Advancing cognate reflex prediction

Incorporating expert evaluations and multi-tiered representations

Tiago **Tresoldi**[1], Freja **Lindgren**[1], Oscar **Billing**[1], John **Huisman**[1] & Fabrício Ferraz **Gerardi**[2]

1 Uppsala Universitet
2 Eberhard Karls Universität Tübingen

56th SLE meeting, Athens, 29.08.2023

UPPSALA UNIVERSITET

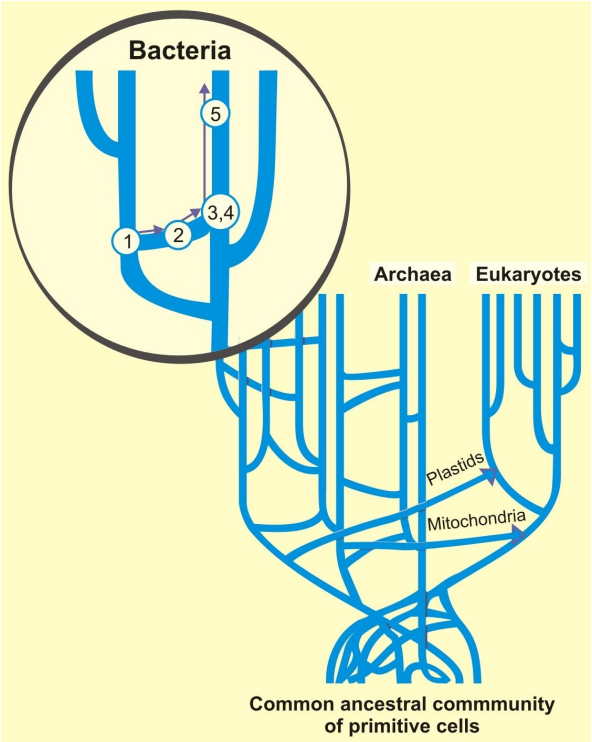
Background

- Different but similar approaches
  - “computational historical linguistics” (Jäger, 2019)
  - “computational phylogenetics” (Bowern, 2018)
  - “phylolinguistics” (Greenhill et al., 2020)
  - “computer-assisted language comparison” (List, 2017)
  - Moscow School of Comparative Linguistics
- Automated steps yielding faster, reproducible, unbiased(?) analyses; however:
  - Focus on tree-building from lexical replacement
  - Minimal commitment to the automation, enhancement, and support of routine tasks in historical linguistics (like inference of sound changes)

DALL-E 2 for “futuristic Tower of Babel”

The Challenge

- Cognate reflex prediction
  - Identifying semantic shifts
  - Cognacy probability and borrowings
  - Imputation of missing reflexes
  - Assistance with reconstructions
  - Building step towards sound change inference
- Different approach
  - To explain and not to predict
  - “interpretable” machine learning



“Horizontal gene transfer” (Wikipedia Commons)

Previous work

Cognate Set	German	English	Dutch
ASH	a ʃ ɛ	æʃ	ɑ s
BITE	b ai s ə n	b ai t	b ɛi t ə
BELLY	b au x	-	b œi k

- Notable earlier attempts:
  - character-based (Beinborn et al., 2013)
  - encoder-decoder (Meloni et al., 2021)
  - low-resource machine-translation (Fourrier et al. (2021)
  - RNN (Dekker and Zuidema, 2021)
  - SYGTYP2022 shared task

- Two contributions from Hammarström et al. (2019)
  - Focus on predicting rather than explaining (Shmueli, 2010)
  - “the most conspicuous **difference** between Computational Historical Linguistics [...] and the classical comparative method is the recognition of **sound changes and their role** in reconstruction and subgrouping. Sound changes are especially powerful for subgrouping as they are typically directional”

Our Approach

- Roots in Chacon & List (2015)
- Novel method combining
  - extended phonological alignments (“multitiers”)
  - expert evaluations
  - focus on explainability
  - ultimate focus on sound changes
- Linguistics groups under study
  - Dravidian
  - Anatolian
  - Japonic
  - Tupi-Guarani

“SumperJumbo black box” (Getty Images)

Handling of phonological data

- Properly modelling sounds/words and sounds changes
  - “properly” means ***not*** treating them as orthographic characters and strings
  - Understand that segments are abstractions and suprasegmentals are essential
- ***`maniphono`*** allows the “symbolic manipulation” of sounds, segments, and sequences, and includes tools for normalizing transcriptions
  - Both human- and machine- representation
  - Building on experience from ***CLTS*** (Anderson et al., 2018)
- ***`alteruphono`*** allows to apply sound changes in forward and backwards direction
  - Necessary for any non-trivial process, such as apophonies and metaphonies
  - Building on experience from ***foma*** (Hulden, 2009)

Alteruphono

802 lines (802 sloc) | 40.1 KB

Raw Blame

Search this file...

ID	RULE	WEIGHT	TEST_ANTE	TEST_POST
0	V s > @1 z @1 / # p|b r _ t|d	1	p r e s t o	p r e z e t o
1	C N > @1 / _ #	1	a g r o g ŋ	a g r o g
2	C > :null: / _ #	1	a d j a: d	a d j a:
5	C > :null: / r _	1	i k ʔ e r j a	i k ʔ e r a
4	s C > :null: / _ #	1	g e g s i s k	g e g s i
5	s|k C > @1 / _ #	1	a k a n k m i k s	a k a n k m i k
6	K > ʔ / _ s	2	a k a n k m i k s	a k a n k m i ʔ s
7	L > d / # _	1	l a b j o p l ɔ l	d a b j o p l ɔ l
8	N S > h @2	3	r i m b s u	r i h b s u
9	N > m / _ #	2	a k θ u n	a k θ u m
10	N > n / _ #	5	a p m b i r o m	a p m b i r o n
11	N V > n @2	1	a k a n k m i k s	a k a n k n i k s
12	N d|l > n @2	1	i m l i m k o	i n l i m k o
13	N v|k|s > n @2	1	i m l i m k o	i m l i n k o

Alteruphono

802 lines (802 sloc) | 40.1 KB

Raw Blame

Search th

1	ID				
2	0				
3	1				
4	2				
5	3				
6	4				
7	5				
8	6				
9	7				
10	8				
11	9				
12	10				
13	11				
14	12	N d|l > n @2	1	i m l i m k o	i n l i m k o
15	13	N v|k|s > n @2	1	i m l i m k o	i m l i n k o

/ p a r b a/

p -> b / _ V

/b a r b a/

Alteruphono

802 lines (802 sloc) | 40.1 KB

Raw Blame

Search th

1	ID				
2	0				
3	1				
4	2				
5	3				
6	4				
7	5				
8	6				
9	7				
10	8				
11	9				
12	10				
13	11				
14	12	N d|l > n @2	1	i m l i m k o	i n l i m k o
15	13	N v|k|s > n @2	1	i m l i m k o	i m l i n k o

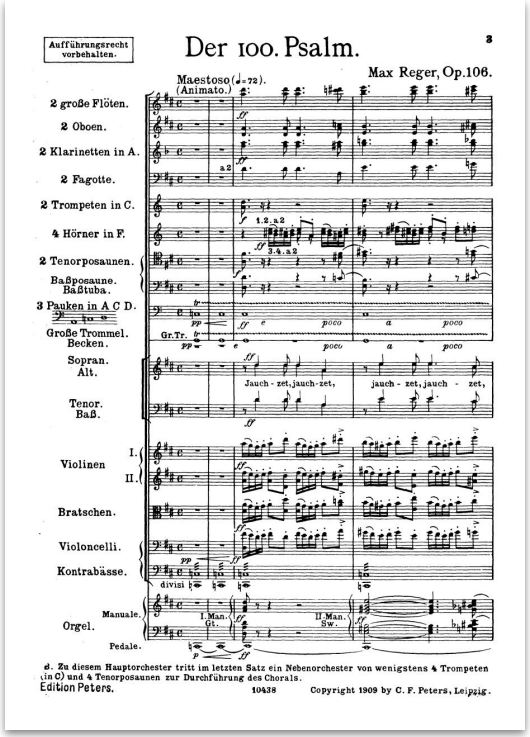
/b a r b a/, /b a r p V/, /p V r b a/, /p V r p V/

p -> b / _ V

/b a r b a/

Multitiered representations

- The “sheet music” of alignments (List)
- Not aligned linear components, but dataset records
- The dataset becomes a 2D-matrix for machine learning
- Designed for improved explainability
  - Identify which tiers (“features”) contribute positively or negatively towards an outcome, how much they contribute, and their interaction effect



Max Reger, Der 100 Psalm
(Wikimedia Commons)

Language	Site #1	Site #2	Site #3	Site #4	Site #5	Site #6
English	f	ɑː	ð	-	ə	ɹ
German	f	ɑː	tʰ	-	ɐ	-
Italian	p	a	d	ɾ	e	-
Spanish	p	a	ð̞	ɾ	e	-

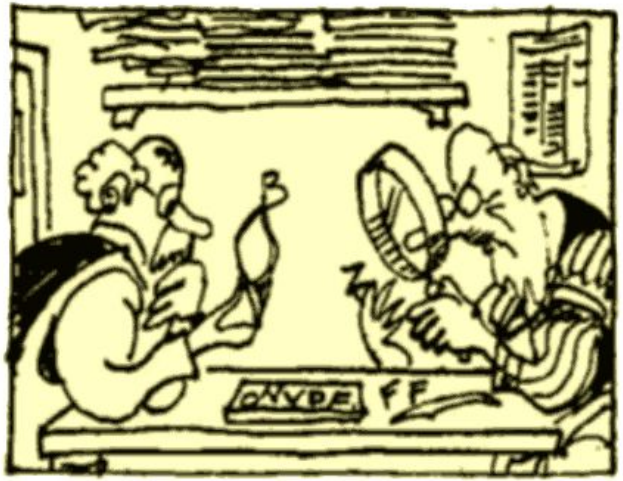
Tier	Site #1	Site #2	Site #3	Site #4	Site #5	Site #6
Segment_EN	f	ɑː	ð	-	ə	ɹ
Segment_DE	f	ɑː	tʰ	-	ɐ	-
Segment_IT	p	a	d	ɾ	e	-
Segment_ES	p	a	ð̞	ɾ	e	-

Tier	Site #1	Site #2	Site #3	Site #4	Site #5	Site #6
Segment_EN	f	ɑː	ð	-	ə	ɹ
Segment_DE	f	ɑː	tʰ	-	ɐ	-
Segment_IT	p	a	d	ɾ	e	-
Segment_ES	p	a	ð̞	ɾ	e	-

Tier	Site #1	Site #2	Site #3	Site #4	Site #5	Site #6
Segment_EN	f	ɑː	ð	-	ə	ɹ
SC_EN	B	A	D	-	A	R
Segment_DE	f	ɑː	tʰ	-	ɐ	-
SC_DE	B	A	T	-	E	-
Segment_IT	p	a	d	ɾ	e	-
SC_IT	B	A	D	R	E	-
Segment_ES	p	a	d̪	ɾ	e	-
SC_ES	B	A	D	R	E	-

Tier	Site #1	Site #2	Site #3	Site #4	Site #5	Site #6
Segment_EN	f	ɑː	ð	-	ə	ɹ
Segment_EN_L1	-	f	ɑː	ð	-	ə
SC_EN	B	A	D	-	A	R
SC_EN_R1	A	D	-	A	R	-
Segment_DE	f	ɑː	tʰ	-	ɐ	-
Segment_DE_L1	-	f	ɑː	tʰ	-	ɐ
SC_DE	B	A	T	-	E	-
SC_DE_R1	-	B	A	T	-	E
Segment_IT	p	a	d	r	e	-
Segment_IT_L1	-	p	a	d	r	e
…	…	…	…	…	…	…

Methodology



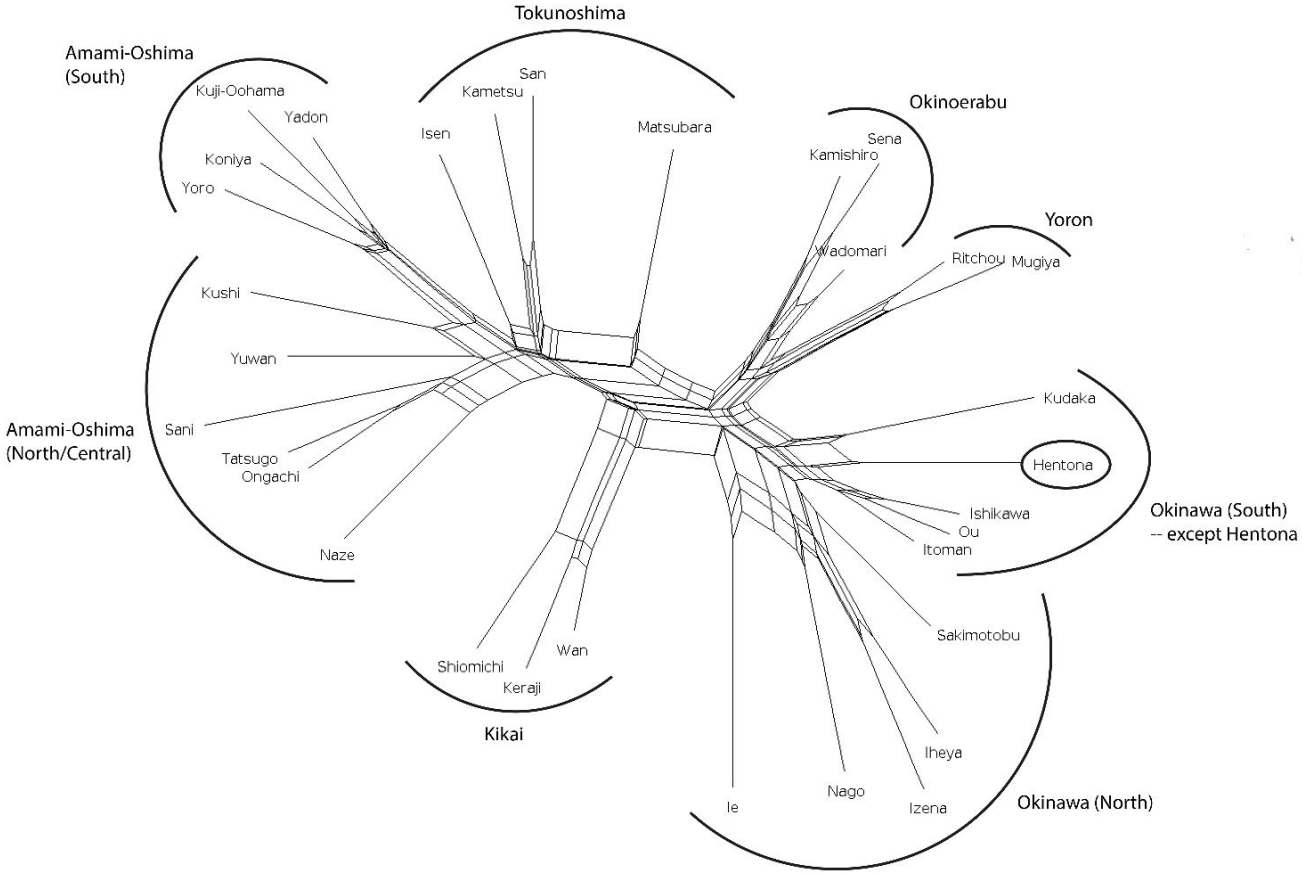
- “Leave-one out”: each reflex is dropped from the dataset before training
  - One training round for each reflex: lots of computation!
- Different models are trained to obtain the word prediction
  - Predictions have probabilities
- Dual evaluation approach
  - Normal statistical evaluation of machine learning
  - Expert evaluations, later incorporated to improve explanations
- Building blocks for sound change inference
- ***Still on-going!***

ID	Dutch	English	German
*bainan	beːn	bəʊn	bain
*balgiz	bɑlx	bɛlɪ	balk
*fadeer	vɑːdər	fɑːðer	faːtər
*fasteenan	vɑstə	fɑːst	fastən
*fuloon		fəʊl	foːlən
*kalbaz	kɑlf	kɑːf	kalp
…	…	…	…

ID	Dutch	English	German
*bainan	beːn	bəʊn	bain
*balgiz	bɑlx	bɛlɪ	balk
*fadeer	vɑːdər	fɑːðer	
*fasteenan	vɑstə	fɑːst	fastən
*fuloon		fəʊl	foːlən
*kalbaz	kɑlf	kɑːf	kalp
…	…	…	…

	Site #1	Site #2	Site #3	Site #4	Site #5
Reference	f	a:	t	ə	r
Best	f (0.96)	a: (0.83)	t (0.88)	ə (0.92)	r (0.97)
Second best	v (0.03)	a (0.08)	d (0.10)	e (0.02)	- (0.02)
…	…	…	…	…	…

*fadeer	vɑːdər	fɑːðer	
*fasteenan	vɑstə	fɑːst	fastən
*fuloon		fəʊl	foːlən
*kalbaz	kɑlf	kɑːf	kalp
…	…	…	…


Tokunoshima
Amami-Oshima
(South)
Kuji-Oohama
Yadon
Koniya
Yoro
San
Kametsu
Isen
Matsubara
Okinoerabu
Sena
Kamishiro
Yoron
Wadomari
Ritchou
Mugiya
Kushi
Yuwan
Amami-Oshima
(North/Central)
Sani
Tatsugo
Ongachi
Naze
Kudaka
Hentona
Ishikawa
Ou
Itoman
Okinawa (South)
-- except Hentona
Sakimotobu
Shiomichi
Wan
Keraji
Kikai
Iheya
Nago
Ie
Izena
Okinawa (North)


Kikai
Amami-Oshima
Tokunoshima
Okinoerabu
Yoron
Okinawa

Doculect	CONCEPT	IPA Alignment	Correspondences
Chabana	NOSE	p a n a	92 414 215 1881
Hentona	NOSE	ɸ a n a	92 414 215 1881
Izena	NOSE	ɸ a n aː	92 414 215 1881
Kametsu	NOSE	h a n a	92 414 215 1881
Chabana	WIND	h a d i	2971 414 298 299
Hentona	WIND	h a dʒ i	2971 414 298 299
Izena	WIND	h a dʒ iː	2971 414 298 299
Kametsu	WIND	k a d ɨ	2971 414 298 299

Chabana	p
Hateruma	p
Iheya	h
Ikema	h
Kudaka	p
Kuroshima	p
Matsubara	h
…	…

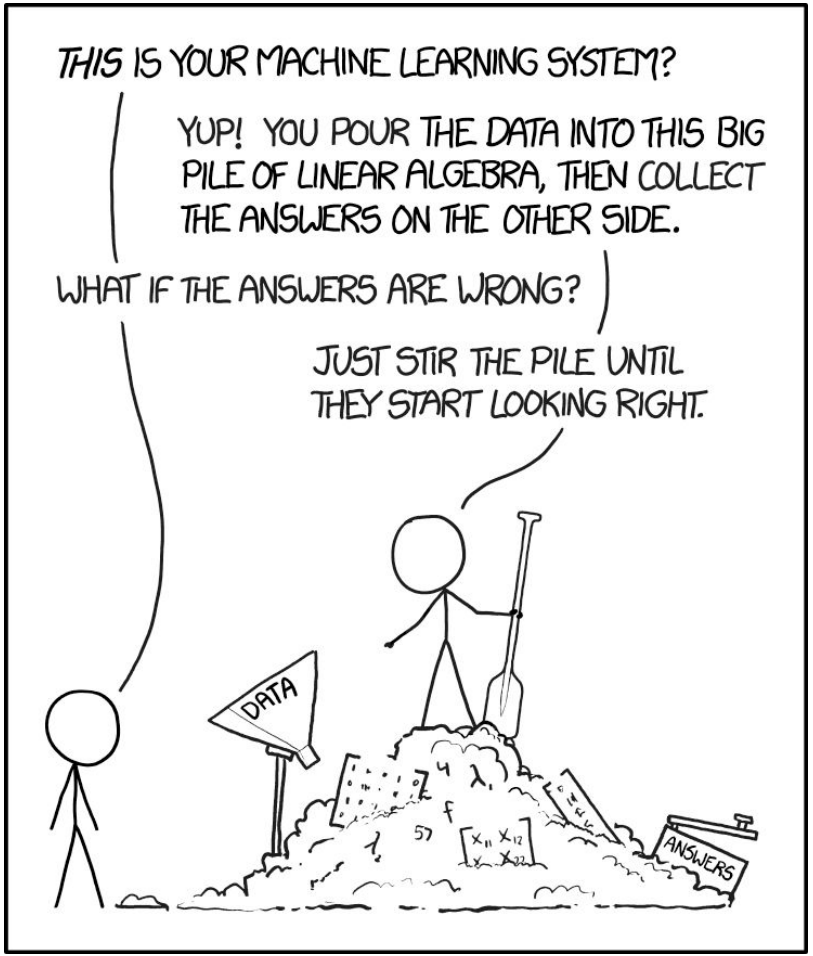
Docu		IPA Alignment	Correspondences
Chaba		p a n a	92 414 215 1881
Hento		ɸ a n a	92 414 215 1881
Izena		ɸ a n a:	92 414 215 1881
Kame		h a n a	92 414 215 1881
Chaba		h a d i	2971 414 298 299
Hento	WIND	h a dʒ i	2971 414 298 299
Izena	WIND	h a dʒ i:	2971 414 298 299
Kametsu	WIND	k a d ɨ	2971 414 298 299

Doculect	CONCEPT	IPA Alignment	Correspondences
Chabana	NOSE	p a n a	92 414 215 1881
		ɸ a n a	92 414 215 1881
		ɸ a n aː	92 414 215 1881
		h a n a	92 414 215 1881
		h a d i	2971 414 298 299
		h a dʒ i	2971 414 298 299
		h a dʒ iː	2971 414 298 299
		k a d i	2971 414 298 299

Chabana	a
Hateruma	a
Iheya	a
Ikema	a
Kudaka	a
Kuroshima	a
Matsubara	a
…	…

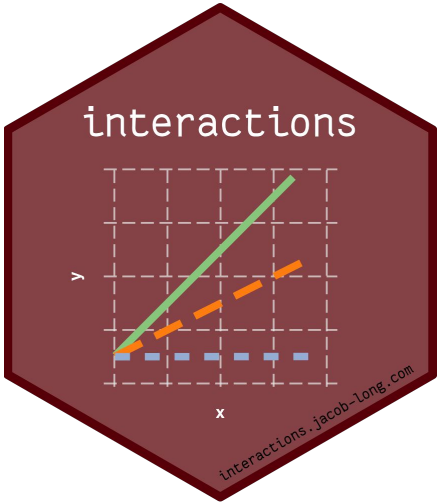
Chabana	d
Hateruma	tʃ
Iheya	dʒ
Ikema	d
Kudaka	r
Kuroshima	dʒ
Matsubara	dʒ
…	…

Docul		IPA Alignment	Correspondences
Chaba		p a n a	92 414 215 1881
Hentor		ɸ a n a	92 414 215 1881
Izena		ɸ a n aː	92 414 215 1881
Kamet		h a n a	92 414 215 1881
Chaba		h a d i	2971 414 298 299
Hentor		h a dʒ i	2971 414 298 299
Izena		h a dʒ iː	2971 414 298 299
Kametsu	WIND	k a d ɨ	2971 414 298 299



“Machine learning” (XKCD 1838)

Approach



- Prediction first (computers), explanation second (experts)
- Advanced statistical methods
  - Probabilities
  - Interaction effects
- A novel approach to cognate detection
- Ultimate result of reflex prediction: probabilities for segments and words
- Ultimate result of sound change inference: probabilities for sequences of sound changes, given in standard notation

Conclusions

- To explain and not to predict
- Not a black-box approach
- Agnostic about the machine learning solutions

- Encourage collaboration and a computer-assisted approach
- Paving the way for the Graal of Computational Historical Linguistics: inference of sound change history (relative chronology)