

N. E. Schweikhard

Max Planck Institute for the Science of Human History
Department of Linguistic and Cultural Evolution
CALC Project

June 28th, 2019



MAX-PLANCK-GESELLSCHAFT



Example of an Annotated Wordlist

ID	Language_ID	Parameter_ID	Value	Form	Segments	Source
Tibetan_Old_Tibetan-1741-1	Tibetan_Old_Tibetan	1741	<i>steŋ</i>	<i>steŋ</i>	<i>s t e ŋ</i>	Huang1992
rGyalrong_Japhug-1741-1	rGyalrong_Japhug	1741	<i>w-taʁ</i>	<i>w-taʁ</i>	<i>w + t a ʁ</i>	Jacques2015b
Tibetan_Old_Tibetan-98-1	Tibetan_Old_Tibetan	98	<i>t^hams.tɕad</i>	<i>t^hams.tɕad</i>	<i>t^h a m s + tɕ a d</i>	Huang1992
Kiranti_Khaling-98-1	Kiranti_Khaling	98	<i>k^høle</i>	<i>k^høle</i>	<i>k^h ø l e</i>	Jacques2017FN
Kiranti_Limbu-98-1	Kiranti_Limbu	98	<i>kak</i>	<i>kak</i>	<i>k a k</i>	Jacques2017FN
rGyalrong_Japhug-98-1	rGyalrong_Japhug	98	<i>%t^hamtɕrt</i>	<i>%t^hamtɕrt</i>	<i>t^h a m tɕ r t</i>	Jacques2015b
Tangut-98-1	Tangut	98	<i>zji¹</i>	<i>zji¹</i>	<i>z j j i¹</i>	Li1997
Tibetan_Old_Tibetan-1292-1	Tibetan_Old_Tibetan	1292	<i>ŋan</i>	<i>ŋan</i>	<i>ŋ a n</i>	Huang1992
rGyalrong_Japhug-1292-1	rGyalrong_Japhug	1292	<i>%ŋyn</i>	<i>%ŋyn</i>	<i>ŋ r n</i>	Jacques2015b
Tibetan_Old_Tibetan-1422-1	Tibetan_Old_Tibetan	1422	<i>gson + po</i>	<i>gson + po</i>	<i>g s o n + p o</i>	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	<i>Tibetan_Old_Tibetan</i>	1741	<i>steŋ</i>	<i>steŋ</i>	<i>s t e ŋ</i>	Huang1992
rGyalrong_Japhug-1741-1	<i>rGyalrong_Japhug</i>	1741	<i>w-taʁ</i>	<i>w-taʁ</i>	<i>w + t a ʁ</i>	Jacques2015b
Tibetan_Old_Tibetan-98-1	<i>Tibetan_Old_Tibetan</i>	98	<i>t^hams.tɕad</i>	<i>t^hams.tɕad</i>	<i>t^h a m s + tɕ a d</i>	Huang1992
Kiranti_Khaling-98-1	<i>Kiranti_Khaling</i>	98	<i>k^høle</i>	<i>k^høle</i>	<i>k^h ø l e</i>	Jacques2017FN
Kiranti_Limbu-98-1	<i>Kiranti_Limbu</i>	98	<i>kak</i>	<i>kak</i>	<i>k a k</i>	Jacques2017FN
rGyalrong_Japhug-98-1	<i>rGyalrong_Japhug</i>	98	<i>%t^hamtɕɹt</i>	<i>%t^hamtɕɹt</i>	<i>t^h a m tɕ ɹ t</i>	Jacques2015b
Tangut-98-1	<i>Tangut</i>	98	<i>zji¹</i>	<i>zji¹</i>	<i>z j i¹</i>	Li1997
Tibetan_Old_Tibetan-1292-1	<i>Tibetan_Old_Tibetan</i>	1292	<i>ŋan</i>	<i>ŋan</i>	<i>ŋ a n</i>	Huang1992
rGyalrong_Japhug-1292-1	<i>rGyalrong_Japhug</i>	1292	<i>%ŋɣn</i>	<i>%ŋɣn</i>	<i>ŋ ɣ n</i>	Jacques2015b
Tibetan_Old_Tibetan-1422-1	<i>Tibetan_Old_Tibetan</i>	1422	<i>gson + po</i>	<i>gson + po</i>	<i>g s o n + p o</i>	Huang1992

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

Glottolog
Languages
Families
Language Search
References
Reference Search
About

Family: Indo-European

Glottocode:

Classification

- Indo-European (583)
 - Albanian (4)
 - Arbëreshe Albanian
 - Calabrian Albanian
 - Campo Marino Albanian
 - Central Mountain Albanian
 - Sicilian Albanian
 - Arvanitika Albanian
 - Northern Tosk Albanian
 - Gheg Albanian
- Anatolian (10)
- Armenic (3)
- Balto-Slavic (23)
- Celtic (16)
- Dacian
- Germanic (105)
- Graeco-Phrygian (10)
- Indo-Iranian (320)
- Italic (86)
- Lusitanian
- Messapic
- Thracian
- Tskharian (?)

Map

Leaflet | © OpenStreetMap contributors

Links

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

Cross-Links to Reference Catalogs: Concepticon

ID	Language_ID	<i>Parameter_ID</i>	Value	Form	Segments	Source
Tibetan_Old_Tibetan-1741-1	Tibetan_Old_Tibetan	<i>1741</i>	<i>steŋ</i>	<i>steŋ</i>	<i>s t e ŋ</i>	Huang1992
rGyalrong_Japhug-1741-1	rGyalrong_Japhug	<i>1741</i>	<i>w-taʁ</i>	<i>w-taʁ</i>	<i>w + t a ʁ</i>	Jacques2015b
Tibetan_Old_Tibetan-98-1	Tibetan_Old_Tibetan	<i>98</i>	<i>t^hams.tɕad</i>	<i>t^hams.tɕad</i>	<i>t^h a m s + tɕ a d</i>	Huang1992
Kiranti_Khaling-98-1	Kiranti_Khaling	<i>98</i>	<i>k^høle</i>	<i>k^høle</i>	<i>k^h ø l e</i>	Jacques2017FN
Kiranti_Limbu-98-1	Kiranti_Limbu	<i>98</i>	<i>kak</i>	<i>kak</i>	<i>k a k</i>	Jacques2017FN
rGyalrong_Japhug-98-1	rGyalrong_Japhug	<i>98</i>	<i>%t^hamtɕrt</i>	<i>%t^hamtɕrt</i>	<i>t^h a m tɕ r t</i>	Jacques2015b
Tangut-98-1	Tangut	<i>98</i>	<i>zji¹</i>	<i>zji¹</i>	<i>z j i¹</i>	Li1997
Tibetan_Old_Tibetan-1292-1	Tibetan_Old_Tibetan	<i>1292</i>	<i>ŋan</i>	<i>ŋan</i>	<i>ŋ a n</i>	Huang1992
rGyalrong_Japhug-1292-1	rGyalrong_Japhug	<i>1292</i>	<i>%ŋɣn</i>	<i>%ŋɣn</i>	<i>ŋ r n</i>	Jacques2015b
Tibetan_Old_Tibetan-1422-1	Tibetan_Old_Tibetan	<i>1422</i>	<i>gson + po</i>	<i>gson + po</i>	<i>g s o n + p o</i>	Huang1992

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



Concept set BARKING

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

Showing 1 to 11 of 11 entries

← Previous

1

Next →



Id	Concept in source	Conceptlist
<input type="text" value="Search"/>	<input type="text" value="Search"/>	<input type="text" value="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-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 gø [danish]; blaffen [dutch]; bark [english]; haukkua [finnish]; bellen [german]; γαυγιζει [greek]; linbo'ax [hebrew]; ugat [hungarian]; gelta [icelandic]; (ag) tafann [irish]; abbaiare [italian]; loti [lithuanian];	Luniewska 2016 299

Metadata

MRC Psycholinguistic Database

KUCERA FRANCIS FREQUENCY 2

MRC WORD BARKING

Mapping to OmegaWiki

OMEGAWIKI ID 5444

Edinburgh Associative Thesaurus

DEGREE 23

EAT WORD BARKING

WEIGHTED DEGREE 105.00

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	<i>steŋ</i>	<i>steŋ</i>	<i>s t e ŋ</i>	Huang1992
rGyalrong_Japhug-1741-1	rGyalrong_Japhug	1741	<i>w-taʁ</i>	<i>w-taʁ</i>	<i>w + t a ʁ</i>	Jacques2015b
Tibetan_Old_Tibetan-98-1	Tibetan_Old_Tibetan	98	<i>t^hams.tɕad</i>	<i>t^hams.tɕad</i>	<i>t^h a m s + tɕ a d</i>	Huang1992
Kiranti_Khaling-98-1	Kiranti_Khaling	98	<i>k^høle</i>	<i>k^høle</i>	<i>k^h ø l e</i>	Jacques2017FN
Kiranti_Limbu-98-1	Kiranti_Limbu	98	<i>kak</i>	<i>kak</i>	<i>k a k</i>	Jacques2017FN
rGyalrong_Japhug-98-1	rGyalrong_Japhug	98	<i>%t^hamtɕrt</i>	<i>%t^hamtɕrt</i>	<i>t^h a m tɕ r t</i>	Jacques2015b
Tangut-98-1	Tangut	98	<i>zji¹</i>	<i>zji¹</i>	<i>z j j i¹</i>	Li1997
Tibetan_Old_Tibetan-1292-1	Tibetan_Old_Tibetan	1292	<i>ŋan</i>	<i>ŋan</i>	<i>ŋ a n</i>	Huang1992
rGyalrong_Japhug-1292-1	rGyalrong_Japhug	1292	<i>%ŋɣn</i>	<i>%ŋɣn</i>	<i>ŋ ɣ n</i>	Jacques2015b
Tibetan_Old_Tibetan-1422-1	Tibetan_Old_Tibetan	1422	<i>gson + po</i>	<i>gson + po</i>	<i>g s o n + p o</i>	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	<i>ʃ p ɪ n ə</i>	<i>ʃ p ɪ n + ə</i>	SPIN _e-suff	1 2
341	German	spider web	Spinnwebe	<i>ʃ p ɪ n v eː b ə</i>	<i>ʃ p ɪ n + v eː b + ə</i>	SPIN WEAVE _e-suff	1 3 2
342	German	spider web	Spinnennetz	<i>ʃ p ɪ n ə n n ɛ t s</i>	<i>ʃ p ɪ n + ə n + n ɛ t s</i>	SPIN _en-fuge NET	1 4 5
753	German	spin	spinnen	<i>ʃ p ɪ n ə n</i>	<i>ʃ p ɪ n + ə n</i>	SPIN _inf	1 6

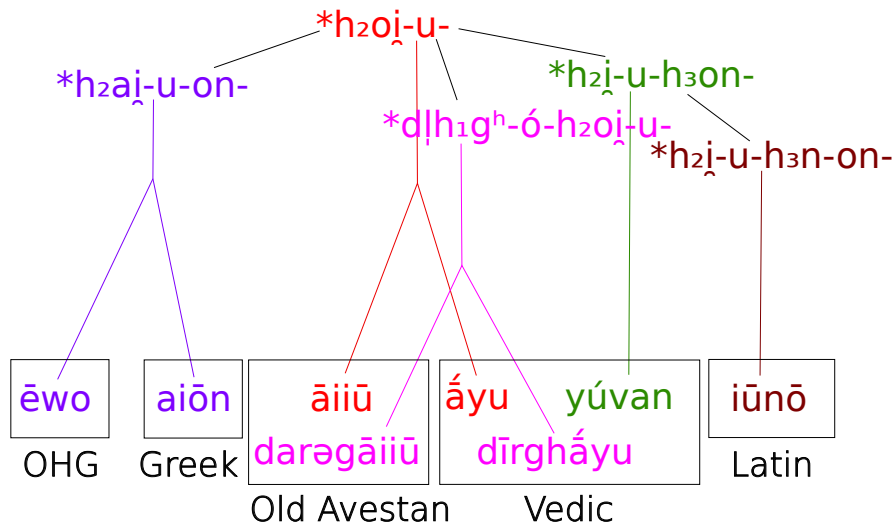
Data based on the Intercontinental Dictionary Series (Key and Comrie 2016)

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

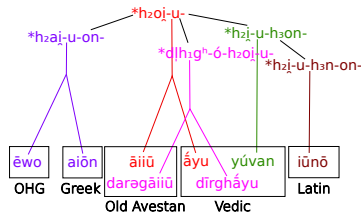


A family tree of **h₂ei-u-* (based on Wodtko et al. 2008 and Mallory/Adams 2006)

Annotation of Word Formation Processes I

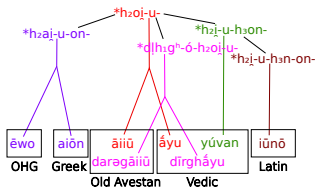
ID	LANGUAGE	CONCEPT	FORM	MORPHEMES	COGNATES	ROOTS
1	Old High German	eternity	ēwo	ēw o	1 2	1 2
2	Ancient Greek	life	aiōn	ai ōn	1 2	1 2
3	Old Avestan	life	āiū	āiū	3	1
4	Old Avestan	long-living	darəgāiū	darəg a āiū	4 5 3	3 4 1
5	Vedic	life	āyu	āyu	3	1
6	Vedic	long-living	dīrghāyu	dīrgh ā āyu	4 5 3	3 4 1
7	Vedic	young	yūvan	yūv an	6 7	1 5
8	Latin	(deity name)	iūnō	iū n ō	6 8 2	1 5 2
9	Indo-European	life	*h ₂ aj-u-on-	h ₂ aju on	3 2	1 2
10	Indo-European	life	*h ₂ oj-u-	h ₂ ojū	1	1
11	Indo-European	long-living	*dlh ₁ g ^h -ó-h ₂ oj-u-	dlh ₁ g ^h ó h ₂ ojū	4 5 1	3 4 1
12	Indo-European	young	*h ₂ j-u-h ₃ on-	h ₂ jū h ₃ on	6 7	1 5
13	Indo-European	the young one	*h ₂ j-u-h ₃ n-on-	h ₂ jū h ₃ n on	6 8 2	1 5 2

Source	Source-ID	Target	Target-ID	Change
*h ₂ aj-u-on-	1	aiōn	2	sound change
*h ₂ oj-u-	3	*h ₂ aj-u-on-	1	e-grade, on-suffix
*h ₂ oj-u-	3	*dlh ₁ g ^h -ó-h ₂ oj-u-	4	compound with *dlh ₁ g ^h -ó-
*dlh ₁ g ^h -ó-h ₂ oj-u-	7	dīrghāyu	8	sound change
...



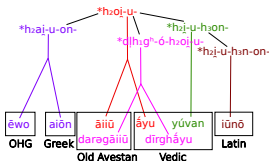
Annotation of Word Formation Processes II

ID	LANGUAGE	CONCEPT	FORM	MORPHEMES	COGNATES	ROOTS
1	Old High German	eternity	<i>ēwo</i>	<i>ēw o</i>	1 2	1 2
2	Ancient Greek	life	<i>aiōn</i>	<i>ai ōn</i>	1 2	1 2
3	Old Avestan	life	<i>āiiū</i>	<i>āiiū</i>	3	1
4	Old Avestan	long-living	<i>darəgāiiū</i>	<i>darəg a āiiū</i>	4 5 3	3 4 1
5	Vedic	life	<i>āyu</i>	<i>āyu</i>	3	1
6	Vedic	long-living	<i>dirghāyu</i>	<i>dirgh á āyu</i>	4 5 3	3 4 1
7	Vedic	young	<i>yúvan</i>	<i>yúv an</i>	6 7	1 5
8	Latin	(deity name)	<i>iūnō</i>	<i>iū n ō</i>	6 8 2	1 5 2
9	Indo-European	life	<i>*h₂ai-u-on-</i>	<i>h₂aiu on</i>	3 2	1 2
10	Indo-European	life	<i>*h₂oi-u-</i>	<i>h₂oiu</i>	1	1
11	Indo-European	long-living	<i>*dlh₁g^h-ó-h₂oi-u-</i>	<i>dlh₁g^h ó h₂oiu</i>	4 5 1	3 4 1
12	Indo-European	young	<i>*h₂i-u-h₃on-</i>	<i>h₂iu h₃on</i>	6 7	1 5
13	Indo-European	the young one	<i>*h₂i-u-h₃n-on-</i>	<i>h₂iu h₃n on</i>	6 8 2	1 5 2



Source	Source-ID	Target	Target-ID	Change
<i>*h₂ai-u-on-</i>	1	<i>aiōn</i>	2	sound change
<i>*h₂oi-u-</i>	3	<i>*h₂ai-u-on-</i>	1	e-grade, on-suffix
<i>*h₂oi-u-</i>	3	<i>*dlh₁g^h-ó-h₂oi-u-</i>	4	compound with <i>*dlh₁g^h-ó-</i>
<i>*dlh₁g^h-ó-h₂oi-u-</i>	7	<i>dirghāyu</i>	8	sound change
...

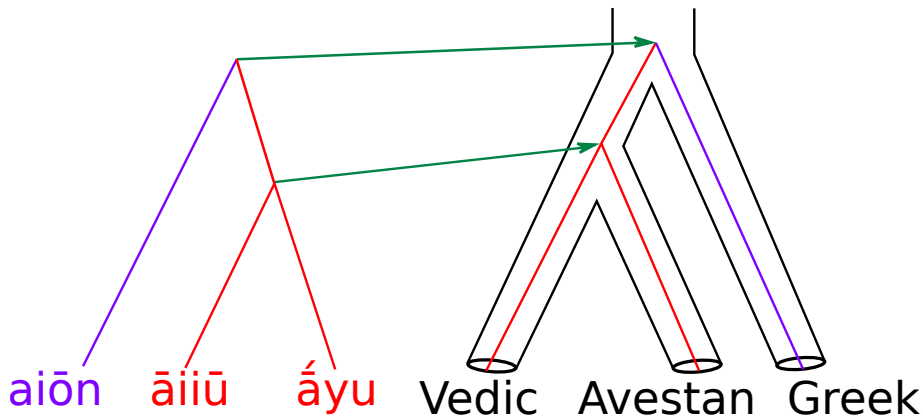
Annotation of Word Formation Processes III



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	āiū	āiū	3	1
4	Old Avestan	long-living	darəgāiū	darəg a āiū	4 5 3	3 4 1
5	Vedic	life	āyū	āyū	3	1
6	Vedic	long-living	dirghāyū	dirgh ā āyū	4 5 3	3 4 1
7	Vedic	young	yūvan	yūv an	6 7	1 5
8	Latin	(deity name)	iūnō	iū n ō	6 8 2	1 5 2
9	Indo-European	life	*h ₂ aj-u-on-	h ₂ ajū on	3 2	1 2
10	Indo-European	life	*h ₂ oj-u-	h ₂ ojū	1	1
11	Indo-European	long-living	*d̪h ₁ gʰ-ó-h ₂ oj-u-	d̪h ₁ gʰ ó h ₂ ojū	4 5 1	3 4 1
12	Indo-European	young	*h ₂ j-u-h ₃ on-	h ₂ jū h ₃ on	6 7	1 5
13	Indo-European	the young one	*h ₂ j-u-h ₃ n-on-	h ₂ jū h ₃ n on	6 8 2	1 5 2

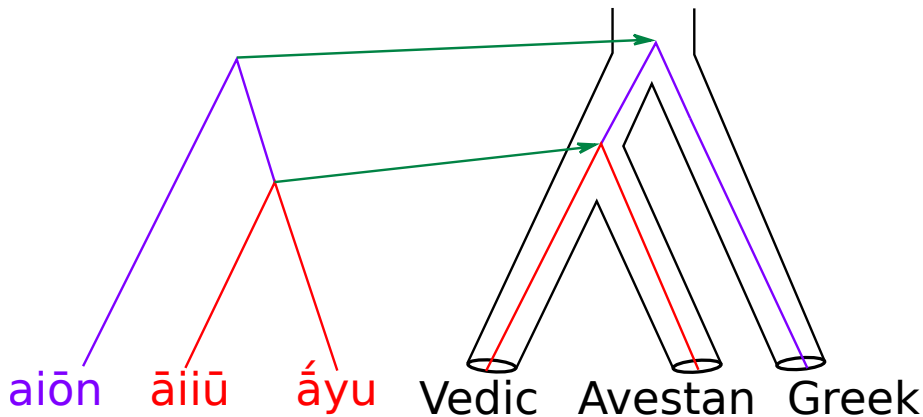
Source	Source-ID	Target	Target-ID	Change
*h ₂ aj-u-on-	1	aiōn	2	sound change
*h ₂ oj-u-	3	*h ₂ aj-u-on-	1	e-grade, on-suffix
*h ₂ oj-u-	3	*d̪h ₁ gʰ-ó-h ₂ oj-u-	4	compound with *d̪h ₁ gʰ-ó-
*d̪h ₁ gʰ-ó-h ₂ oj-u-	7	dirghāyū	8	sound change
...

Modelling Language History I



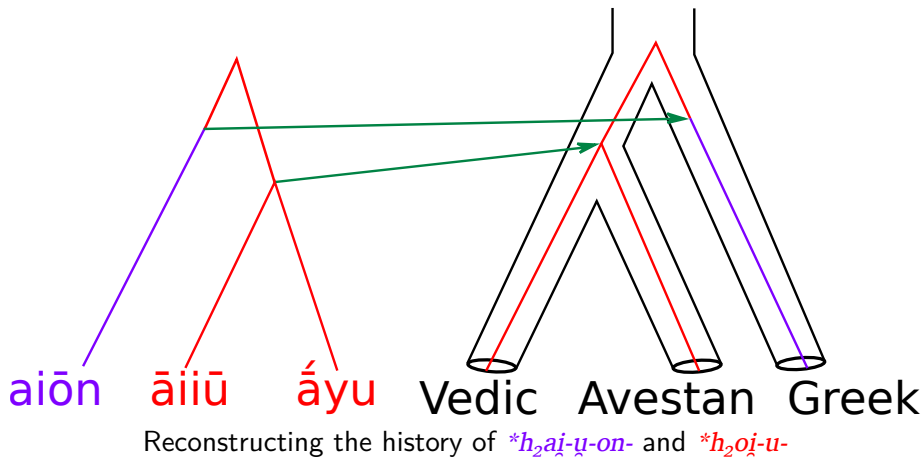
Reconstructing the history of **h₂ai̯-u-on-* and **h₂oi̯-u-*

Modelling Language History II



Reconstructing the history of **h₂ai̯-u-on-* and **h₂oi̯-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.

Thank you for your attention!

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

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)

