Trimming Phonetic Alignments Improves the Inference of Sound Correspondence Patterns from Multilingual Wordlists

Frederic Blum
MPI-EVA

Johann-Mattis List
University of Passau & MPI-EVA

The 5th Workshop on Research in
Computational Linguistic Typology and Multilingual NLP
06.05.2023

Goals

Trimming Phonetic Alignments

- Improve the regularity of automatically inferred correspondence patterns among cognate sets from related languages
- Eliminate noisy data: morphemes and non-cognate elements
- Shorten long-tail distribution of correspondence patterns with few occurrences

Correspondence Patterns in Linguistics

	I		II		II		I	
Language A	t	a	h	e	h	i	t	u
Language B	t^h	a	x	e	x	u	t^h	i
Language C	t	a	x	e	x	u	t	i
Language D	ts	a	x	e	x	u	ts	i

Figure: Corresponding alignment sites in a set of four fictitious languages.

Correspondence Patterns

- Patterns are formed by a set of sound correspondences
- Shared between multiple languages, not language pairs
- Recurring correspondence patterns form the basis for the reconstruction of proto-languages

Trimming in Historical Linguistics

Figure: Trimming morphemes in Quechua. The root is combined with different morphemes in some varieties.

Trimming in Historical Linguistics

Figure: Trimming morphemes in Quechua. The root is combined with different morphemes in some varieties.

Examples for trimming

- Trimming is often practiced without being made explicit
- Explicit examples are Payne (1991) and Cayón & Chacon (2022)

Trimming of Alignment Sites in Computational Biology

How does the trimming proceed?

- Trimming DNA sequence alignments
- Reduce noise in the data for improved phylogenetic analysis (Talavera & Castresana 2007)

Trimming of Alignment Sites in Computational Biology

How does the trimming proceed?

- Trimming DNA sequence alignments
- Reduce noise in the data for improved phylogenetic analysis (Talavera & Castresana 2007)

Methods for trimming

- Removing sites with many gaps (Capella-Gutiérrez et al. 2009)
- Removing sites based on entropy values (Criscuolo & Gribaldo 2010)

Datasets Used in the Study

Data set	Lang.	Conc.	Cog.-Sets	Words	Source
constenlachibchan	25	106	213	1216	Constenla Umaña (2005)
crossandean	20	150	223	2789	Blum et al. (forthcoming)
dravlex	20	100	179	1341	Kolipakam et al. (2018)
felekesemitic	21	150	271	2622	Feleke (2021)
hattorijaponic	10	197	235	1710	Hattori (1973)
houchinese	15	139	228	1816	Hóu (2004)
leekoreanic	15	206	233	2131	Lee (2015)
robinsonap	13	216	253	1424	Robinson & Holton (2012)
walworthpolynesian	20	205	383	3637	Walworth (2018)
zhivlovobugrian	21	110	182	1974	Zhivlov (2011)

Table: Number of languages, concepts, non-singleton cognate sets and total entries across the different datasets

Datasets Used in the Study

Data set	Lang.	Conc.	Cog.-Sets	Words	Source
constenlachibchan	25	106	213	1216	Constenla Umaña (2005)
crossandean	20	150	223	2789	Blum et al. (forthcoming)
dravlex	20	100	179	1341	Kolipakam et al. (2018)
felekesemitic	21	150	271	2622	Feleke (2021)
hattorijaponic	10	197	235	1710	Hattori (1973)
houchinese	15	139	228	1816	Hóu (2004)
leekoreanic	15	206	233	2131	Lee (2015)
robinsonap	13	216	253	1424	Robinson & Holton (2012)
walworthpolynesian	20	205	383	3637	Walworth (2018)
zhivlovobugrian	21	110	182	1974	Zhivlov (2011)

Table: Number of languages, concepts, non-singleton cognate sets and total entries across the different datasets

Standardized datasets

- Lexibank-datasets (List et al. 2022) are openly available
- Cognacy annotated manually by dataset creators

Trimming Strategies

Language	Core-oriented								Gap-oriented						
Language A	s	-	t	e	r	b	-		s	-	t	e	r	b	-
Language B	m	e	t^h	e	-	-	-		m	e	t^h	e	-	-	-
Language C	-	a	t	e	-	b	u		-	a	t	e	-	b	u
Language D	-	-	t	e	-	b	-		-	-	t	e	-	b	-
Gap proportion	0.5	0.5	0.0	0.0	0.75	0.25	0.75		0.5	0.5	0.0	0.0	0.75	0.25	0.75

Figure: Artificial example for the computation of gap profiles followed by trimming using the *core-oriented* (left) and the *gap-oriented* strategy (right).

Computational Details

- Minimal CV/VC skeleton is preserved in all settings
- Sites with more than 50% gaps are trimmed

Regularity thresholds

How do we measure regularity?

- Count number of occurrences for each correspondence pattern
- All patterns with at least three occurrences are considered to be ‘regular’

Regularity thresholds

How do we measure regularity?

- Count number of occurrences for each correspondence pattern
- All patterns with at least three occurrences are considered to be 'regular'
- How many patterns in a cognate set are above this threshold?

Regularity thresholds

How do we measure regularity?

- Count number of occurrences for each correspondence pattern
- All patterns with at least three occurrences are considered to be 'regular'
- How many patterns in a cognate set are above this threshold?
- All words with more than 75% of regular patterns are analyzed as 'regular'

Results

	Original		Core		Gap	
Dataset	P	W	P	W	P	W
constenlachibchan	0.71	0.50	0.69	0.46	**0.76**	**0.51**
crossandean	0.73	0.58	0.74	0.60	**0.75**	**0.64**
dravlex	0.56	0.23	0.57	0.27	**0.61**	**0.31**
felekesemitic	0.55	0.22	0.58	0.25	**0.62**	**0.29**
hattorijaponic	0.58	0.33	0.57	0.33	**0.59**	**0.38**
houchinese	0.65	0.40	0.65	0.42	**0.69**	**0.45**
leekoreanic	0.44	0.21	0.47	0.20	**0.52**	**0.22**
robinsonap	0.64	0.36	0.65	0.37	**0.67**	**0.41**
walworthpolynesian	0.66	0.40	0.66	0.40	**0.72**	**0.48**
zhivlovobugrian	0.57	0.24	0.58	0.26	**0.61**	**0.28**

Table: Proportion of regular correspondence patterns (P) and regular words (W) across all datasets after trimming.

Results

	Original		Core		Gap	
Dataset	P	W	P	W	P	W
constenlachibchan	0.71	0.50	0.69	0.46	**0.76**	**0.51**
crossandean	0.73	0.58	0.74	0.60	**0.75**	**0.64**
dravlex	0.56	0.23	0.57	0.27	**0.61**	**0.31**
felekesemitic	0.55	0.22	0.58	0.25	**0.62**	**0.29**
hattorijaponic	0.58	0.33	0.57	0.33	**0.59**	**0.38**
houchinese	0.65	0.40	0.65	0.42	**0.69**	**0.45**
leekoreanic	0.44	0.21	0.47	0.20	**0.52**	**0.22**
robinsonap	0.64	0.36	0.65	0.37	**0.67**	**0.41**
walworthpolynesian	0.66	0.40	0.66	0.40	**0.72**	**0.48**
zhivlovobugrian	0.57	0.24	0.58	0.26	**0.61**	**0.28**

Table: Proportion of regular correspondence patterns (P) and regular words (W) across all datasets after trimming.

Summary

- Gap-oriented trimming shows the best results for all datasets
- Datasets with low internal diversity show the fewest improvements

Comparing the Random Model

Dataset	Core	Gap
constenlachibchan	0.58	0.00
crossandean	0.02	0.00
dravlex	0.00	0.00
felekesemitic	0.17	0.01
hattorijaponic	0.40	0.00
houchinese	0.05	0.00
leekoreanic	0.54	0.06
robinsonap	0.34	0.00
walworthpolynesian	0.11	0.00
zhivlovobugrian	0.12	0.05

Table: Percentage of models with random deletion of alignment sites that achieved higher regularity than the respective trimming model.

Successful Removal of Irregular Patterns

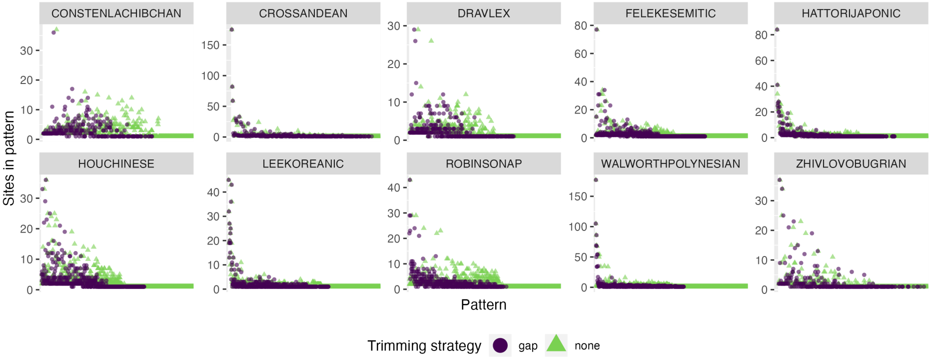

Figure: Distribution of alignment sites per pattern with gap-oriented trimming and without. Each point on the x-axis represents one correspondence pattern.

Example I: Successful Trimming in Chibchan

Boruca	-	-	b	ɾ	u	-	ŋ	-	-	-
Cabecar	-	-	b	-	u	-	ɹ	i	t	u
Chimila	-	-	b	-	u	h	ŋ	a	ʔ	-
Malayo	-	-	b	-	ɨ	-	n	-	-	-
Ngabere	ŋ	ɯ	b	ɾ	ɯ	-	-	-	-	-
Proto-Chibchan			ᵐb		ũ		ⁿd			

Figure: Gap-oriented trimming for the cognate words of ashes

Evaluation

- Reconstruction provided by Pache (2018)
- Trimming identifies problematic alignment sites and removes them

Example II: Problematic Trimming in Chibchan

Boruca	d	i	ʔ
Bribri	d	i	ʔ
Buglere	tʃ	i	-
Cogui	n	i	-
Ngabere	ɲ	ɣ	-
Proto-Chibchan	^{n}d	i	ʔ

Figure: Trimming for the cognate words of `water`

Evaluation

- Reconstruction provided by Pache (2018)
- Our strategy erroneously eliminates a site that includes reconstructed segments

Outlook

What we have

- Trimming improves the regularity of inferred correspondence patterns
- Shortening of the distribution tail of patterns with few alignment sites
- Promising transfer of trimming to historical linguistics

Outlook

What we have

- Trimming improves the regularity of inferred correspondence patterns
- Shortening of the distribution tail of patterns with few alignment sites
- Promising transfer of trimming to historical linguistics

Where we want to go

- Find the best thresholds for gaps and regularity
- Use inferred correspondence patterns for sound reconstruction

References I

Blum, Frederic, Carlos Barrientos, Adriano Ingunza & Zoe Poirier. Forthcoming. A phylolinguistic classi cation of the Quechua language family. Indiana 0(0). 1 20. https://doi.org/https://doi.org/10.31235/osf.io/twu6a .

Capella-Gutiérrez, Salvador, José M. Silla-Martínez & Toni Gabaldón. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15). 1972 1973. https://doi.org/10.1093/bioinformatics/btp348 .

Cayón, Luis & Thiago Chacon. 2022. Diversity, multilingualism and inter-ethnic relations in the long-term history of the Upper Rio Negro region of the Amazon. Interface Focus 13(1). https://doi.org/10.1098/rsfs.2022.0050 .

References II

Constenla Umaña, Adolfo. 2005. ¾Existe relación genealógica entre las lenguas misumalpas y las chibchenses? Estudios de Lingüística Chibcha 14. 7 85.

Criscuolo, Alexis & Simonetta Gribaldo. 2010. BMGE (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evolutionary Biology 10(1). 210. https://doi.org/10.1186/1471-2148-10-210 .

Feleke, Tekabe Legesse. 2021. Ethiosemitic languages: classi cations and classi cation determinants. Ampersand . 100074. https://doi.org/10.1016/j.amper.2021.100074 .

References III

Hattori, Shiro. 1973. Japanese dialects. In Henry M. Hoenigswald & Robert H. Langacre (eds.), Diachronic, areal and typological linguistics (Current Trends in Linguistics 11), 368 400. The Hague & Paris: Mouton. https://doi.org/10.1515/9783111418797-017.

Hóu, J ngy (ed.). 2004. Xiàndài hànyu fangyán ynkù [phonological database of chinese dialects]. Shànghai: Shànghai Jiàoyù.

Kolipakam, Vishnupriya, Fiona M. Jordan, Michael Dunn, Simon J. Greenhill, Remco Bouckaert, Russell D. Gray & Annemarie Verkerk. 2018. A bayesian phylogenetic study of the Dravidian language family. Royal Society Open Science 5(3). 171504. https://doi.org/10.1098/rsos.171504.

References IV

Lee, Sean. 2015. A sketch of language history in the Korean Peninsula. PLOS ONE 10(5). e0128448. https://doi.org/10.1371/journal.pone.0128448. https://doi.org/10.1371%5C%2Fjournal.pone.0128448.

List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch & Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. Scienti c Data 9(316). 1 31. https://doi.org/10.1038/s41597-022-01432-0.

Pache, Matthias. 2018. Contributions to Chibchan Historical Linguistics. Universiteit Leiden dissertation. https://hdl.handle.net/1887/67094.

References V

Payne, David L. 1991. A Classification of Maipuran (Arawakan) Languages Based on Shared Lexical Retentions. In Desmond C. Derbyshire & Geoffrey K. Pullum (eds.), *Handbook of amazonian languages*. Berlin, New York: Mouton De Gruyter. https://doi.org/10.1515/9783110854374.

Robinson, Laura C & Gary Holton. 2012. Internal classification of the Alor-Pantar language family using computational methods applied to the lexicon. *Language Dynamics and Change* 2(2). 123–149. https://doi.org/10.1163/22105832-20120201.

Talavera, Gerard & Jose Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* 56(4). 564–577. https://doi.org/10.1080/10635150701472164.

References VI

Walworth, Mary. 2018. *Polynesian Segmented Data*. Version v1. https://doi.org/10.5281/zenodo.1689909.

Zhivlov, Mikhail. 2011. Annotated Swadesh wordlists for the Ob-Ugrian group (Uralic family). In George S. Starostin (ed.), *The Global Lexicostatistical Database*. Moscow: RGGU. http://starling.rinet.ru/new100/oug.xls.