# Quantifying the Comparative Method

COMPUTATIONAL APPROACHES TO HISTORICAL LINGUISTICS

JOSEPH RHYNE

# The Comparative Method

The traditional workflow of historical and comparative linguistics

Accounts for similarities that cannot be chance

- Establishs genetic relationships among languages through commonly inherited forms

# The Comparative Method

A number of different steps (Trask 2000:64-67):

1. Establish genetic relationship *prima facie*
   - Easy to do for closely related languages, such as Romance
2. Identify cognate sets through systematic correspondences of sounds in words of similar meaning
3. Set up proto-forms from the correspondence sets
   - This allows us to reconstruct the proto-language and detect the sound changes that occurred from mother to daughter languages

# Comparative Method in Action

Fortson 2004: 131

| Language | Word |
|---|---|
| Latin | *centum* |
| Greek | *hekaton* |
| Tocharian B | *kante* |
| Old Irish | *cét* |
| Middle Welsh | *cant* |
| Gothic | *hund* |
| Sanskrit | *śatám* |
| Avestan | *satəm* |
| Lithuanian | *šimtas* |
| Old Church Slavic | *sŭto* |

# Comparative Method in Action

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Latin** | | | c | e | n | t | u | m |
| **Greek** | (h | e) | k | a | -- | t | o | n |
| **Tocharian B** | | | k | a | n | t | e | -- |
| **Old Irish** | | | c | é | -- | t | -- | -- |
| **Middle Welsh** | | | c | a | n | t | -- | -- |
| **Gothic** | | | h | u | n | d | -- | -- |
| **Sanskrit** | | | ś | a | -- | t | á | m |
| **Avestan** | | | s | a | -- | t | ə | m |
| **Lith** | | | š | i | m̃ | t | a | s |
| **OCS** | | | s | ŭ | -- | t | o | |

# Comparative Method

From these correspondence sets, we can reconstruct a proto-form: PIE **ḱm̥tóm*

This process requires expert knowledge of the languages involved

Easy with a limited data set

- How can we do something like the Austronesian language family (1200+ languages)?

# Computational Approaches

Quantitative and computational methods are being used more and more in historical linguistics

- More objective, transparent, and easily replicable (List & Moran 2013)

Built from evolutionary phylogeny

- Concerned with the evolutionary history of species, genes, and morphological characteristics
- Compare to historical linguistics: investigates evolution of language, grammatical features, and words
- Data structure is similar—sequence of characters (DNA, etc.)

# Computational Approaches

Previous methods:

- Phonetic alignment algorithms (Kondrak 2000)
- Tests for genealogical relatedness (Kessler 2001)
- Phylogenetic reconstruction (Holman et al. 2001)
- Automatic cognate detection (Steiner et al. 2011)
- Automatic borrowing detection (Nelson-Sathi et al. 2011)
- Automatic proto-form reconstruction (Bouchard-Cote et al. 2013)

# LingPy (List & Forkel 2016)

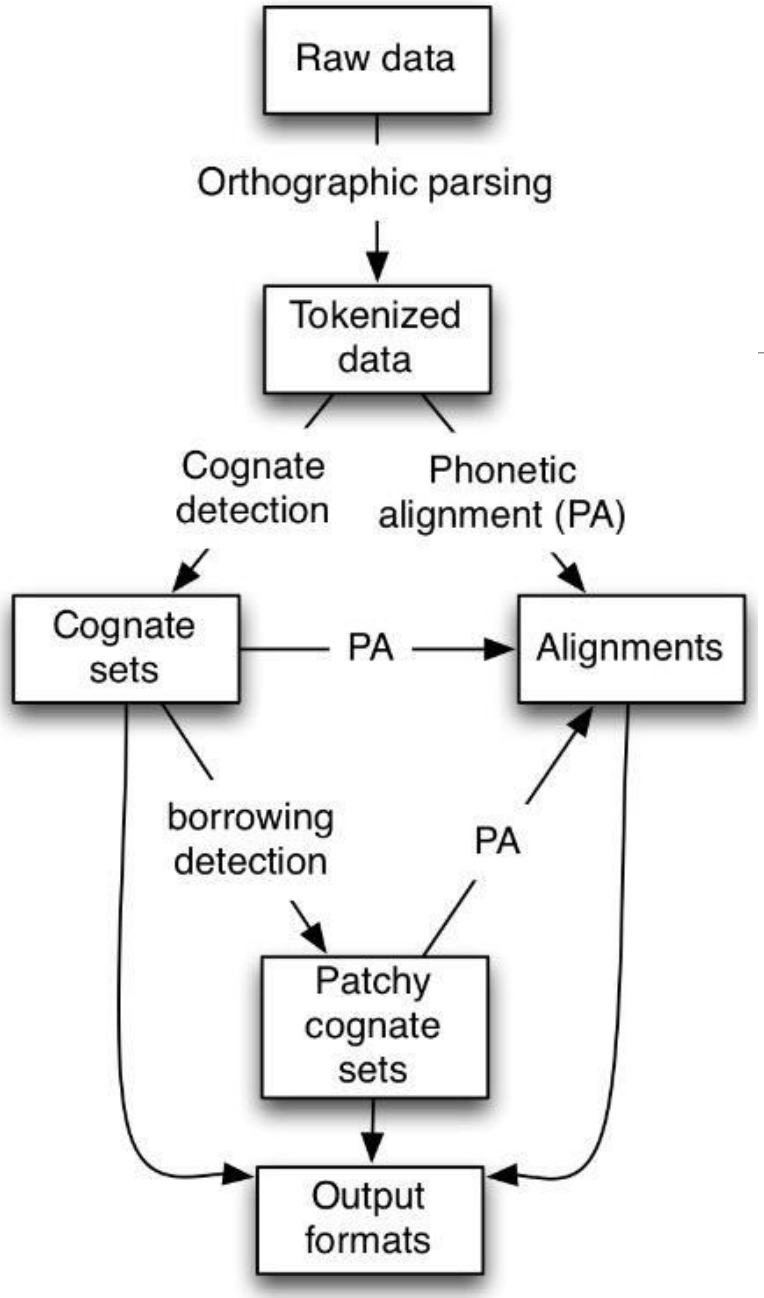
An open-source python library

- Source code is readily available online (lingpy.org)

It implements many computational methods in a general workflow mimicking the Comparative Method

# LingPy

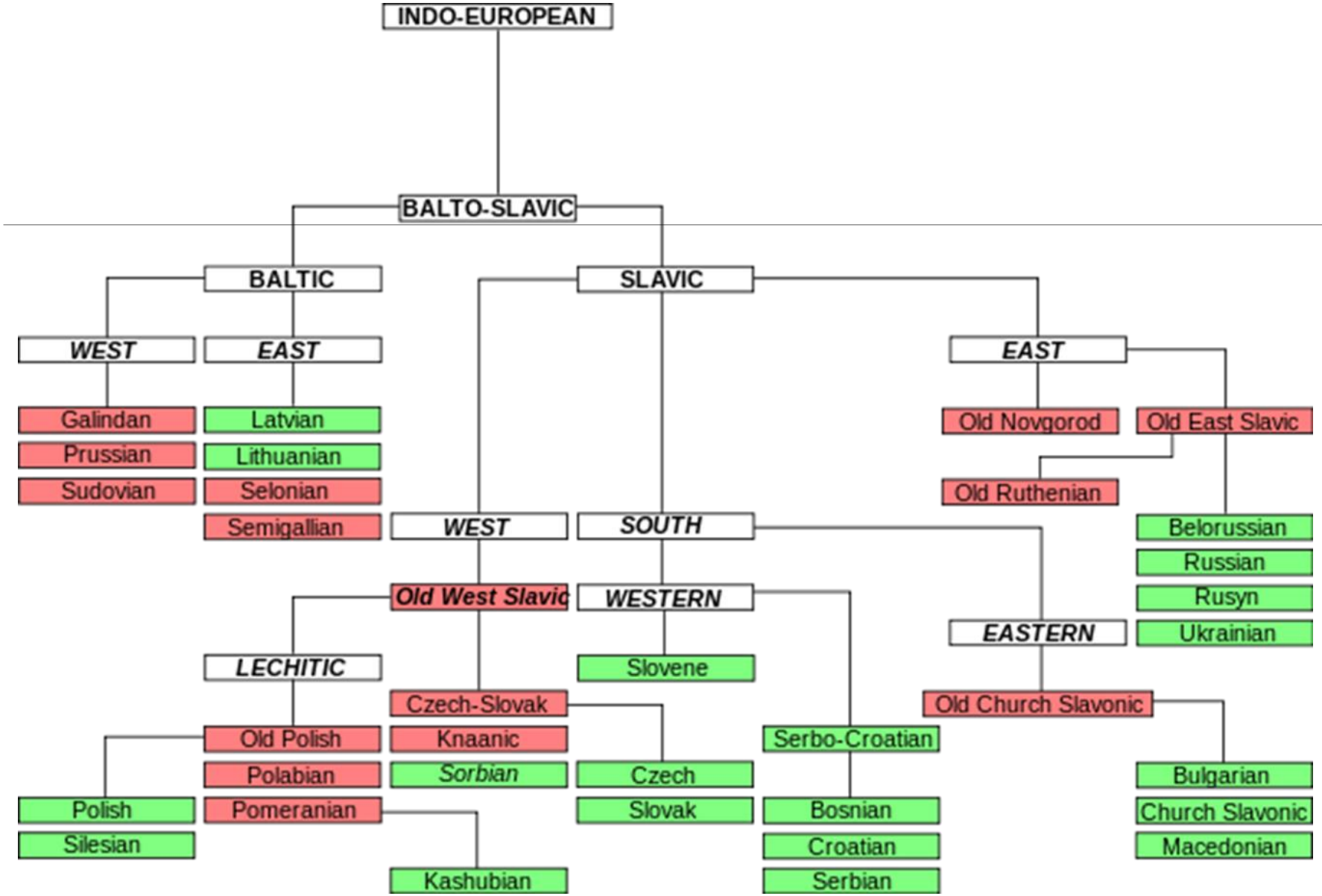
The basic workflow of LingPy (List & Moran 2013)



# Current Study

Apply the LingPy methods to Baltic and Slavic data

- Look at cognate judgements
- Establish a phylogenetic tree
- Create rough reconstructions
- Look at borrowing networks


INDO-EUROPEAN
BALTO-SLAVIC
BALTIC
SLAVIC
WEST
EAST
EAST
Galindan
Prussian
Sudovian
Latvian
Lithuanian
Selonian
Semigallian
Old Novgorod
Old East Slavic
Old Ruthenian
WEST
SOUTH
Belorussian
Russian
Rusyn
Ukrainian
Old West Slavic
WESTERN
EASTERN
LECHITIC
Slovene
Czech-Slovak
Old Church Slavonic
Old Polish
Knaanic
Serbo-Croatian
Polabian
Sorbian
Czech
Bulgarian
Polish
Pomeranian
Slovak
Bosnian
Church Slavonic
Silesian
Croatian
Macedonian
Kashubian
Serbian

# Balto-Slavic Languages

Branch of Indo-European

- The specific relation between them is controversial
  - A single branch, like Indo-Iranian?
  - Two separate branches?
- Large number of words shared exclusively by Baltic and Slavic (Trautmann 1923)
- no major isogloss that separates the two branches
- Relatively lately attested:
  - Slavic ca. 9$^{th}$ century
  - Baltic cs. 12$^{th}$ century

# Data

Swadesh lists for 6 Slavic languages and 3 Baltic languages

- Bulgarian, Czech, Croatian, OCS, Polish, Russian
- Latvian, Lithuanian, Old Prussian
- 172 words in each list

Lists taken from the Indo-European Lexical Cognacy Database (IELex, http://ielex.mpi.nl/)

- Compiled from various etymological dictionaries

# Data

Input data: Wordlist

- Tab-delimited text file organized into rows and columns with headers

Balto-Slavic.qlc - Notepad

File Edit Format View Help

```
# Wordlist

# DATA
ID      CONCEPT IPA     DOCULECT        TOKENS
#
1540    I       as      Bulgarian       a s
1541    I       ja      Russian         j a
1542    I       ja      Polish          j a
1543    I       ja:     Czech           j a:
1544    I       as      Old Prussian    a s
1545    I       es      Latvian         e s
1546    I       eʃ      Lithuanian      e ʃ
1547    I       jâ:     Serbo-Croatian          j â:
1548    I       ɑzŭ     Old Church Slavic       ɑ z ŭ
#
```

# Implementing the data

Import the wordlist file

LingPy can manipulate the data

- Find specific entries for concepts
- Return entries for specific languages
- Add new entries

The IPA entries need to be tokenized and aligned

# Cognate Judgements

After tokenization, cognate judgements can be determined

Follows the STARLING approach

- Cognate words are assigned the same cognate ID

Accomplished through the LexStat method (List 2012)

- Other methods (Turchin, NED, and SCA) are also available in LingPy

# Results: Example Alignment

| Language | Alignments | | | |
|---|---|---|---|---|
| Bulgarian | d | -- | v | a |
| Croatian | d | -- | v | â: |
| Czech | d | -- | v | a |
| Polish | d | - | v | a |
| Russian | d | - | v | a |
| OCS | d | ŏ | v | a |
| Latvian | d | i | v | i |
| Lithuania | d | -- | v | i |
| Old Prussian | d | -- | w | ai |

Plays a crucial role in automatic approaches

Gets at the idea of sound correspondence sets

# LexStat

Language-specific: no predefined scoring function

Uses an expanded version of Dolgopolsky's (1964) sound classes

Computes cognate distance scores through pairwise alignments, following Bouchard-Cote et al. (2013)

- Close to the idea of sound correspondence sets

Words drawn from randomized sample

- Repeatedly aligned with each other
- Creates a distribution of sound transitions
- Compared to the actual distribution from aligned words in the wordlist

# LexStat

Sequence conversion
- Input converted to sound classes; sonority profiles determined

Scoring-scheme creation
- Language specific; created through a permutation method

Distance calculation
- Pairwise distance between all words are computed

Sequence clustering
- Sequences clustered into cognate sets whose average distance is beyond a certain threshold
- Flat cluster variant of the UPGMA algorithm

# LexStat output

Balto-Slavic_lexstat.qlc - Notepad

File Edit Format View Help

```
# Wordlist

# META
@vowels:T V _
@json: {"params": {"cluster": "lexstat_upgma_0.50", "cscorer'

# DATA
ID      CONCEPT IPA     DOCULECT        TOKENS  LEXSTATID
#
1540    I       as      Bulgarian       a s     1
1541    I       ja      Russian         j a     2
1542    I       ja      Polish          j a     2
1543    I       ja:     Czech           j a:    2
1544    I       as      Old Prussian    a s     1
1545    I       es      Latvian         e s     1
1546    I       eʃ      Lithuanian      e ʃ     1
1547    I       jâ:     Serbo-Croatian  j â:    2
1548    I       ɑzŭ     Old Church Slavic       ɑ z ŭ   1
#
```

# Results: Cognate Judgements

Cognate words are assigned a CogID

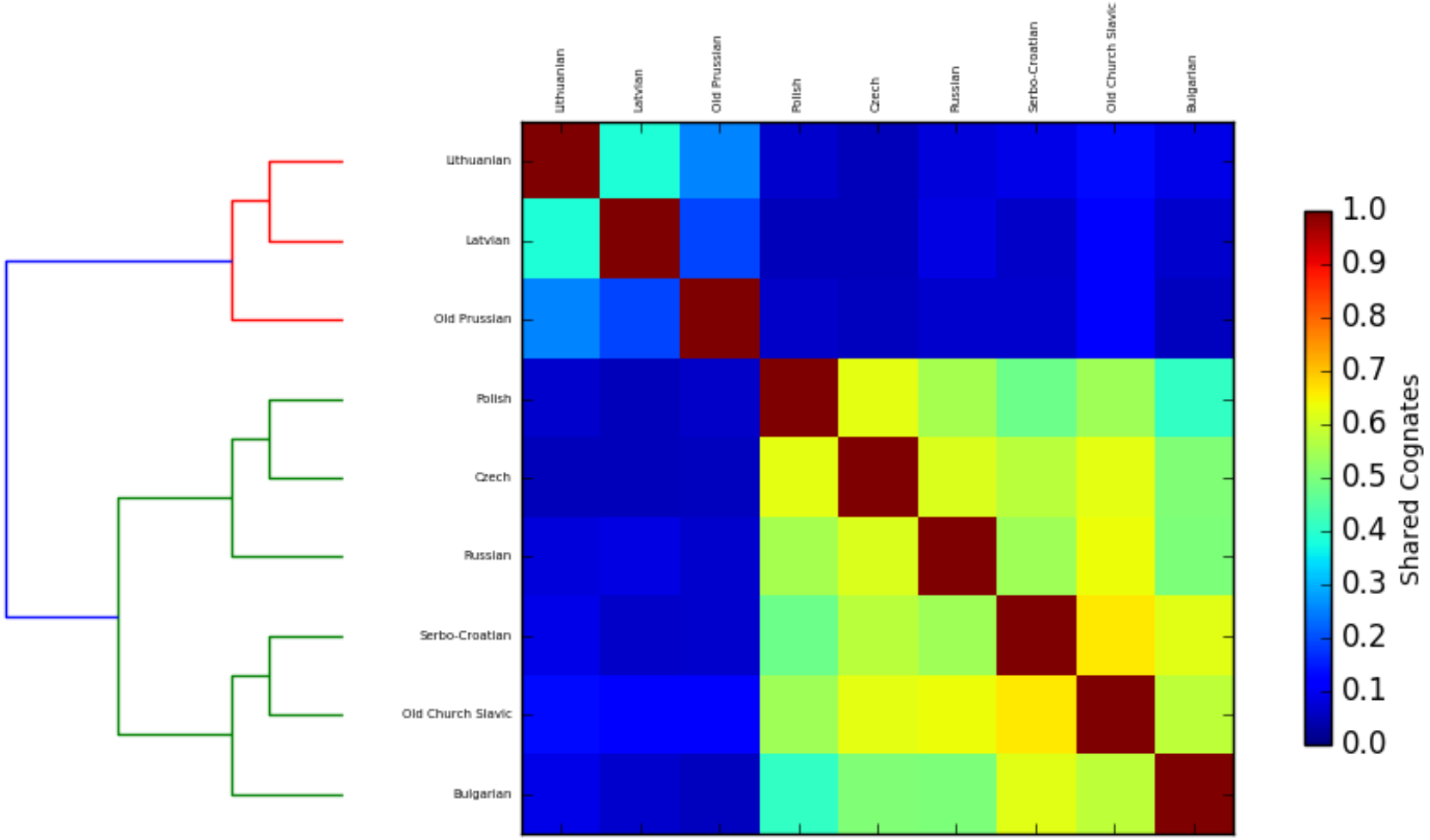
For example, every word for “two” has a CogID of 971

Not foolproof

- Some missed cognates: ‘I’, ‘full’, etc.
  - Actual cognates can be misidentified because of sound classifications, alignments, etc.

Can display the percentage of cognates shared by languages in a heat map

# Percentage Shared Cognates



# Consensus Reconstruction

From this, we can create “quick and dirty” reconstructions

- Consensus strings are calculated from all alignments
- Selects the most frequent characters
- Typically around 2 edit operations from expert reconstructions

# Results: Reconstructions

Examples:

For ‘two’, we get *dva
- Cf. PSl *dъva, PBSl *duwō

For ‘day’, we get *dein-
- Cf. PBSl *dein-/*din-

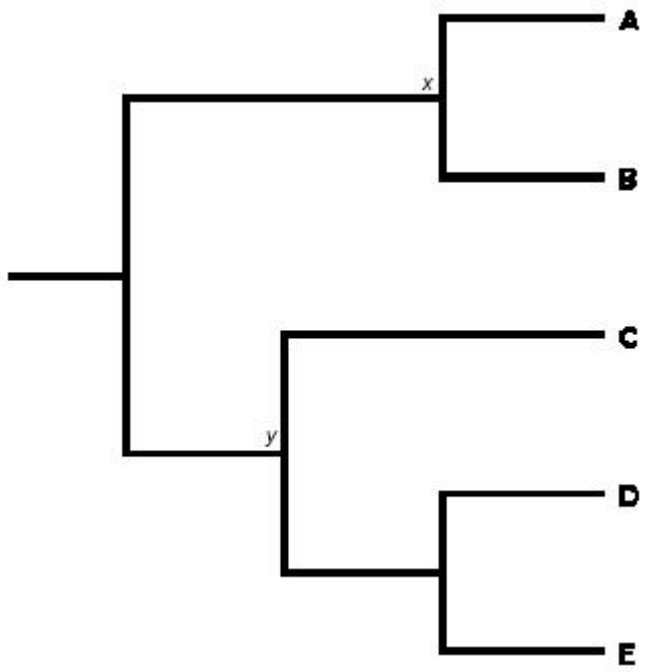
For ‘stone’, we get Slavic *kamen-, Baltic *akmens
- Cf. PSl *kamen~kamy, PB *akm̄o

For ‘good’, we get Slavic *dɔbr, Baltic *labs
- Cf. PSl *dobrъ, PB *labas

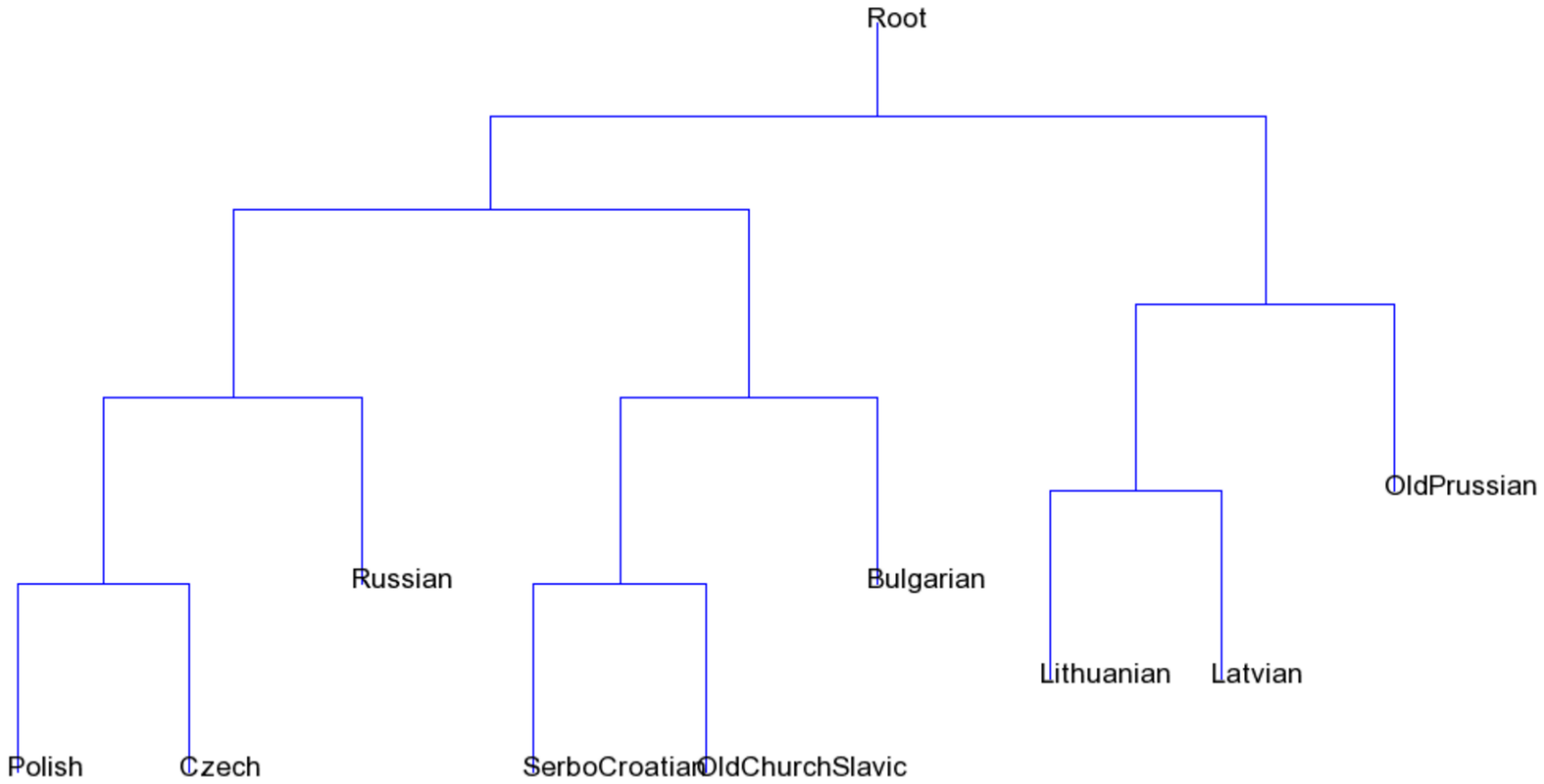
# Phylogenetic Trees

Also from this, we can create a simple bifurcating tree for the languages

- Use either Neighbor-joining or UPGMA
- Distance matrices=number of shared cognates
- Outputs simple Newick tree format
  - ((A,(B,C,),(D,E));



# Results: Phylogenetic Tree



# Borrowing Detection

Evolution of language is both a vertical and horizontal process

- Vertical=inheritance
- Horizontal=borrowing

Follows the method of Nelson-Sathi et al. (2011)

- Apply phylogenetic networks to recover frequency of hidden borrowings

# Borrowing Detection

Minimal Lateral Network (MLN)

- Networks=mathematical structures used to model pairwise relations between entities
  - Entities=vertices
  - Edges=interactions between vertices
- Applies the technique of gain-loss mapping to presence-absence patterns of cognate sets
- Searches for cognate sets incompatible with a reference tree typology
  - Points to borrowing

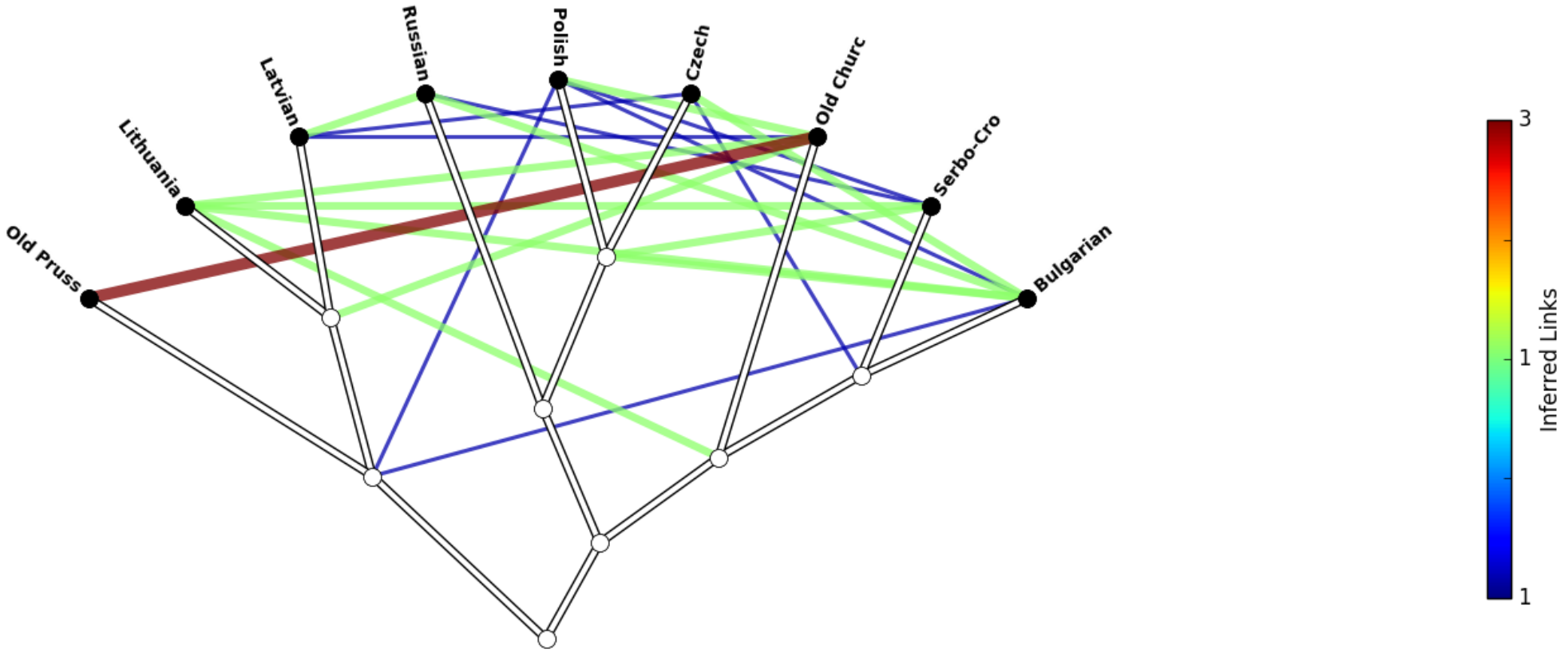
# Results: Borrowing Detection

Use MLN to capture the inferred horizontal relationships

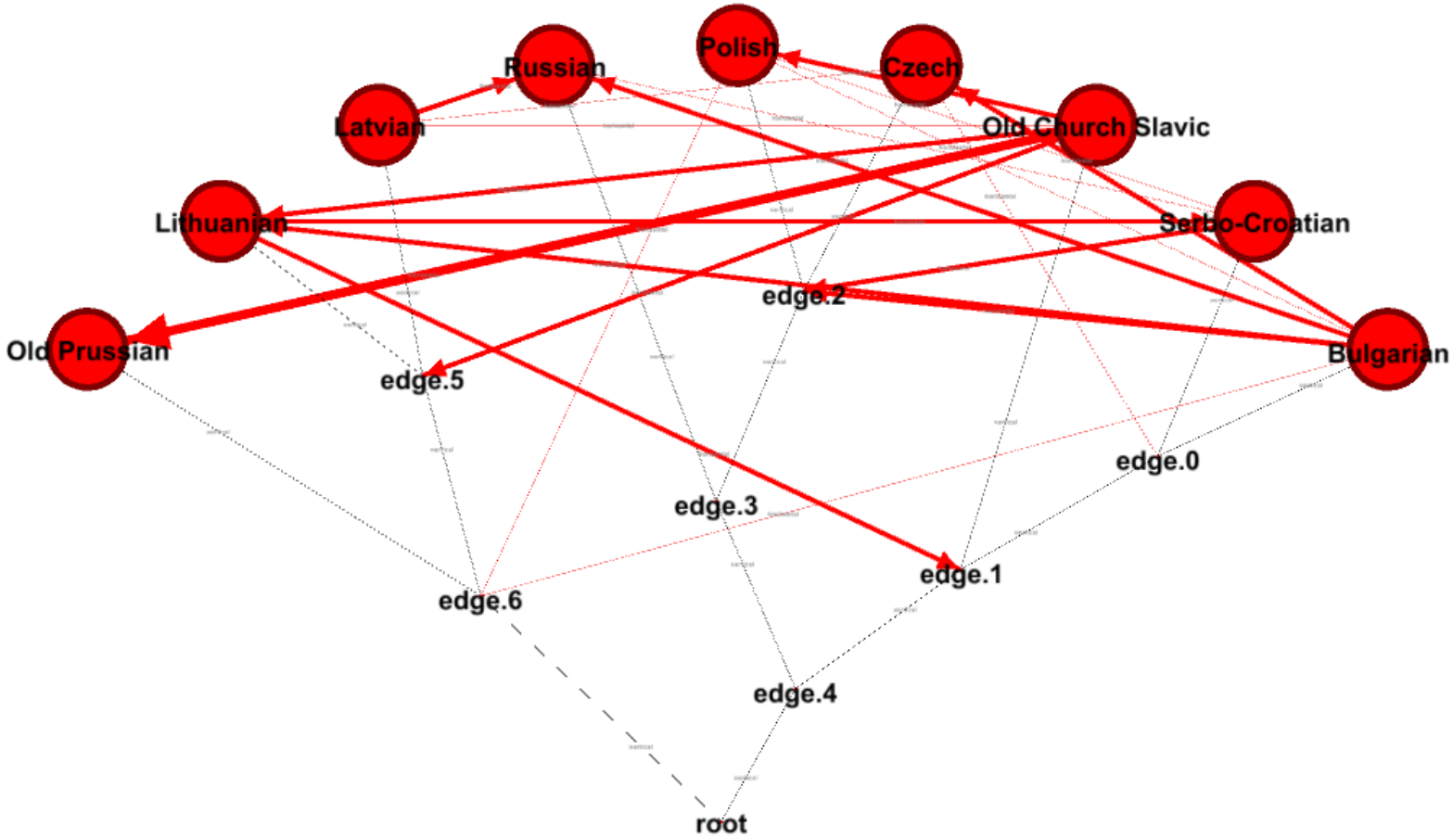
- Example: Old Prussian nage ‘foot’
  - Cf. Lith. koja, Lat. kãja; Rus. noga, OCS nɔgɑ

Plot the results against our reference tree

# Results: Borrowing Detection



# Results: Borrowing Detection w/Direction



# Conclusions:

Useful, but not infallible

- Best combined with expert knowledge

Needs refinement in cognate judgements and reconstructions

Baltic and Slavic:

- Still uncertain about their exact relationship
  - Need to examine it further within a wider Indo-European context
- Extensive borrowing into Baltic from Slavic
- Latvian and Lithuanian are more closely related than Old Prussian

# Further Study

Cognate judgements

- Low B-Cubed scores (Bouchard-Côté et al. 2013)
- Expand on sound classes that are used to establish the cognate sets
- Implement expert judgements

IPA transcription

- This still has to be done by hand
- Letter-to-phoneme conversion as Machine Translation (Rama & Gali 2009)

# Further Study

Track the development of individual words through the language network

- Both inheritance and borrowing
- Examine intermediate stages of words

Implement more data

- More languages
- Longer wordlists
- Examine Balto-Slavic within a wider Indo-European context

# Bibliography


Bouchard-Côté, D. Hall, T. L. Griffiths, and D. Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. PNAS, 110(11):4224–4229.

A. B. Dolgopolsky. 1964. Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točky zrenija [A probabilistic hypothesis concerning the oldest relationships among the language families of Northern Eurasia]. Voprosy Jazykoznanija, 2:53–63.

B. Fortson. 2004. *Indo-European Language and Culture.* Wiley-Blackwell: Oxford.

E. W. Holman, C. H. Brown, S. Wichmann, A. Müller, V. Velupillai, H. Hammarström, S. Sauppe, H. Jung, D. Bakker, P. Brown, O. Belyaev, M. Urban, R. Mailhammer, J.-M. List, and D. Egorov. 2011. Automated dating of the world's language families based on lexical similarity. Current Anthropology, 52(6):841–875

B. Kessler. 2001. The significance of word lists. Statistical tests for investigating historical connections between languages. CSLI Publications, Stanford.

J.M List and R. Forkel. 2016 : **LingPy. A Python library for historical linguistics**. Version 2.5. URL: http://lingpy.org, DOI:https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy.

J.M. List and S. Moran. 2013. An Open Source Toolkit for Quantitative Historical Linguistics. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,* 13-18.

J.-M. List. 2012a. LexStat. Automatic detection of cognates in multilingual wordlists. In Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, pages 117–125. Association for Computational Linguistics.

T. Rama and K Gali. Modeling machine transliteration as a phrase based statistical machine translation problem. *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration,* 124-127.

S. Nelson-Sathi, J.-M. List, H. Geisler, H. Fangerau, R. D. Gray, W. Martin, and T. Dagan. 2011. Networks uncover hidden lexical borrowing in IndoEuropean language evolution. Proceedings of the Royal Society B, 278(1713):1794–1803.

L. Steiner, P. F. Stadler, and M. Cysouw. 2011. A pipeline for computational historical linguistics. Language Dynamics and Change, 1(1):89–127.

R. L. Trask. 2000. The dictionary of historical and comparative linguistics. Edinburgh University Press, Edinburgh.

R. Trautmann. 1923. *Baltisch-Slavisches Worterbuch.* Gottingen.