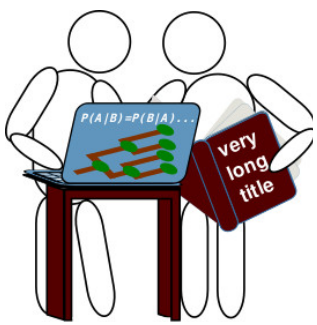
N. E. Schweikhard

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

June 28th, 2019



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	*zji*^{1}	*zji*^{1}	*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

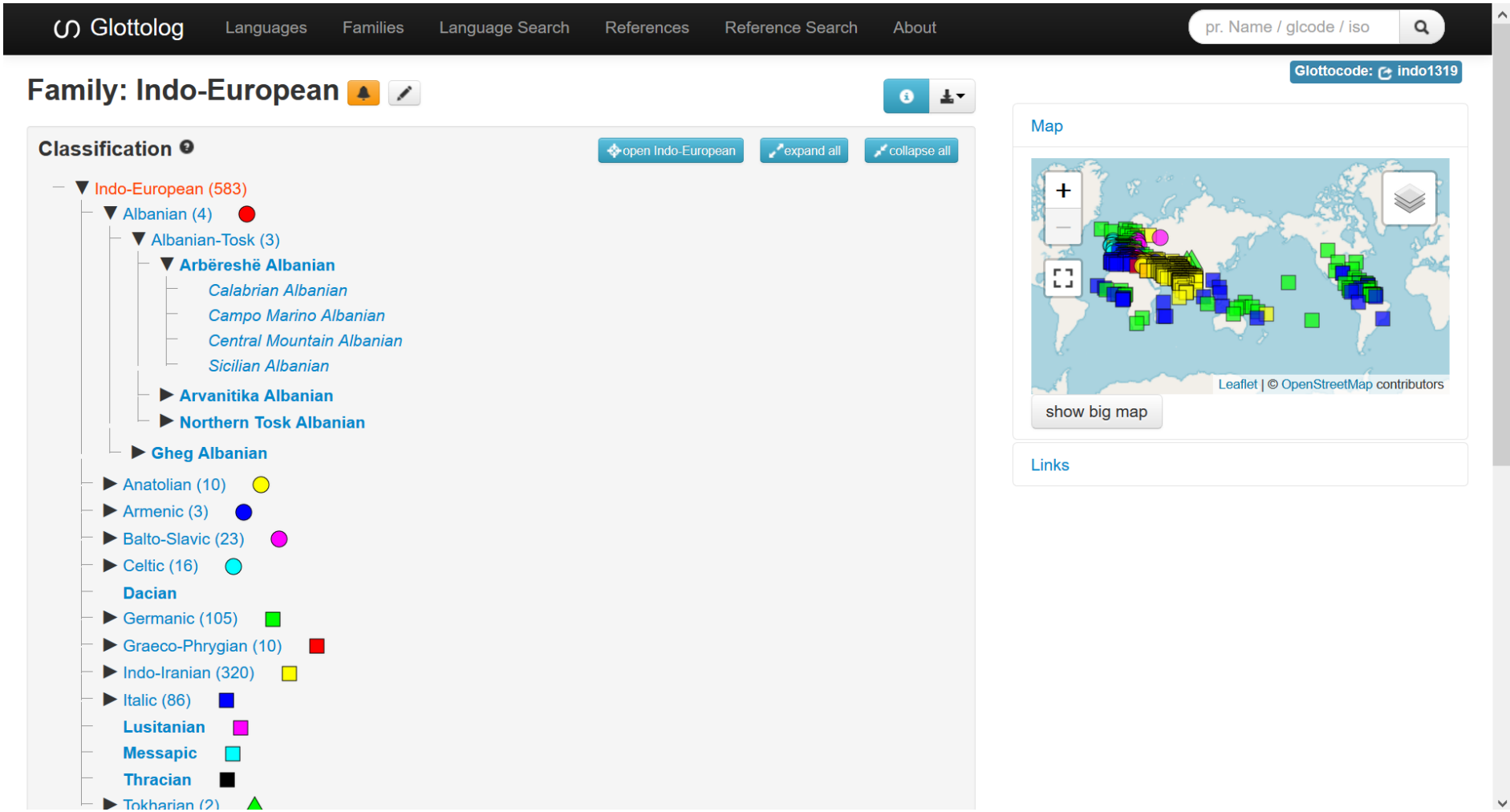
Excerpt from Sino-Tibetan Database of Lexical Cognates (Sagart et al. 2019).

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	*zji*[1]	*zji*[1]	*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



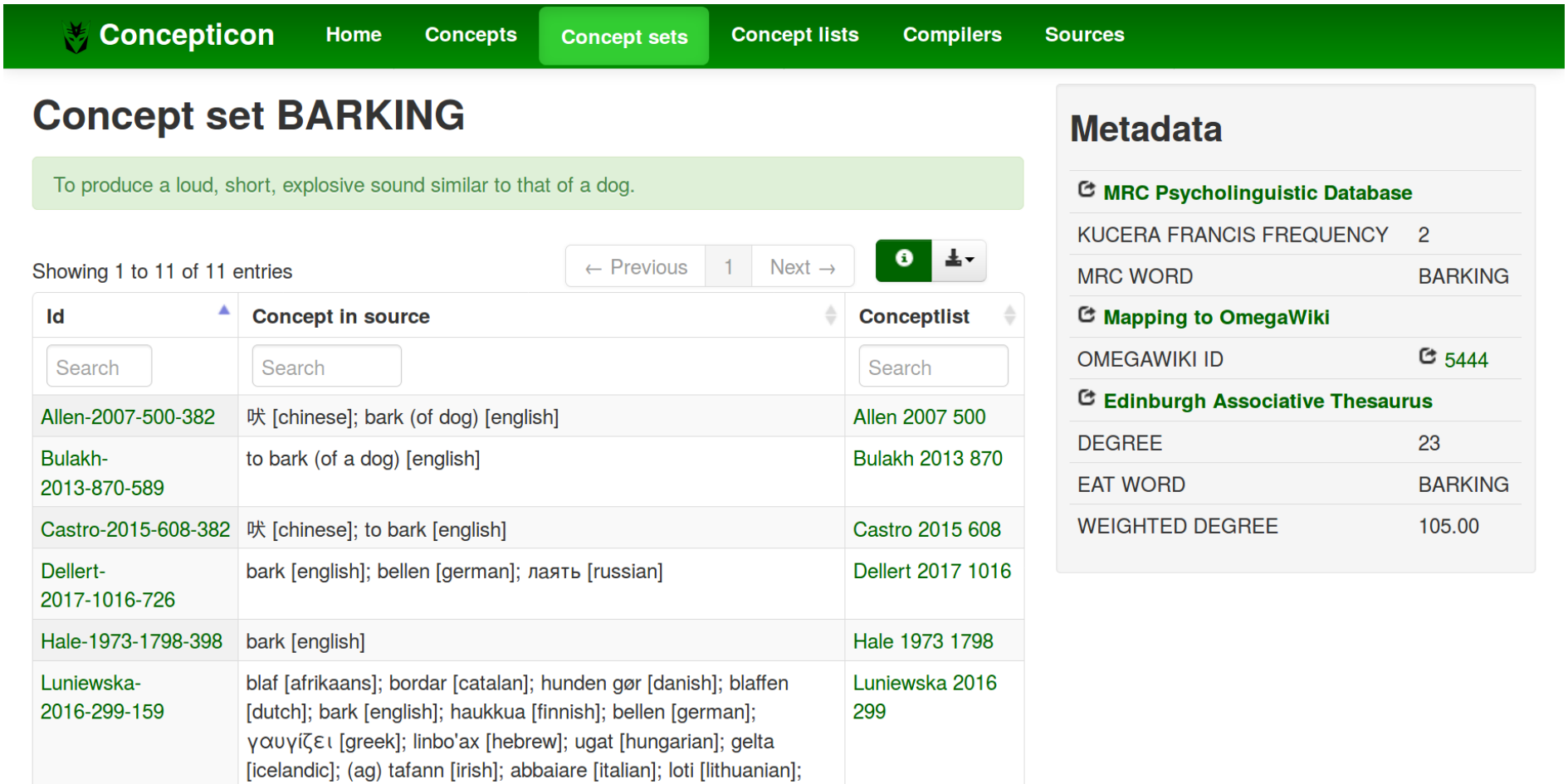
Glottolog, a reference database of languages and their genealogical relations (Hammarström et al. 2019).

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-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).

Concepticon



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-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	*zji*[1]	*zji*[1]	*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).

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 ts	ʃ p ɪ n + ə n + n ɛ ts	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

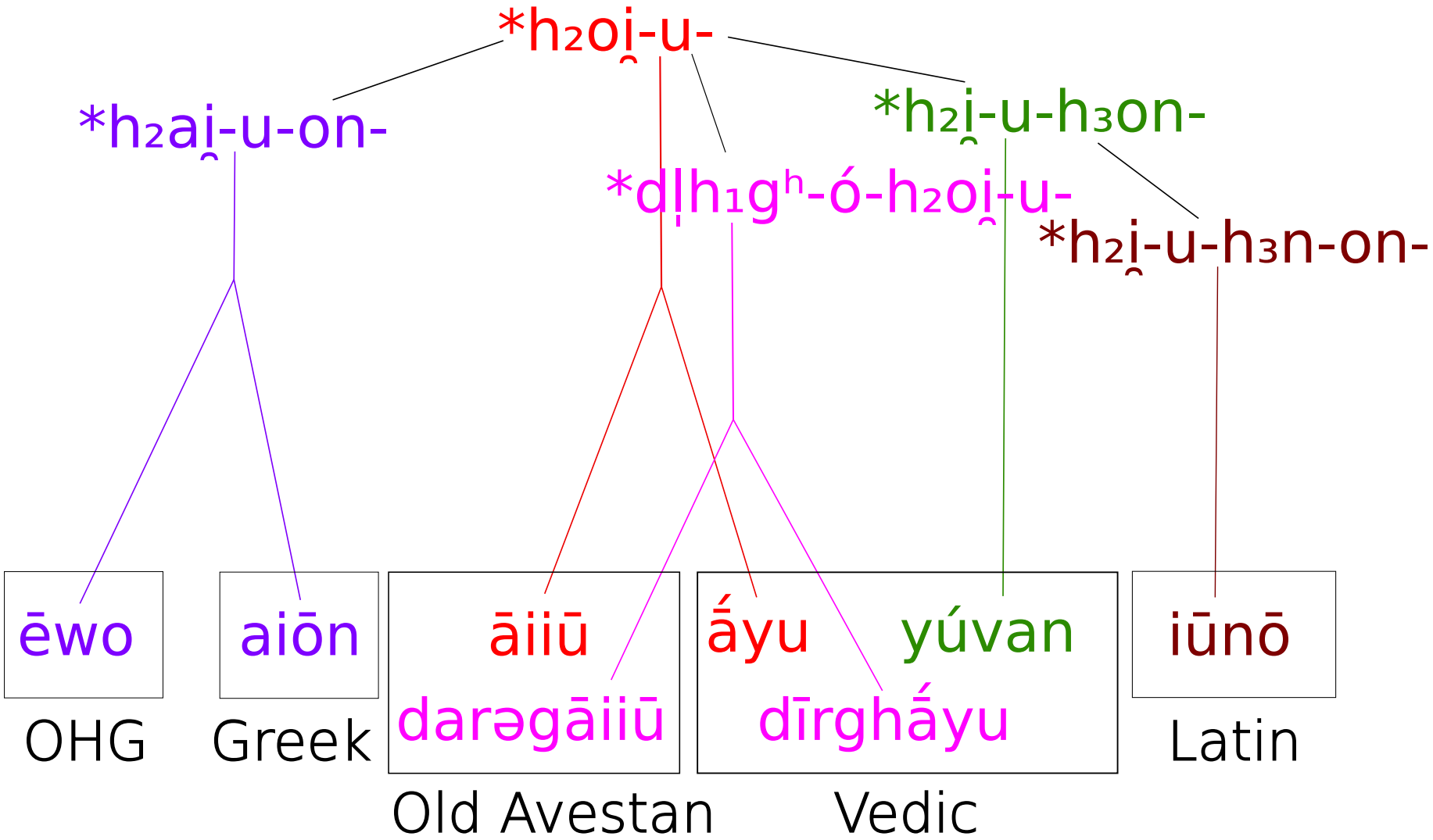
Data based on the Intercontinental Dictionary Series (Key and Comrie 2016)

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

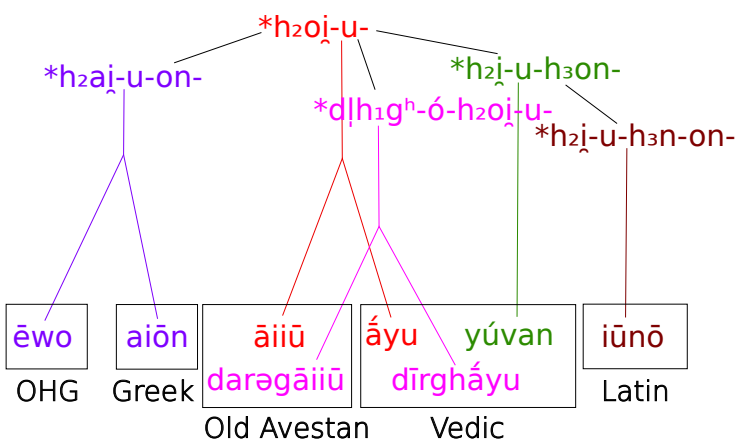


A family tree of *h_2ei-u- (based on Wodtko et al. 2008 and Mallory/Adams 2006)

Annotation of Word Formation Processes I

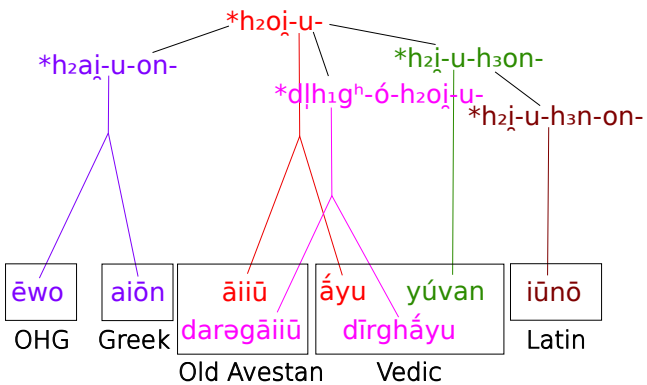
ID	LANGUAGE	CONCEPT	FORM	MORPHEMES	COGNATES	ROOTS
1	Old High German	eternity	*ēwo*	*ēw o*	1 2	1 2
2	Ancient Greek	life	*aiōn*	*ai ōn*	1 2	1 2
3	Old Avestan	life	*āiiū*	*āiiū*	3	1
4	Old Avestan	long-living	*darəgāiiū*	*darəg a āiiū*	4 5 3	3 4 1
5	Vedic	life	*áyu*	*áyu*	3	1
6	Vedic	long-living	*dīrghā́yu*	*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	**h₂aḭ-u-on-*	*h₂aḭu on*	3 2	1 2
10	Indo-European	life	**h₂oḭ-u-*	*h₂oḭu*	1	1
11	Indo-European	long-living	**dḷh₁gʰ-ó-h₂oḭ-u-*	*dḷh₁gʰ ó h₂oḭu*	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
**h₂aḭ-u-on-*	1	*aiōn*	2	sound change
**h₂oḭ-u-*	3	**h₂aḭ-u-on-*	1	e-grade, on-suffix
**h₂oḭ-u-*	3	**dḷh₁gʰ-ó-h₂oḭ-u-*	4	compound with **dḷh₁gʰ-ó-*
**dḷh₁gʰ-ó-h₂oḭ-u-*	7	*dīrghā́yu*	8	sound change
...	...	...	...	...



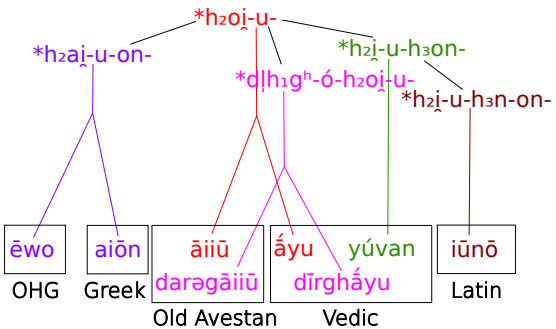
Annotation of Word Formation Processes II

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	*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
...	...	...	...	...

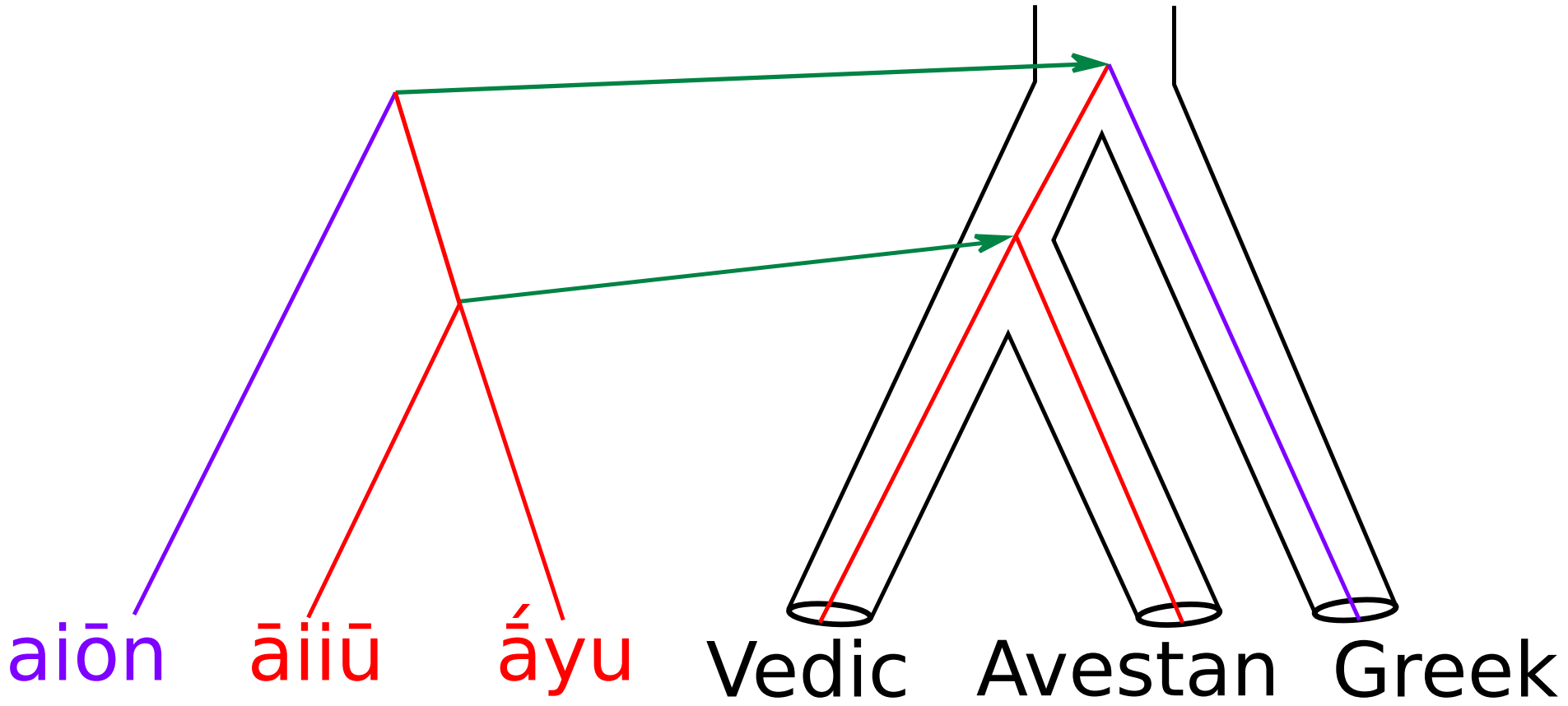
Annotation of Word Formation Processes III



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	*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

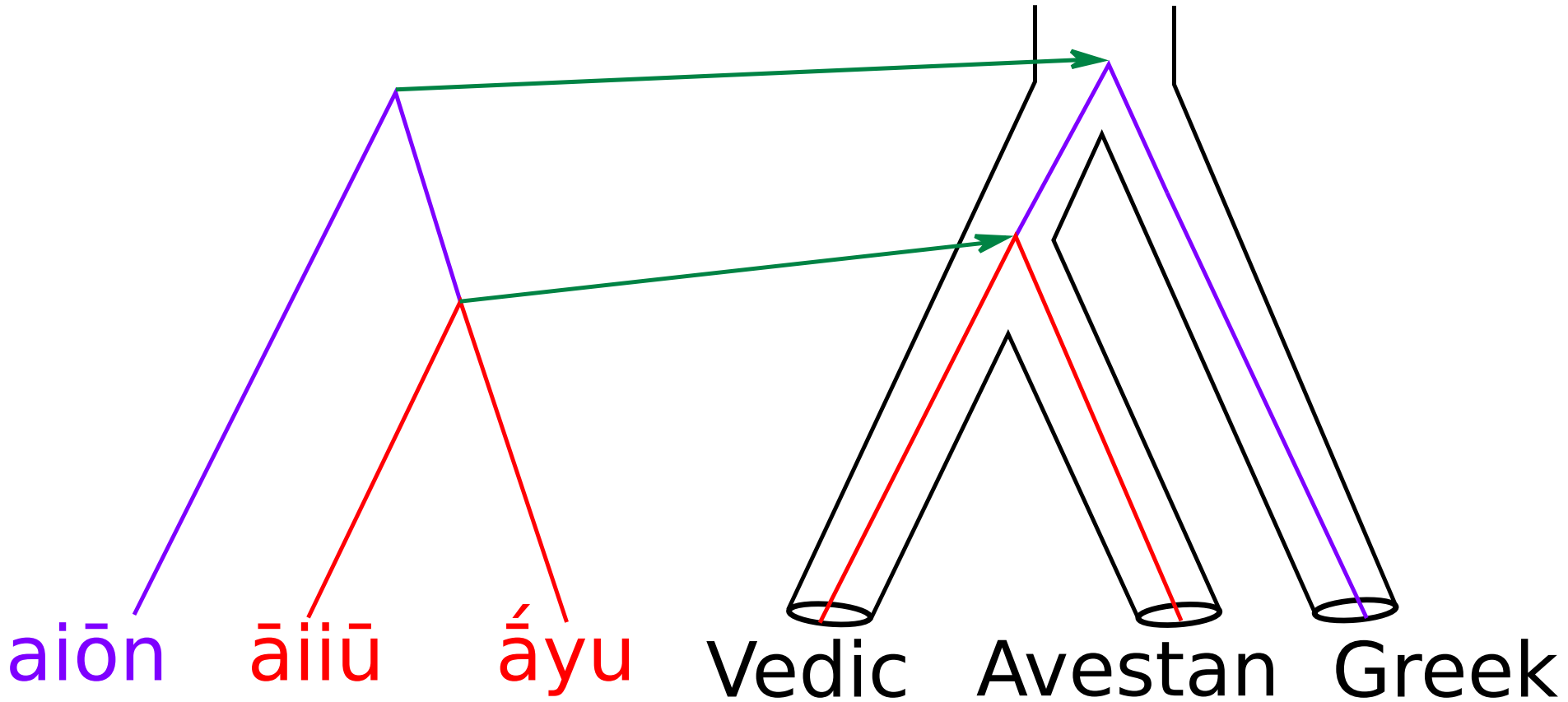
Source	Source-ID	Target	Target-ID	Change
**h₂ai̯-u-on-*	1	*aiōn*	2	sound change
**h₂oi̯-u-*	3	**h₂ai̯-u-on-*	1	e-grade, on-suffix
**h₂oi̯-u-*	3	**dl̥h₁gʰ-ó-h₂oi̯-u-*	4	compound with **dl̥h₁gʰ-ó-*
**dl̥h₁gʰ-ó-h₂oi̯-u-*	7	*dīrghấyu*	8	sound change
...	...	...	...	...

Modelling Language History I



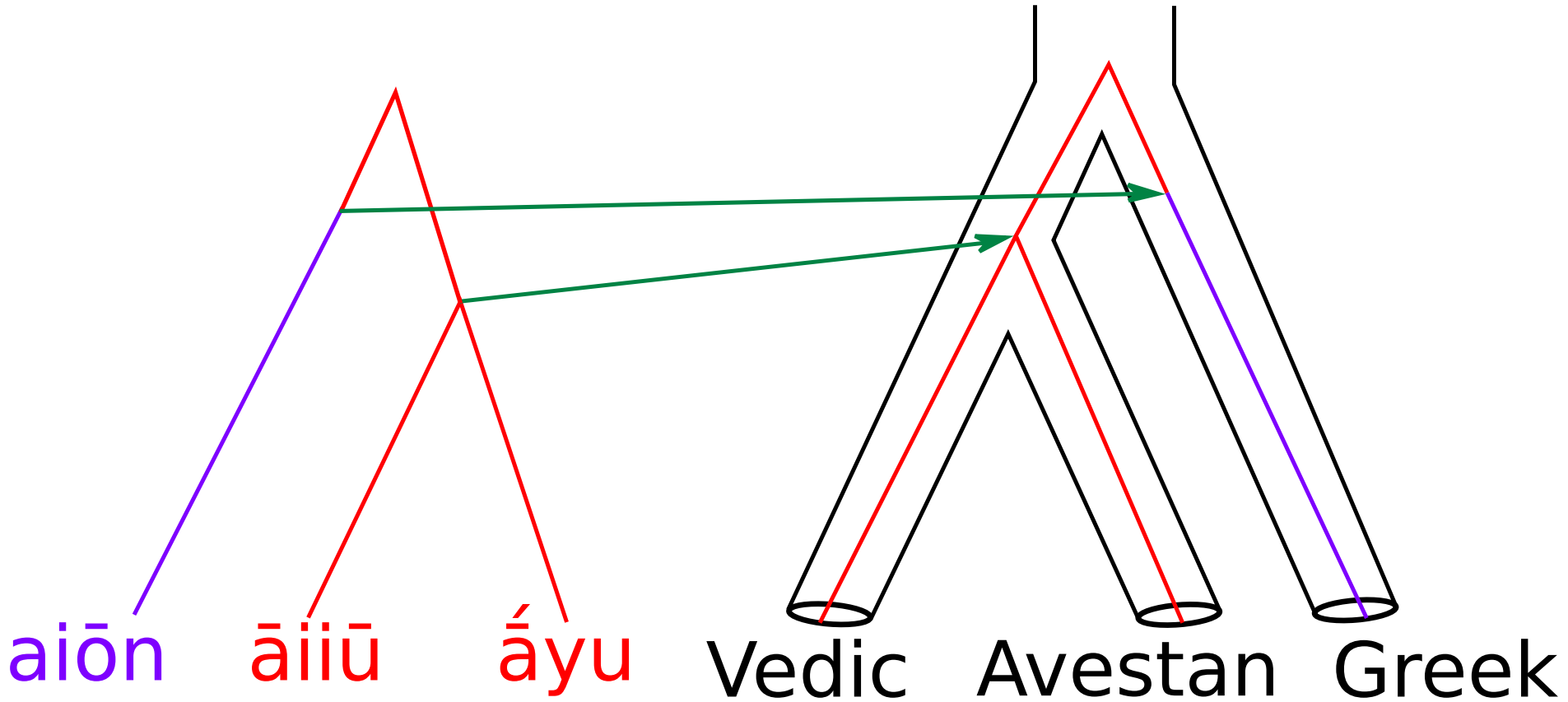
Reconstructing the history of *$h_2ai̯$-$u̯$-*on*- and *$h_2oi̯$-*u*-

Modelling Language History II



Reconstructing the history of *$h_2ai̯$-$u̯$-*on*- and *$h_2oi̯$-*u*-

Modelling Language History III



Reconstructing the history of *$h_2ai̯$-$u̯$-on- and *$h_2oi̯$-u-

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)