Assisted Reconstruction

Lydia Steiner & Michael Cysouw
University of Leipzig, Ludwig Maximilian University München

Our approach

Our approach

- Computer-assisted language reconstruction
  - Beyond the ‘Black Box’ approach

Our approach

- Computer-assisted language reconstruction
  - Beyond the ‘Black Box’ approach
- Mimic historical-comparative methodology
  - Though using brute computer force

Our approach

- Computer-assisted language reconstruction
  - Beyond the 'Black Box' approach
- Mimic historical-comparative methodology
  - Though using brute computer force
- Produce intermediate results that (hopefully) make linguistic sense
  - cognate sets, sound correspondences, meaning shifts

‘Black box’ comparison

‘Black box’ comparison

- Produce a tree without linguistic argumentation

‘Black box’ comparison

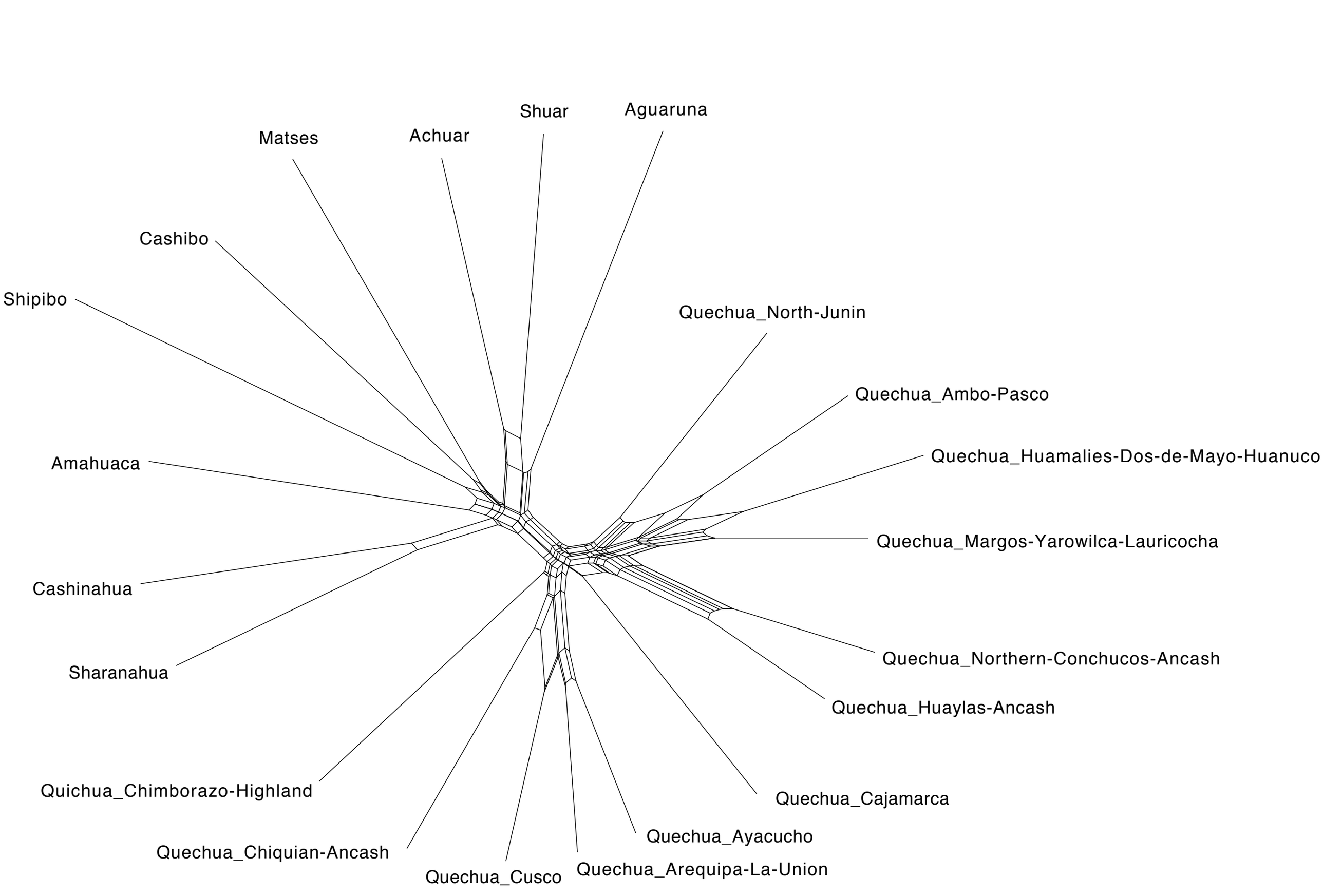
- Produce a tree without linguistic argumentation
- Based on written words
  - Possibly some form of orthographic normalization (e.g. IPA)
  - Levenshtein, n-grams, zipping

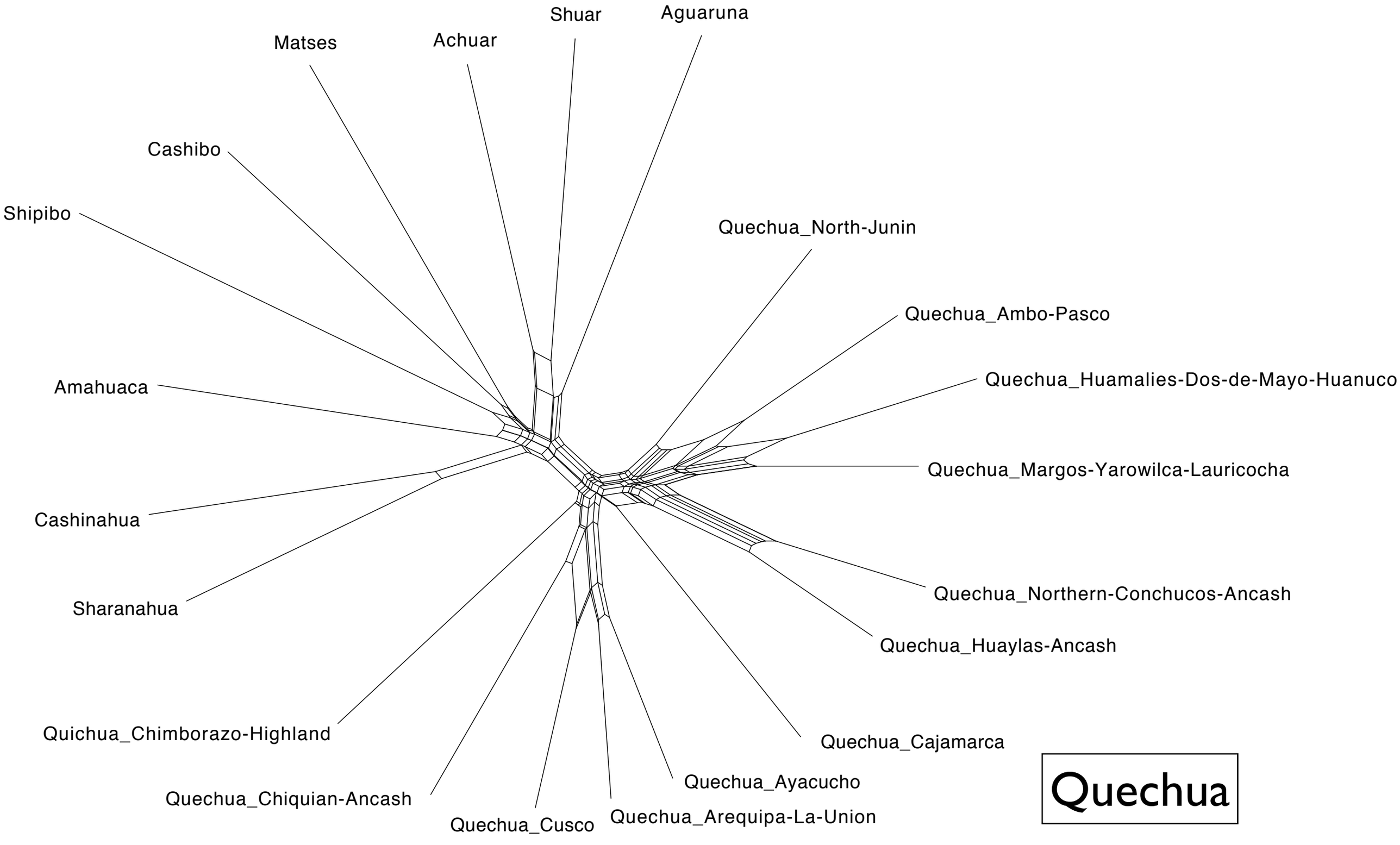
‘Black box’ comparison

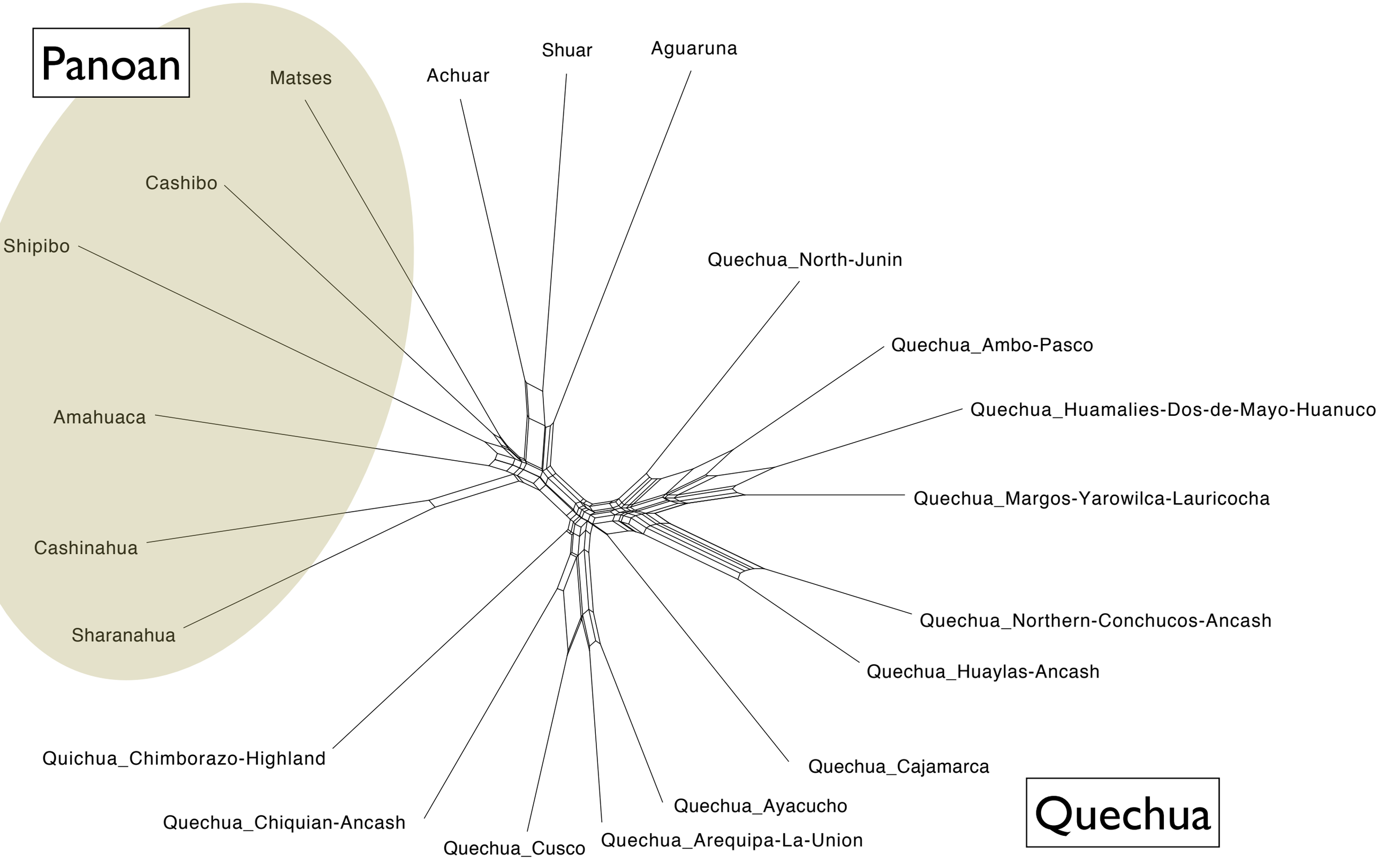
- Produce a tree without linguistic argumentation
- Based on written words
  - Possibly some form of orthographic normalization (e.g. IPA)
  - Levenshtein, n-grams, zipping
- Based on distribution of cognates over meanings
  - ‘Swadesh approach’
  - Maximum Parsimony, Maximum Likelihood, Bayesian MCMC

‘Black box’ comparison

- Produce a tree without linguistic argumentation
- Based on written words
  - Possibly some form of orthographic normalization (e.g. IPA)
  - Levenshtein, n-grams, zipping
- Based on distribution of cognates over meanings
  - ‘Swadesh approach’
  - Maximum Parsimony, Maximum Likelihood, Bayesian MCMC
- Based on typological characteristics


Matses
Achuar
Shuar
Aguaruna
Cashibo
Shipibo
Quechua_North-Junin
Quechua_Ambo-Pasco
Amahuaca
Quechua_Huamalies-Dos-de-Mayo-Huanuco
Quechua_Margos-Yarowilca-Lauricocha
Cashinahua
Quechua_Northern-Conchucos-Ancash
Sharanahua
Quechua_Huaylas-Ancash
Quichua_Chimborazo-Highland
Quechua_Cajamarca
Quechua_Ayacucho
Quechua_Chiquian-Ancash
Quechua_Arequipa-La-Union
Quechua_Cusco


Matses
Achuar
Shuar
Aguaruna
Cashibo
Shipibo
Quechua_North-Junin
Quechua_Ambo-Pasco
Amahuaca
Quechua_Huamalies-Dos-de-Mayo-Huanuco
Quechua_Margos-Yarowilca-Lauricocha
Cashinahua
Quechua_Northern-Conchucos-Ancash
Sharanahua
Quechua_Huaylas-Ancash
Quichua_Chimborazo-Highland
Quechua_Cajamarca
Quechua_Ayacucho
Quechua_Chiquian-Ancash
Quechua_Arequipa-La-Union
Quechua_Cusco
Quechua


Panoan
Matses
Achuar
Shuar
Aguaruna
Cashibo
Shipibo
Amahuaca
Cashinahua
Sharanahua
Quechua_North-Junin
Quechua_Ambo-Pasco
Quechua_Huamalies-Dos-de-Mayo-Huanuco
Quechua_Margos-Yarowilca-Lauricocha
Quechua_Northern-Conchucos-Ancash
Quechua_Huaylas-Ancash
Quechua_Cajamarca
Quechua_Ayacucho
Quechua_Arequipa-La-Union
Quechua_Cusco
Quechua_Chiquian-Ancash
Quichua_Chimborazo-Highland
Quechua

Jivaroan

Panoan

Achuar

Shuar

Aguaruna

Matses

Cashibo

Shipibo

Amahuaca

Cashinahua

Sharanahua

Quechua_North-Junin

Quechua_Ambo-Pasco

Quechua_Huamalies-Dos-de-Mayo-Huanuco

Quechua_Margos-Yarowilca-Lauricocha

Quechua_Northern-Conchucos-Ancash

Quechua_Huaylas-Ancash

Quechua_Cajamarca

Quechua_Ayacucho

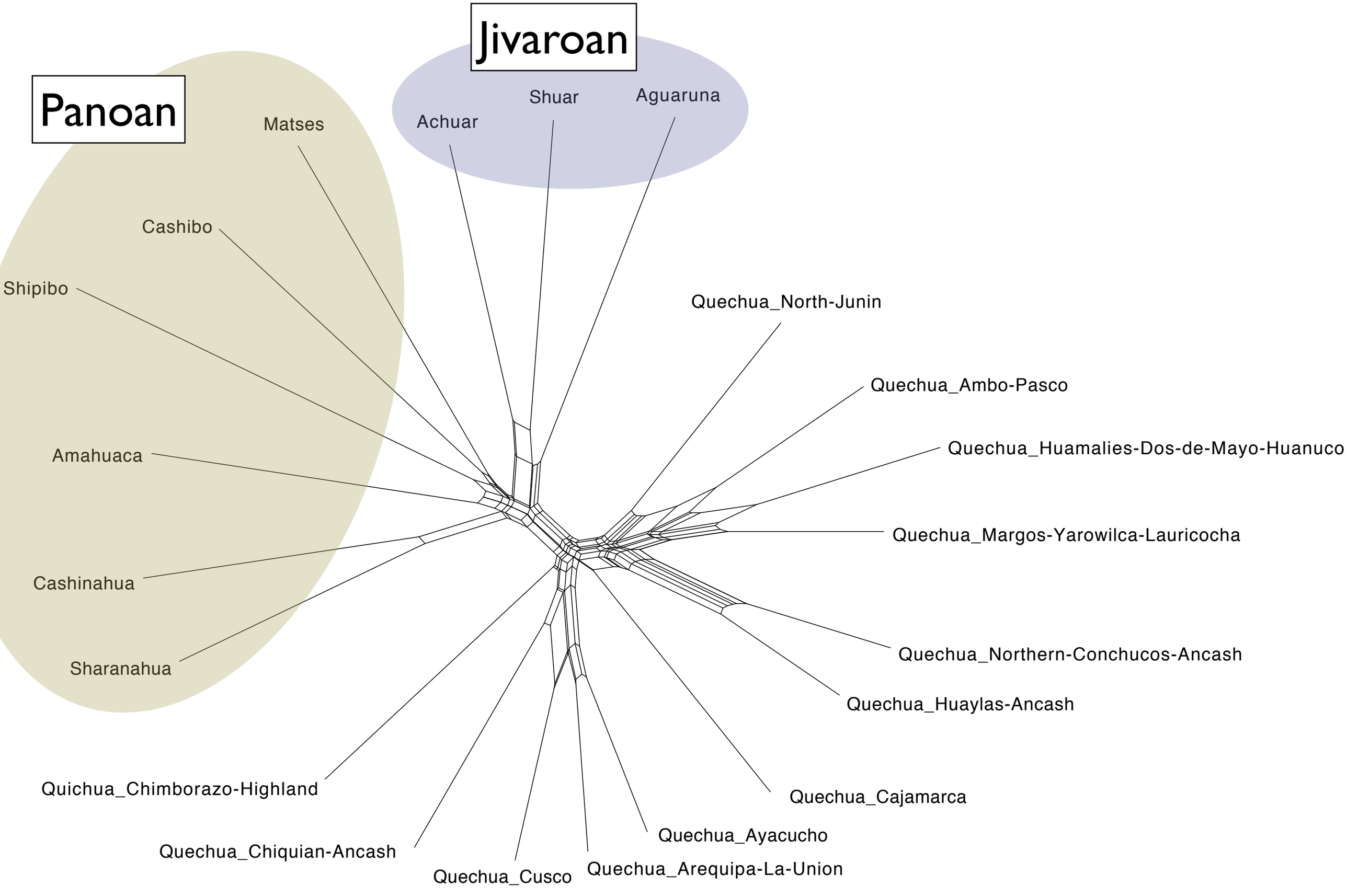
Quechua_Arequipa-La-Union

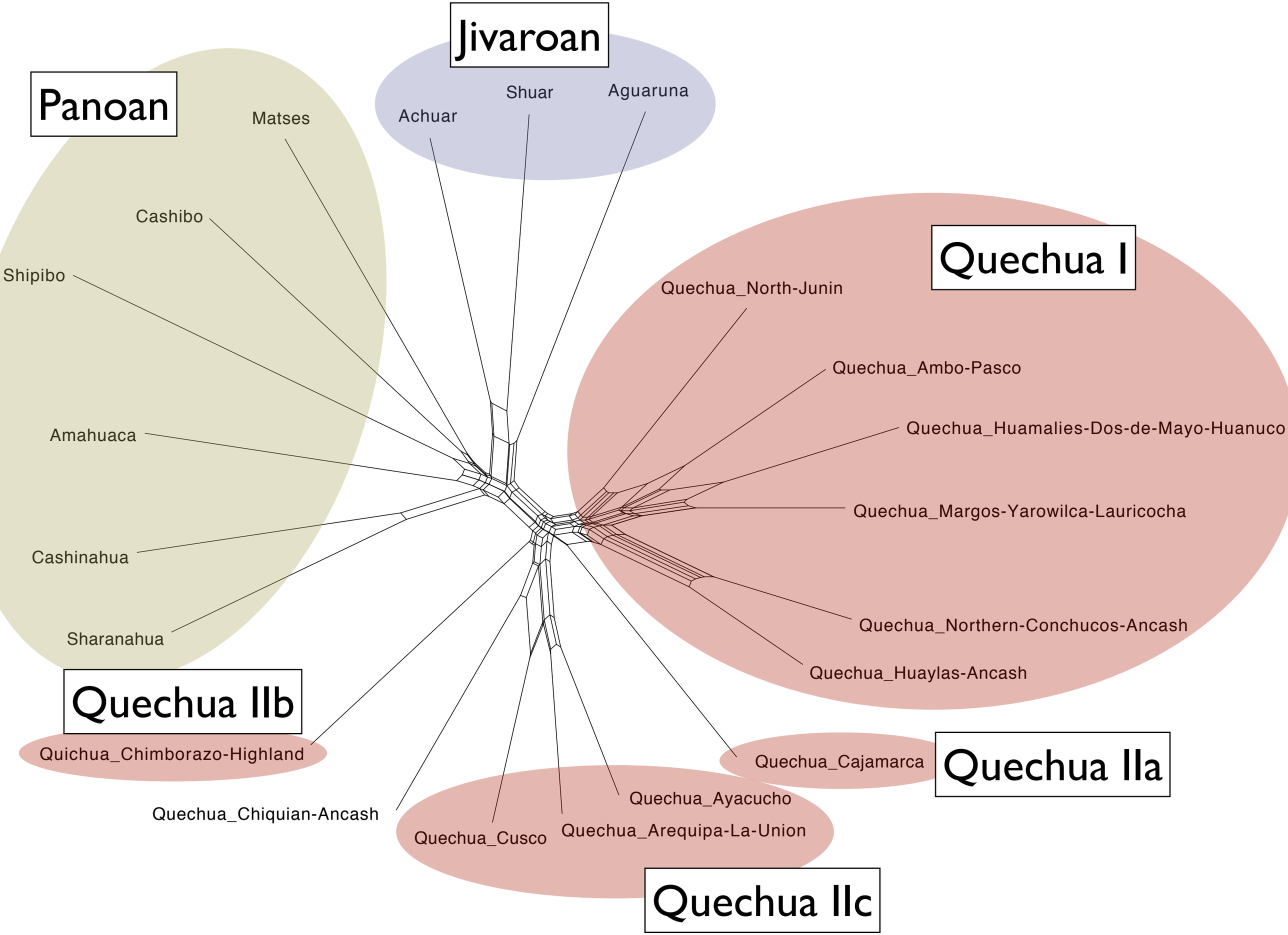
Quechua_Cusco

Quechua_Chiquian-Ancash

Quichua_Chimborazo-Highland

Quechua


Jivaroan
Panoan
Achuar
Shuar
Aguaruna
Matses
Cashibo
Shipibo
Amahuaca
Cashinahua
Sharanahua
Quechua_North-Junin
Quechua_Ambo-Pasco
Quechua_Huamalies-Dos-de-Mayo-Huanuco
Quechua_Margos-Yarowilca-Lauricocha
Quechua_Northern-Conchucos-Ancash
Quechua_Huaylas-Ancash
Quechua_Cajamarca
Quechua_Ayacucho
Quechua_Arequipa-La-Union
Quechua_Cusco
Quechua_Chiquian-Ancash
Quichua_Chimborazo-Highland


Jivaroan
Panoan
Achuar
Shuar
Aguaruna
Matses
Cashibo
Shipibo
Amahuaca
Cashinahua
Sharanahua
Quechua I
Quechua_North-Junin
Quechua_Ambo-Pasco
Quechua_Huamalies-Dos-de-Mayo-Huanuco
Quechua_Margos-Yarowilca-Lauricocha
Quechua_Northern-Conchucos-Ancash
Quechua_Huaylas-Ancash
Quechua IIb
Quichua_Chimborazo-Highland
Quechua_Chiquian-Ancash
Quechua IIa
Quechua_Cajamarca
Quechua_Ayacucho
Quechua_Arequipa-La-Union
Quechua_Cusco
Quechua IIc

IDS Main Page
Simple Browsing
Advanced Browsing
Download Data
Technical Background

The Intercontinental Dictionary Series

Founding Editor:
Mary Ritchie Key (University of California, Irvine)
General Editor:
Bernard Comrie (Max Planck Institute for Evolutionary Anthropology, Leipzig)

Purpose: The purpose of the IDS is to establish a database where lexical material across the continents is organized in such a way tha comparisons can be made. Historical studies, comparative, and theoretical linguistic research can be based on this documentation. This is a long term cooperative project that will go on for the next generation or so and will involve linguists all over the world. It is aimed towards international understanding and cooperation. This is a pioneering effort that will have global impact. The purpose also contributes to preserving information on the little-known and "non-prestigious" languages of the world, many of which are becoming extinct.

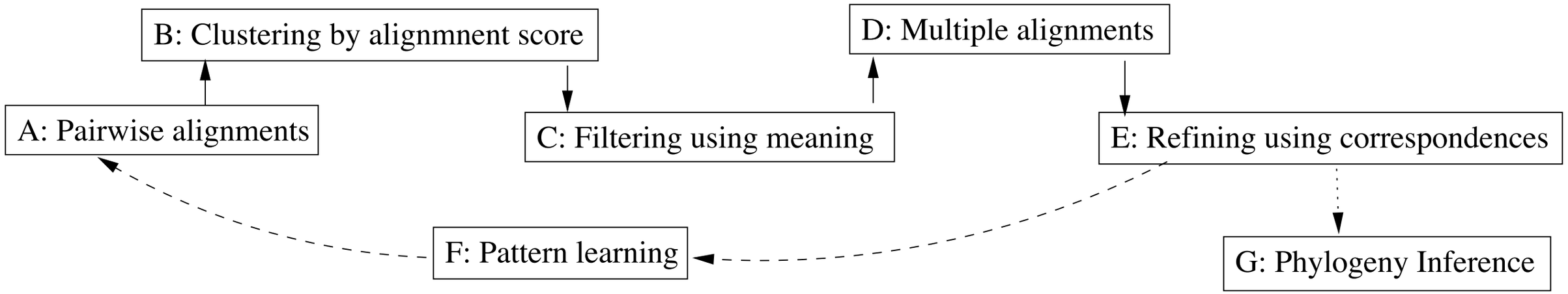
Rationale: Information on languages of the world is scattered over all the continents and islands and published in dozens of languages and scripts. There is need of a database where one can find comparable material to formulate hypotheses and test and validate those theories. For example, theories on intercontinental connections have been proposed on the basis of the distribution of 'sweet potato' and yet there is no single source, where words with this meaning can be found in many languages. Good quantitative and statistical studies are almost impossible to do now in non-Western languages. The IDS will provide a quantitative base for a scientific approach to language analysis and comparisons. The IDS will provide the research tools necessary for expanding studies such as phonological theory, word formation, language change, lexical distribution, symbolism and onomatopoeia, classification, and other ideas that have to do with history of people and migrations. The IDS will serve not only as a synonym dictionary but as an index to meaning and to cultures of various people around the earth.

Plan of Series: The IDS series may appear as 1) a volume with 25 or more languages recorded; 2) a fascicle with 5 to 10 languages recorded 3) in single WordLists, which are archived until enough are gathered to make up a fascicle or volume. A list of fascicles and volumes in progress is available from the general editor.

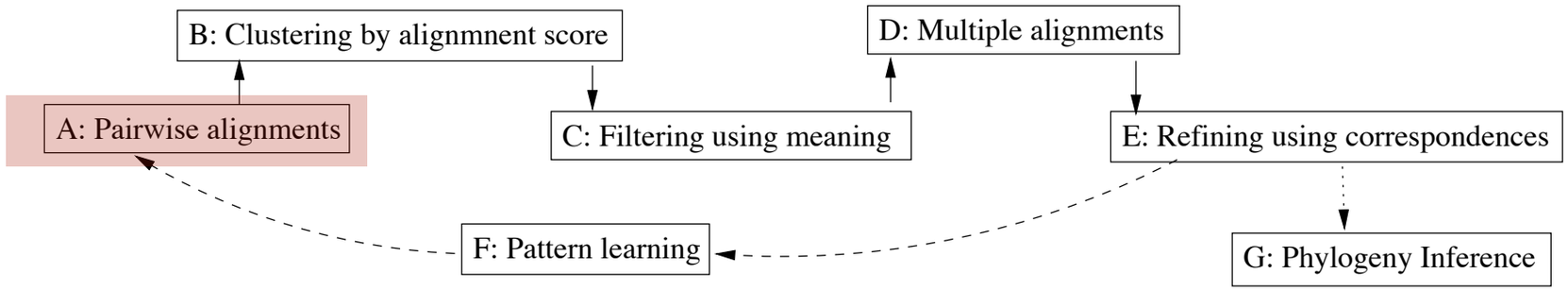
Procedure: The IDS is developed in cooperation and complementation with other research projects. Throughout the world there are linguistic activities from establishing of databases in universities and think-tanks to publishing grammar series and literacy materials, to individual projects such as the Tibetan dictionary project. Many projects seek to make linguistic data accessible in a format that will allow generalizations to be made. The computer now gives us the potential for tying together linguistic databases. The IDS editors will continue to monitor linguistic activity around the world, both for choosing the languages for forthcoming compilations and for collaboration with other research teams.

http://lingweb.eva.mpg.de/ids/

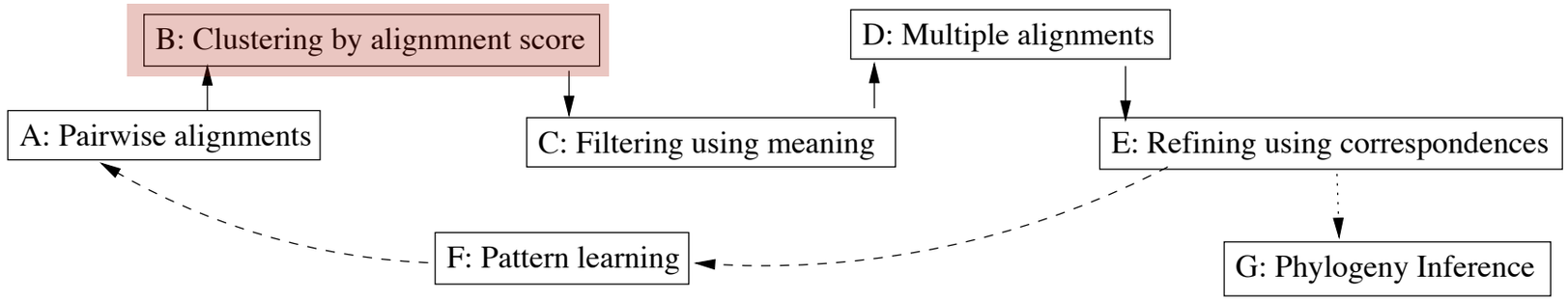
Pipeline



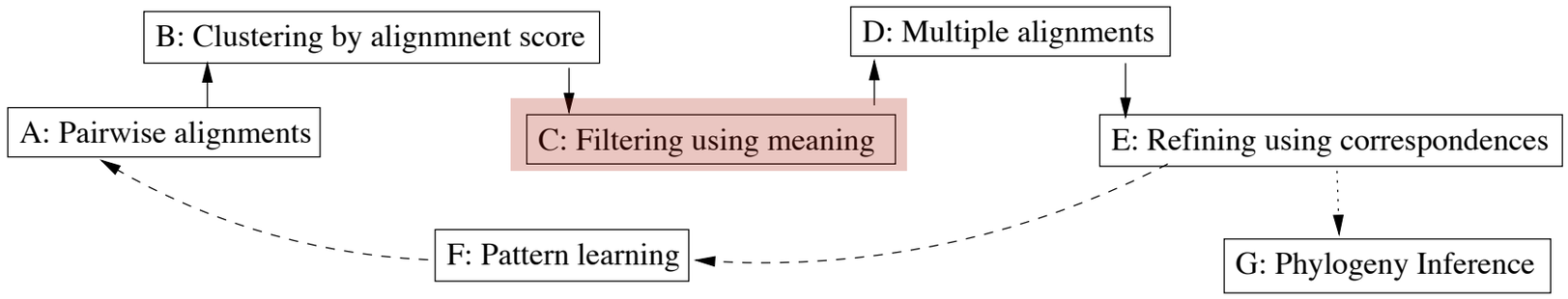
Steiner, Lydia, Peter Stadler, Michael Cysouw (2011).
A Pipeline for Computational Historical Linguistics. *Language Dynamics and Change*, 1(1).



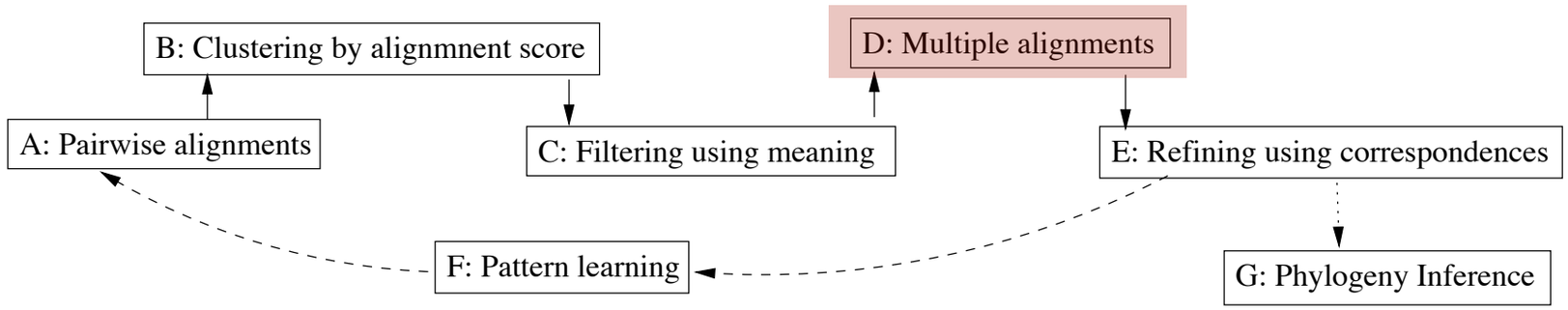
- For all words (across all meanings) compute pairwise similarities
  - basically Levenshtein distance
    counting the number of changes
    necessary to get from one word to the next



- Cluster similar words into groups
  - These groups are purely based on similarity in form

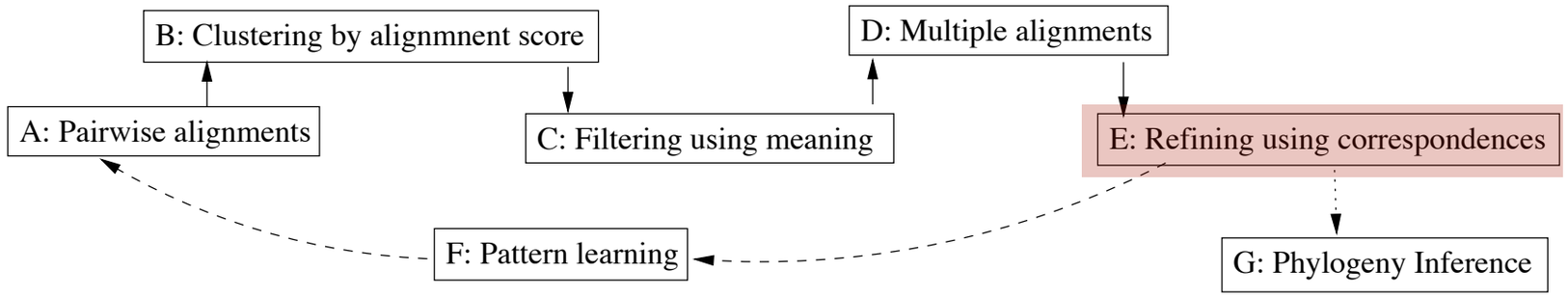


- Quantify similarities in meaning
  - Basically, we use the finding that
    on average over many languages
    similar forms indicate similar meanings
- Split clusters until the meanings are sufficiently similar
- Results in hypotheses of cognate sets



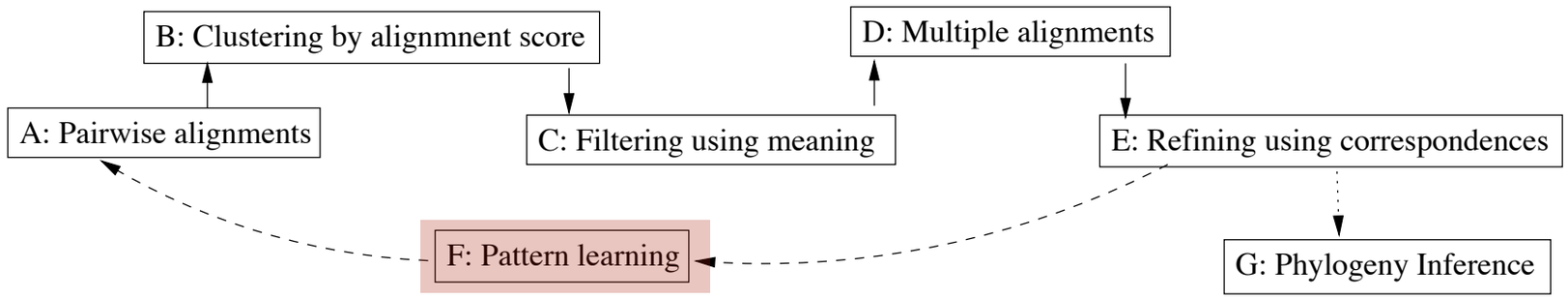
- Align sounds within cognate sets
  - This is technically a difficult problem (NP complete)
  - Fortunately, words are relatively short, so the problem is solvable

Language	IDS	meaning	alignment										
Pilagá	15.810/15.820	heavy/light	d	e	s	a	l	i					
Toba	15.810	heavy	d	e	s	a	l^y	i					
Mocoví	15.810/15.820	heavy/light	r	e	s	a	l^y	i					
Pilagá	9.440	build	n	ʔ	o	ʁ	o	–	s	e	g	e	m
Toba	9.440	build	n	ʔ	o	ʁ	o	o	š	i	g	e	m
Mocoví	9.440	build	n	o	ʔ	ʁ	o	n	š	i	g	i	m

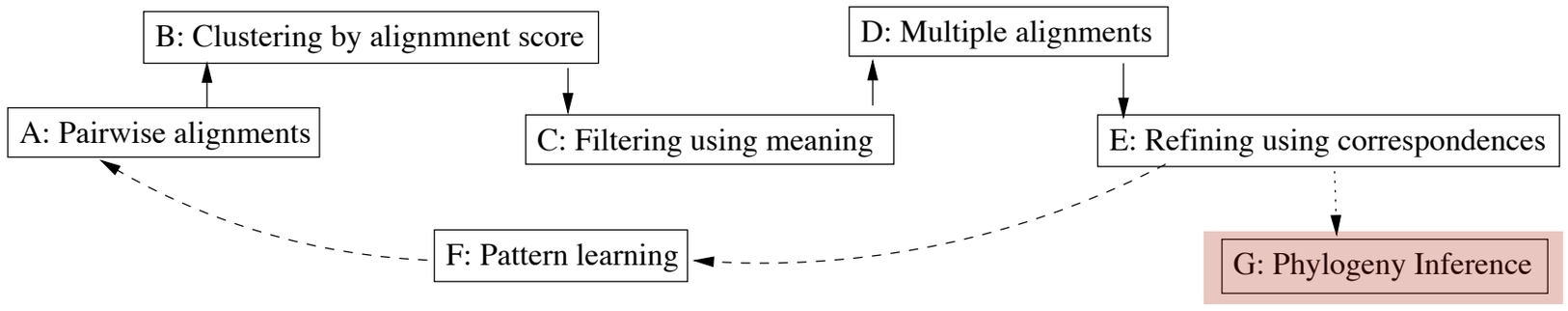


- Columns in alignment are correspondences
  - Recurrent correspondences are considered ‘better’
  - Words that are only in a cognate set with incidentally occurring correspondences are removed

Language	IDS	meaning	alignment										
Pilagá	15.810/15.820	heavy/light	d	e	s	a	l	i					
Toba	15.810	heavy	d	e	s	a	l^y	i					
Mocoví	15.810/15.820	heavy/light	r	e	s	a	l^y	i					
Pilagá	9.440	build	n	ʔ	o	ʁ	o	–	s	e	g	e	m
Toba	9.440	build	n	ʔ	o	ʁ	o	o	š	i	g	e	m
Mocoví	9.440	build	n	o	ʔ	ʁ	o	n	š	i	g	i	m

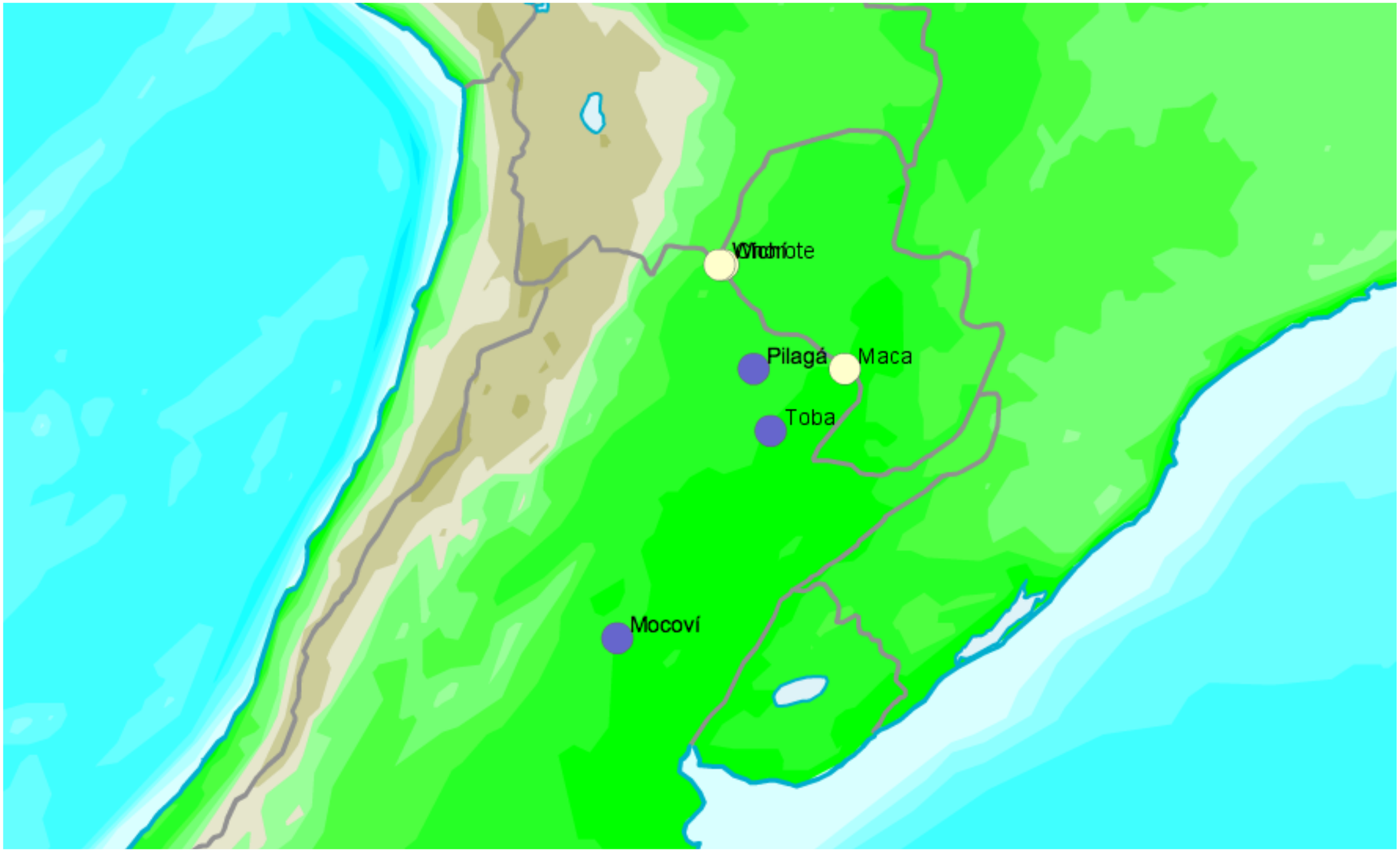


- Learn recurrent patterns of change
  - Use these to for a better starting notion of when are words similar to each other in form
- Start over with the analysis
  - After two cycles there was no significant change anymore



- Use the resulting data to build trees
  - Statistical ‘black box’ methods
  - Using distribution of cognate sets
    (i.e. which meanings do still have cognates)
  - Using distribution of sound correspondences
    (i.e. classical sound change analysis)

Is there a link between Mataco and Guiacuruan?



Language	IDS	meaning	alignment						
Nivaclé	5.124	unripe	n	i	y	i	ʔ	y	a
Pilagá	5.125	rotten	n	i	č	i	ʔ	y	a
Wichí	10.240	drip	n	i	t	o	n		
Wichí	4.591	dribble	n	i	t	u	n		
Pilagá	5.130	drink	n	i	y	o	m		
Wichí	5.370/3.655	spoon/shell	l	a	n	e	k		
Toba	5.370	spoon	l	e	m	e	k		
Pilagá	5.370	spoon	l	e	m	e	k		
Nivaclé	15.830	wet	w	a	ʔ	a	i		
Pilagá	1.329/1.320	ocean/sea	w	a	ʁ	a	i		
Wichí	7.330/10.710	chimney/road	n	o	y	i	h		
Mocoví	10.710	road	n	a	ʔ	i	k		
Nivaclé	14.332	for a long time	k	a	x	u	ʔ		
Mocoví	14.332	for a long time	k̤	a	w	a	ʔ		
Wichí	1.520/14.530	sun/clock	h^{w}	a	l	a	ʔ		
Toba	1.520	sun	n	a	l	a	ʔ		
Wichí	9.220	cut	y	i	s	e	t		
Pilagá	9.110	do/make	y	i	ʔ	e	t		
Wichí	9.210/4.760	hit/kill	i	l	o	n			
Toba	18.420	call by name	i	l	o	n			
Wichí	17.172	imitate	i	t	e	n			
Mocoví	16.510	dare	i	t	e	n			
Wichí	4.858	scar	l	a	h	ɲ	i		
Pilagá	4.374	footprint	l	i	i	ɲ	i		
Maca	3.585	hawk	m	i	y	o			
Toba	3.950	frog	m	i	y	o			
Nivaclé	19.590	prevent	f	a	ʔ	m	a	t	a	n
Mocoví	16.670	tell lies	n	a	ʔ	m	a	h	a	n
Chorote	5.123	ripe	y	o	w	e	ʔ		
Toba	5.123	ripe	y	a	m	o	k̤		

Some regular changes in Mataco

Language	IDS	meaning	alignment							
Nivaclé	8.680	tobacco	f	i	n	ɔ	k			
Maca	8.680	tobacco	f	i	n	a	k			
Nivaclé	6.310	to spin	ɔ	f	t	i	ɬ			
Maca	6.310	to spin	a	f	t	i	ɬ			
Nivaclé	8.690	to smoke	w	a	n	k	a	ɬ	ɔ	n
Maca	8.690	to smoke	w	a	n	ḳ	a	ɬ	a	n
Maca	10.613	carry-on-shoulder	t	i	ɬ	o	χ			
Wichí	10.613	carry-on-shoulder	t	i	ɬ	o	h			
Maca	9.220	cut	i	s	a	χ	i			
Wichí	9.222	chop	i	h^w	a	h	i			

Demo

Meaning change

- šapo
  - Shipibo Conibo: *light in weight*
  - Chacobo: *to weave*
  - Cashibo, Yaminahua: *cotton*
- βiško
  - Shipibo Conibo: *wound*
  - Yaminahua: *sling*

Conclusions

Conclusions

- Historical-comparative linguistics is a great method, though labour-intensive

Conclusions

- Historical-comparative linguistics is a great method, though labour-intensive
- Automated procedures can help

Conclusions

- Historical-comparative linguistics is a great method, though labour-intensive
- Automated procedures can help
- Controlling meaning change is possible, though produces still many false positives

Conclusions

- Historical-comparative linguistics is a great method, though labour-intensive
- Automated procedures can help
- Controlling meaning change is possible, though produces still many false positives
- Regular sound changes are really difficult to find: most change seems to be non-regular