EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

PHILOSOPHISCHE
FAKULTÄT

EVOLAEMP
Language Evolution:
The Empirical Turn

Ancestral State Reconstruction and Loanword Detection

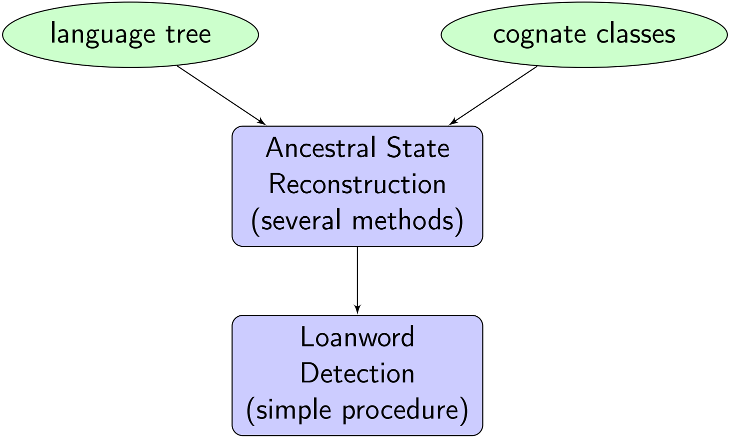
Marisa Koellner and Johannes Dellert

University of Tuebingen

28. October 2015

This research has been supported by the ERC Advanced Grant 324246 EVOLAEMP.

Model


language tree

cognate classes

Ancestral State Reconstruction (several methods)

Loanword Detection (simple procedure)

IELex and Language Tree

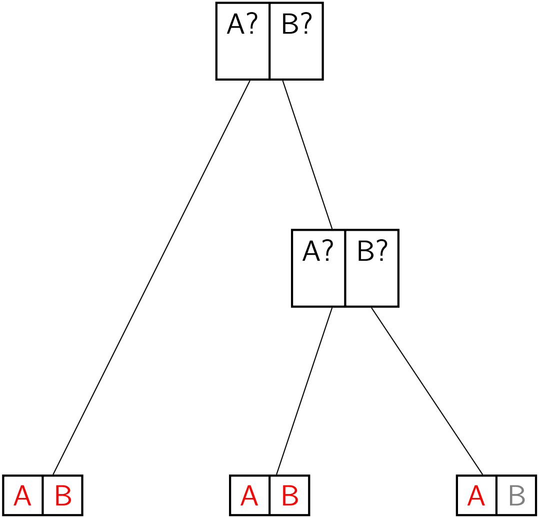
- language sample (IELex):
  consists out of 207 concepts across 95 languages
- language tree:
  classifications from Ethnologue for living languages
  classifications from Glottolog for extinct languages
- cognate classes (IELex):
  represented at the leaves of the tree
- loanword judgements (IELex):
  binary annotation
  - 1 indicates loanwords judgement
  - 0 indicates either the absence of borrowing or incomplete data

Ancestral State Reconstruction

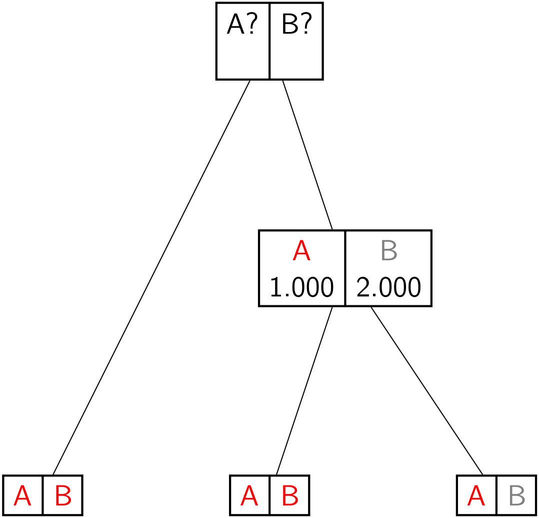
Two versions of ancestral state reconstruction:

- the **Sankoff algorithm** for maximum parsimony, as implemented as part of the PAUP* software
(assumes that there should be exactly one cognate class for each concept at each node, only allows multiple reconstructions if both variants lead to maximum parsimony)
- an alternative **threshold-based method**
built on a recursively computed confidence measure
(considers cognate classes separately, no bias against multiple reconstructed classes)

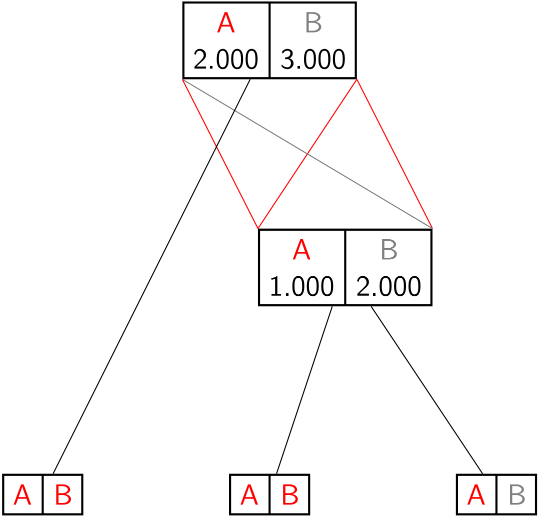
ASR: Sankoff Algorithm as in PAUP*



ASR: Sankoff Algorithm as in PAUP*



ASR: Sankoff Algorithm as in PAUP*



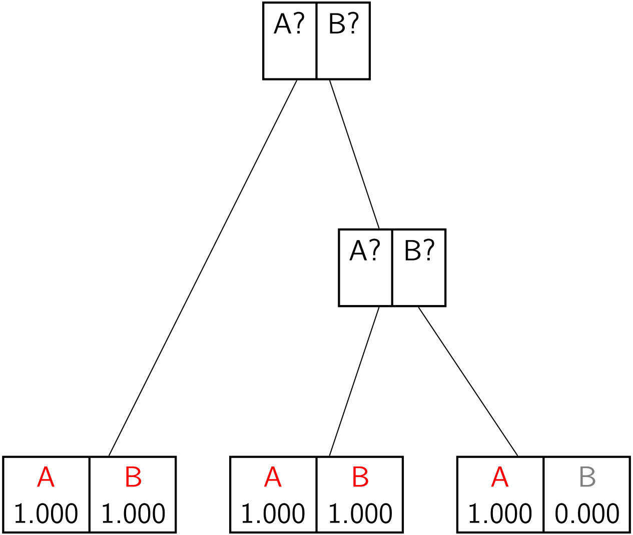
ASR: Computing the Confidence Measure

- assign confidence $cn(v, c)$ to each class c at each node v
- for attested languages, $cn(v, c) := 1$ or $cn(v, c) := 0$
- for non-leaves (i.e. reconstructed nodes), we recursively compute confidence values as follows ($Ch(v)$ = children of v):

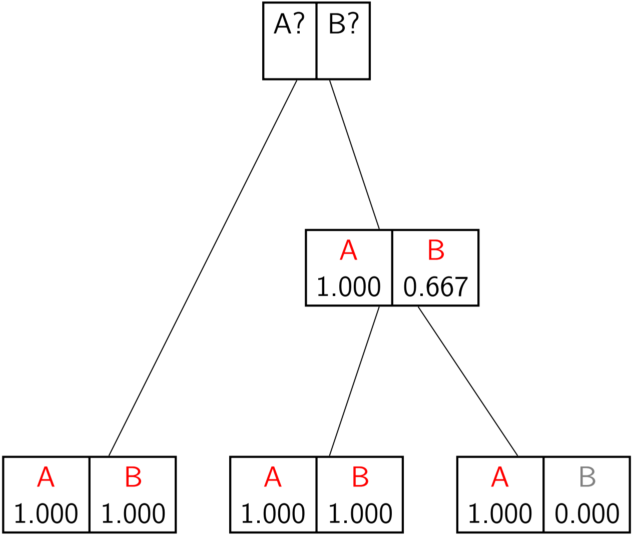
$$cn(v, c) := \max \left\{ 0, 1 - \frac{1 - \frac{\sum_{v_i \in Ch(v)} cn(v_i, c)}{|Ch(v)|}}{\sum_{v_i \in Ch(v)} cn(v_i, c) + 0.5} \right\}$$

- intuition: close to 1 if average of child confidences is high, even closer to 1 if attested across many branches
- threshold: reconstruct class c for v whenever $cn(v, c) > 0.4$

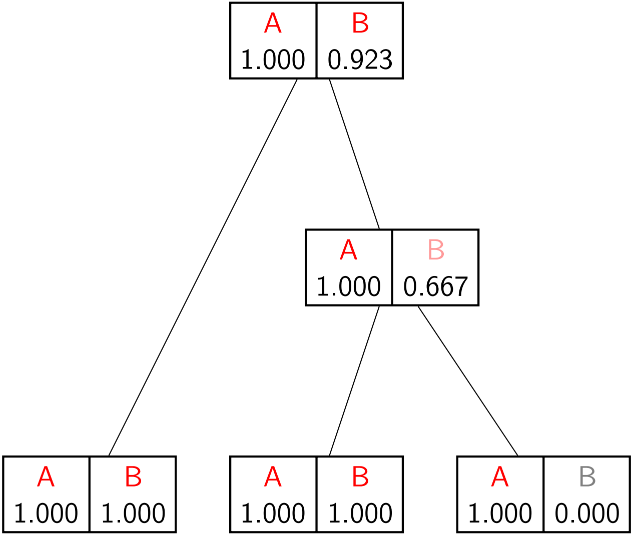
ASR: Our Reconstruction Method



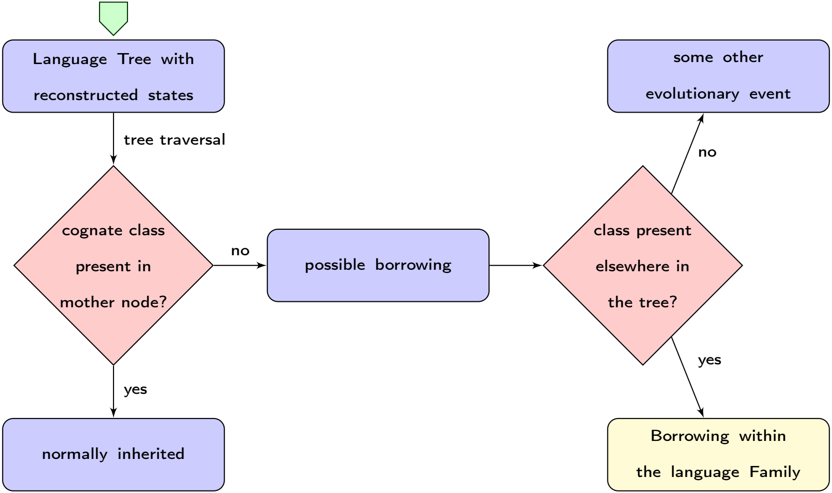
ASR: Our Reconstruction Method



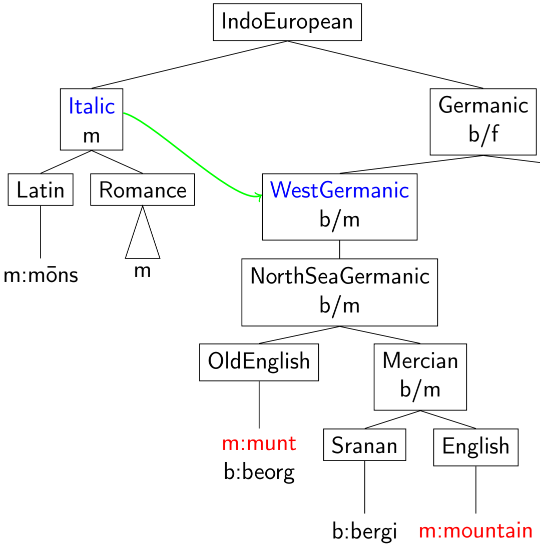
ASR: Our Reconstruction Method



Loanword Detection

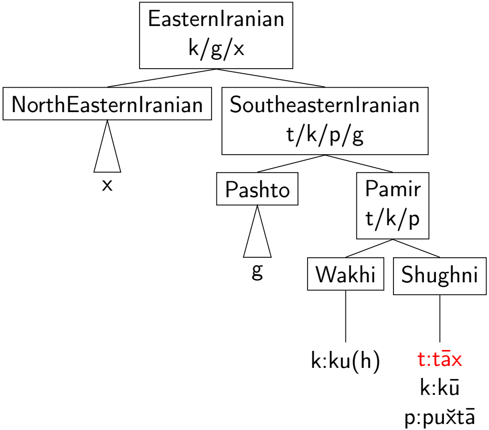

Language Tree with reconstructed states
tree traversal
cognate class present in mother node?
no
possible borrowing
class present elsewhere in the tree?
no
some other evolutionary event
yes
normally inherited
yes
Borrowing within the language Family

Internal Borrowing (concept *mountain*)



Cognate Classes:
m = mountain
b = berg
f = fjäll

External Borrowing (concept *mountain*)



EasternIranian
k/g/x

NorthEasternIranian

x

SoutheasternIranian
t/k/p/g

Pashto

g

Pamir
t/k/p

Wakhi

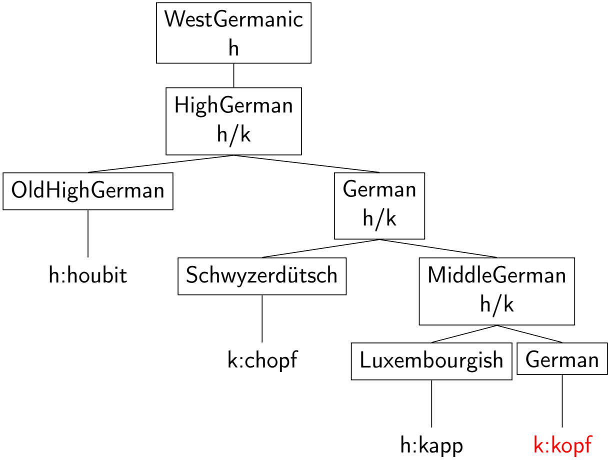
k:ku(h)

Shughni

t:tāx
k:kū
p:pux̌tā

Cognate Classes:
t = tāx
k = kū
p = pux̌tā
g = gar/gora
x = xox

Semantic Evolution (concept *head*)



Cognate Classes:
k = kopf
h = head

Evaluation

	detected loanwords	true loanwords	Precision	Recall
Confidence	1409	239/1100	16%	22%
Sankoff	4532	477/1100	10%	46%

- quite low overall performance
- loanword detection method highly depends on the quality of the hypothetical cognate classes at the internal nodes

Evaluation

Limits of borrowing detection:

- binary annotation → true loanwords might not be annotated
- directionality → only target languages can be compared
- data → no gold standard including source languages

Method Comparison

Performance of the methods shown on the concept *mountain*:

	detected loanwords	true loanwords
Confidence	2	2
Sankoff	39	3

Method Comparison

Performance of the methods shown on the concept *spit*:

	detected lonwords	true loanwords
Confidence	1	1
Sankoff	8	0

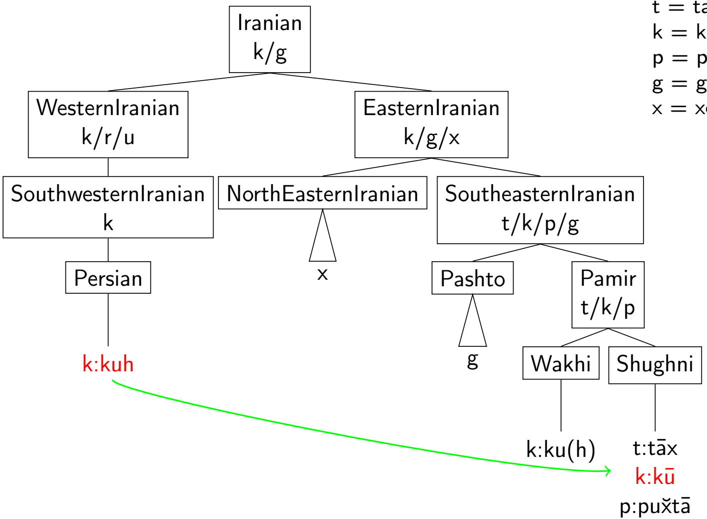
Additional Limits of loanword detection

(1) external borrowing:
the cognate class is not present in the tree

(2) semantic evolution:
the word changes its meaning over time

(3) borrowing within a cognate class:
the borrowing took place within one cognate class

Borrowing within a cognate class (concept *mountain*)



Cognate Classes:
t = tāx
k = kū
p = puxtā
g = gor/gora
x = xox

Further work

(1) reconstruction:
try more ancestral state reconstruction methods

(2) data simulation:
alternative evaluation of the model on much more data

(3) more complex model of borrowing:
detect more complex linguistic cases
(e.g. within cognate classes)

(4) collect expert loanword judgements:
building a gold standard which includes source language

(5) model directionality:
getting a clearer picture of language contact

Thank you for your attention!

Dunn, M. (Ed.). (2015). *Indo-European Lexical Cognacy Database.* http://ielex.mpi.nl/.

Hammarstroem, H., Forkel, R., Haspelmath, M., & Bank, S. (2015). *Glottolog 2.5.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `http://glottolog.org`

Lewis, M. P. (2009). Ethnologue: Languages of the world sixteenth edition. *Dallas, Tex.: SIL International. Online version: http://www. ethnologue. com*.

Sankoff, D. (1975). Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, *28*(1), 35–42.

Swafford, D. (2002). *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.* Sunderland, Mass.: Sinauer Associates.