2) i   i iː
3) a   a aː
4) u   u uː
5) o   o oː
6) ɛ   ɛ
7) ø   ø øː
8) Nasal Vowel   ɑ̃ ĩ ɛ̃ ɛ̃j̃ ɐ̃j̃ ɔ̃ õ ɛ̃ ɛ̃w̃ ẽ ũ œ̃
9) Vowel   ɐ œ ɔ ə o̯a ɨ aj ej ɐj au̯ ij ɑ y ɨw ɔj oj e̯a ɛj ɪ ʊ
10) Nasal Consonant   m n ɲ ŋ ȷ̃
11) Non Sylabic Sonorant   j l r w
12) Consonant   f b ɾ x ʎ ʁ d t tʃ s p z ʃ ð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

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.

1. e
2. i
3. a
4. u
5. o
6. ε
7. ø
8. Nasal Vowel
9. Vowel
10. Nasal Consonant
11. Non Sylabic Sonorant
12. Consonant

?