

Understanding Language Evolution Using an Event-Based Model

John Huelsenbeck

Department of Integrative Biology

University of California, Berkeley

johnh@berkeley.edu

Introduction

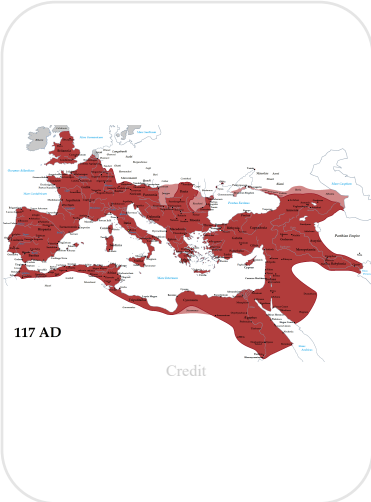
Modern languages are related to one another through a complicated history of divergence and word borrowing. The divergence of languages is caused by the slow change in spoken language as it is passed from parents to offspring. Over time, divergence causes languages to become increasingly different from 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	/ma:no/	1
Russian	/rɔka/	2
Polish	/rɛŋka/	2

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

Languages

Latin



French



Spanish



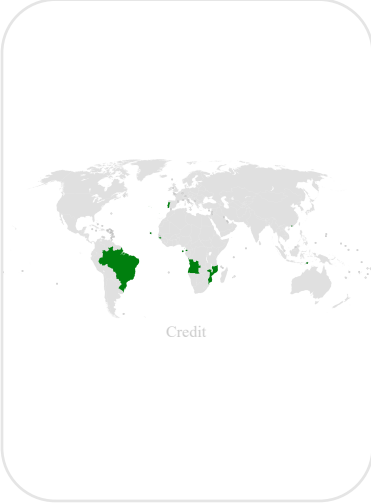
Italian



Brazilian Portuguese



Portuguese



Catalan



Walloon



Friulian

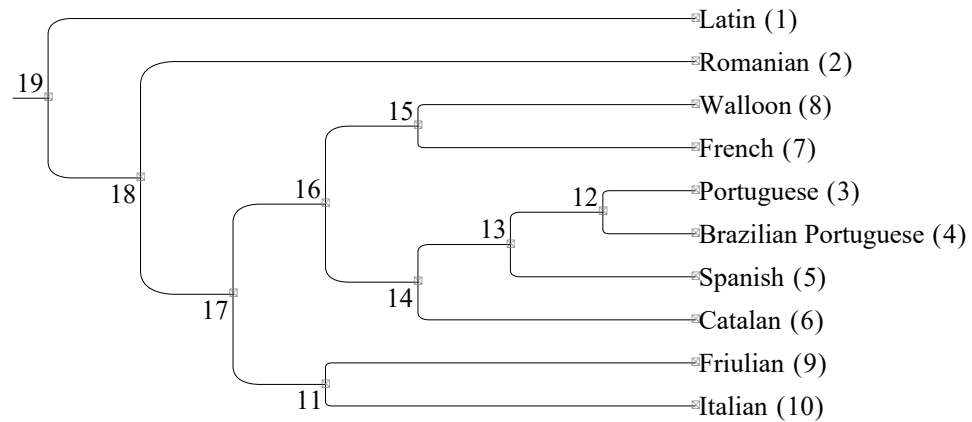


Romanian



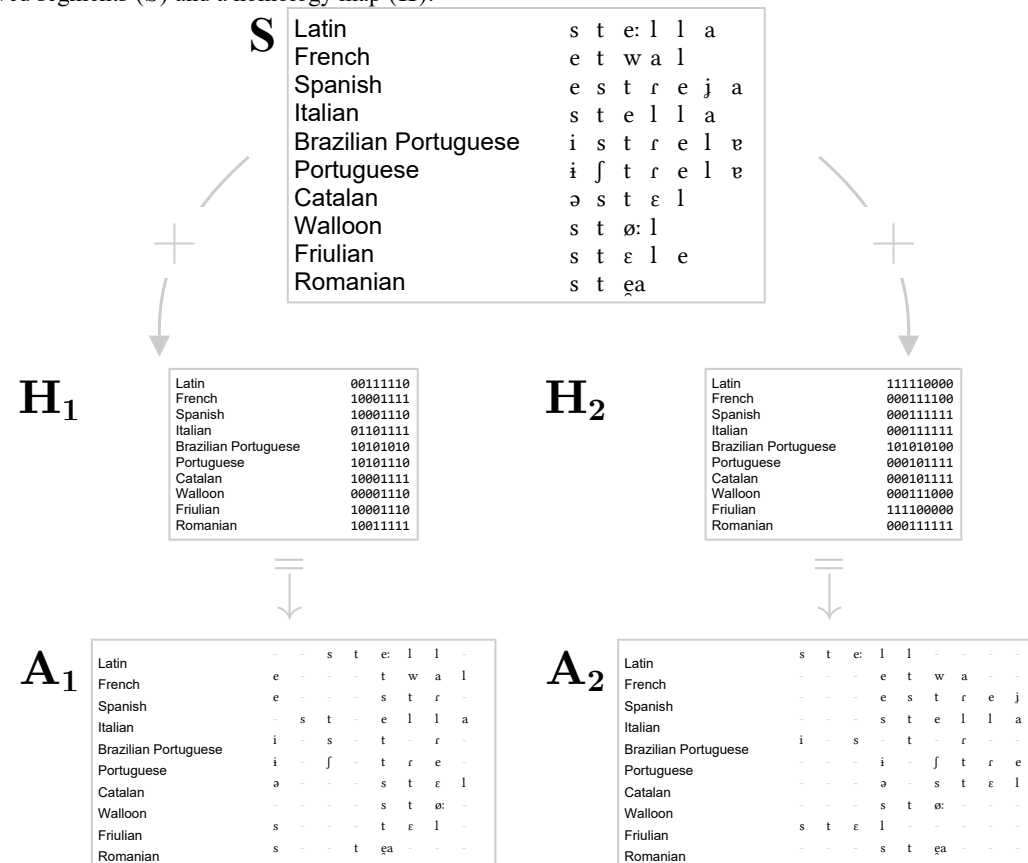
Example Tree

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	o:	4	ʒ	5	ə	6	j	7	o	8	i	9	w	10	ɔ	11	dʒ	12	ɪ	13	j	14	n	15	s	16	u	17	ʃ	18	oj	19	v
20	b	21	i:	22	ʌ	23	l	24	r	25	ɪ	26	h	27	k	28	t	29	a	30	ɐ	31	u:	32	y	33	d	34	f	35	e:	36	m				
37	r	38	œ	39	ɛ	40	x	41	ɾ	42	ɲ	43	ɔa	44	ẽj	45	õ	46	ã	47	ĩ	48	ẽ	49	tʃ	50	p	51	z	52	ð	53	ts				
54	a:	55	ẽj	56	ɾ	57	aj	58	ɛ:	59	θ	60	ej	61	ɐj	62	õ	63	ẽ	64	ẽw	65	ẽ	66	c	67	β	68	ɬ	69	ɥ						
70	dʒ	71	ø:	72	g	73	g ^w	74	ɳ	75	õ:	76	au	77	ø	78	ɛa	79	k ^j	80	tʃ ^j	81	ɟ	82	ɳ ^w	83	k:	84	ç	85	ɣ						
86	ẽ:	87	ɑ	88	ɔ:	89	ɛj	90	k ^w	91	ẽɥ	92	o(w)	93	ɳ	94	tʃ	95	ts	96	ũ	97	ɭ	98	ɑ:	99	ʝ	100	œ								
101	d:	102	t:	103	ɥ	104	ã	105	ɐ	106	ʊ																										

Partition Assignments

Model: “Linguistically Informed”

1 Nasal Vowel

ẽ ĵ ȳ ã ĩ ẽr ẽj õ ɛ ẽw ẽ õ: ẽ: ẽu ũ œ â

2 Vowel

e o: ə o i ɔ ɪ u o j i: i a ʋ u: y e: œ ɛ ɔ̞ a: a j ɛ: e j ʋ j ø: a u̯ ø ɛ̞ a ɑ ɔ: ɛ j o(w) ɑ: u̯ ɛ̞ ʋ

3 Nasal Consonant

$$\mathfrak{n} \ \mathfrak{m} \ \mathfrak{j} \ \mathfrak{y} \ \mathfrak{y}^{\mathfrak{w}} \ \mathfrak{n}_{\overline{\mathfrak{a}}} \ \tilde{\mathfrak{j}}$$

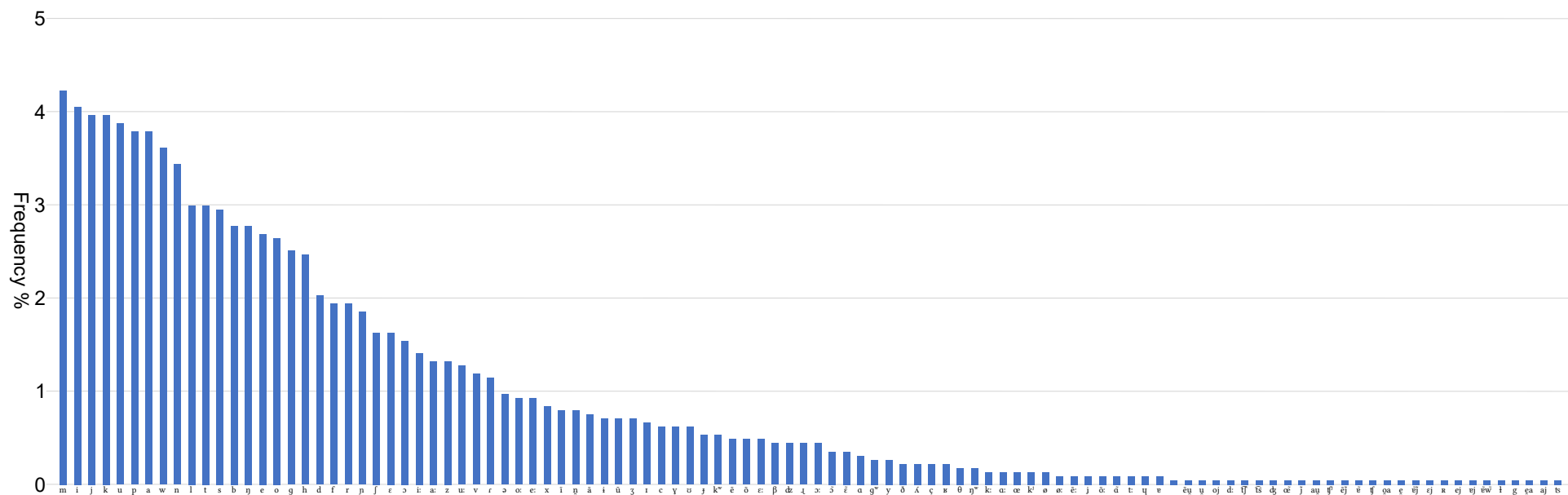
4 Non Sylabic Sonorant

w j l r

5 Consonant

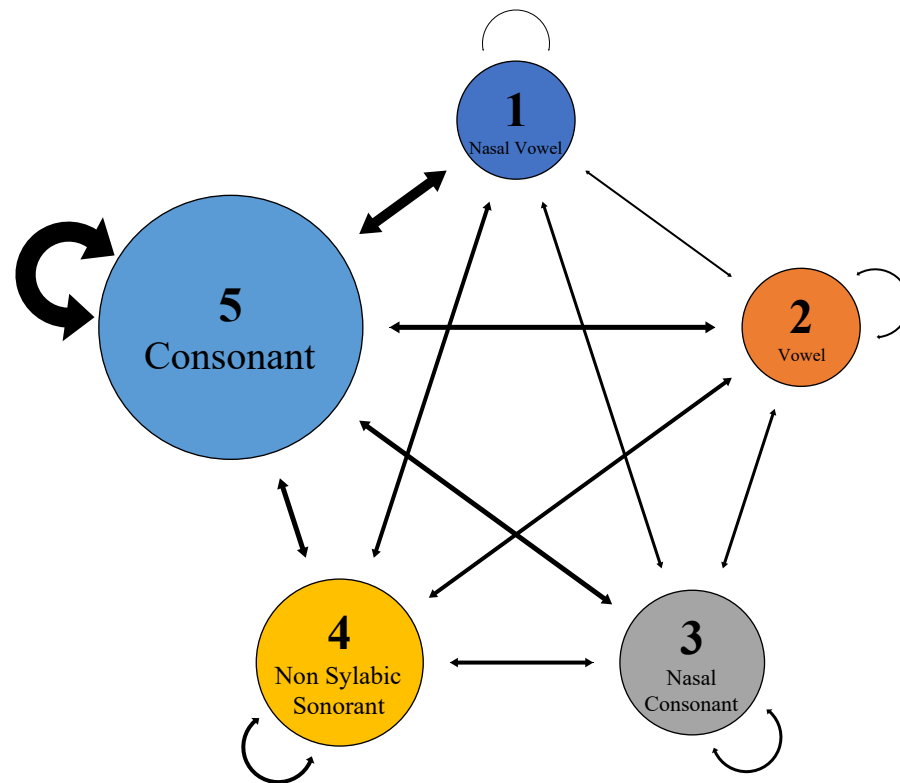
g z j d s f v b A h k t d f r x B t f p z d t s R t c b l u d z g g^w k^j t^j j k: c y k^w t f t s . l d: t:

Prior Segment Frequencies

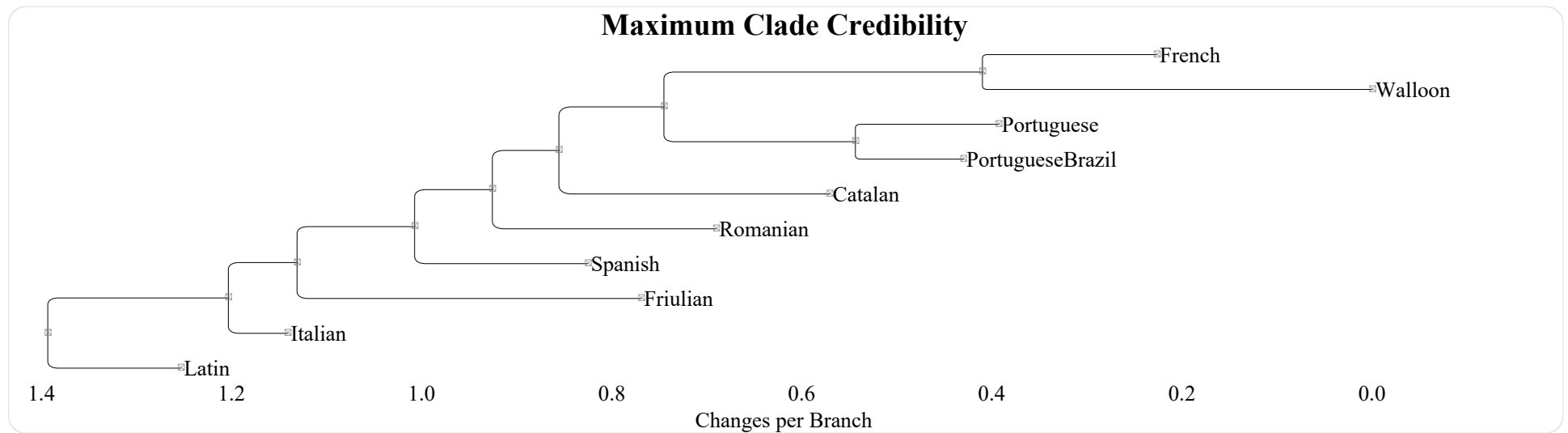


Transition Rates Between Partitions

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



Questions

