Fundamentals of Computer-Assisted Language Comparison

National Taiwan University

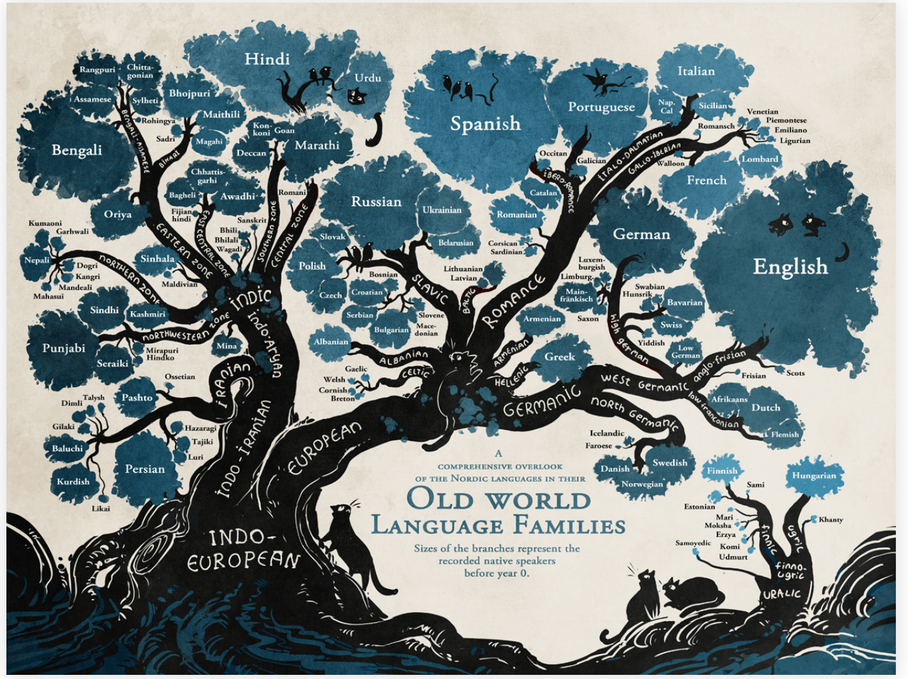
2019.06.28

Introduction

Tiago Tresoldi

Historical linguistics

- HL is the general scientific study of linguistic change and evolution in time
- HL is frequently taken as a synonym for "comparative linguistics", or even for "Indo-European studies"
- Laymen are more familiar with family trees and proto-forms
  - English "water", from Proto-Germanic *watōr, from PIE *wódr̥
  - Mandarin 水 shuǐ, from Old Chinese *s.turʔ ("that which flows"), from Proto-Sino-Tibetan *lwi(j) ("flow, stream")

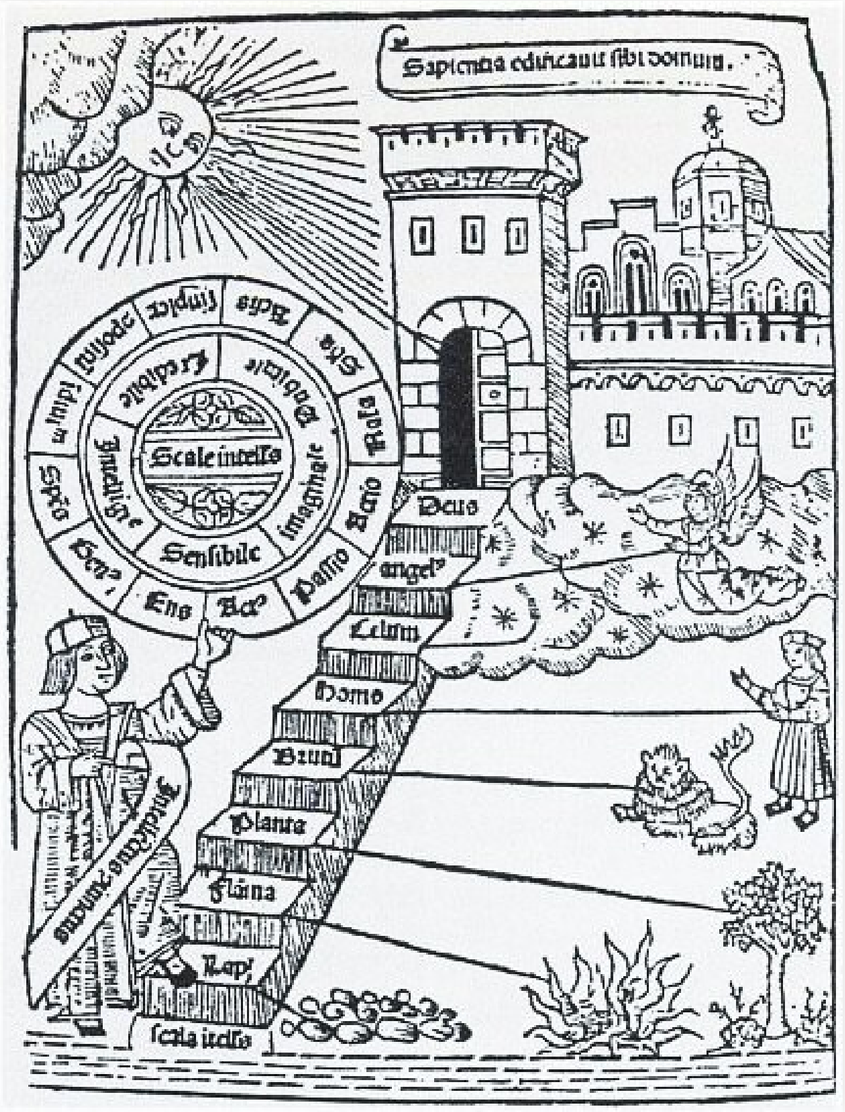


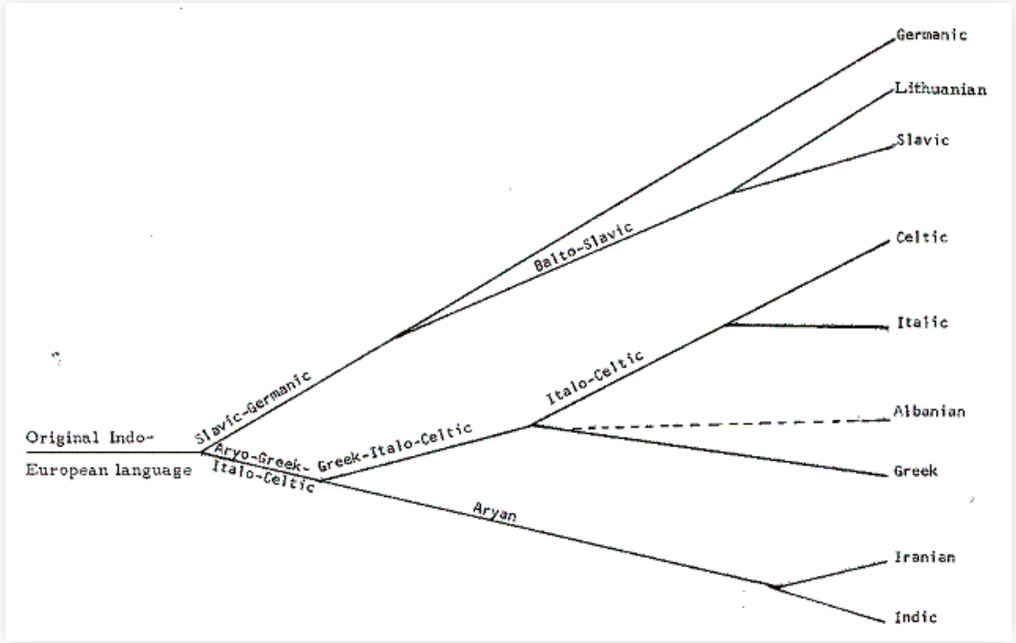
History of the comparative method

- Philosophers in Europe and Asia have debated for millenia how:
  - Languages show similarities that cannot be explained by chance alone
  - Languages change
- As a branch of philology, historical linguistics was born as a "hot" science in the 17th century
  - Colonial enterprises, e.g. the analyses of Van Boxhorn (1612-1653) and the reconstructions of William Wotton (1713)
  - Religious missions, especially Jesuitic, e.g. Matteo Ricci and Xu Guangqi 徐光啓 (16th-17th century) and Lorenzo Hervás (1735-1809)
  - "Orientalism" as in William Jones' discourse to the Asiatic Society (1786)

Comparative method -I

Mental model of "stair" replaced by that of "tree"

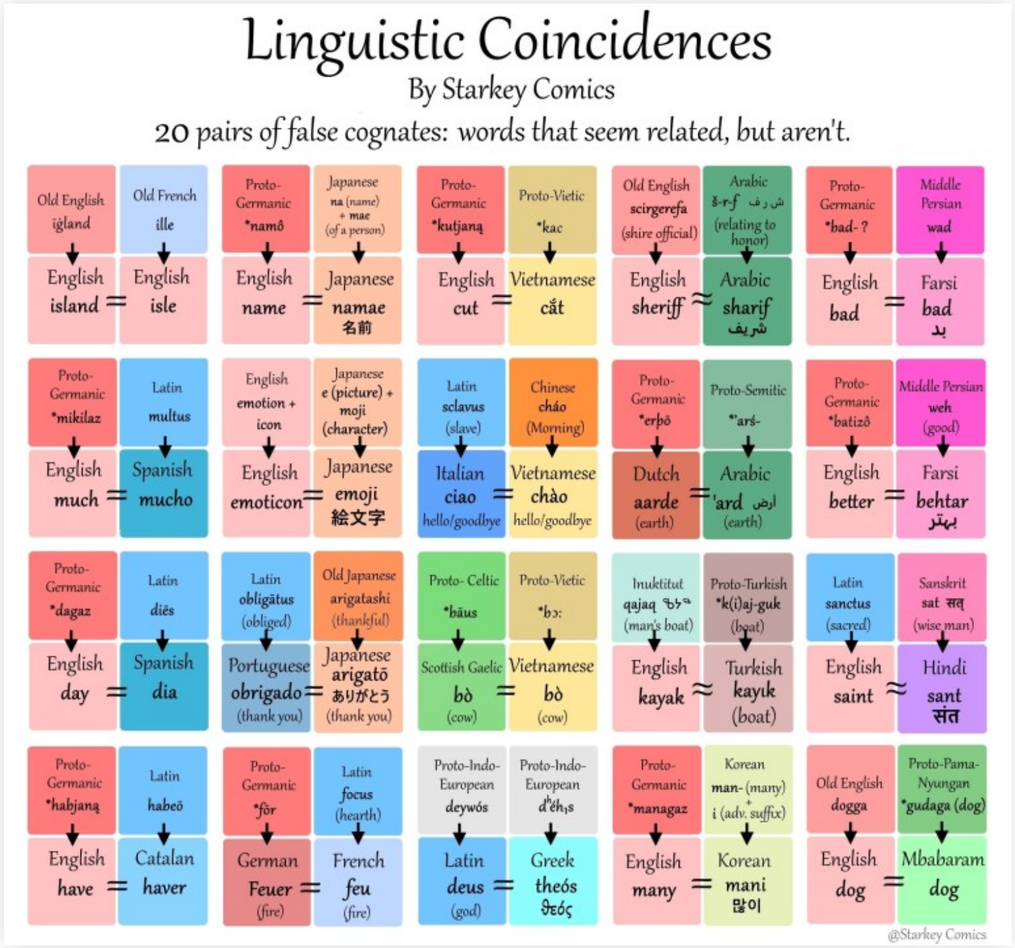




Comparative method -II

- Progressive influence of Darwin and biological analogies
- German promotion of "Indo-Germanic" studies, leading to the Neogrammarian tenets including:
  - Regularity of sound changes
  - Immediate and total effect of sound changes

Traditional workflow



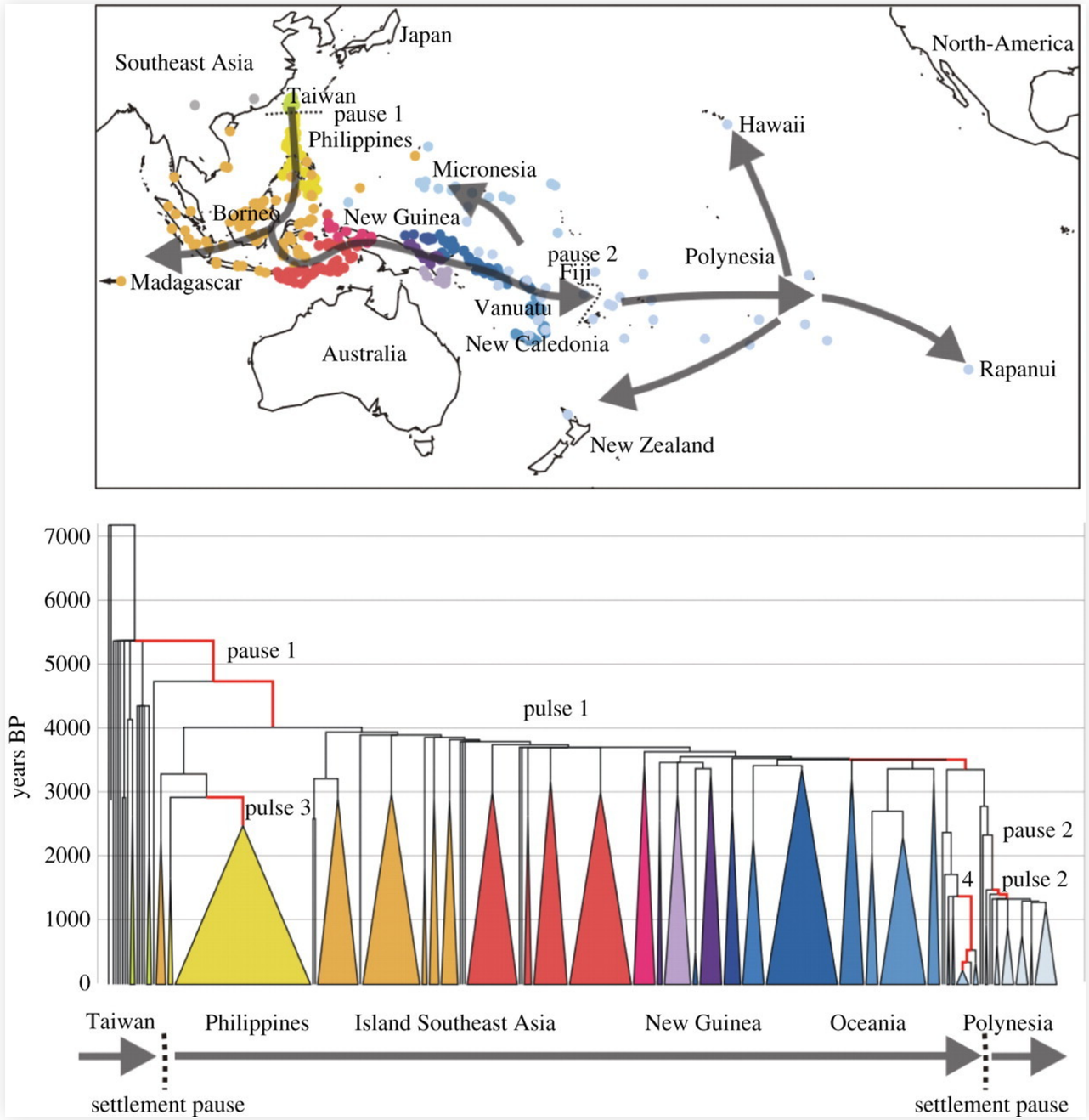
- Collection of data
- Identification of cognates
- Study of correspondences
- Reconstruction of sound changes
- Analysis of typology
- Correction of errors and repetition

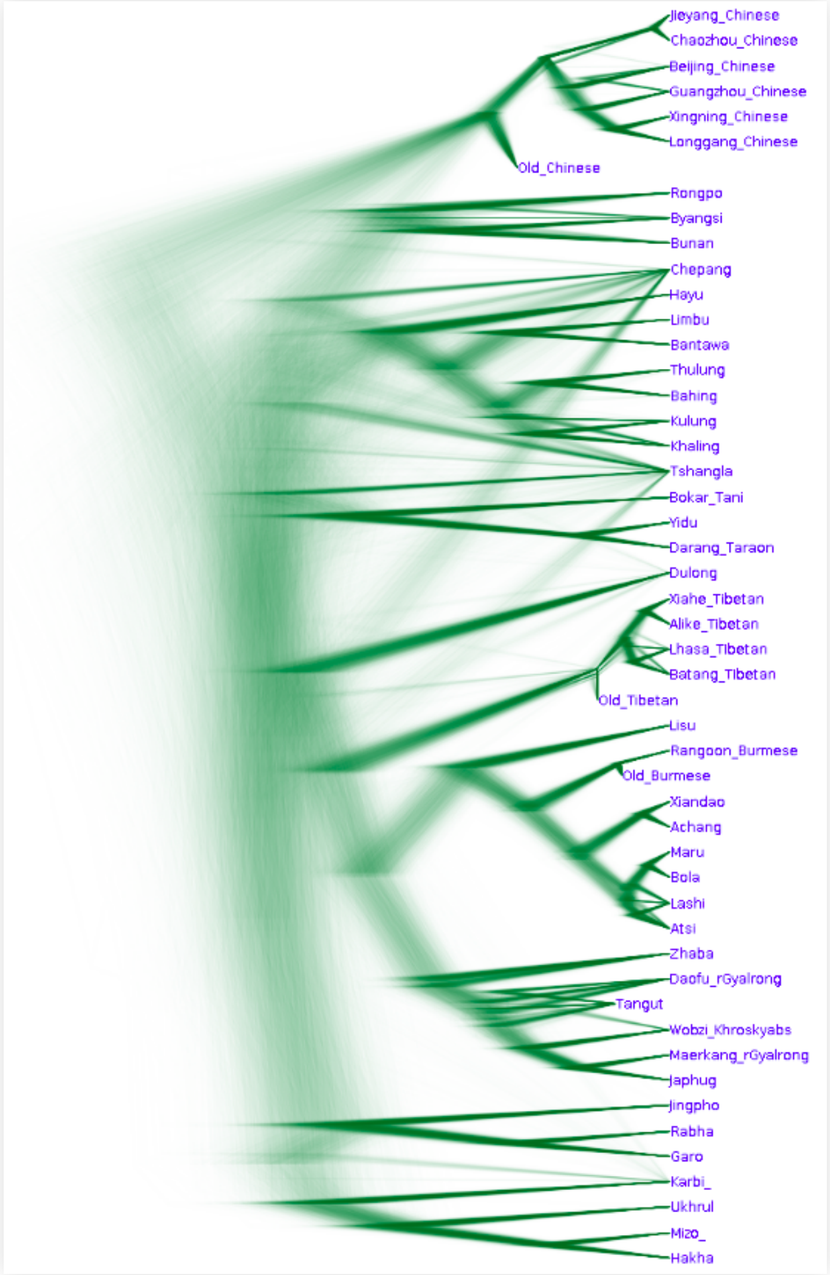
Quantitative turn

- Statistical approaches have always been common, as in Sapir (1916)
- Computational methods begin in the 1950s with lexicostatistics and glottochronology
  - Morris Swadesh
  - Joseph Greenberg
  - Sergei Starostin and the Moscow School

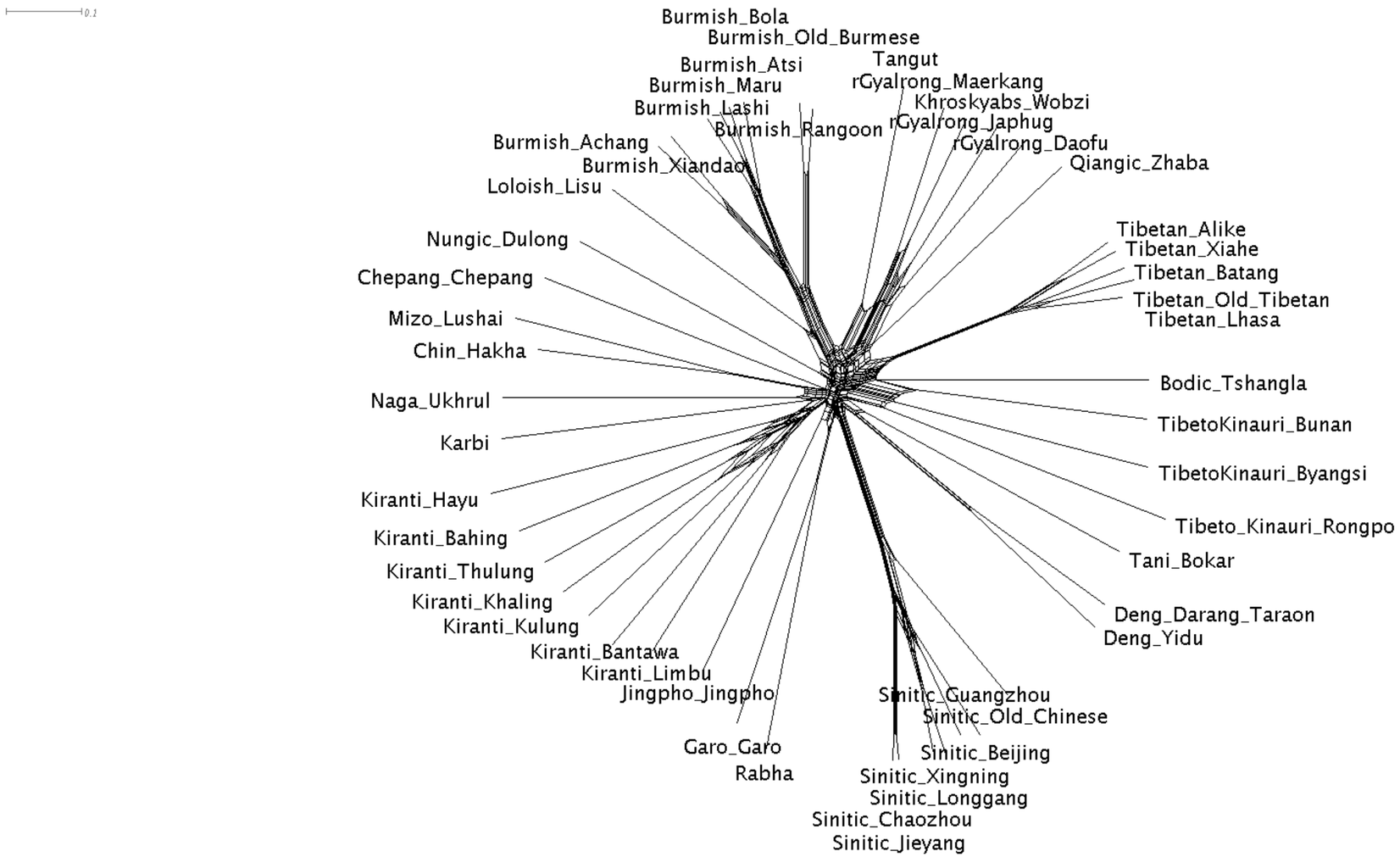
Cladistics and phylogenetics

- Computational phylogenetic approaches begin in the early 1990s with works such as Donald Ringe
- Impressive media coverage for Gray & Atkinson (2003)
  - Initial opposition by many traditional practitioners
  - Progressively more phylogenetic analyses are being published, such as Sagart et al. (2019)


Japan
Southeast Asia
North-America
Taiwan
pause 1
Philippines
Hawaii
Micronesia
Borneo
New Guinea
pause 2
Fiji
Polynesia
Madagascar
Vanuatu
New Caledonia
Australia
Rapanui
New Zealand
7000
6000
5000
4000
3000
2000
1000
0
years BP
pause 1
pulse 1
pulse 3
pause 2
4
pulse 2
Taiwan
Philippines
Island Southeast Asia
New Guinea
Oceania
Polynesia
settlement pause
settlement pause


Jieyang_Chinese
Chaozhou_Chinese
Beijing_Chinese
Guangzhou_Chinese
Xingning_Chinese
Longgang_Chinese
Old_Chinese
Rongpo
Byangsi
Bunan
Chepang
Hayu
Limbu
Bantawa
Thulung
Bahing
Kulung
Khaling
Tshangla
Bokar_Tani
Yidu
Darang_Taraon
Dulong
Xiahe_Tibetan
Alike_Tibetan
Lhasa_Tibetan
Batang_Tibetan
Old_Tibetan
Lisu
Rangoon_Burmese
Old_Burmese
Xiandao
Achang
Maru
Bola
Lashi
Atsi
Zhaba
Daofu_rGyalrong
Tangut
Wobzi_Khroskyabs
Maerkang_rGyalrong
Japhug
Jingpho
Rabha
Garo
Karbi_
Ukhrul
Mizo_
Hakha

(Sagart, 2019)



Cognate data is drawn from (Sagart, 2019)

Computer-Assisted Language Comparison

Tiago Tresoldi

Computer-Assisted Language Comparison

In the scenario of increasing digital data, open access, and interdisciplinarity, the comparative method must expand:

- Not only major families, but also minority ones
- Not only small laboratories with closed data, but a global collaboration on "fair" data
- Avoid "black-boxes", favoring results that help us understand human languages
- Not only fascination with proto-forms, but collaboration with history, biology, psychology...

Computer-Assisted Language Comparison

- Methods: alignment, cognate detection, correspondence detection
- Tools: LingPy, edictor

LingPy

Programming library for historical linguistics, state of the art:

- multiple phonetic alignment: 98% (pair score, List, 2014)
- automatic cognate detection: 89% (B-Cubed scores, List et al., 2017)
- phylogenetic reconstruction: 0.08 (Gen. Quart. Dist, Rama et al., 2018)
- correspondence pattern identification: NP-hard (no human attempts, List, 2019)

Alignment

Given cognates for 水 such as Hakha "tîi", Bunan "tɕʰu", Burmish (Rangoon) "je^{22}", Beijing "ʂuəi^{214}", Guangzhou "søy^{35}", Jieyang "tsui^{31}", Kiranti "ti", rGyalrong (Daofu) "ɣrə", how can we align?

Language	Alignments				
Hakha	–	t	–	iː	–
Bunan	–	tɕʰ	–	u	–
Burmish (Rangoon)	–	j	–	e	22
Beijing	–	ʂ	u	əi	214
Guangzhou	–	s	ø	y	35
Jieyang	–	ts	u	i	31
Kiranti	–	t	–	i	–
rGyalrong (Daofu)	ɣ	r	–	ə	–

Alignment methods

Sequence alignment algorithms from bioinformatics such as Needleman-Wunsch and Smith-Waterman, implemented in LingPy as described in List (2014).

Needleman-Wunsch

match = 1 mismatch = -1 gap = -1

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

Cognate detection

A problem of partitioning/clustering based in the correspondence of alignment sites according to implied evolutionary models.

- *Edit Distance*
  - *Linguistic extensions (Dolgopolsky, SCA)*
- Flat clustering (hierarchical or graph-based)
- *LexStat*
- Machine learning (PMI similarity, Support Vector Machines)

Edit distance - I

Comparing Jieyang "tsui^{31}" to Kiranti "ti", there are three changes over four alignment positions, thus a score of 1.0 - (3/4) = 0.75.

Edits	Rule	Alignment
0		ts
1	Delete tone	ts
2	Delete vowel	ts
3	Change initial	t

Edit distance – II

- Two words are considered cognates if their edit distance score is above a given value (threshold), which can be decided from the distribution of pair scores.
- Serious limits in a na"ive approach: Beijing "ʂuəi^{214}" and Guangzhou "søy^{35}" have a score of 0.0
  - The initial, the medial, the nucleus, the coda, and tone are different

Extensions to edit distance

- Early solutions compared not *sounds*, but *sound classes*
  - In the SCA model, Beijing "ʂuəi^{214}" is "SYE06" and Guangzhou "søy^{35}" is "SUY02".
  - Classes can be based on articulatory features or global patterns of sound change.
  - More advanced models involve additional information, such as SCA which incorporates prosodic strings.

LexStat

- LexStat is an advanced method that emulates the reasoning behind human judgement for cognacy
- The method involves multiple *permutations* that allow to compute individual segment similarities
  - The expected similarities allow a specific and instructed alignment, whose score is used for cognacy judgment.

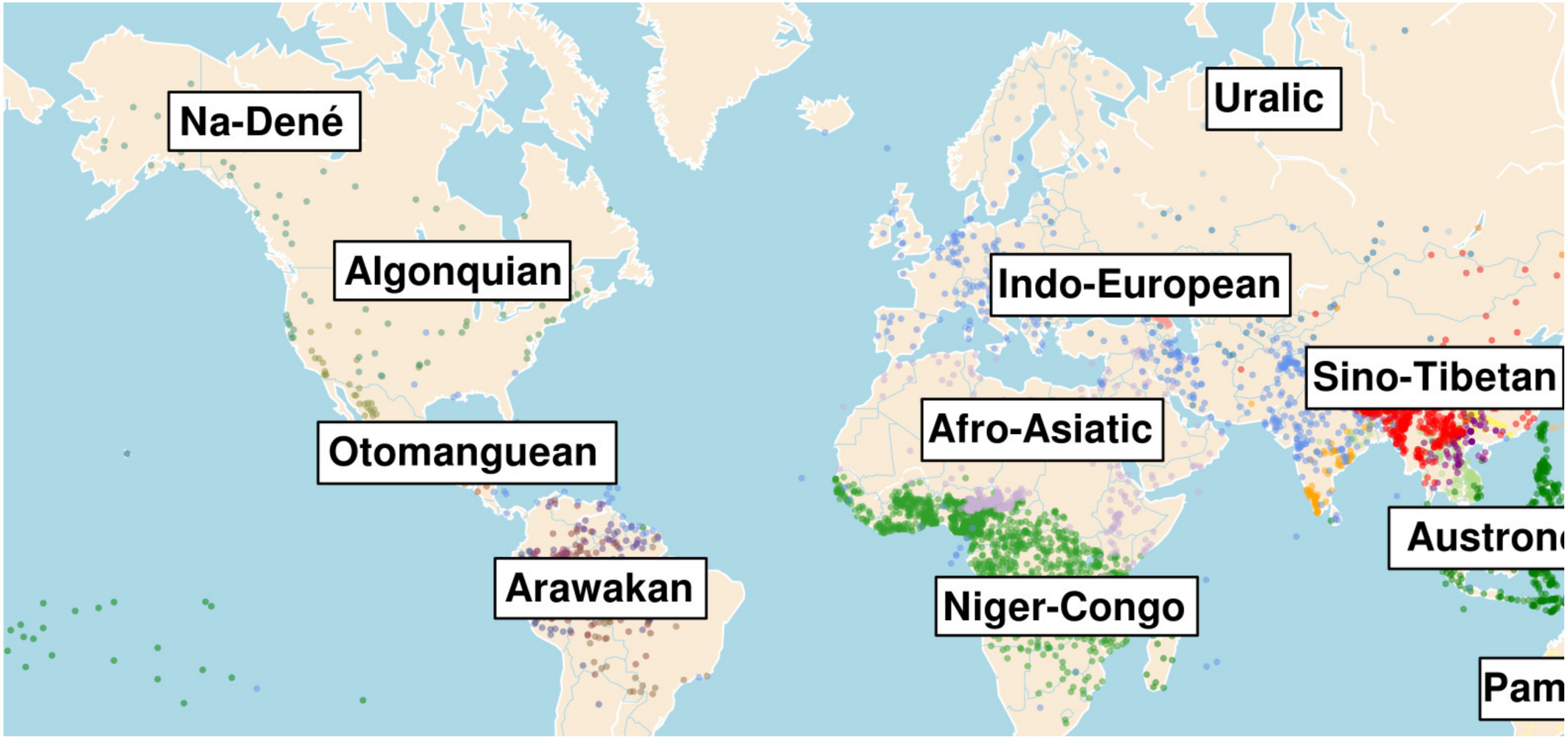
Correspondences

- New network approach for the inference of sound correspondence patterns across multiple languages.
- Columns in aligned cognate sets are the nodes, the compatibility between nodes are the edge weights
  - Compatible correspondence sets are detected by "minimum clique cover problem"

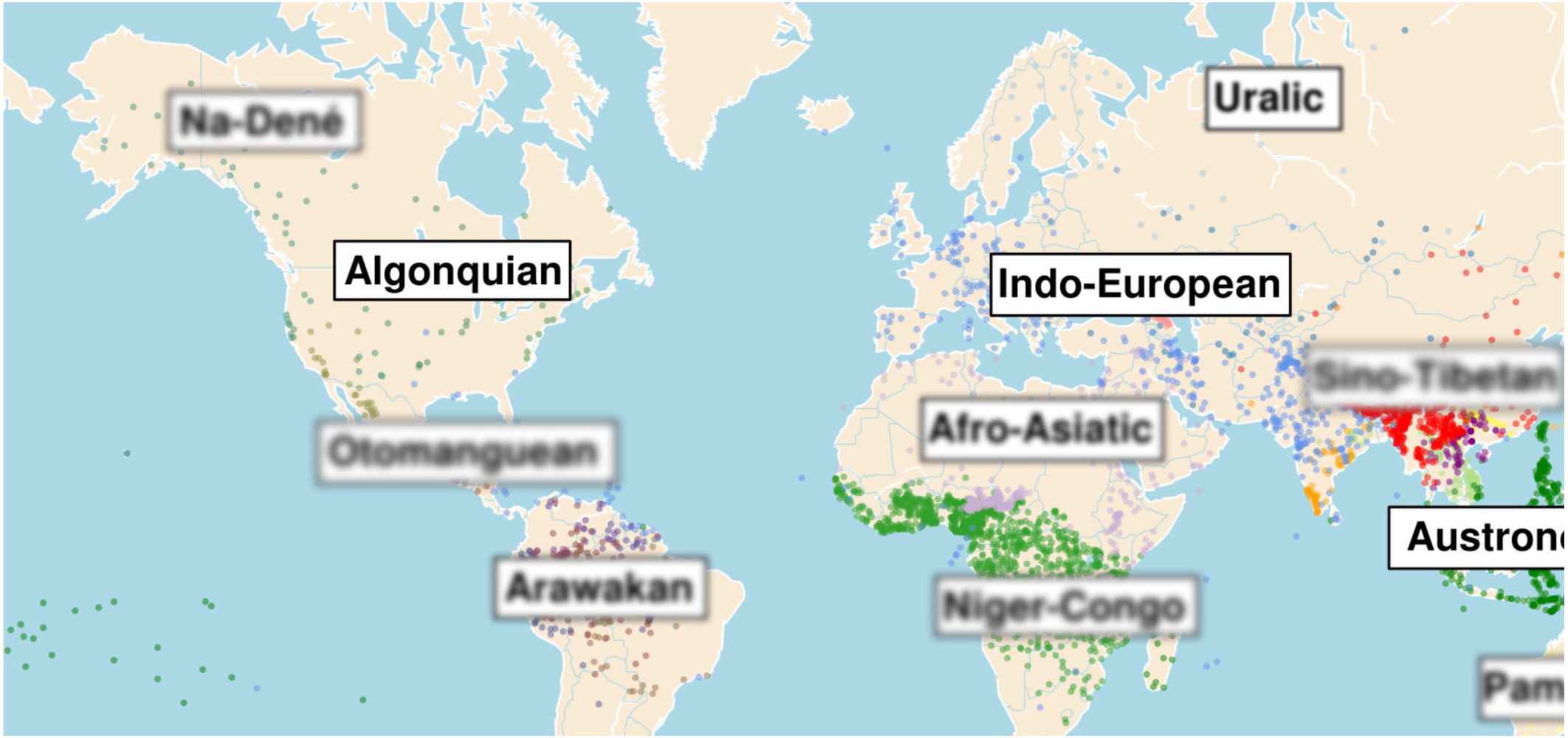
CALC workflows

Mei-Shin Wu

The Gap Between Computational and Traditional Historical Linguistics



The Gap Between Computational and Traditional Historical Linguistics



A computer-assisted approach

To allow humans and machines to work together successfully, it is important that:

- our data is both human- and machine-readable,
- we follow transparent guidelines when handling linguistic datasets,
- we offer interfaces that allow humans and machines to access the data at the same time.

CALC workflow

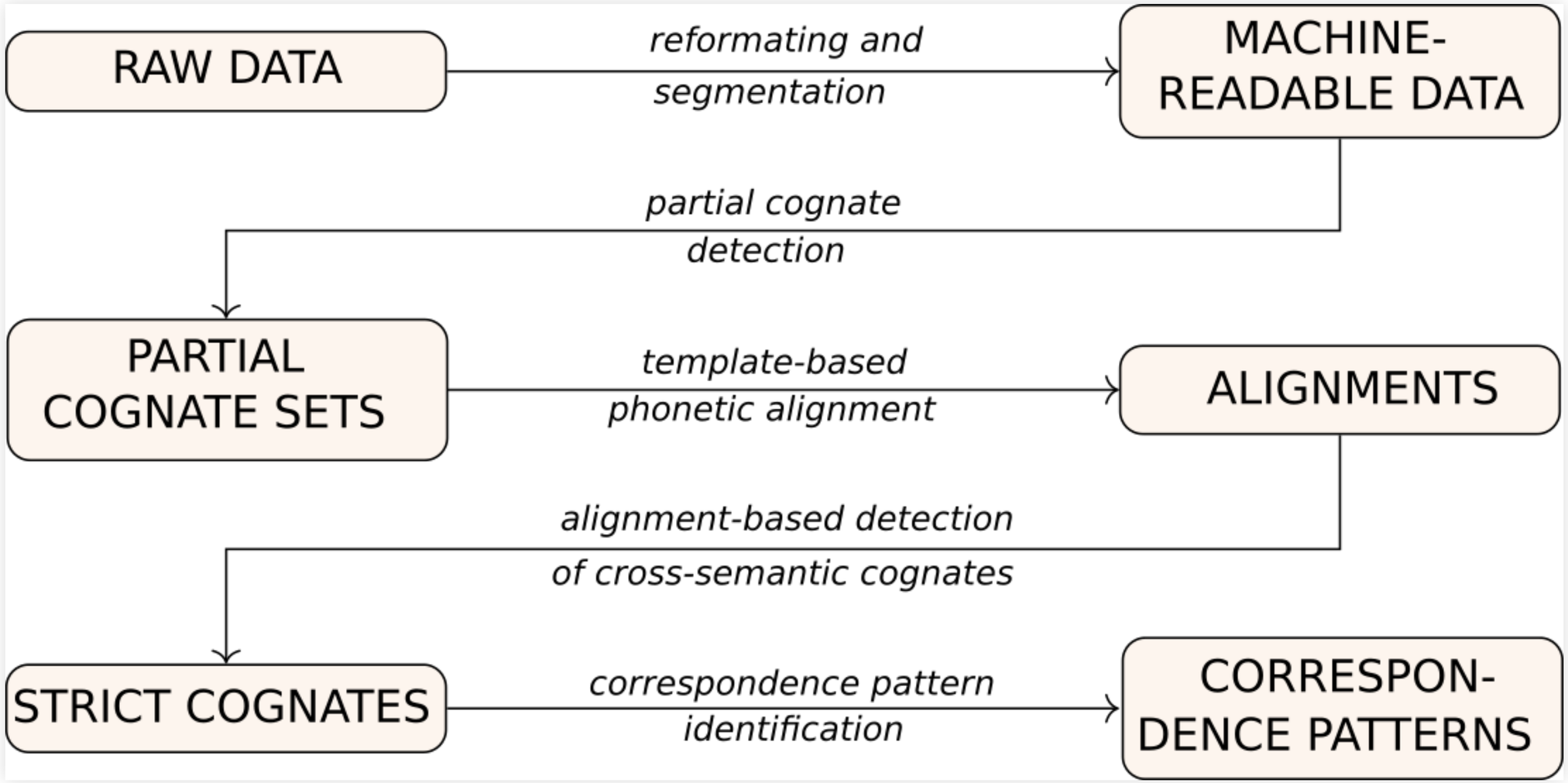
DE GRUYTER OPEN

Yearbook of the Poznań Linguistic Meeting 3 (2017), pp. 47–76
DOI: 10.1515/yplm-2017-0003

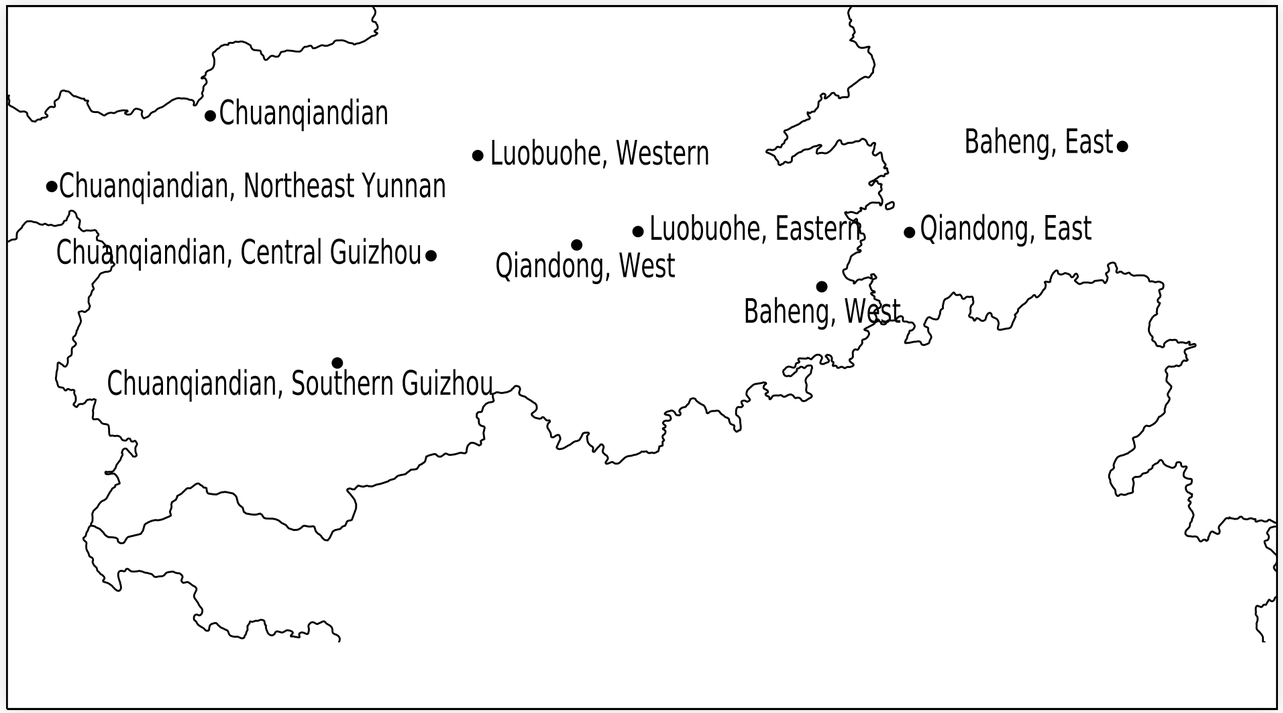
Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages

Nathan W. Hill[a] and Johann-Mattis List[b]
[a]*SOAS, London*
[b]*Max-Planck-Institute for the Science of Human History, Jena*
[b]*mattis.list@shh.mpg.de*

Details of the workflows

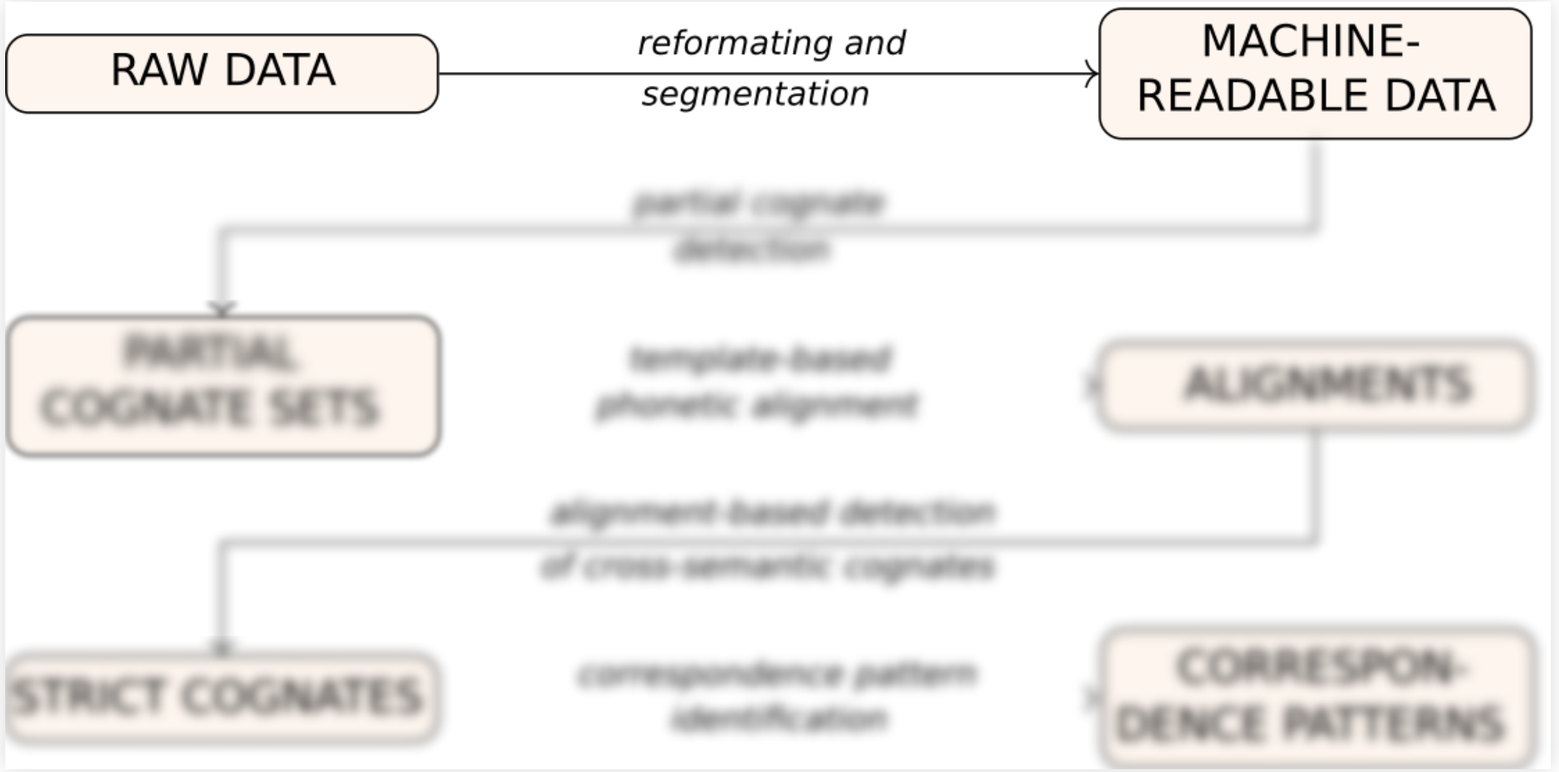


Materials and methods



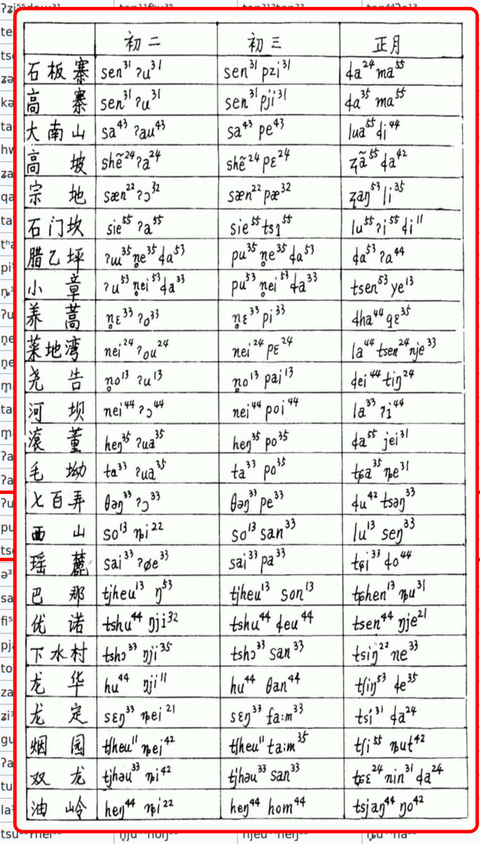
- Chén 陳其光 (2012). Miao and Yao language. 苗瑤语文
- 25 Hmong-Mien languages in the original (10 in our selection)
- 885 concepts in the original (313 in our selection, compatible with the Burmish Etymological dictionary project)

From raw data to machine-readable data



From raw data to machine-readable data

前面	pʰje²⁴tei⁵⁵zei³³	tɕi⁵⁵ði³¹	tau²⁴nte³¹	sõ¹³plɛ̃³¹	tɕaŋ¹³mplæn⁵³	ʔi⁵⁵ʰa¹¹	kɯ⁴⁴nə²²				qa²¹tue⁵⁵	ʔə⁰³moŋ³
后面	pʰje²⁴tei⁵⁵qɯen³³	tə⁰²qwen³¹	tau²⁴qaŋ⁴³	sə¹³qə⁵⁵	ɳtaŋ⁴⁴hua³²	ʔa⁵⁵qɯ⁵⁵	kɯ⁴⁴tei⁵³				taŋ²¹qæ⁴⁴	ta³¹tjeu⁴²
左边	pʰje²⁴ʔə⁰²ʒo³³	ʔə⁰²ja³¹	pʰaŋ⁴³lou¹³	pɛ³¹qə⁰²tɕi⁵¹	tɕaŋ¹³lu⁵³	saɯ⁵⁵fə¹¹	pɑ⁴⁴nɨ⁴²				tɕi²¹tɕa²¹	pu³¹qʰaŋ
右边	pʰje²⁴ʔə⁰²nen²⁴	ʔə⁰²ŋe³⁵	pʰaŋ⁴³si³³	pɛ³¹qə⁰²ntʂã²²	tɕaŋ¹³ʒaŋ²²	saɯ⁵⁵si¹¹	pɑ⁴⁴ta³¹				tɕi²¹te⁵⁵	pu³¹kwe³
上边	pə⁰⁵sa³³	χə⁵⁵sɑ³¹	ʂou⁴⁴	sʰe¹³ʒe³¹	tɕaŋ¹³pe¹¹	ʔa⁵⁵ʂa⁵⁵	kɯ⁴⁴lo⁵³				xʰæ⁴⁴vei⁵⁵	ʔə⁰³waŋ³
下边	tə⁰²qʰaŋ⁵⁵	χə⁵⁵ti³¹	tɕe³¹	kɛ¹³to²²	ha⁴⁴to¹³	pi⁵⁵dau⁵³	kɯ⁴⁴hɑŋ⁴⁴				xʰæ⁴⁴ɻæ⁵⁵	ʔə⁰³tɛ⁴²ɳ
里边	tə⁰²χwɛ⁵⁵	χə⁵⁵ɳtoŋ³¹	hou⁵⁵ɳtau¹³	ntʂõ²²ntʂõ²²	lə⁴⁴	ndlo⁵³	kɯ⁴⁴ɳɑŋ⁴²				xʰæ⁴⁴tjuŋ⁴⁴	ʔə⁰³qɔ⁴²
外边	χo³¹ŋɢwaŋ²⁴	lu³¹tɕoŋ³¹	ɳtou⁴⁴	ʂõ¹³ntʂo⁴²	tɕaŋ¹³ɳtɔ⁴⁴	ɢhau³¹	kɯ⁴⁴tei⁵³				xʰæ⁴⁴ku⁵⁵	ʔə⁰³nɔ³¹ŋ
中间	pə⁰⁵ɳtuŋ³¹	pə⁰⁵ɳtoŋ³¹	ki²⁴ɳtaŋ⁴³	kã³¹ntʂã²⁴	ɳtaŋ³²	pi⁵⁵ɳtau⁵⁵	tɑ³⁵ɳtoŋ³⁵				qa³³tjuŋ⁴⁴	qo⁰³seŋ⁵⁵
旁边	ʔi⁵⁵pʰje²⁴	tə⁰²ɳtɑŋ⁵⁵	ntɔ¹³	qə⁰²mpõ⁴²	tɕaŋ¹³ntoŋ¹³	ʔa⁵⁵ntu⁵³	qo³⁵cɑ⁴⁴				qa³³xæ³³	ʔə⁰³jaŋ⁴⁴
今天	na³¹naŋ⁵⁵	ŋə³¹ʔnoŋ⁵⁵	ŋo⁴³na⁵⁵	ŋã²⁴nã¹³	noŋ²³²na⁴²	mu³³na³³	tʰa⁵³nen⁴⁴				tuẽ³³na¹³	ɱ³¹ŋei³⁵
明天	pʰəu³¹na³¹	pʰe³¹ne³¹je³⁵	tɕa⁴⁴ki²¹	pe³¹ntʂʰe⁴²	ha⁴⁴ɳte¹³	pi³¹dʑi³¹	ɕi³⁵ŋe³⁵				tæ³³seɯ¹³	po³⁵ŋei³⁵
昨天	mo⁵⁵na³¹	tɑ³¹ŋin⁵⁵ŋa³¹	ʔa⁴³naŋ¹³	qə⁰²la³¹ŋã²⁴	ɳtɔ¹¹noŋ²³²	ʔa³¹naɯ³¹	nɨ³¹ŋe³⁵				nei⁴⁴naŋ²¹	qə⁰³ŋei³⁵
后天	si⁵⁵nɑ³¹	si⁵⁵ŋɑ³¹	naŋ¹³ki²¹	ŋã²⁴tʂo⁴²	noŋ²³²ɳtɔ⁴⁴	ʔa³¹naɯ³¹dʑɦi¹¹	ʔu³⁵pɛ⁴⁴ŋe³⁵				ɕi⁴⁴ʒa²¹	tɕi⁵⁵ŋei³⁵
上午		ti⁵⁵je³¹	taŋ⁵⁵nau³¹tʂʰai³³		tæn⁴²ɳte¹³	sey⁵⁵ntso⁵⁵	te³⁵te³⁵ntu³⁵				kua²²tæ⁴⁴	qə⁰³tei⁵⁵
中午	tə⁰²si³¹ɳa²⁴ji³³	ɳtoŋ³¹ŋɑ³¹	taŋ⁵⁵nau³¹ʂo⁴⁴	ntʂã²⁴ŋã²⁴	pje¹³soŋ³⁵	nɦau³⁵ʂu³³	ŋe³⁵ntu³⁵				tjuŋ⁴⁴nei⁴⁴	ma⁵⁵ɳoŋ
下午	tə⁰²si³¹ɳa²⁴pjo³³	ɳɑ³¹sɑŋ³⁵	taŋ⁵⁵nau³¹ɱau⁴⁴	sʰi²⁴lã¹³sʰõ⁴²	he⁵³mɔ³⁵	nɦau³⁵gu³³tʰau³³	ŋe³⁵ɕi³⁵				tju³³sa⁵³	qə⁰³ma¹¹
早晨	tei⁵⁵ji³³	tei⁵⁵je³¹	ʂeu⁵⁵ntso⁵⁵	ntsʰe⁴²	tæn⁴²mpɯ¹¹	sey⁵⁵ntso⁵⁵bfɯ¹¹	qa⁴⁴ntso⁴⁴				kua²¹seɯ¹³	qə⁰³tei⁵⁵
晚上	tə⁰²muŋ²⁴	tə⁰²ɱoŋ³⁵	ɱau⁴⁴nto³¹	ɱõ⁴²ŋqã⁵⁵	tæn⁴²mɔ³⁵	ŋo³³ndɦu³⁵	tɕi³⁵ɱɑŋ⁵³				ʑi³³maŋ³³	qə⁰³ma¹¹
半夜	ʔi⁵⁵ndun²⁴muŋ²⁴	ɳtoŋ³¹ɱoŋ³⁵	taŋ¹³ɱau⁴⁴	ʔi²⁴tõ³¹ɱõ⁴²	toŋ³³mɔ³⁵	dau³¹ɱo³³	tu⁵³mu²²				taŋ²¹maŋ³³	ma⁵⁵ɳoŋ
初一	sen³¹ʔi²⁴	sen³¹ʔi³⁵	sa⁴³ʔi⁴³	sʰẽ²⁴ʔi²⁴	sæn²²ʔei³²	sie⁵⁵ʔi⁵⁵	ʔɑ⁴⁴ŋe³⁵ɬɑ⁵³				nei⁴⁴ʔi⁴⁴	heŋ³⁵jei³
初二	sen³¹ʔu³¹	sen³¹ʔu³¹	sa⁴³ʔau⁴³	sʰẽ²⁴ʔa²⁴	sæn²²ʔɔ³²	sie⁵⁵ʔa⁵⁵	ʔɯ³⁵ŋe³⁵ɬɑ⁵³				nei⁴⁴ʔɔ⁴⁴	heŋ³⁵ʔua
初三	sen³¹pzi³¹	sen³¹pji³¹	sa⁴³pe⁴³	sʰẽ²⁴pɛ²⁴	sæn²²pæ³²	sie⁵⁵tsɪ⁵⁵	pu³⁵ŋe³⁵ɬɑ⁵³				nei⁴⁴poi⁴⁴	heŋ³⁵po
正月	ɬa²⁴ma⁵⁵	ɬɑ³⁵mɑ⁵⁵	lua⁵⁵ɬi⁴⁴	ʐã⁵⁵ɬa⁴²	ʐaŋ⁵³li³⁵	lu⁵⁵ʔi⁵⁵ɬi¹¹	ɬɑ⁵³ʔɑ⁴⁴				la³³ʔi⁴⁴	ɬa⁵⁵jei³¹
二月	ɬa²⁴ʐuŋ³¹	ɬɑ³⁵ʐoŋ³¹	ʐaŋ³¹ɬi⁴⁴	nã²⁴ɬa⁴²	na³²li³⁵	lu⁵⁵ʔa⁵⁵ɬi¹¹	ɬɑ⁵³ʔɯ³⁵				la³³ʔɔ⁴⁴	ɬa⁵⁵ʔua³⁵
三月	ɬa²⁴tɕa³¹	ɬɑ³⁵tɕɑ³¹	pe⁴³ɬi⁴⁴	mĩ³¹ɬa⁴²	men¹¹li³⁵	lu⁵⁵tsɪ⁵⁵ɬi¹¹	ɬɑ⁵³pu³⁵				la³³poi⁴⁴	ɬa⁵⁵po³⁵
四月	ɬa²⁴qɛ³¹	ɬɑ³⁵qe³¹	plou⁴³ɬi⁴⁴	zã⁵⁵ɬa⁴²	ʐaŋ⁵³li³⁵	lu⁵⁵tɬau⁵⁵ɬi¹¹	ɬɑ⁵³pʐei³⁵				la³³ɬɔ⁴⁴	ɬa⁵⁵tji³⁵
五月	ɬa²⁴qlei⁵⁵	ɬɑ³⁵qlei⁵⁵	tʂi⁴³ɬi⁴⁴	lẽ²⁴ɬa⁴²	læn³²li³⁵	lu⁵⁵pɯ⁵⁵ɬi¹¹	ɬɑ⁵³pʐɑ³⁵				la³³pja⁴⁴	ɬa⁵⁵tja³⁵
六月	ɬa²⁴mpzi²⁴	ɬɑ³⁵mpji³⁵	ʈou⁴⁴ɬi⁴⁴	qe²⁴ɬa⁴²	he³²li³⁵	lu⁵⁵tɬau¹¹ɬi³³	ɬɑ⁵³ʈɔ⁵³				la³³tju³³	ɬa⁵⁵tcu⁵⁵
七月	ɬa²⁴nen⁵⁵	ɬɑ³⁵nen⁵⁵	ɕaŋ⁴⁴ɬi⁴⁴	tɬɛ¹³ɬa⁴²	ɬæ⁴²li³⁵	lu⁵⁵ɕau¹¹ɬi³³	ɬɑ⁵³tɕoŋ⁴²				la³³cuŋ²²	ɬa⁵⁵tɕaŋ⁴
八月	ɬa²⁴ʐu⁵⁵	ɬɑ³⁵ʐu⁵⁵	ʑi²⁴ɬi⁴⁴	mpa⁴²ɬa⁴²	mpa⁵⁵li³⁵	lu⁵⁵ʑɦi³¹ɬi³³	ɬɑ⁵³ʑi²²				la³³za⁵³	ɬa⁵⁵ji⁴²
九月	ɬa²⁴tsəu⁵⁵	ɬɑ³⁵tsu⁵⁵	tɕua³³ɬi⁴⁴	nã³¹ɬa⁴²	nə¹¹li³⁵	lu⁵⁵dʑɦa³⁵ɬi¹¹	ɬɑ⁵³tɕo³¹				la³³tɕa⁵⁵	ɬa⁵⁵ko³³
十月	ɬa²⁴ʔla⁵⁵	ɬɑ³⁵ʔlɑ⁵⁵	kou²⁴ɬi⁴⁴	ŋã⁵⁵ɬa⁴²	ŋu⁵³li³⁵	lu⁵⁵gɦau³¹ɬi³³	ɬɑ⁵³ku²²				la³³tɕu⁵³	ɬa⁵⁵ku⁴²
十一月	ɬa²⁴juŋ³¹	ɬɑ³⁵joŋ³¹	ɳo³¹ɬi⁴⁴	sə¹³ɬa⁴²	so⁴²li³⁵	lu⁵⁵gɦau³¹ʔi⁵⁵ɬi³³	ɬɑ⁵³toŋ³⁵				la³³ce³³ʐu²¹	ɬa⁵⁵ku⁴²je
十二月	ɬa²⁴ʔnen³¹	ɬɑ³⁵ʔnen³¹	tʂo⁵⁵ɬi⁴⁴	lu¹³⁴ɬa⁴²	la⁴²li³⁵	lu⁵⁵gɦau³¹ʔa⁵⁵ɬi³³	ɬɑ⁵³zo²²				la³³mja⁵³	ɬa⁵⁵ku⁴²ʔ
今年	nen³¹naŋ⁵⁵	nen ʔnoŋ⁵⁵	ɕoŋ⁴⁴na⁵⁵	sʰõ⁴²nã¹³	tɕaŋ⁵⁵na⁴²	ɕau³³na⁵⁵	tɕu⁵³nen⁴⁴	tsu⁵⁵ɳei⁵⁵	ŋju⁵⁵ɳoŋ⁵⁵	njeu⁵⁵ɳei⁵⁵	ŋu⁵⁵ɳa⁵⁵	ne⁵⁵na¹³	ɱ³¹tɕaŋ⁵
去年	muŋ⁵⁵nen³¹	nen³¹mu³¹ʔnoŋ⁵⁵	ʔa⁴³tʂe⁴³na⁵⁵	qə⁰²la³¹sʰõ⁴²	tɕaŋ⁵⁵pzẽ³²	ʔa³³pu⁵⁵na³³	tɕu⁵³pa³⁵	ŋ³⁵pa⁵³tsu³³	ŋju⁴⁴ʔɛ³³	njeu⁴⁴so¹³	ŋu⁴⁴ʔo¹³	ne⁵⁵kua³³	qa³³tɕaŋ
明年	pʰə³¹nen³¹	pə⁰⁵nen³¹	ə²⁴ɕoŋ⁴⁴	ta⁵⁵sʰõ⁴²nã¹³	tɕaŋ⁵⁵sæn²²	ɳɦau¹¹na⁵⁵	tɯ³¹tɕu⁵³	səɯ⁵³ku⁵⁵tsu³³	ŋju⁴⁴qaŋ³³	po¹³njeu⁴⁴	pe²²ŋu⁴⁴	ne⁵⁵qæ⁴⁴	pə³³tɕaŋ
现在	tʰaŋ³³naŋ⁵⁵	ɕe³¹ʔnoŋ⁵⁵	taŋ⁵⁵na⁵⁵	sʰa³¹nã¹³	sei²²na⁴²	na⁵⁵ɳi⁵⁵	mɑ³¹nen⁴⁴	ʑa¹³ʔnei⁵⁵	ɕaŋ³¹noŋ³⁵	ha⁵³neŋ³³	sei³¹na⁵³	poi⁵³na¹³	ɱ³¹ke⁵⁵



	初二	初三	正月
石板寨	$sen^{31}\ ʔu^{31}$	$sen^{31}\ pzi^{31}$	$ɬa^{24}ma^{55}$
高寨	$sen^{31}\ ʔu^{31}$	$sen^{31}\ pji^{31}$	$ɬa^{35}\ ma^{55}$
大南山	$sã^{43}\ ʔau^{43}$	$sã^{43}\ pẽ^{43}$	$lua^{55}\ ɬi^{44}$
高坡	$shẽ^{24}ʔa^{24}$	$shẽ^{24}\ pɛ^{24}$	$ʐã^{55}\ ɬa^{42}$
宗地	$sæn^{22}\ ʔɔ^{32}$	$sæn^{22}\ pæ^{32}$	$ʐaŋ^{53}\ li^{35}$
石门坎	$sie^{55}\ ʔa^{55}$	$sie^{55}\ tsɪ^{55}$	$lu^{55}\ ʔi^{55}\ ɬi^{11}$
腊乙坪	$ʔɯ^{35}\ ŋe^{35}\ ɬa^{53}$	$pu^{35}\ ŋe^{35}\ ɬa^{53}$	$ɬa^{53}\ ʔa^{44}$
小章	$ʔu^{53}\ ŋei^{53}\ ɬa^{33}$	$pu^{53}\ ŋei^{53}\ ɬa^{33}$	$tsen^{53}\ ye^{13}$
养蒿	$ŋɛ^{33}\ ʔo^{33}$	$ŋɛ^{33}\ pi^{33}$	$ɬha^{44}\ qɛ^{35}$
莱地湾	$nei^{24}\ ʔou^{24}$	$nei^{24}\ pɛ^{24}$	$la^{44}\ tsen^{24}\ ŋje^{33}$
尧告	$ŋo^{13}\ ʔu^{13}$	$ŋo^{13}\ pai^{13}$	$ɬei^{44}\ tiŋ^{24}$
河坝	$nei^{44}\ ɔ^{44}$	$nei^{44}\ poi^{44}$	$la^{33}\ ʔi^{44}$
凌董	$heŋ^{35}\ ʔua^{35}$	$heŋ^{35}\ po^{35}$	$ɬa^{55}\ jei^{31}$
毛坳	$ta^{33}\ ʔuə^{35}$	$ta^{33}\ po^{35}$	$tɕa^{35}\ ŋe^{31}$
又百弄	$ɓəŋ^{33}\ ɔ^{33}$	$ɓəŋ^{33}\ pe^{33}$	$ɬu^{42}\ tsəŋ^{33}$
西山	$so^{13}\ ŋi^{22}$	$so^{13}\ san^{33}$	$lu^{13}\ seŋ^{33}$
瑶麓	$sai^{33}\ ʔɑe^{33}$	$sai^{33}\ pa^{33}$	$tɕi^{33}\ ɬo^{44}$
巴那	$tɧeu^{13}\ ŋ^{53}$	$tɧeu^{13}\ son^{13}$	$tɕhen^{13}\ ɳu^{31}$
优诺	$tshu^{44}\ ŋji^{32}$	$tshu^{44}\ ɬeu^{44}$	$tsen^{44}\ ŋje^{21}$
下水村	$tshɔ^{33}\ ŋji^{35}$	$tshɔ^{33}\ san^{33}$	$tsiŋ^{22}\ ne^{33}$
尤华	$hu^{44}\ ŋji^{11}$	$hu^{44}\ ɓan^{44}$	$tɧiŋ^{53}\ ɬe^{35}$
尤定	$seŋ^{33}\ ɳei^{21}$	$seŋ^{33}\ ta:m^{33}$	$tsi^{31}\ ɬa^{24}$
烟囤	$tɧeu^{11}\ ɳej^{42}$	$tɧeu^{11}\ ta:m^{35}$	$tɧi^{55}\ ɳut^{42}$
双龙	$tɧəu^{33}\ ɳi^{42}$	$tɧəu^{33}\ san^{33}$	$tɕɛ^{24}\ nin^{31}\ ɬa^{24}$
油岭	$heŋ^{44}\ ɳi^{22}$	$heŋ^{44}\ hom^{44}$	$tsjaŋ^{44}\ ŋo^{42}$

From raw data to machine-readable data

	A	B	C	D	E
1		Baheng,e	Baheng, \	Qiandonɡ	Qiandonɡ
2	七	tsha^{31},tsjɯ	tshang^{44}	shung^{53}	shung^{22}
3	月亮	la^{03}lha^{55}	ʔa^{03}lha^{55}	la^{44}la^{44}	pau^{11}la^{33}
4	星星	la^{03}qang^{3}	qa^{03}qang$^{}$	qei^{24}qei^{24}	tei^{44}qei^{44}

	A	B	C	D	E	F	G	H
1	ID	DOCU	CONC	ENGL	VALU	FORM	TOKE	NOTE
2	1	Bahei	七	SEVE	tsja^{31},	tsja^{31}		
3	2	Bahei	七	SEVE	tsja^{31},	tsjunɡ		variar
4	2	Bahei	七	SEVE	tsjanɡ	tsjanɡ		
5	3	Qianc	七	SEVE	sjungˡ	sjungˡ		
6	4	Qianc	七	SEVE	sjungˀ	sjungˀ		
7	5	Bahei	月亮	MOOI	la^{03}lha	la^{03}lha		
8	6	Bahei	月亮	MOOI	ʔa^{03}lh	ʔa^{03}lh		
9	7	Qianc	月亮	MOOI	la^{44}laˀ	la^{44}laˀ		
10	8	Qianc	月亮	MOOI	pau^{11}l	pau^{11}l		
11	9	Bahei	星星	STAR	la^{03}qa	la^{03}qa		
12	10	Bahei	星星	STAR	qa^{03}q	qa^{03}q		
13	11	Qianc	星星	STAR	qei^{24}q	qei^{24}q		
14	12	Qianc	星星	STAR	tei^{44}q	tei^{44}q		

From raw data to machine-readable data

We recommend *Orthography Profiles* as a way to:

- Convert arbitrary input data to IPA:
  tsj ----> tɕ
  ng ----> ŋ
- And to segment the input data:
  tsja^{31} ----> tɕa^{31} ----> tɕ a ^{31}

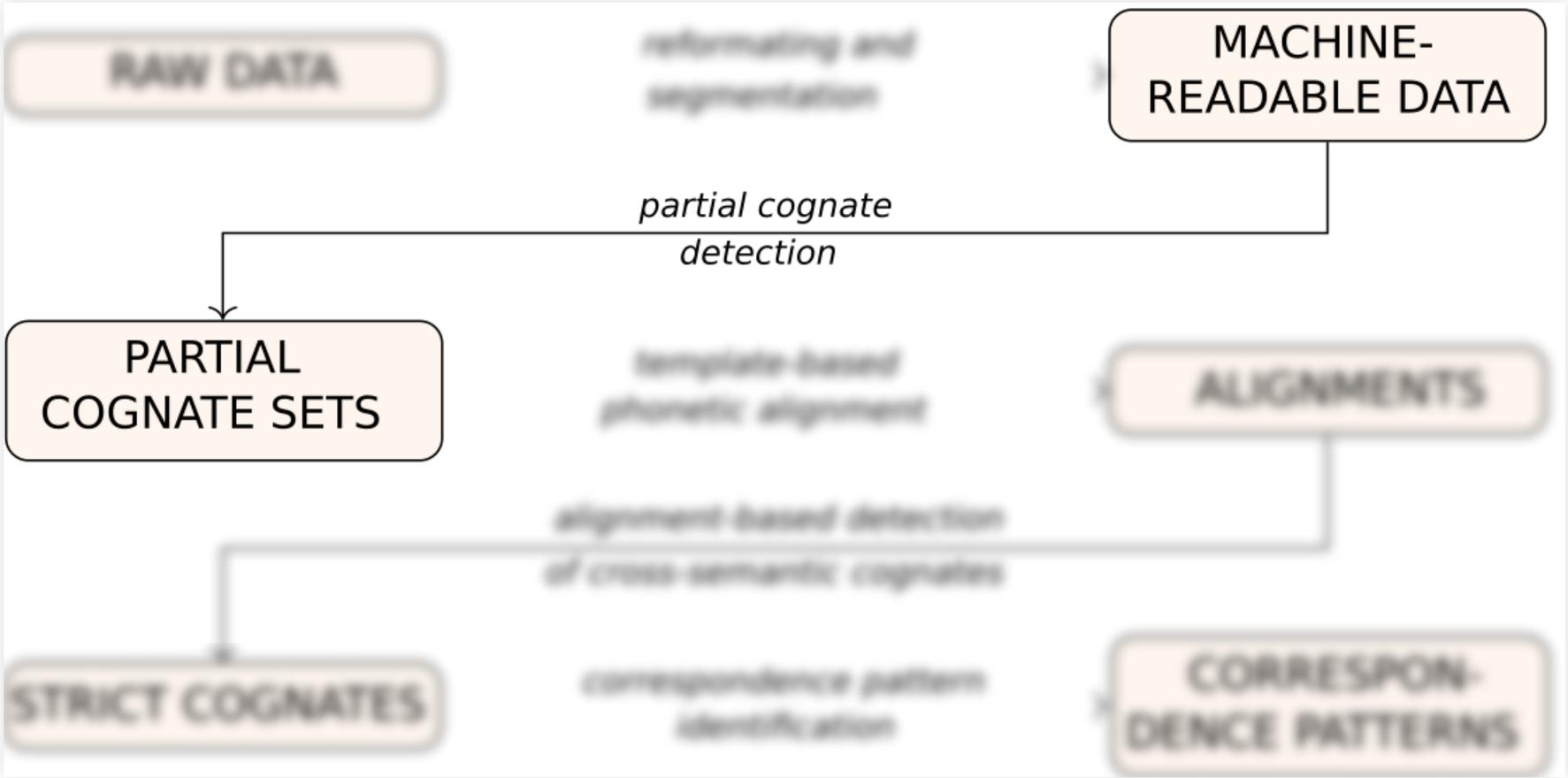
From raw data to machine-readable data

	A	B
1	Graphe	IPA
2	č	tʃ
3	ž	dʒ
4	th	tʰ
5	dh	d̪
6	sh	ʃ
7	a	a
8	aa	aː
9	tsj	tɕ
10	la	l a

From raw data to machine-readable data

	A	B	C	D	E	F	G	H
1	ID	DOCULECT	CONCEPT	ENGLISH	VALUE	FORM	TOKENS	COGIDS
2	1	Baheng, east	七	SEVEN	tsja^{31},tsjung^{44}	tsja^{31}	tɕ a ^{31}	
3	2	Baheng, east	七	SEVEN	tsja^{31},tsjung^{44}	tsjung^{44}	tɕ u ŋ ^{44}	
4	3	Baheng, west	七	SEVEN	tsjang^{44}	tsjang^{44}	tɕ a ŋ ^{44}	
5	4	Qiandong, east	七	SEVEN	sjung^{53}	sjung^{53}	ɕ u ŋ ^{53}	
6	5	Qiandong, wesst	七	SEVEN	sjung^{22}	sjung^{22}	ɕ u ŋ ^{22}	
7	6	Baheng, east	月亮	MOON	la^{03}lha^{55}	la^{03}lha^{55}	l a $^{3/0}$ + ɬ a ^{55}	
8	7	Baheng, west	月亮	MOON	ʔa^{03}lha^{55}	ʔa^{03}lha^{55}	ʔ a $^{3/0}$ + ɬ a ^{55}	
9	8	Qiandong, east	月亮	MOON	la^{44}la^{44}	la^{44}la^{44}	l a ^{44} + l a ^{44}	
10	9	Qiandong, wesst	月亮	MOON	pau^{11}la^{33}	pau^{11}la^{33}	p ɔ ^{11} + l a ^{33}	
11	10	Baheng, east	星星	STAR	la^{03}qang^{35}	la^{03}qang^{35}	l a $^{3/0}$ + q a ŋ ^{35}	
12	11	Baheng, west	星星	STAR	qa^{03}qang^{35}	qa^{03}qang^{35}	q a $^{3/0}$ + q a ŋ ^{35}	
13	12	Qiandong, east	星星	STAR	qei^{24}qei^{24}	qei^{24}qei^{24}	q ei ^{24} + q ei ^{24}	
14	13	Qiandong, wesst	星星	STAR	tei^{44}qei^{44}	tei^{44}qei^{44}	t ei - ^{44} + q ei ^{44}	

From segmented words to computer-inferred cognates

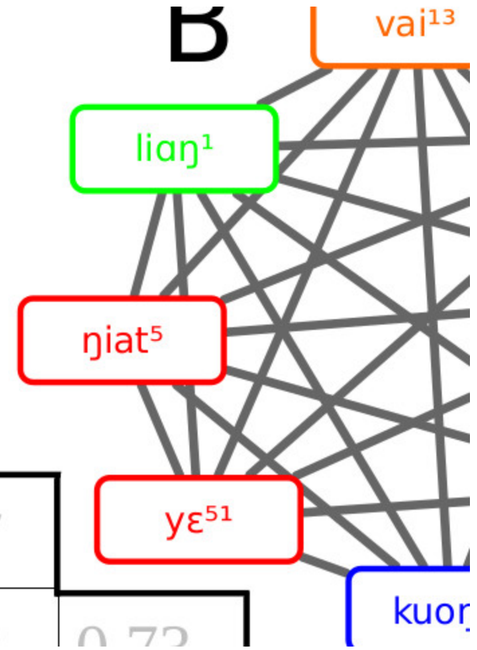


From segmented words to computer-inferred cognates

Variety	Form	Character	Cognacy
Fúzhōu	ŋuoʔ^{5}	月	1
Měixiàn	ŋiat^{5} kuoŋ^{44}	月光	1 2
Wēnzhōu	ŋ̊y^{21} kuɔ^{35} vai^{13}	月光佛	1 2 3
Běijīng	yɛ^{51} liɑŋ^{1}	月亮	1 4

From segmented words to computer-inferred cognates

		A				
Fúzhōu	ŋuoʔ^{5}					
Měixiàn	ŋiat^{5}	0.44				
	kuoŋ^{44}	0.78	0.78			
Wēnzhōu	n̥y^{21}	0.30	0.35	0.67		
	kuɔ^{35}	0.80	0.85	0.27	0.67	
	vai^{13}	0.85	0.85	0.82	0.72	0.72

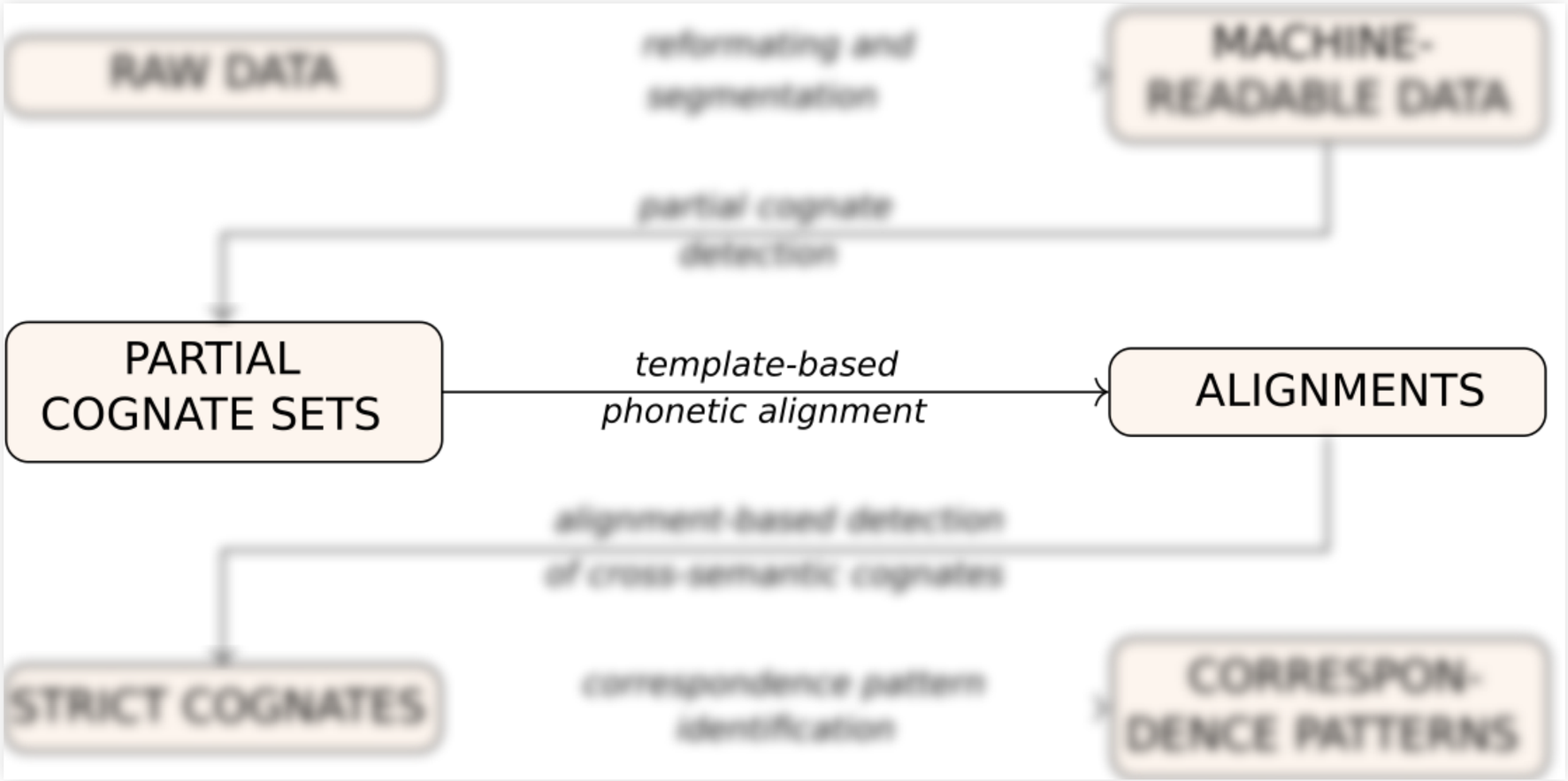


List et al. (2016). Using sequence similarity networks to identify partial cognates in multilingual wordlists. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 2, pp. 599-605).

From segmented words to computer-inferred cognates

	A	B	C	D	E	F	G	H
1	ID	DOCULECT	CONCEPT	ENGLISH	VALUE	FORM	TOKENS	COGIDS
2	1	Baheng, east	七	SEVEN	tsja^{31},tsjung^{44}	tsja^{31}	tɕ a ^{31}	3
3	2	Baheng, east	七	SEVEN	tsja^{31},tsjung^{44}	tsjung^{44}	tɕ u ŋ ^{44}	3
4	3	Baheng, west	七	SEVEN	tsjang^{44}	tsjang^{44}	tɕ a ŋ ^{44}	3
5	4	Qiandong, east	七	SEVEN	sjung^{53}	sjung^{53}	ɕ u ŋ ^{53}	3
6	5	Qiandong, wesst	七	SEVEN	sjung^{22}	sjung^{22}	ɕ u ŋ ^{22}	3
7	6	Baheng, east	月亮	MOON	la^{03}lha^{55}	la^{03}lha^{55}	l a $^{3/0}$ + ɬ a ^{55}	1908 1907
8	7	Baheng, west	月亮	MOON	ʔa^{03}lha^{55}	ʔa^{03}lha^{55}	ʔ a $^{3/0}$ + ɬ a ^{55}	1909 1907
9	8	Qiandong, east	月亮	MOON	la^{44}la^{44}	la^{44}la^{44}	l a ^{44} + l a ^{44}	1908 1907
10	9	Qiandong, wesst	月亮	MOON	pau^{11}la^{33}	pau^{11}la^{33}	p ɔ ^{11} + l a ^{33}	1910 1907
11	10	Baheng, east	星星	STAR	la^{03}qang^{35}	la^{03}qang^{35}	l a $^{3/0}$ + q a ŋ ^{35}	1874 1870
12	11	Baheng, west	星星	STAR	qa^{03}qang^{35}	qa^{03}qang^{35}	q a $^{3/0}$ + q a ŋ ^{35}	1872 1870
13	12	Qiandong, east	星星	STAR	qei^{24}qei^{24}	qei^{24}qei^{24}	q ei ^{24} + q ei ^{24}	1872 1870
14	13	Qiandong, wesst	星星	STAR	tei^{44}qei^{44}	tei^{44}qei^{44}	t ei - ^{44} + q ei ^{44}	1871 1870

From cognates to alignments



From cognates to alignments

bahengeast «ㄊ» (COGIDS: 436)

DOCULECTS	CONCEPTS	ID: 436			
luobuoheeastern	ㄊ	z	u	ŋ	24
luobuohewestern	ㄊ	z	o	ŋ	35
chuanqiandian	ㄊ	ɕ	a	ŋ	44
chuanqiandiancentralg…	ㄊ	s	ã	-	22
chuanqiandiansouthern…	ㄊ	tɕ	a	ŋ	13
chuanqiandiannortheas…	ㄊ	ɕ	aɯ	-	33
qiandongeast	ㄊ	ɕ	u	ŋ	53
qiandongwest	ㄊ	ɕ	u	ŋ	22
bahengwest	ㄊ	tɕ	a	ŋ	44
bahengeast	ㄊ	tɕ	a	-	31

Phonetic alignment techniques are well-known in historical linguistics and have been applied for quite some time now.

From cognates to alignments

We propose *Template-Based Alignments* as an alternative to semi-automatically computed alignments.

- Languages with a rather restricted syllable structure can usually be aligned in a very consistent way by simply using a template.
- A typical Chinese syllable, for example, consists of *initial*, *medial*, *nucleus*, *coda* and *tone* (Wang 1996). Once we know the individual template of a Chinese word, we can easily align it with any other word, as long as we know the template.

From cognates to alignments

Doculect	Concept	Tokens	Structure		i	m	n	c	t		Alignment			
East Baheng	seven	tɕ a ^{31}	i n t		tɕ	-	a	-	^{31}		tɕ	a	-	^{31}
West Baheng	seven	tɕ a ŋ ^{44}	i n c t	→	tɕ	-	a	ŋ	^{44}	→	tɕ	a	ŋ	^{44}
Chuanqiandian	seven	ɕ a ŋ ^{44}	i n c t		ɕ	-	a	ŋ	^{44}		ɕ	a	ŋ	^{44}
Chuanqiandian (CG)	seven	s ã ^{22}	i n t		s	-	ã	-	^{22}		s	ã	-	^{22}

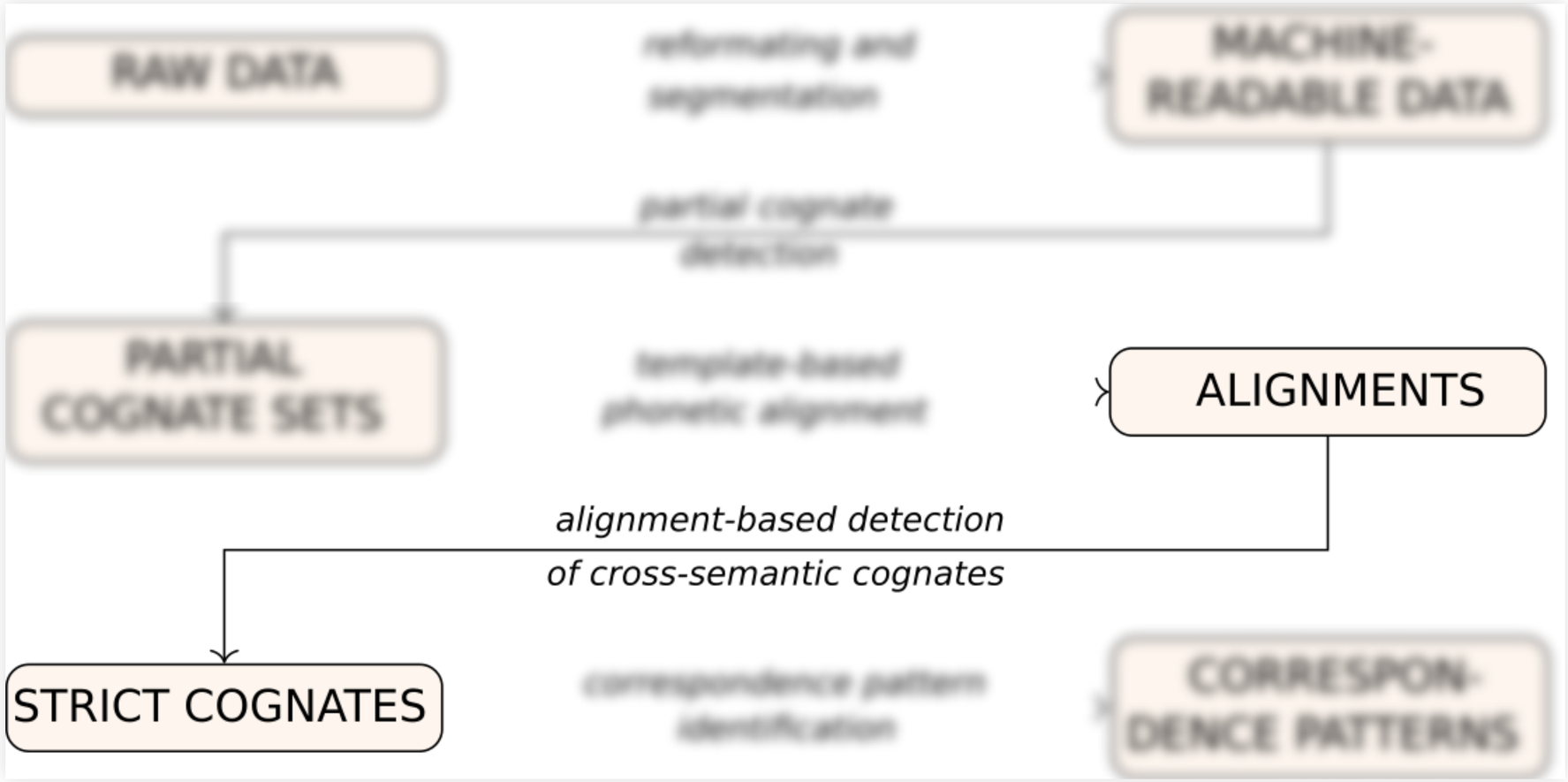
From cognates to alignments

Grapheme	Segmented	IPA	Structure
^tʰ	tʰ	tʰ	i
^v	v	v	i
ɛ	ɛ	ɛ	n
au	au	au	n
iei	i ei	j ei	m n
oŋ	o ŋ	o ŋ	n c
uɑŋ	u ɑ ŋ	w ɑ ŋ	m n c
loŋ	l o ŋ	l o ŋ	m n c
31$	31	31	t

From cognates to alignments

	A	B	C	D	E	F	G
1	ID	DOCULECT	ENGLISH	TOKENS	STRUCTURE	ALIGNMENT	COGIDS
2	1	Baheng, east	SEVEN	tɕ a ^{31}	i n t	tɕ a - ^{31}	3
3	2	Baheng, west	SEVEN	tɕ a ŋ ^{44}	i n c t	tɕ a ŋ ^{44}	3
4	3	Qiandong, east	SEVEN	ɕ u ŋ ^{53}	i n c t	ɕ u ŋ ^{53}	3
5	4	Qiandong, wesst	SEVEN	ɕ u ŋ ^{22}	i n c t	ɕ u ŋ ^{22}	3
6	5	Baheng, east	MOON	l a $^{3/0}$ + ɬ a ^{55}	i n t + i n t	l a $^{3/0}$ + ɬ a ^{55}	1908 1907
7	6	Baheng, west	MOON	ʔ a $^{3/0}$ + ɬ a ^{55}	i n t + i n t	ʔ a $^{3/0}$ + ɬ a ^{55}	1909 1907
8	7	Qiandong, east	MOON	l a ^{44} + l a ^{44}	i n t + i n t	l a ^{44} + l a ^{44}	1908 1907
9	8	Qiandong, wesst	MOON	p ɔ ^{11} + l a ^{33}	i n t + i n t	p ɔ ^{11} + l a ^{33}	1910 1907
10	9	Baheng, east	STAR	l a $^{3/0}$ + q a ŋ ^{35}	i n t + i n c t	l a $^{3/0}$ + q a ŋ ^{35}	1874 1870
11	10	Baheng, west	STAR	q a $^{3/0}$ + q a ŋ ^{35}	i n t + i n c t	q a $^{3/0}$ + q a ŋ ^{35}	1872 1870
12	11	Qiandong, east	STAR	q ei ^{24} + q ei ^{24}	i n t + i n t	q ei ^{24} + q ei - ^{24}	1872 1870
13	12	Qiandong, wesst	STAR	t ei - ^{44} + q ei ^{44}	i n t + i n t	t ei - ^{44} + q ei - ^{44}	1871 1870

From alignments to strict, cross-semantic cognates



From alignments to strict, cross-semantic cognates

- For a realistic analysis, we need to identify cognates not only within the same meaning slot, but across different concepts.
- However, our algorithm for automatic congate detection designed to search words with the same meaning.
- Therefore, we need to find *cross-semantic* partial (=normal) cognates in a second stage.

From alignments to strict, cross-semantic cognates

- For this task, we employ a new algorithm to *merge* cognates in our data into larger groups.
- The basic idea is to check if two alignments are compatible with each other, and to fuse them to form a bigger alignment, if this is the case.
- As a side effect, all words we identify in this way are *strictly* cognate, since our procedure does not allow to identify a morpheme in the same language to be cognate if this does not show the exact same form.

From alignments to strict, cross-semantic cognates

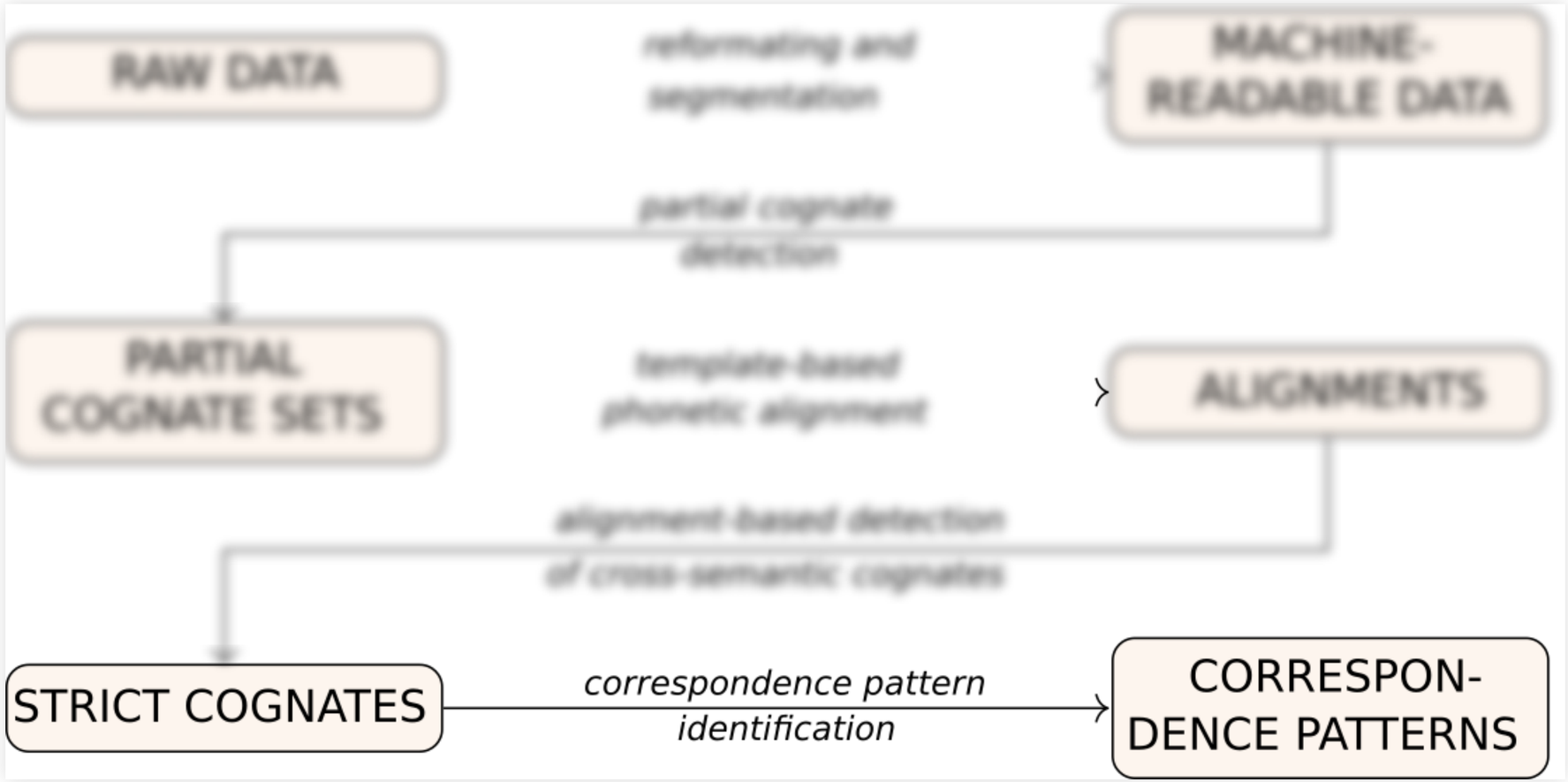
Language	Concept	Form	Morphemes
East Baheng	NOSE	^{n}pjau^{31}	NOSE
East Baheng	NASAL MUCUS	qa$^{3/0}$ + ^{n}pjau^{31}	qa NOSE
West Luobuohe	TWO	ʔu^{31}	TWO
West Luobuohe	TWENTY	ʔu^{31} + ʑo^{31}	TWO zo
West Baheng	SON	ʔa$^{3/0}$ + taŋ^{35}	a SON
West Baheng	SON-IN-LAW	taŋ^{35} + wei^{31}	SON wei
West Baheng	GRANDSON	taŋ^{35} + seŋ^{31}	SON seng
East Qiandong	SUN	q^{h}aŋ^{33} + nei^{24}	qhang SUN
East Qiandong	DAY (NOT NIGHT)	nei^{24}	SUN
West Baheng	FAECES (EXCREMENT)	qa^{31}	SHIT
West Baheng	STOMACH	ʔa$^{3/0}$ + tɕ^{h}i^{35} + qa^{31}	a tci SHIT
West Qiandong	ANT	kæ̃^{44} + mjɔ^{22}	INSECT mjo
West Qiandong	EARTHWORM	kæ̃^{44}+tɕuŋ^{44}	INSECT tsung
East Baheng	BIRD	taŋ^{35} + nuŋ^{31}	BIRD-A BIRD-B
East Baheng	NEST	ʑo^{11} + taŋ^{35} + nuŋ^{31}	zo BIRD-A BIRD-B

From alignments to strict, cross-semantic cognates

Language	Concept	Form	Cognacy	Cross-Semantic
East Baheng	SON	$taŋ^{35}$	1	1
East Baheng	DAUGHTER	$p^{h}je^{53}$	2	2
West Baheng	SON	$ʔa^{3/0}$ + $taŋ^{35}$	3 1	3 1
West Baheng	DAUGHTER	ta^{55} + $qa^{3/0}$ + $t^{h}jei^{53}$	4 5 6	4 5 6
Chuanqiandian	SON	to^{43}	1	1
Chuanqiandian	DAUGHTER	$^{n}ts^{h}ai^{33}$	7	7
Chuanqiandian (Central Guizhou)	SON	$tə^{2/0}$ + $tə̃^{24}$	8 1	8 1
Chuanqiandian (Central Guizhou)	DAUGHTER	$tə̃^{24}$ + $^{n}p^{h}e^{42}$	9 2	1 2
East Qiandong	SON	tei^{24}	1	1
East Qiandong	DAUGHTER	tei^{24} + $p^{h}a^{35}$	9 2	1 2

From alignments to strict, cross-semantic cognates

	A	B	C	D	E	F	G	H
1	ID	DOCULECT	ENGLISH	TOKENS	STRUCTURE	ALIGNMENT	CROSSIDS	COGIDS
2	1	Baheng, east	SEVEN	tɕ a ^{31}	i n t	tɕ a - ^{31}	3	3
3	2	Baheng, west	SEVEN	tɕ a ŋ ^{44}	i n c t	tɕ a ŋ ^{44}	3	3
4	3	Qiandong, east	SEVEN	ɕ u ŋ ^{53}	i n c t	ɕ u ŋ ^{53}	3	3
5	4	Qiandong, wesst	SEVEN	ɕ u ŋ ^{22}	i n c t	ɕ u ŋ ^{22}	3	3
6	5	Baheng, east	MOON	l a $^{3/0}$ + ɬ a ^{55}	i n t + i n t	l a $^{3/0}$ + ɬ a ^{55}	1908 351	1908 1907
7	6	Baheng, west	MOON	ʔ a $^{3/0}$ + ɬ a ^{55}	i n t + i n t	ʔ a $^{3/0}$ + ɬ a ^{55}	41 351	1909 1907
8	7	Qiandong, east	MOON	l a ^{44} + l a ^{44}	i n t + i n t	l a ^{44} + l a ^{44}	1908 351	1908 1907
9	8	Qiandong, wesst	MOON	p ɔ ^{11} + l a ^{33}	i n t + i n t	p ɔ ^{11} + l a ^{33}	1910 351	1910 1907
10	9	Baheng, east	STAR	l a $^{3/0}$ + q a ŋ ^{35}	i n t + i n c t	l a $^{3/0}$ + q a ŋ ^{35}	1874 1834	1874 1870
11	10	Baheng, west	STAR	q a $^{3/0}$ + q a ŋ ^{35}	i n t + i n c t	q a $^{3/0}$ + q a ŋ ^{35}	1872 1834	1872 1870
12	11	Qiandong, east	STAR	q ei ^{24} + q ei ^{24}	i n t + i n t	q ei ^{24} + q ei - ^{24}	1872 1834	1872 1870
13	12	Qiandong, wesst	STAR	t ei - ^{44} + q ei ^{44}	i n t + i n t	t ei - ^{44} + q ei - ^{44}	1234 1834	1871 1870

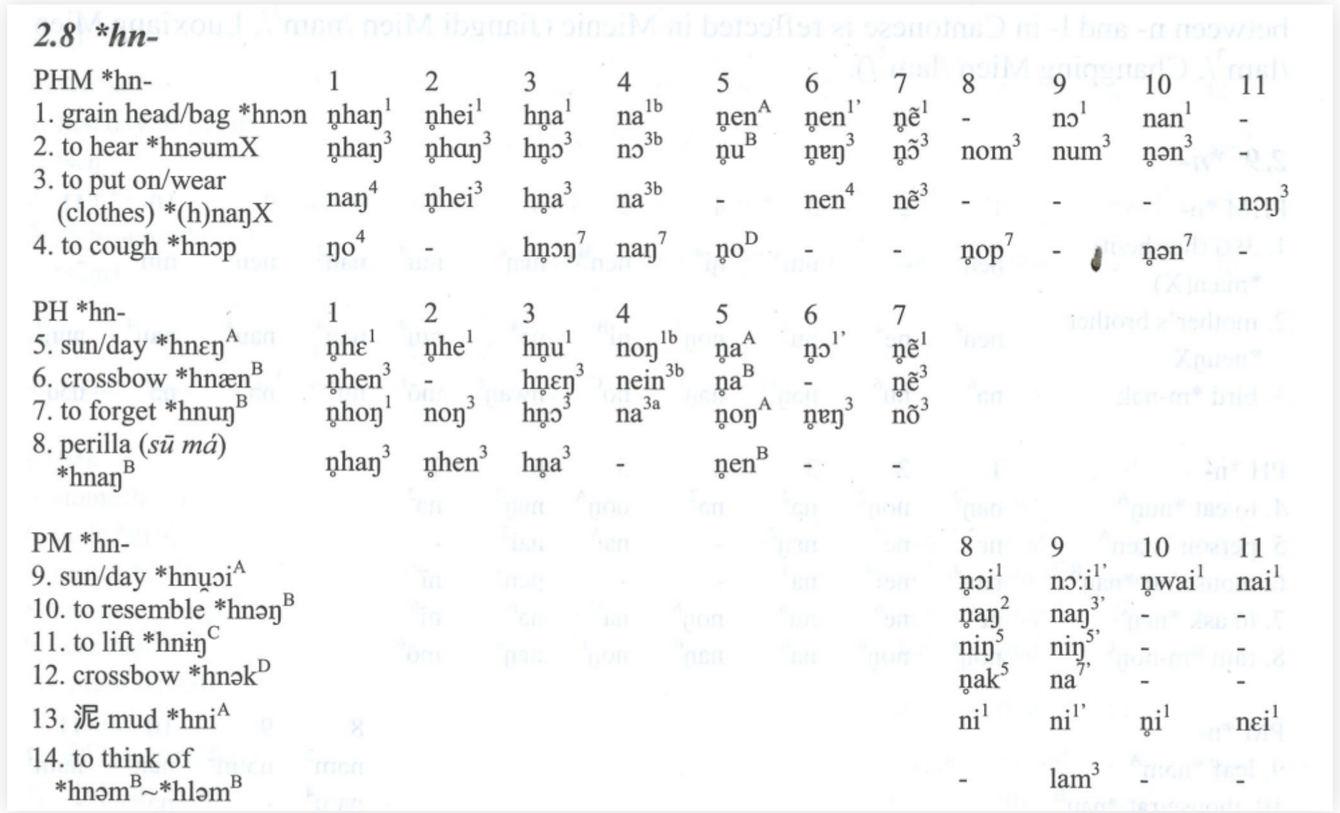

RAW DATA
reformating and segmentation
MACHINE-READABLE DATA
partial cognate detection
PARTIAL COGNATE SETS
template-based phonetic alignment
ALIGNMENTS
alignment-based detection of cross-semantic cognates
STRICT COGNATES
correspondence pattern identification
CORRESPON-DENCE PATTERNS

From strict cognates to sound correspondence patterns

Sound correspondence patterns for Indo-European stops, following Clackson (2007, page 37) .

PIE	Hittite	Sanskrit	Greek	Latin	Gothic	...
*p	p	p	p	p	f b	...
*b	b p	b	b	b	p	...
*b^h	b p	$b^{ɦ}$/bh	p^h/ph	f b	b	...
*t	t	t	t	t	θ/þ d	...
*d	d t	d	d	d	t	...
*d^h	d t	d^h/dh h	t^h/th	f d b	d	...
...	...	...	...	...	...	...
*k^w	k^w/ku	k c	k p t	k^w/qu	h^w/hw g	...
*g^w	k^w/u	g j	g b d	g^w/gu u	q	...
*g^{wh}	k^w/ku g^w/gu	$g^{ɦ}$/gh h	p^h/ph t^h/th k^h/kh	f g^w/gu u	g b	...

From strict cognates to sound correspondence patterns



2.8 *hn-

PHM *hn-	1	2	3	4	5	6	7	8	9	10	11
1. grain head/bag *hnɔn	ŋ̊haŋ^{1}	ŋ̊hei^{1}	hŋ̊a^{1}	na^{1b}	ŋ̊en^{A}	ŋ̊en$^{1'}$	ŋ̊ẽ^{1}	-	nɔ^{1}	nan^{1}	-
2. to hear *hnəumX	ŋ̊haŋ^{3}	ŋ̊hɑŋ^{3}	hŋ̊ɔ^{3}	nɔ^{3b}	ŋ̊u^{B}	ŋ̊ɐŋ^{3}	ŋ̊õ^{3}	nom^{3}	num^{3}	ŋ̊ən^{3}	-
3. to put on/wear (clothes) *(h)naŋX	naŋ^{4}	ŋ̊hei^{3}	hŋ̊a^{3}	na^{3b}	-	nen^{4}	nẽ^{3}	-	-	-	nɔŋ^{3}
4. to cough *hnɔp	ŋo^{4}	-	hŋ̊ɔŋ^{7}	naŋ^{7}	ŋ̊o^{D}	-	-	ŋ̊op^{7}	-	ŋ̊ən^{7}	-

PH *hn-	1	2	3	4	5	6	7
5. sun/day *hnɛŋ^{A}	ŋ̊hɛ^{1}	ŋ̊he^{1}	hŋ̊u^{1}	noŋ^{1b}	ŋ̊a^{A}	ŋ̊ɔ$^{1'}$	ŋ̊ẽ^{1}
6. crossbow *hnæn^{B}	ŋ̊hen^{3}	-	hŋ̊ɛŋ^{3}	nein^{3b}	ŋ̊a^{B}	-	ŋ̊ẽ^{3}
7. to forget *hnuŋ^{B}	ŋ̊hoŋ^{1}	noŋ^{3}	hŋ̊ɔ^{3}	na^{3a}	ŋ̊oŋ^{A}	ŋ̊ɐŋ^{3}	nõ^{3}
8. perilla (*sū má*) *hnaŋ^{B}	ŋ̊haŋ^{3}	ŋ̊hen^{3}	hŋ̊a^{3}	-	ŋ̊en^{B}	-	-

PM *hn-	8	9	10	11
9. sun/day *hnu̯ɔi^{A}	ŋ̊ɔi^{1}	nɔːi$^{1'}$	ŋ̊wai^{1}	nai^{1}
10. to resemble *hnəŋ^{B}	ŋ̊aŋ^{2}	naŋ$^{3'}$	-	-
11. to lift *hniŋ^{C}	niŋ^{5}	niŋ$^{5'}$	-	-
12. crossbow *hnək^{D}	ŋ̊ak^{5}	na$^{7'}$	-	-
13. 泥 mud *hni^{A}	ni^{1}	ni$^{1'}$	ŋ̊i^{1}	nɛi^{1}
14. to think of *hnəm^{B}~*hləm^{B}	-	lam^{3}	-	-

Ratliff et al. (2010). Hmong-Mien language history. Pacific Linguistics (Page 57)

From strict cognates to sound correspondence patterns

COGNATES	INDEX	PATTERN	CONCEPTS	luo	luo	chu	chu	qia	qia	bah	bah	chu	chu	SIZE
8	1	ŋ / 28	太阳,天	n	ŋ	ŋ	ŋ	n	n	ŋ	ŋ	ŋ	n	3.50 / 6
207	1	ŋ / 28	听见	n	ŋ	ŋ	ŋ	n	n	ŋ	Ø	ŋ	n	3.50 / 6
300	1	ŋ / 28	咳嗽	n	Ø	Ø	Ø	Ø	n	Ø	Ø	ŋ	n	3.50 / 6
847	1	ŋ / 28	麻雀	n	Ø	Ø	Ø	n	n	Ø	Ø	Ø	Ø	3.50 / 6
105	1	ŋ / 28	箭	n	ŋ	Ø	Ø	Ø	Ø	ŋ	Ø	Ø	Ø	3.50 / 6
93	1	ŋ / 28	弓	Ø	Ø	ŋ	ŋ	Ø	n	ŋ	Ø	ŋ	n	3.50 / 6

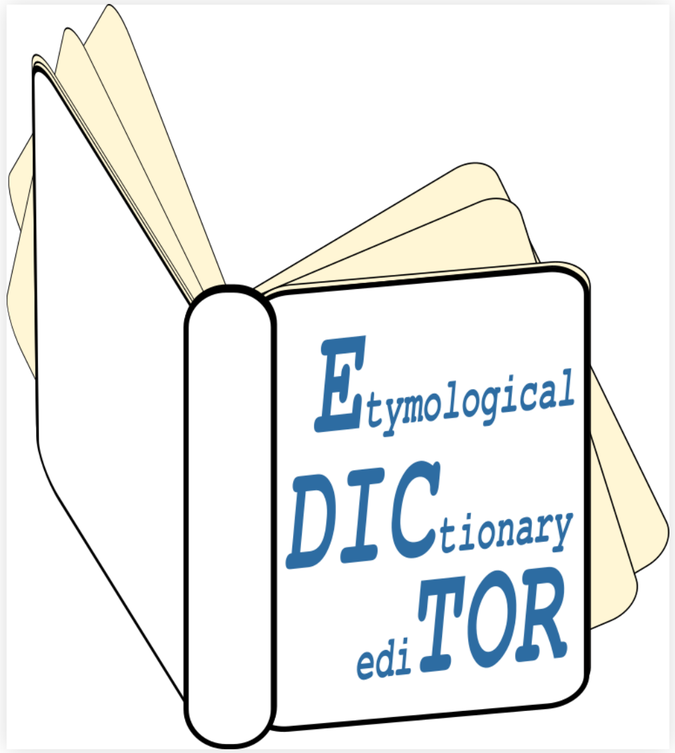
Illustration of the Workflow

Orthography profiles

http://calc.digling.org/profile/

Illustration of the Workflow

EDICTOR: a web-based tool to edit, analyse, and publish etymological data.



Modeling and annotation

Nathanael E. Schweikhard

Example of an Annotated Wordlist

ID	Language_ID	Parameter_ID	Value	Form	Segments	Source
Tibetan_Old_Tibetan-1741-1	Tibetan_Old_Tibetan	1741	*steŋ*	*steŋ*	*s t e ŋ*	Huang1992
rGyalrong_Japhug-1741-1	rGyalrong_Japhug	1741	*ɯ-taʁ*	*ɯ-taʁ*	*ɯ + t a ʁ*	Jacques2015b
Tibetan_Old_Tibetan-98-1	Tibetan_Old_Tibetan	98	*tʰams.tɕad*	*tʰams.tɕad*	*tʰ a m s + tɕ a d*	Huang1992
Kiranti_Khaling-98-1	Kiranti_Khaling	98	*kʰøle*	*kʰøle*	*kʰ ø l e*	Jacques2017FN
Kiranti_Limbu-98-1	Kiranti_Limbu	98	*kak*	*kak*	*k a k*	Jacques2017FN
rGyalrong_Japhug-98-1	rGyalrong_Japhug	98	*%tʰamtɕʏt*	*%tʰamtɕʏt*	*tʰ a m tɕ ʏ t*	Jacques2015b
Tangut-98-1	Tangut	98	*zji¹*	*zji¹*	*z j i ¹*	Li1997
Tibetan_Old_Tibetan-1292-1	Tibetan_Old_Tibetan	1292	*ŋan*	*ŋan*	*ŋ a n*	Huang1992
rGyalrong_Japhug-1292-1	rGyalrong_Japhug	1292	*%ŋʏn*	*%ŋʏn*	*ŋ ʏ n*	Jacques2015b
Tibetan_Old_Tibetan-1422-1	Tibetan_Old_Tibetan	1422	*gson+po*	*gson+po*	*g s o n + p o*	Huang1992

Excerpt from Sino-Tibetan Database of Lexical Cognates (Sagart et al. 2019).

Cross-Links to Reference Catalogs: Glottolog

ID	*Language_ID*	Parameter_ID	Value	Form	Segments	Source
Tibetan_Old_Tibetan-1741-1	*Tibetan_Old_Tibetan*	1741	*steŋ*	*steŋ*	*s t e ŋ*	Huang1992
rGyalrong_Japhug-1741-1	*rGyalrong_Japhug*	1741	*ɯ-taʁ*	*ɯ-taʁ*	*ɯ + t a ʁ*	Jacques2015b
Tibetan_Old_Tibetan-98-1	*Tibetan_Old_Tibetan*	98	*tʰams.tɕad*	*tʰams.tɕad*	*tʰ a m s + tɕ a d*	Huang1992
Kiranti_Khaling-98-1	*Kiranti_Khaling*	98	*kʰøle*	*kʰøle*	*kʰ ø l e*	Jacques2017FN
Kiranti_Limbu-98-1	*Kiranti_Limbu*	98	*kak*	*kak*	*k a k*	Jacques2017FN
rGyalrong_Japhug-98-1	*rGyalrong_Japhug*	98	*%tʰamtɕʏt*	*%tʰamtɕʏt*	*tʰ a m tɕ ʏ t*	Jacques2015b
Tangut-98-1	*Tangut*	98	*zji¹*	*zji¹*	*z j i ¹*	Li1997
Tibetan_Old_Tibetan-1292-1	*Tibetan_Old_Tibetan*	1292	*ŋan*	*ŋan*	*ŋ a n*	Huang1992
rGyalrong_Japhug-1292-1	*rGyalrong_Japhug*	1292	*%ŋʏn*	*%ŋʏn*	*ŋ ʏ n*	Jacques2015b
Tibetan_Old_Tibetan-1422-1	*Tibetan_Old_Tibetan*	1422	*gson+po*	*gson+po*	*g s o n + p o*	Huang1992

Excerpt from Sino-Tibetan Database of Lexical Cognates (Sagart et al. 2019).

Glottolog

Classification

open Indo-European | expand all | collapse all

- ▼ Indo-European (588)
  - ▶ Albanian (4)
  - ▶ Anatolian (10)
  - ▶ Armenic (3)
  - ▶ Balto-Slavic (23)

Family membership references

- Fortson, IV, Benjamin F. 2004
- Petri Kallio and Jorma Koivulehto 2018

Comments on family membership

Fortson, IV, Benjamin F. 2004 , Petri Kallio and Jorma Koivulehto 2018

Comments on subclassification

Don Ringe 2017 James Clackson 2007

show big map

Links

References

This family has more than 500 languages. Please select an appropriate sub-family to get a list of relevant references.

Glottolog 4.0 edited by Hammarström, Harald & Forkel, Robert & Haspelmath, Martin
is licensed under a Creative Commons Attribution 4.0 International License.

Glottolog, a reference database of languages and their genealogical relations (Hammarström et al. 2019).

Cross-Links to Reference Catalogs: Concepticon

ID	Language_ID	*Parameter_ID*	Value	Form	Segments	Source
Tibetan_Old_Tibetan-1741-1	Tibetan_Old_Tibetan	*1741*	*steŋ*	*steŋ*	*s t e ŋ*	Huang1992
rGyalrong_Japhug-1741-1	rGyalrong_Japhug	*1741*	*ɯ-tɐʁ*	*ɯ-tɐʁ*	*ɯ + t ɐ ʁ*	Jacques2015b
Tibetan_Old_Tibetan-98-1	Tibetan_Old_Tibetan	*98*	*tʰams.tɸad*	*tʰams.tɸad*	*tʰ a m s + tɸ a d*	Huang1992
Kiranti_Khaling-98-1	Kiranti_Khaling	*98*	*kʰøle*	*kʰøle*	*kʰ ø l e*	Jacques2017FN
Kiranti_Limbu-98-1	Kiranti_Limbu	*98*	*kak*	*kak*	*k a k*	Jacques2017FN
rGyalrong_Japhug-98-1	rGyalrong_Japhug	*98*	*%tʰamtɸʁt*	*%tʰamtɸʁt*	*tʰ a m tɸ ʁ t*	Jacques2015b
Tangut-98-1	Tangut	*98*	*zji¹*	*zji¹*	*z j i ¹*	Li1997
Tibetan_Old_Tibetan-1292-1	Tibetan_Old_Tibetan	*1292*	*ŋan*	*ŋan*	*ŋ a n*	Huang1992
rGyalrong_Japhug-1292-1	rGyalrong_Japhug	*1292*	*%ŋʁn*	*%ŋʁn*	*ŋ ʁ n*	Jacques2015b
Tibetan_Old_Tibetan-1422-1	Tibetan_Old_Tibetan	*1422*	*gson+po*	*gson+po*	*g s o n + p o*	Huang1992

Excerpt from Sino-Tibetan Database of Lexical Cognates (Sagart et al. 2019).

Concepticon

To produce a loud, short, explosive sound similar to that of a dog.

Showing 1 to 12 of 12 entries

← Previous 1 Next →

Id	Concept in source	Conceptlist
Search	Search	Search
Allen-2007-500-382	吠 [chinese]; bark (of dog) [english]	Allen 2007 500
Bulakh-2013-870-589	to bark (of a dog) [english]	Bulakh 2013 870
Castro-2010-540-382	吠(吠叫) [chinese]; to bark [english]	Castro 2010 540
Castro-2015-608-382	吠 [chinese]; to bark [english]	Castro 2015 608
Dellert-2017-1016-726	bark [english]; bellen [german]; лаять [russian]	Dellert 2017 1016
Hale-1973-1798-398	bark [english]	Hale 1973 1798
Luniewska-2016-299-159	blaf [afrikaans]; bordar [catalan]; hunden gør [danish]; blaffen [dutch]; bark [english]; haukkua [finnish]; bellen [german]; γαυγίζει [greek]; linbo'ax [hebrew]; ugat [hungarian]; gelta [icelandic]; (ag) tafann [irish]; abbaiare [italian]; loti [lithuanian]; billen [luxembourgish]; tinbaħ [maltese]; szczekać [polish]; гавкать [russian]; lajati [serbian]; štekať [slovak]; bark [southafricanenglish]; ladrar [spanish]; skälla [swedish]; havlamak [turkish]; khonkotha [xhosa]	Luniewska 2016 299
Mann-1998-406-82	bark [english]	Mann 1998 406
Mitterhofer-2013-300-231	bark (dog) [english]	Mitterhofer 2013 300
Mitterhofer-2013-355-231	bark (dog) [english]	Mitterhofer 2013 355
Robinson-2012-398-	to bark [english]	Robinson 2012 398

MRC Psycholinguistic Database	
KUCERA FRANCIS FREQUENCY	2
MRC WORD	BARKING
Mapping to OmegaWiki	
OMEGAWIKI ID	5444
Edinburgh Associative Thesaurus	
EAT WORD	BARKING
WEIGHTED DEGREE	105.00
DEGREE	23

The concept ’barking’ in the Concepticon database (List et al. 2019).

A Morpheme-Segmented Wordlist

ID	Language_ID	Parameter_ID	Value	Form	Segments	Source
Tibetan_Old_Tibetan-1741-1	Tibetan_Old_Tibetan	1741	*steŋ*	*steŋ*	*s t e ŋ*	Huang1992
rGyalrong_Japhug-1741-1	rGyalrong_Japhug	1741	*ɯ-tɐʁ*	*ɯ-tɐʁ*	*ɯ + t ɐ ʁ*	Jacques2015b
Tibetan_Old_Tibetan-98-1	Tibetan_Old_Tibetan	98	*tʰams.tɕad*	*tʰams.tɕad*	*tʰ a m s + tɕ a d*	Huang1992
Kiranti_Khaling-98-1	Kiranti_Khaling	98	*kʰɵle*	*kʰɵle*	*kʰ ɵ l e*	Jacques2017FN
Kiranti_Limbu-98-1	Kiranti_Limbu	98	*kak*	*kak*	*k a k*	Jacques2017FN
rGyalrong_Japhug-98-1	rGyalrong_Japhug	98	*%tʰamtɵʁt*	*%tʰamtɵʁt*	*tʰ a m tɵ ʁ t*	Jacques2015b
Tangut-98-1	Tangut	98	*zji¹*	*zji¹*	*z j i ¹*	Li1997
Tibetan_Old_Tibetan-1292-1	Tibetan_Old_Tibetan	1292	*ŋan*	*ŋan*	*ŋ a n*	Huang1992
rGyalrong_Japhug-1292-1	rGyalrong_Japhug	1292	*%ŋʁn*	*%ŋʁn*	*ŋ ʁ n*	Jacques2015b
Tibetan_Old_Tibetan-1422-1	Tibetan_Old_Tibetan	1422	*gson+po*	*gson+po*	*g s o n + p o*	Huang1992

Excerpt from Sino-Tibetan Database of Lexical Cognates (Sagart et al. 2019).

Compositionality

- Compositionality is a basic feature of human language (Zeige 2015). Language consists of re-combinable elements.
- This entails an unlimited amount of expressions from a limited amount of elements.
- Different words may therefore share some of their morphemes.
- With morpheme annotation we can study the structure of the lexicon and even language history.

Automated Morpheme Segmentation

- Morphemes (List 2019)
  - are recurring combinations of form and meaning
  - and abstraction of relations within the lexicon
  - which reflect language history
  - and are often bound to phonotactic restrictions
  - while being sometimes marked orthographically (space, dash, different character).
- Many approaches search only for recurring letter strings.
- The quality of an approach depends on language and amount of data.
- There is no standard for testing new methods.
- Morpheme-segmented wordlists could be used for testing purposes.

Glossed morphemes

ID	DOCULECT	CONCEPT	FORM	TOKENS	SEGM-TOKENS	MORPHEMES	COGNATES
339	German	spider	Spinne	ʃ p ɪ n ə	ʃ p ɪ n + ə	SPIN _e-suff	1 2
341	German	spider web	Spinnwebe	ʃ p ɪ n v eː b ə	ʃ p ɪ n + v eː b + ə	SPIN WEAVE _e-suff	1 3 2
342	German	spider web	Spinnennetz	ʃ p ɪ n ə n n ɛ ʦ	ʃ p ɪ n + ə n + n ɛ ʦ	SPIN _en-fuge NET	1 4 5
753	German	spin	spinnen	ʃ p ɪ n ə n	ʃ p ɪ n + ə n	SPIN _inf	1 6

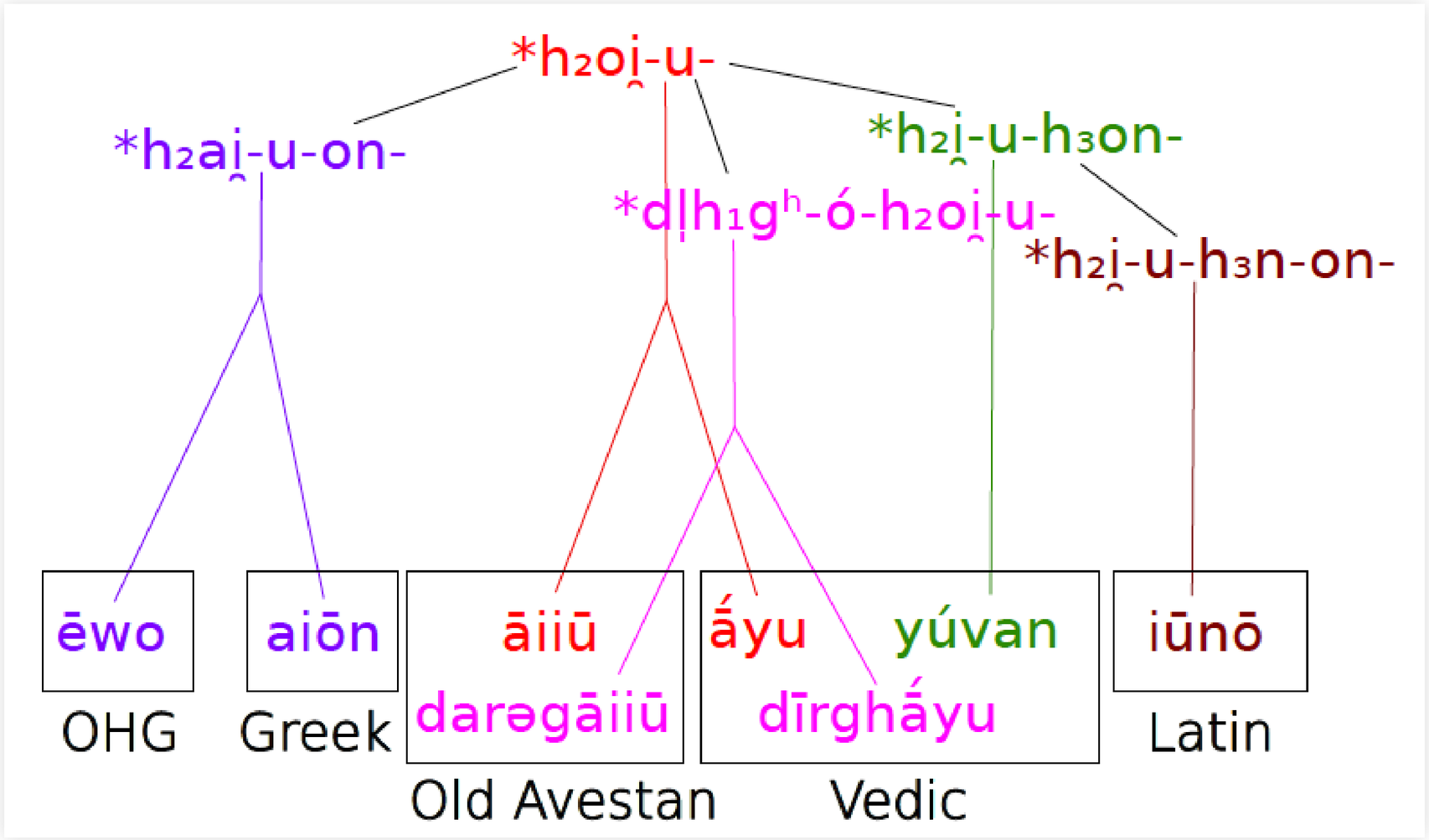
Data based on the Intercontinental Dictionary Series (Key and Comrie 2016)

Word Formation

Basic Type	Process	Example
concatenative	compounding	fish + tank → fish tank
	affixation	fish + er → fisher
	full reduplication	Malay: bunga ('flower') → bungabunga ('flowers')
	conversion	fish (noun) → fish (verb)
allomorphic	pattern-based	German: Apfel ('apple') → Äpfel ('apples')
	blending	breakfast + lunch → brunch
	infixation	Tagalog: basag (to write') → bumasag ('wrote')
	reanalysis	sculptor → sculpt
	partial reduplication	Mangab-mbula: kan ('to eat') → kanan ('be eating')
shortening	acronyms	radio detection and ranging → radar
	clippings	discotheque → disco

Types of word formation, based on Haspelmath 2001 and Trask 2000.

Word Formation in Indo-European

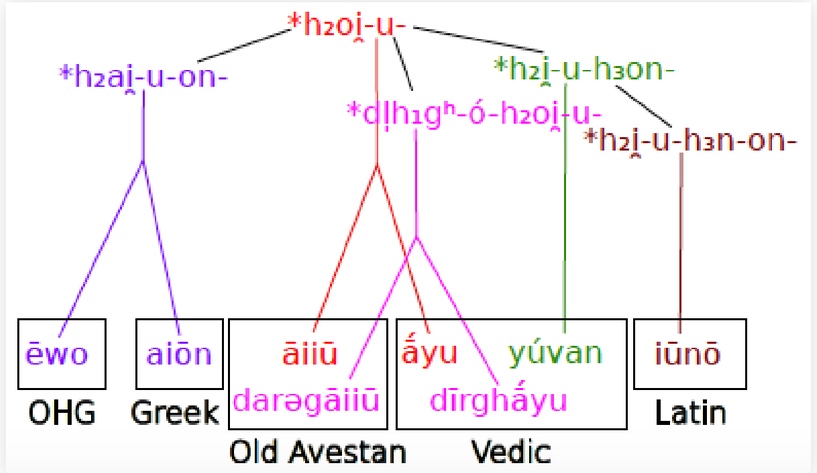


A family tree of *h₂ei-u-* (based on Wodtko et al. 2008 and Mallory/Adams 2006)

Annotation of Word Formation Process I

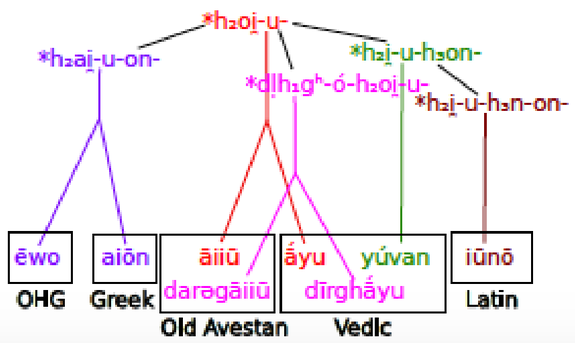
ID	LANGUAGE	CONCEPT	FORM	MORPHEMES	COGNATES	ROOTS
1	Old High German	eternity	*ēwo*	*ēw o*	1 2	1 2
2	Ancient Greek	life	*aiōn*	*ai ōn*	1 2	1 2
3	Old Avestan	life	*āiiū*	*āiiū*	3	1
4	Old Avestan	long-living	*darəgāiiū*	*darəg a āiiū*	4 5 3	3 4 1
5	Vedic	life	*áyu*	*áyu*	3	1
6	Vedic	long-living	*dīrghā́yu*	*dīrgh á ā́yu*	4 5 3	3 4 1
7	Vedic	young	*yúvan*	*yúv an*	6 7	1 5
8	Latin	(deity name)	*iūnō*	*iū n ō*	6 8 2	1 5 2
9	Indo-European	life	**h₂ai̯-u-on-*	*h₂ai̯u on*	3 2	1 2
10	Indo-European	life	**h₂oi̯-u-*	*h₂oi̯u*	1	1
11	Indo-European	long-living	**dl̥h₁gʰ-ó-h₂oi̯-u-*	*dl̥h₁gʰ ó h₂oi̯u*	4 5 1	3 4 1
12	Indo-European	young	**h₂i̯-u-h₃on-*	*h₂i̯u h₃on*	6 7	1 5
13	Indo-European	the young one	**h₂i̯-u-h₃n-on-*	*h₂i̯u h₃n on*	6 8 2	1 5 2

Source	Source-ID	Target	Target-ID	Change
**h₂ai̯-u-on-*	1	*aiōn*	2	sound change
**h₂oi̯-u-*	3	**h₂ai̯-u-on-*	1	e-grade, on-suffix
**h₂oi̯-u-*	3	**dl̥h₁gʰ-ó-h₂oi̯-u-*	4	compound with **dl̥h₁gʰ-ó-*
**dl̥h₁gʰ-ó-h₂oi̯-u-*	7	*dīrghā́yu*	8	sound change
...	...	...	...	...



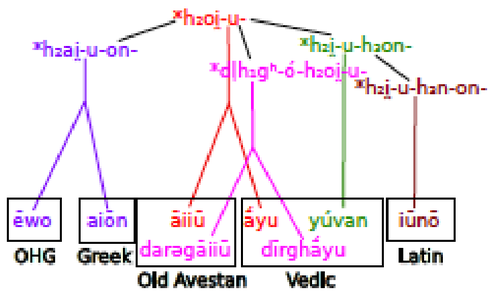
Annotation of Word Formation Processes II

ID	LANGUAGE	CONCEPT	FORM	MORPHEMES	COGNATES	ROOTS
1	Old High German	eternity	*ēwo*	*ēw o*	1 2	1 2
2	Ancient Greek	life	*aiōn*	*ai ōn*	1 2	1 2
3	Old Avestan	life	*āiiū*	*āiiū*	3	1
4	Old Avestan	long-living	*darəgāiiū*	*darəg a āiiū*	4 5 3	3 4 1
5	Vedic	life	*áyu*	*áyu*	3	1
6	Vedic	long-living	*dīrghấyu*	*dīrgh á ấyu*	4 5 3	3 4 1
7	Vedic	young	*yúvan*	*yúv an*	6 7	1 5
8	Latin	(deity name)	*iūnō*	*iū n ō*	6 8 2	1 5 2
9	Indo-European	life	**h₂ai̯-u-on-*	*h₂ai̯u on*	3 2	1 2
10	Indo-European	life	**h₂oi̯-u-*	*h₂oi̯u*	1	1
11	Indo-European	long-living	**dl̥h₁gʰ-ó-h₂oi̯-u-*	*dl̥h₁gʰ ó h₂oi̯u*	4 5 1	3 4 1
12	Indo-European	young	**h₂i̯-u-h₃on-*	*h₂i̯u h₃on*	6 7	1 5
13	Indo-European	the young one	**h₂i̯-u-h₃n-on-*	*h₂i̯u h₃n on*	6 8 2	1 5 2



Source	Source-ID	Target	Target-ID	Change
**h₂ai̯-u-on-*	1	*aiōn*	2	sound change
**h₂oi̯-u-*	3	**h₂ai̯-u-on-*	1	e-grade, on-suffix
**h₂oi̯-u-*	3	**dl̥h₁gʰ-ó-h₂oi̯-u-*	4	compound with **dl̥h₁gʰ-ó-*
**dl̥h₁gʰ-ó-h₂oi̯-u-*	7	*dīrghấyu*	8	sound change
...	...	...	...	...

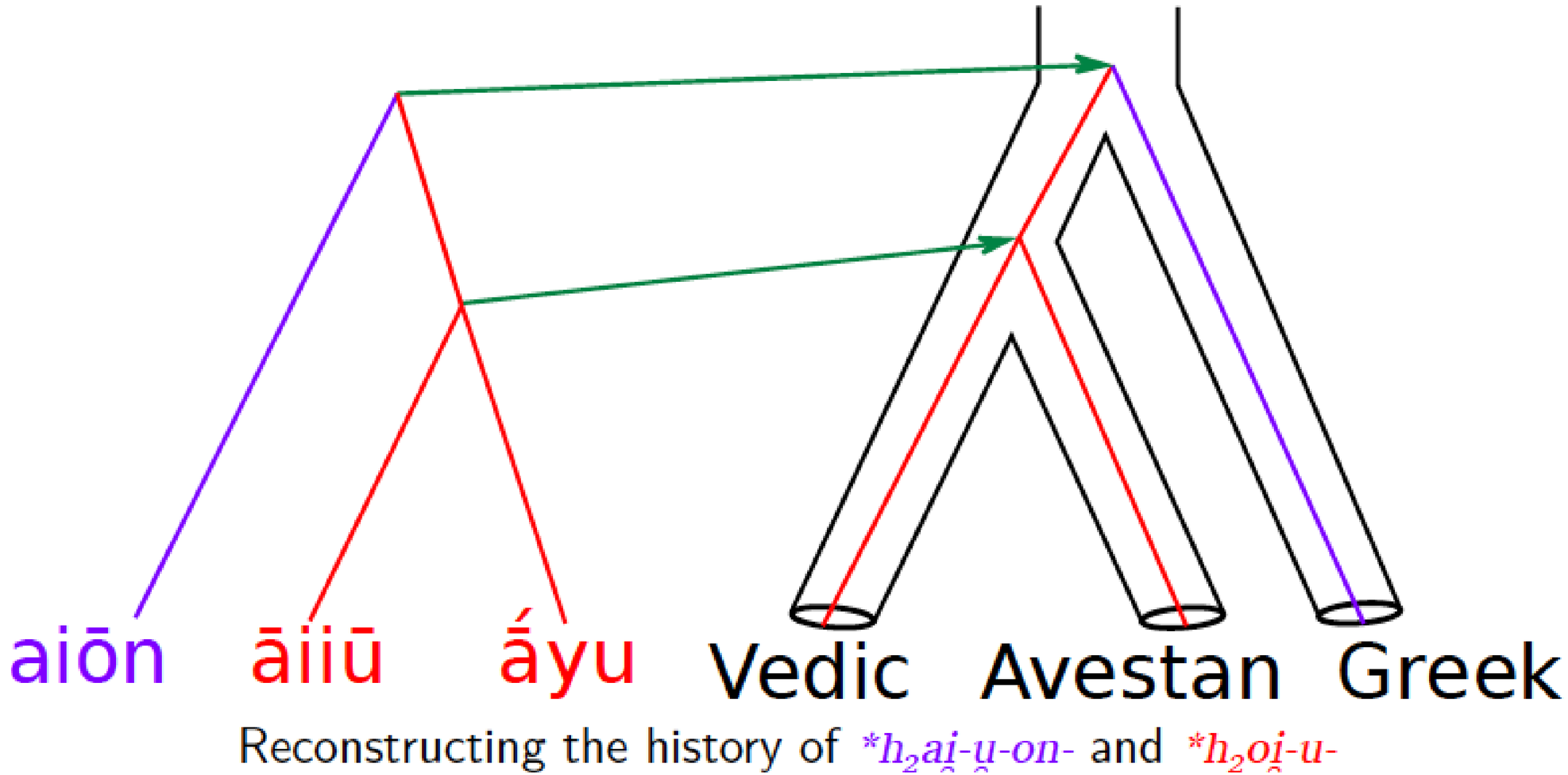
Annotation of Word Formation Processes III



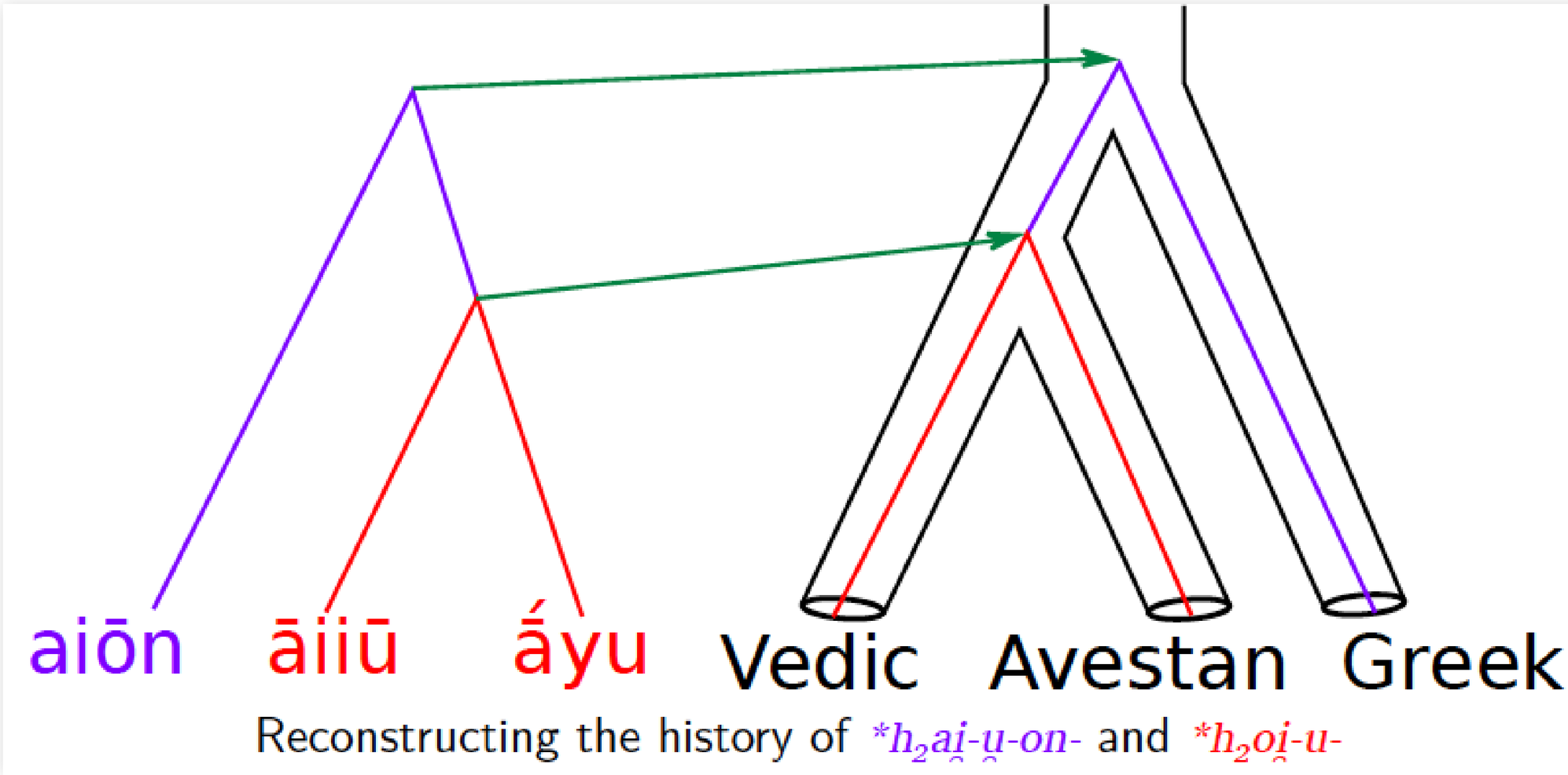
ID	LANGUAGE	CONCEPT	FORM	MORPHEMES	COGNATES	ROOTS
1	Old High German	eternity	*ēwo*	*ēw o*	1 2	1 2
2	Ancient Greek	life	*aiōn*	*ai ōn*	1 2	1 2
3	Old Avestan	life	*āiiū*	*āiiū*	3	1
4	Old Avestan	long-living	*darəgāiiū*	*darəg a āiiū*	4 5 3	3 4 1
5	Vedic	life	*áyu*	*áyu*	3	1
6	Vedic	long-living	*dīrghā́yu*	*dīrgh á ā́yu*	4 5 3	3 4 1
7	Vedic	young	*yúvan*	*yúv an*	6 7	1 5
8	Latin	(deity name)	*iūnō*	*iū n ō*	6 8 2	1 5 2
9	Indo-European	life	**h₂ai̯-u-on-*	*h₂ai̯u on*	3 2	1 2
10	Indo-European	life	**h₂oi̯-u-*	*h₂oi̯u*	1	1
11	Indo-European	long-living	**dḷh₁gʰ-ó-h₂oi̯-u-*	*dḷh₁gʰ ó h₂oi̯u*	4 5 1	3 4 1
12	Indo-European	young	**h₂i̯-u-hₑon-*	*h₂i̯u hₑon*	6 7	1 5
13	Indo-European	the young one	**h₂i̯-u-hₑn-on-*	*h₂i̯u hₑn on*	6 8 2	1 5 2

Source	Source-ID	Target	Target-ID	Change
$*h_2ai̯\text{-}u\text{-}on\text{-}$	1	*aiōn*	2	sound change
$*h_2oi̯\text{-}u\text{-}$	3	$*h_2ai̯\text{-}u\text{-}on\text{-}$	1	e-grade, on-suffix
$*h_2oi̯\text{-}u\text{-}$	3	$*dḷh_1g^h\text{-}ó\text{-}h_2oi̯\text{-}u\text{-}$	4	compound with $*dḷh_1g^h\text{-}ó\text{-}$
$*dḷh_1g^h\text{-}ó\text{-}h_2oi̯\text{-}u\text{-}$	7	*dīrghā́yu*	8	sound change
…	…	…	…	…

Modelling Language History I

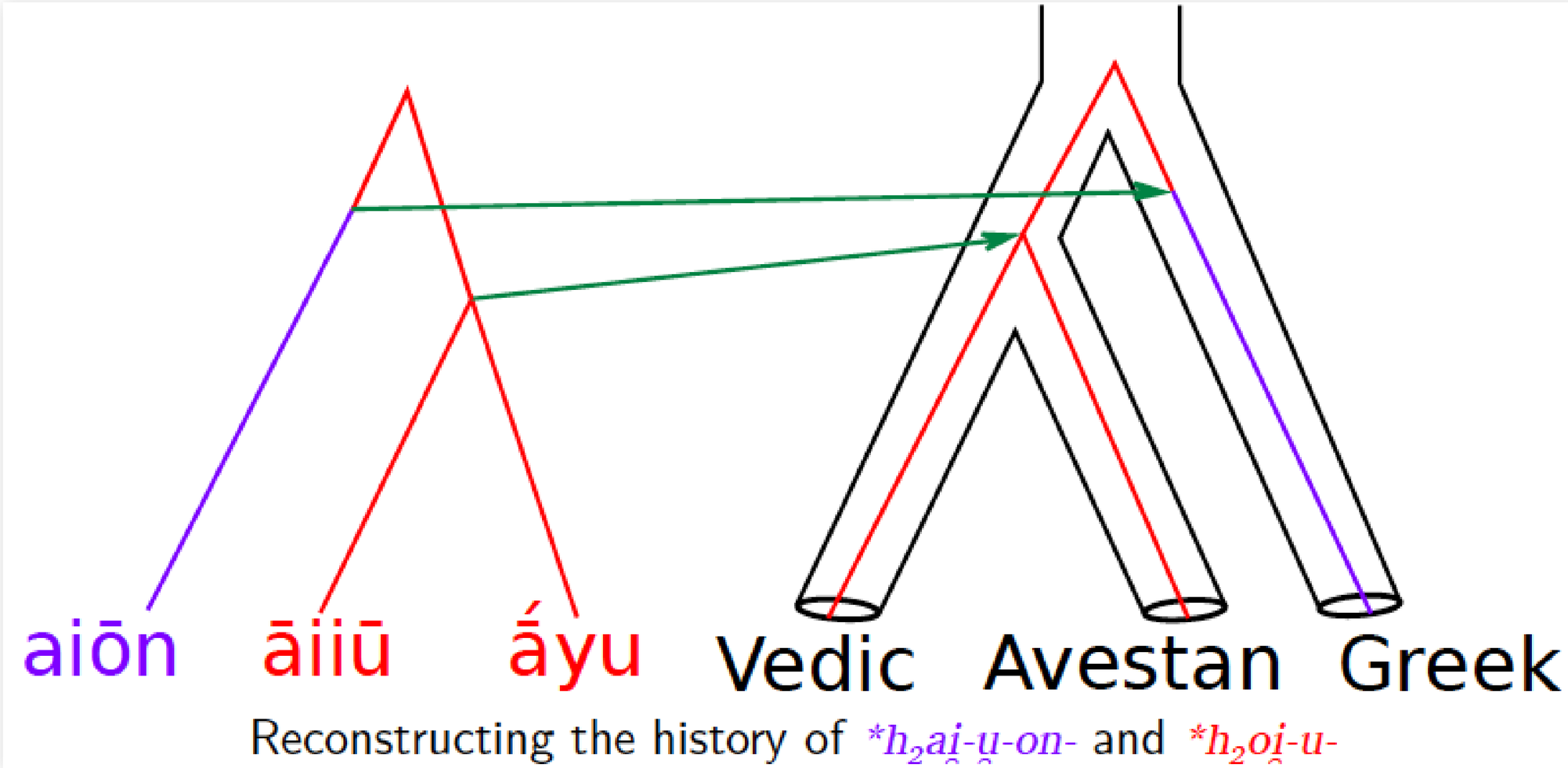


Modelling Language History II



Reconstructing the history of *$h_2$ai̯-u̯-on-* and *$h_2$oi̯-u-*

Modelling Language History III



Modelling Language History IV

By annotating word formation in a machine-readable manner, we will ultimately be able to compare different hypotheses of the language history and calculate their probability.

Summary

The computer-assisted approach can help linguists to

- collaborate,
- handle big data,
- test models and theories, and
- integrate traditional and modern methods and insights with each other.

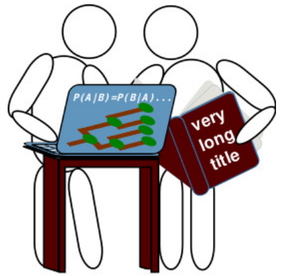
The tools we introduced were

Home | News | Resources | Publications | Talks | Tutorials | Events | People

Welcome to the CALC Project

The ERC-funded research project *CALC* (Computer-Assisted Language Comparison, see here for the official research proposal) establishes a computer-assisted framework for historical linguistics. We pursue an interdisciplinary approach that adapts methods from computer science and bioinformatics for the use in historical linguistics. While purely computational approaches are common today, the project focuses on the communication between classical and computational linguists, developing interfaces that allow historical linguists to produce their data in machine readable formats while at the same time presenting the results of computational analyses in a transparent and human-readable way.

[READ MORE]



Last updated on 2019-07-31.

This website by Johann-Mattis List is licensed under a Creative Commons Attribution 4.0 International License.

IMPRINT

Thank you for your attention!

CALC members:

- Dr. Johann-Mattis List (Group leader)
- Dr. Yunfan Lai (Post-Doc)
- Dr. Tiago Tresoldi (Post-Doc)
- Mei-Shin Wu (Doctorate student)
- Nathanael E. Schweikhard (Doctorate student)

Contact: http://calc.digling.org/