Open Data in Computational Historical Linguistics: Upcycling of Chibchan wordlists

Frederic Blum

09/05/2021

Section 1

Open Data in Linguistics

My motivation

How can I find adequate data in times of a pandemic?

My motivation

How can I find adequate data in times of a pandemic?

- Problem:
  - fieldwork is somewhere between extremely difficult and impossible right now

My motivation

How can I find adequate data in times of a pandemic?

- Problem:
  - fieldwork is somewhere between extremely difficult and impossible right now
- Methodological aims:
  - training computational skills
  - open data

My motivation

How can I find adequate data in times of a pandemic?

- Problem:
  - fieldwork is somewhere between extremely difficult and impossible right now
- Methodological aims:
  - training computational skills
  - open data
- Linguistically:
  - phylolinguistic research
  - the interrelation of languages and history in Central and South America

Goals of my talk

I want to...

- motivate you to work with pre-existing data
- introduce some topics of Computational Historical Linguistics as well as important cross-linguistic formats
- show how computational methods can aid our hypothesis in questions about the history of languages

Goals of my talk

I want to...

- motivate you to work with pre-existing data
- introduce some topics of Computational Historical Linguistics as well as important cross-linguistic formats
- show how computational methods can aid our hypothesis in questions about the history of languages

Workflow:

- cldfbench (Forkel & List 2020)
- cognate assignment
- phylolinguistic analysis

Open Data in Linguistics

Data needs to be FAIR (Wilkinson *et al.* 2016):

- Findable
  - metadata
  - persistent identifier
- Accessible
  - publish data
  - publish code
- Interoperable
  - comparability
- Reusable
  - data license
  - domain-relevant standards

Cross-Linguistic Linked Data

Huge efforts have been made in recent years to improve interoperability in linguisics.

- Glottolog: languages and their genealogies (Hammarström *et al.* 2020)
- WALS: World atlas of language structures (Dryer & Haspelmath 2013)
- Concepticon: resource for linking of concepts (List *et al.* 2020)
- CLTS: Cross-Linguistic Transcription Systems (List *et al.* 2021)

Is this even necessary?

1	c	voiceless alveolar sibilant affricate consonant
2	ts	voiceless alveolar sibilant affricate consonant
3	tˢ	voiceless alveolar sibilant affricate consonant
4	t͡s	voiceless alveolar sibilant affricate consonant
5	t̪s̪	voiceless alveolar sibilant affricate consonant
6	tz	voiceless alveolar sibilant affricate consonant

Figure 1: Sounds

Is this even necessary?

1	c	voiceless alveolar sibilant affricate consonant
2	ts	voiceless alveolar sibilant affricate consonant
3	tˢ	voiceless alveolar sibilant affricate consonant
4	t͡s	voiceless alveolar sibilant affricate consonant
5	t͜s	voiceless alveolar sibilant affricate consonant
6	tz	voiceless alveolar sibilant affricate consonant

Figure 1: Sounds

Yup, it is.

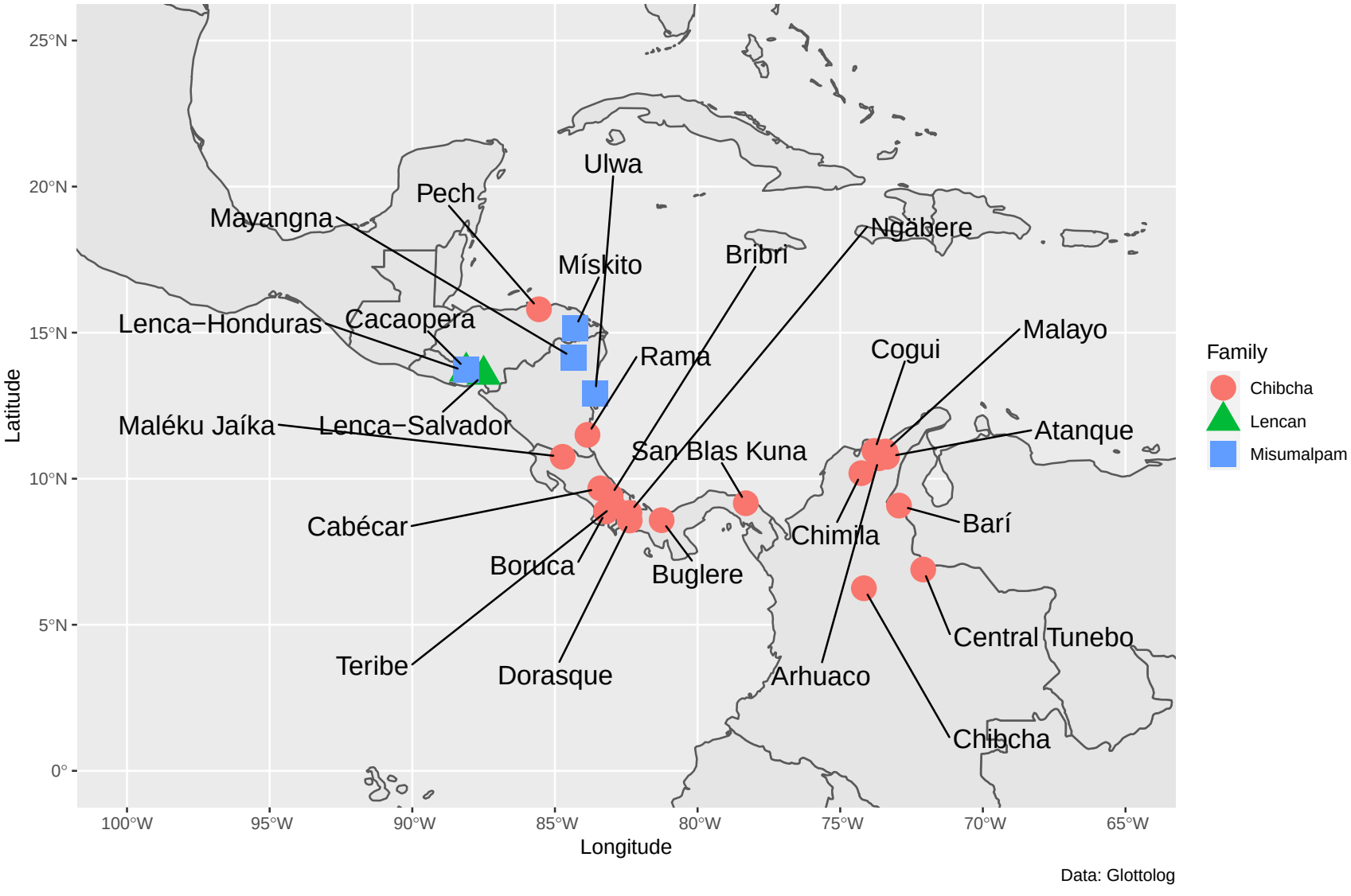
Back to open data

- publish data for others to participate in the research
- publish code to check for errors
- this helps in resolving mistakes and improves all aspects of science: analysis, understanding and replication
- together we are stronger

Section 2

Upcycling of Chibchan Data

Map of Chibchan, Misumalpam and Lencan languages



Historical Data

- The data I used comes from Constenla Umaña (2005)
- 110-item concept list
- analyzing cognates across languages

Historical Data

- The data I used comes from Constenla Umaña (2005)
- 110-item concept list
- analyzing cognates across languages

1. Agua

LS wal (a), **LH** was (a), **Ca** li (b),**Su** was (a),**Ul** was (a), **Mi** li (b), **Pa** àsò, **Ra** si: (c), **Gua** ti: (c), **Bor** díʔ (c), **Bri** dĭʔ (c), **Cab** díklú (c), **Te** dí (c), **Mo** ɲɤ (c), **Boc** tʃi (c), **Dor** *yi* (c), **Cu** ti: (c), **Co** ni (c), **Ica** dʒe (c), **Da** 'dʒɪra (c), **Chi** ditake: (c), **Mu** sie (c), **Tun** 'rɪʔa (c), **Ba** sĩ: mã (c)

2. Amarillo

LS ku, **LH** suninga, **Ca** maju, **Su** lalahni (a), **Ul** lalahka (a), **Mi** lalalhni (a), **Pa** sè: wa, **Ra** nuknukŋa, **Gua** ɬaxa:ɾa ɬu:tu iŋa:, **Bor** ʃòsát,**Bri** tsikidĭdĭ (b), **Cab** tsikidí (b), **Te** ʃòinóɾ, **Mo** ɯbɾɯɾe, **Boc** moloi, **Dor** *utká*, **Cu** kollokwa, **Co** kaʃiku'ama (c), **Ica** 'tʃəmmi (ch), **Da** kɨʃkwama (c), **Chi** tʃoŋgragwattu, **Mu** tɨβan (ch), **Tun** ta'w̃aja (ch), **Ba** kanikã: siũkdu˚.

Figure 2: Comma-seperated list of lexical items

What to do with this?

- Con's:
  - cognates are not analyzed across concepts
  - no phylolinguistic analysis possible for ID's within concepts
  - not machine readable

What to do with this?

- Con's:
  - cognates are not analyzed across concepts
  - no phylolinguistic analysis possible for ID's within concepts
  - not machine readable
- Pro's:
  - it is there! Transparent publishing allows me to do what I am doing
  - (nearly) IPA forms
  - transparent decision making on many cognate sets

First Step: Importing to CSV

- the most boring task
- copy-paste everything into a CSV-file
- automatization is difficult because the text is not machine-readable and IPA signs lack precision

First Step: Importing to CSV

- the most boring task
- copy-paste everything into a CSV-file
- automatization is difficult because the text is not machine-readable and IPA signs lack precision
- Obviously not recommended: Doing this during classes with professors you don't pay attention to anyway

Individual Cognate-ID's for each cognate set

```
raw <- read_csv("../constenlachibchan.git/raw/constenla2005

cogid_numbers <- raw %>%
  mutate(Cogid = ifelse(is.na(Cogid), ID, Cogid)) %>%
  mutate(pattern = paste(Concept, Cogid, sep = "_")) %>%
  distinct(pattern) %>%
  add_column(value = c(1:1418))

comb <- raw %>%
  mutate(CU= Cogid,Cogid = ifelse(is.na(Cogid), ID, Cogid))
  mutate(pattern = paste(Concept, Cogid, sep = "_")) %>%
  left_join(cogid_numbers) %>%
  mutate(Cogid = value) %>%
  select(-value, -pattern)

# write_csv(comb, file = "../constenlachibchan.git/raw/cons
```

CLDF-Transformation

- uniform format for lexical datasets (Forkel *et al.* 2018)
- Linking with Glottolog, Concepticon and CLTS
- running already existing workflows

Concepticon and Glottolog

Linking the languages to glottolog:

- create a csv-file with a list of languages
- adding their glottolog identifier
- adding coordinates
- this table was also used in order to create the map shown earlier

Concepticon and Glottolog

Linking the languages to glottolog:

- create a csv-file with a list of languages
- adding their glottolog identifier
- adding coordinates
- this table was also used in order to create the map shown earlier

Uploading the concept list to Concepticon:

- made easy by this blogpost
- disambiguating concepts
- semantic transparency

The final dataset

Language_ID	Parameter_ID	Segments	Cognacy	Source
LencaHonduras	33_dormir	s a	419	Constenl
Cogui	59_mano	’k/k o k a l a	739	Constenl
Cabecar	56_llorar	h ĩ	699	Constenl
Arhuaco	21_ceniza	’b/b u n z ə g a	270	Constenl
Ulwa	19_cantar	u n b a u	243	Constenl
Cacaopera	60_matar	k u l i n a	752	Constenl
MalekuJaika	48_hombre	o t a p a k a	592	Constenl
LencaSalvador	101_tu	m a n a n i	1293	Constenl

Section 3

Computationally aided Historical Linguistics

Automatic Cognate Dection

- Automated cognate detection works good, but not perfect
- LingPy library (List *et al.* 2018)
- computer-*aided* Historical Linguistics
- different algorithms for different situations
- GitHub tutorial from the LingPy-authors: https://github.com/lingpy/lingpy-tutorial

Cognate detection in the Chibchan data

- the expert assignments are not perfect
- other authors disagree on some cognates and reconstructions (Pache 2018)
- automated cognates need manual correction, but can take over some of the work

Cognate detection in the Chibchan data

- the expert assignments are not perfect
- other authors disagree on some cognates and reconstructions (Pache 2018)
- automated cognates need manual correction, but can take over some of the work How much do the judgements differ?

```
**************************
* B-Cubed-Scores         *
* ---------------------- *
* Precision:     0.7720 *
* Recall:        0.7273 *
* F-Scores:      0.7490 *
**************************'
**************************
* B-Cubed-Scores         *
* ---------------------- *
* Precision:     0.9276 *
* Recall:        0.6491 *
* F-Scores:      0.7637 *
**************************'
```

Figure 3: B-Cubed Scores for two different algorithms

Using the Edictor tool

- visualisation of alignments, cognacy and correspondences (List 2021)
- is part of the LingPy tutorial
- useful for all kinds of tasks, but especially assigning cognacy within a dataset

Using the Edictor tool

- visualisation of alignments, cognacy and correspondences (List 2021)
- is part of the LingPy tutorial
- useful for all kinds of tasks, but especially assigning cognacy within a dataset

Example workflow for other datasets:

- creating automated cognates with one of the LingPy algorithms
- visualize your data with Edictor and improve cognacy by manual assessments with your knowledge about the language family

Example in Edictor

https://lingulist.de/edictor/

Section 4

Phylolinguistic Analysis

What is phylolinguistics?

- Tools of computational biology for inferring distance in genealogy, time and space of DNA sequences

What is phylolinguistics?

- Tools of computational biology for inferring distance in genealogy, time and space of DNA sequences
- Failed ancestor: Glottochronology
- since then: “Linguists don’t do dates” (McMahon, McMahon, & others 2005)

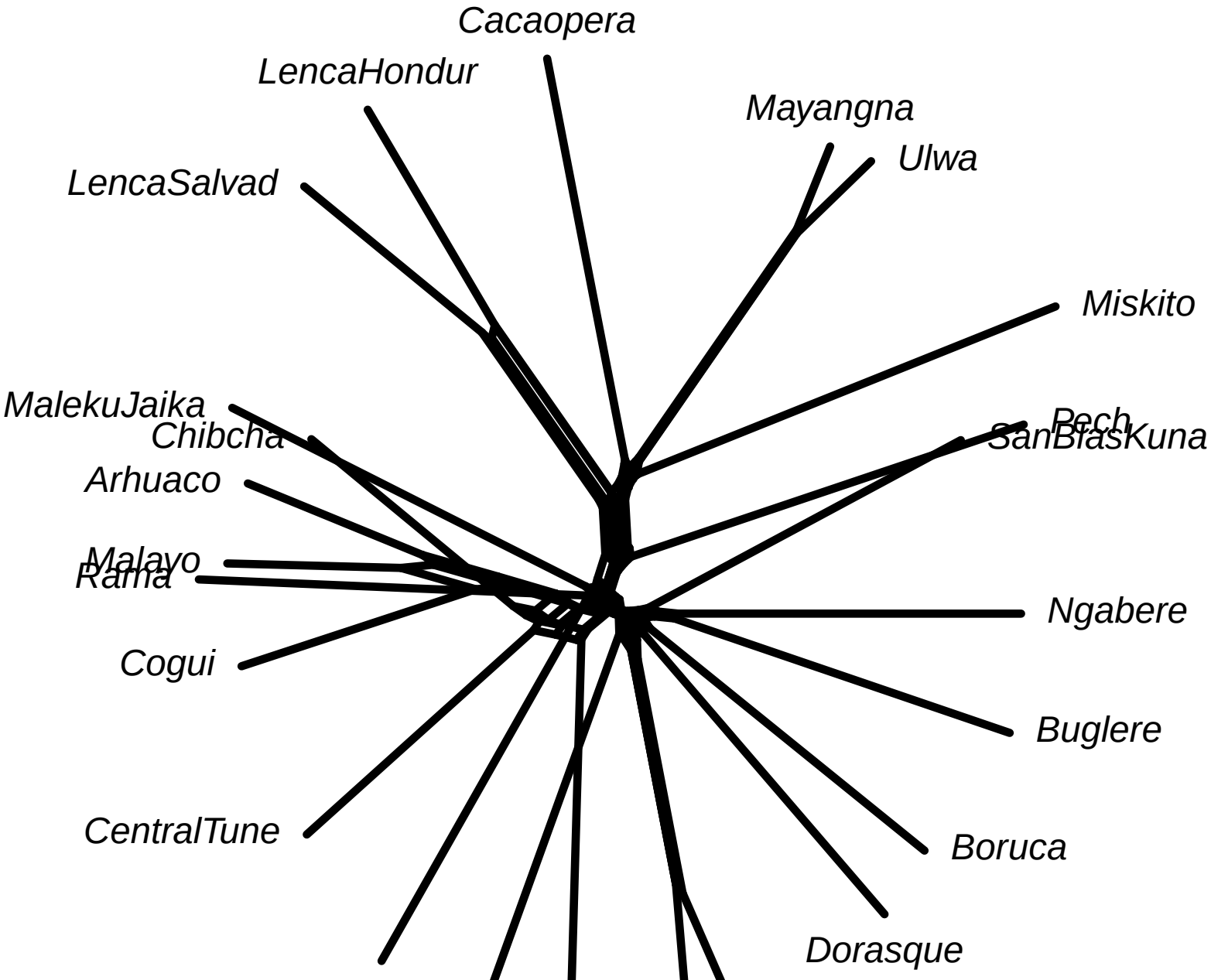
What is phylolinguistics?

- Tools of computational biology for inferring distance in genealogy, time and space of DNA sequences
- Failed ancestor: Glottochronology
- since then: "Linguists don't do dates" (McMahon, McMahon, & others 2005)
- used for linguistic models about genealogy of languages (Michael & Chousou-Polydouri 2019), time of their separation (Gray & Atkinson 2003) and space of dispersals (Neureiter *et al.* 2020)
- some take binary input of cognates, others use phonological inventories (Chacon & List 2015)
- models for time and space are still being refined and seen quite polemically

Applying phylolinguistic methods to the present dataset

- All models used the v0.2 of the dataset
- open questions about the history of the language
- can we assess things about the regional history and movement dispersals?
- Chibchan languages in the bottleneck between North and South America

NeighborNetwork



NeighborNetwork

Well, this doesn't help much. What are we supposed to read out of this?

NeighborNets were common at the beginning of the century, but can't really tell us anything about the genealogy of a linguistic family.

Bayesian Phylolinguistics

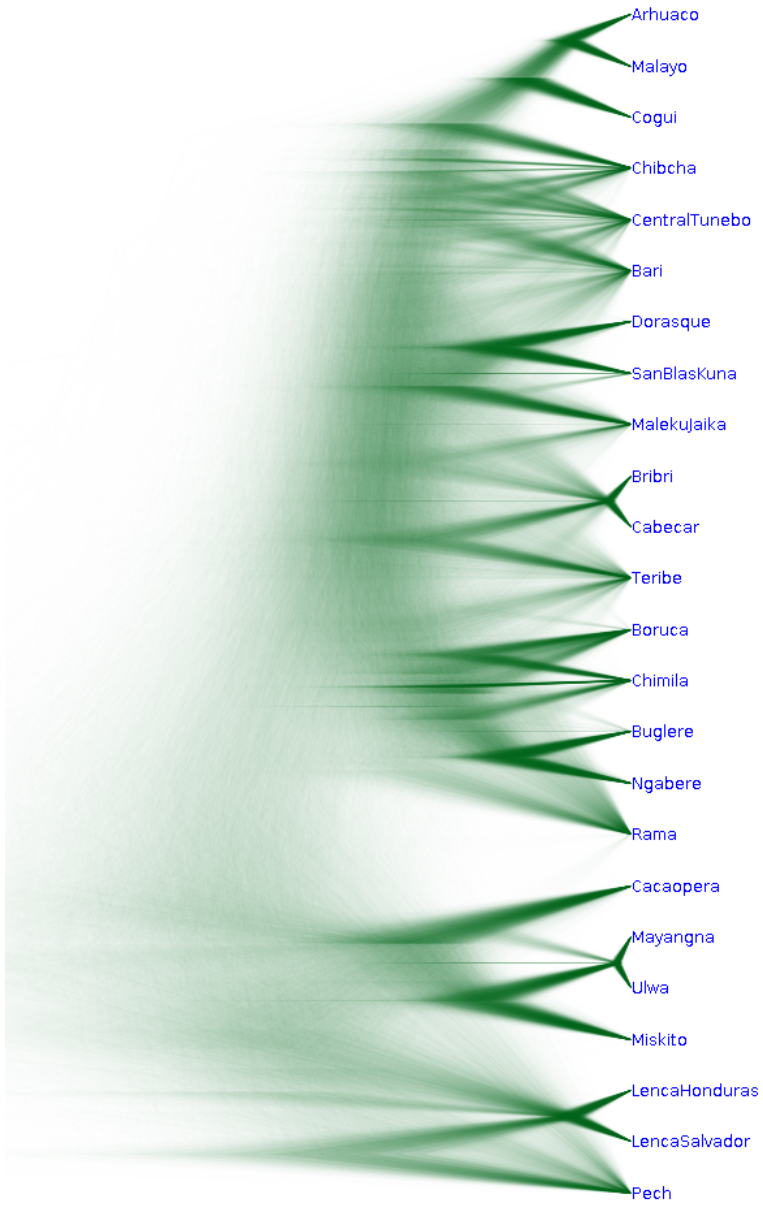
- Bayesian Phylolinguistics incorporates new methods to linguistics
- Bayesian Statistics: quantifying uncertainty (McElreath 2020)
- phylogenetics: Inferences about genealogies from alignment sequences

Bayesian Phylolinguistics

- Bayesian Phylolinguistics incorporates new methods to linguistics
- Bayesian Statistics: quantifying uncertainty (McElreath 2020)
- phylogenetics: Inferences about genealogies from alignment sequences

I can point you to introductions, but cannot teach it.

Preliminary results - densitree



Preliminary results - consensus tree

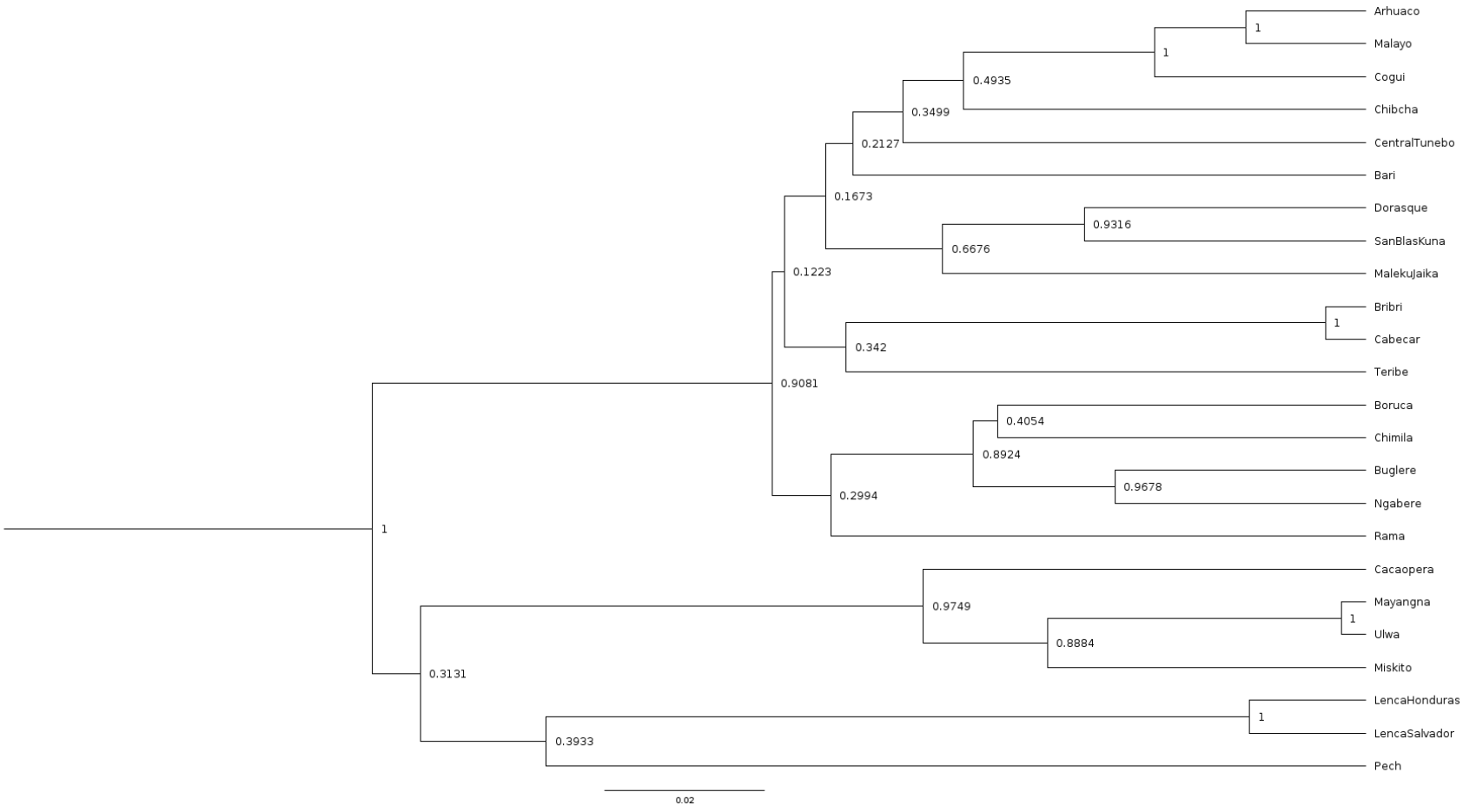


Figure 5: ConsensusTree

Publishing the data

- the dataset is uploaded to GitHub
- https://github.com/lexibank/constenlachibchan/
- the files for phylolinguistic analysis have been added as a zip-file to the description of this talk

References I

Chacon, Thiago Costa & Johann-Mattis List. 2015. Improved computational models of sound change shed light on the history of the tukanoan languages. *Journal of Language Relationship* 3.177–203.

Constenla Umaña, Adolfo. 2005. ?'Existe relación genealógica entre las lenguas misumalpas y las chibchenses? *Estudios de Lingüística Chibcha* 23.7–85.

Dryer, Matthew S. & Martin Haspelmath (Eds.). 2013. *WALS online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

References II

Forkel, Robert & Johann-Mattis List. 2020. CLDFBench. Give your cross-linguistic data a lift. *Proceedings of the twelfth international conference on language resources and evaluation* ed. by 6997-7004. Luxembourg: European Language Resources Association (ELRA).

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, & Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* 5:1.1–10.

Gray, Russell D. & Quentin D. Atkinson. 2003. Language-tree divergence times support the anatolian theory of indo-european origin. *Nature* 426:6965.435–439.

References III

Hammarström, Harald, Robert Forkel, Martin Haspelmath, & Sebastian Bank. 2020. Glottolog/glottolog: Glottolog database 4.3.

List, Johann-Mattis. 2021. *EDICTOR. A web-based tool for creating, editing, and publishing etymological datasets.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

List, Johann-Mattis, Cormac Anderson, Tiago Tresoldi, & Robert Forkel. 2021. CLTS. Cross-linguistic transcription systems.

List, Johann Mattis, Christoph Rzymski, Simon Greenhill, Nathanael Schweikhard, Kristina Pianykh, Annika Tjuka, Mei-Shin Wu, Carolin Hundt, Tiago Tresoldi, & Robert Forkel (Eds.). 2020. *Concepticon 2.4.0*. Jena: Max Planck Institute for the Science of Human History.

References IV

List, Johann-Mattis, Mary Walworth, Simon J. Greenhill, Tiago Tresoldi, & Robert Forkel. 2018. Sequence comparison in computational historical linguistics. *Journal of Language Evolution* 3:2.130–144.

McElreath, Richard. 2020. *Statistical rethinking: A bayesian course with examples in r and stan*. CRC press.

McMahon, April, Robert McMahon, & others. 2005. *Language classification by numbers*. Oxford University Press.

Michael, Lev & Natalia Chousou-Polydouri. 2019. Computational phylogenetics and the classification of south american languages. *Language and Linguistics Compass* 13:12.e12358.

References V

Neureiter, Nico, Peter Ranacher, Rik van Gijn, Balthasar Bickel, & Robert Weibel. 2020. Can bayesian phylogeography reconstruct migrations and expansions in linguistic evolution? *Royal Society Open Science* 8:1.201079.

Pache, Matthias J. 2018. Contributions to chibchan historical linguistics. Unpublished thesis, Leiden University.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, & others. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data* 3:1.1–9.