Do Roots Really Grow Trees?

Quantitative Root-Based Approaches in Historical Linguistics

Hans Geisler, Johann-Mattis List

August 26, 2010

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Structure of the Talk

Introduction
- Comparison and Reconstruction
- The Root Concept in Historical Linguistics
- Lexicostatistics vs. Root-Based Approaches

Two Models of Language Evolution
- The Separation Base Method
- Etymostatistics
- Phylogenetic Reconstruction
- Comparison of the Models

Testing the Models of Language Evolution
- Simulations of the Evolutionary Models
- Testing the Models on Real Data

Conclusion
- Model-Internal Problems
- Models and Reality

Introduction

- Comparison and Reconstruction
- The Root Concept in Historical Linguistics
- Lexicostatistics vs. Root-Based Approaches

Comparison and Reconstruction

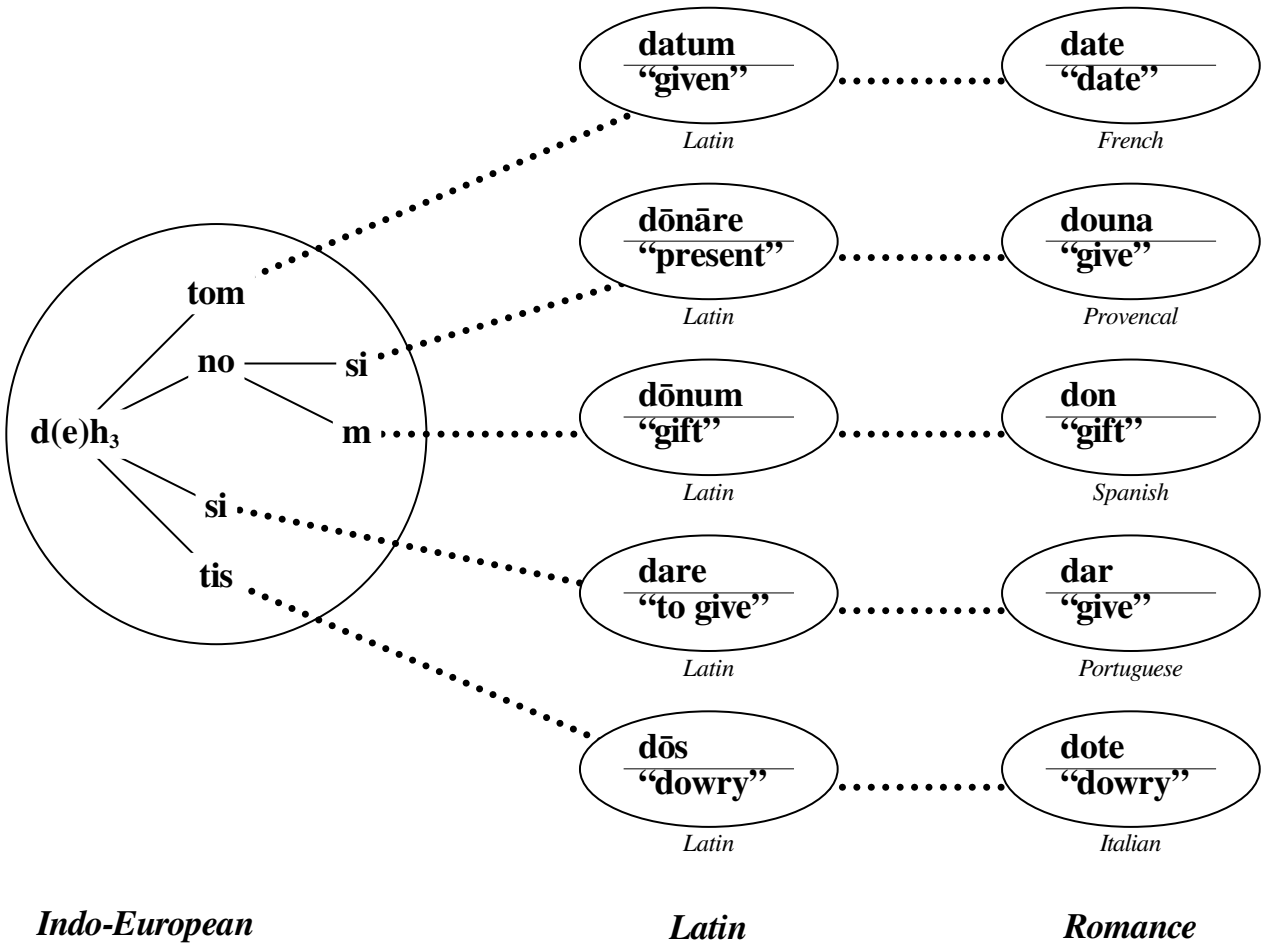
Goal of Comparison

One major goal of comparison in historical linguistics is to reconstruct the way genetically related languages evolved from a common ancestor language.

Characters of Comparison

The characters of comparison differ in the different approaches in historical linguistics. The leading question in character selection is always, whether a specific sample of characters is meaningful for phylogenetic reconstruction.

The Root Concept in Historical Linguistics



Lexicostatistics vs. Root-Based Approaches

	Lexicostatistics	Root-Based-Approaches
Evolutionary Model	replacement of words denoting basic concepts in semantic meaning slots	gain and loss of roots
Comparanda	words denoting the same basic concepts	words which can be traced back to a single root ("word families")
Method of comparison	comparative method	comparative method
Characters	basic concepts	roots (proto-forms)

Lexicostatistics vs. Root-Based Approaches

Concept	Italian	Romanian	Spanish	French	Latin
BIRD	-	pasăre	pássaro	-	passer
	ucello	-	ave	oiseau	avis

Table: The Lexicostatistical Analysis for the Concept BIRD

Root	Meaning	Italian	Romanian	Spanish	French
passer	“sparrow”	passero	pasăre	pássaro	passereau
avis	“bird”	ucello	-	ave	oiseau

Table: Root-Based Analysis for Latin *passer* “sparrow” and *avis* “bird”

Lexicostatistics vs. Root-Based Approaches

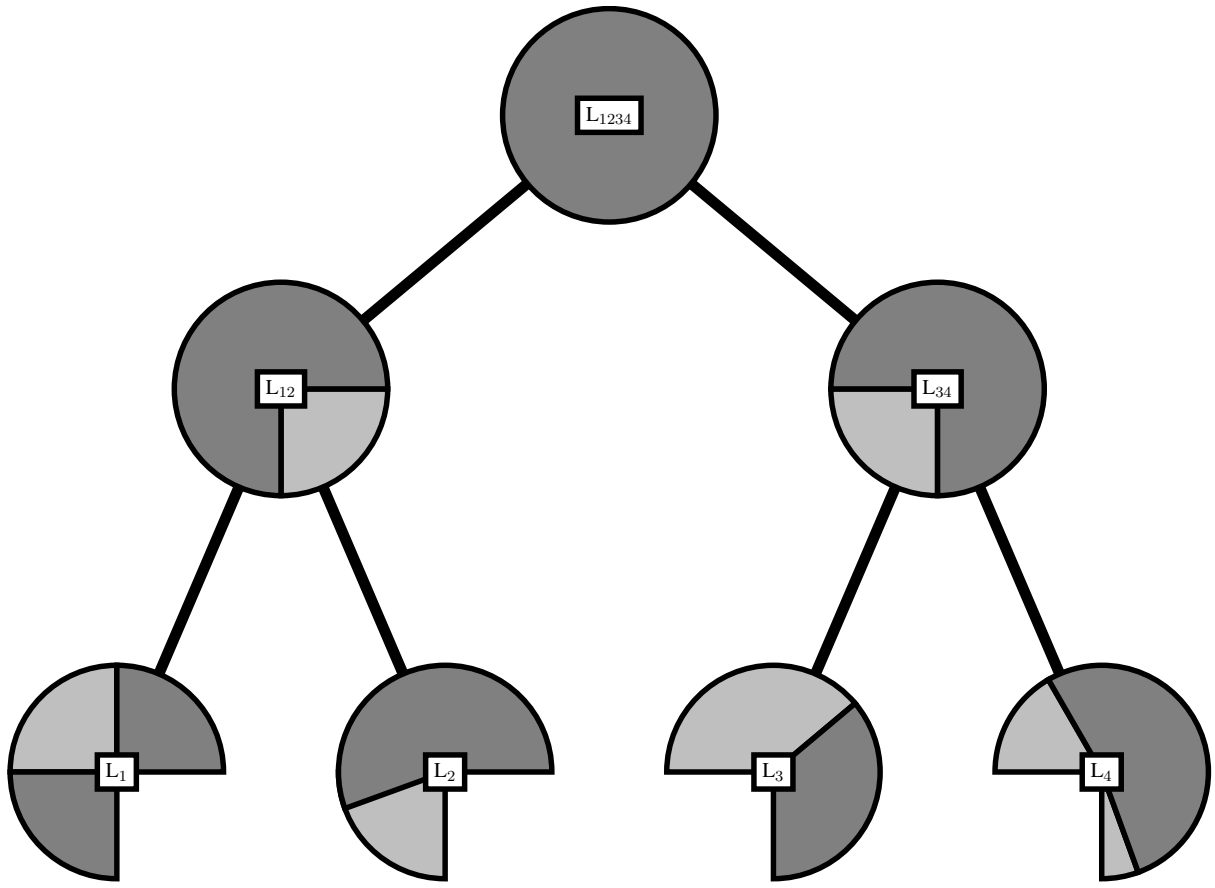
Apparent Advantages of Root-Based Approaches

- Root-based approaches do not depend on the basic vocabulary assumption.
- Dataset is not restricted to the realm of basic vocabulary.
- Use of roots (proto-forms) as primary characters of comparison comes closer to the framework of the comparative method.

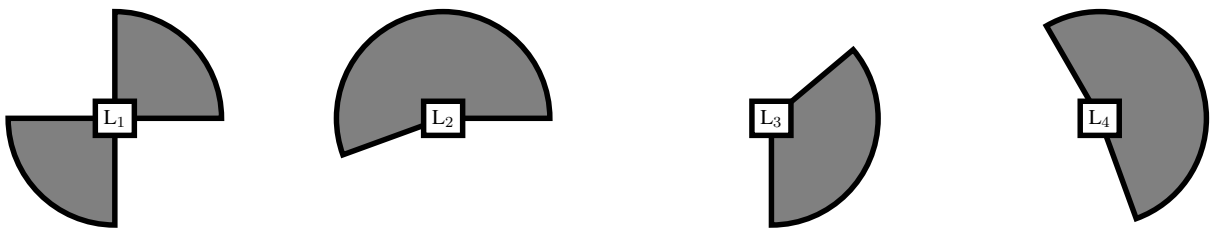
Two Models of Language Evolution

- The Separation Base Method (Holm 2000 & 2008)
- Etymostatistics (Starostin 2000[1989])
- Phylogenetic Reconstruction
- Comparison of the Models

Evolutionary Model of the Separation Base Method



Evolutionary Model of the Separation Base Method

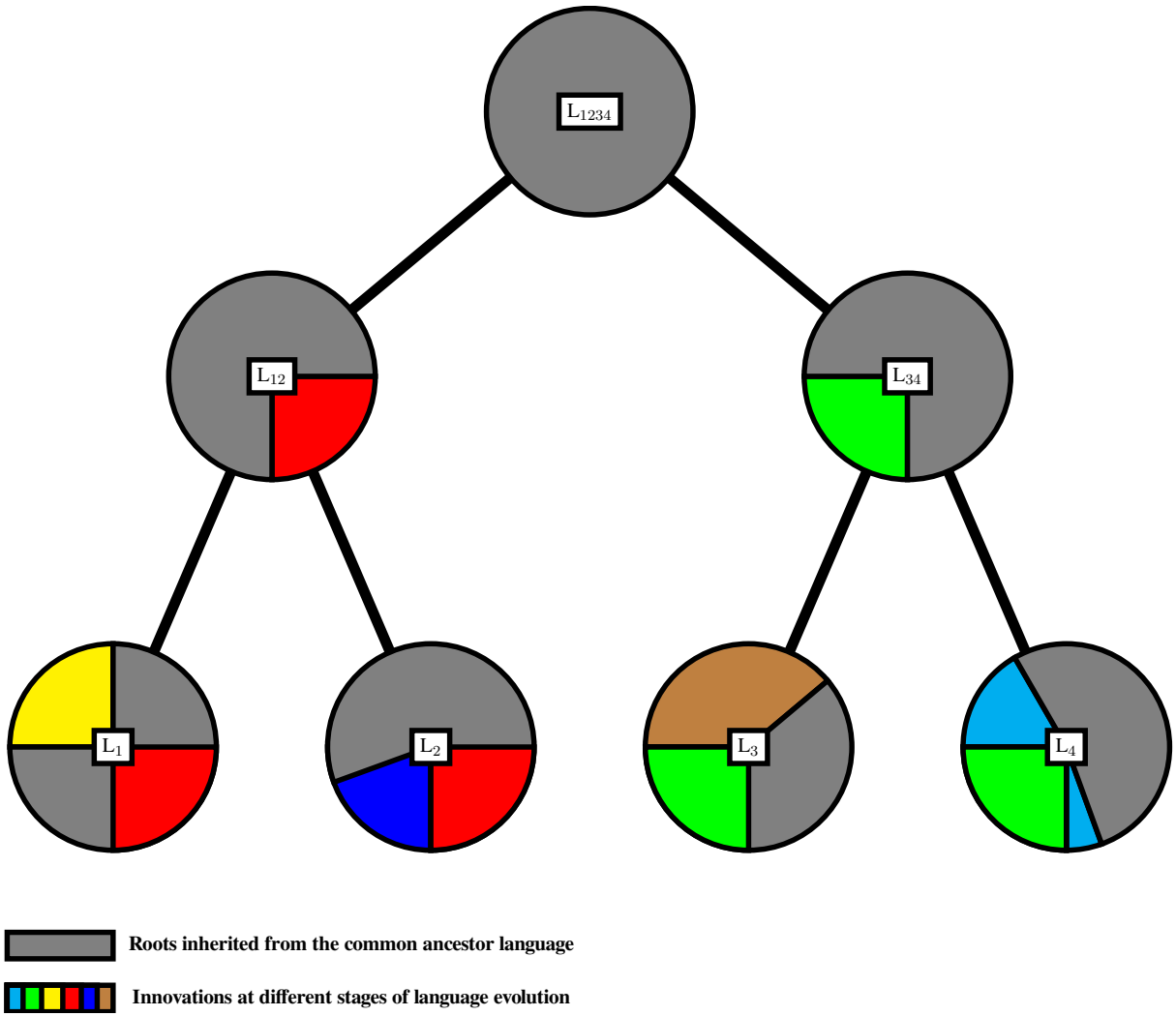


Datasets for the Separation Base Method

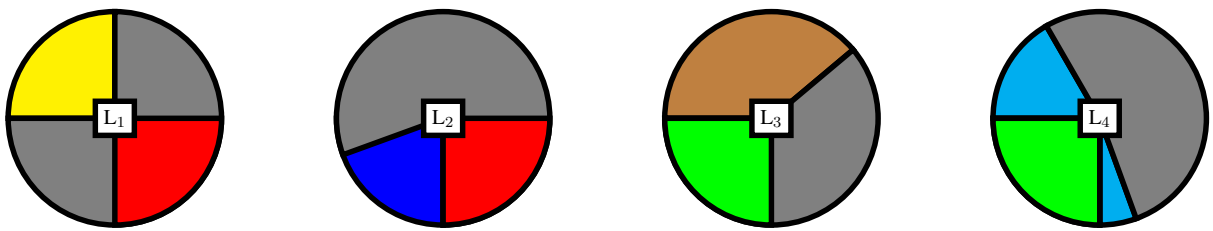
Language	Value	Coding
Proto	*h_2*ent-*	1
Hittite	hant-	1
Old Indian	ánti	1
Avestan	-	0
Armenian	-	0
Greek	antí	1
Slavic	-	0
Baltic	ãnt-i	1
Germanic	*anθ-ia	1
Latin	ante	1
Celtic	*antono	1
Albanian	-	0
Tokharian	ānt	1

Table: Coding of data according to the *Separation Base Method*

Evolutionary Model of Etymostatistics



Evolutionary Model of Etymostatistics



Datasets for Etymostatistics

1. Take whatever text you like for a given language and select from it all non-borrowed lexical roots.
2. Exclude all prefixes, suffixes and proper names and count each root only once.
3. Take this set of roots and look, with help of etymological dictionaries, for each root, whether it has a reflex in other genetically related languages you want to investigate.
4. Compute the similarity of the text-language to the other languages by calculating the percentage of roots reflected in the other languages.
5. Repeat the procedure for the other languages you want to investigate by changing the text-language and selecting different texts for the investigation.

Datasets for Etymostatistics

"Das kräftige Wirtschaftswachstum [...] [hat] die Stimmung der Verbraucher [...] weiter aufgehellt." (Spiegel ONLINE, 2010/08/26)[1]

Word	Meaning	"Lemma"	Root	Reflex	Coding
Das	"that"	das	*þat	that	1
kräftige	"strong"	Kraft	*kraftiz	craft	1
Wirtschaftswachstum	"economic growth"	Wirt	*werđuz	-	0
hat	"has"	haben	*xaƀēnan	to have	1
[die]	= das				
Stimmung	"mood"	Stimme	*stemnō	-	0
[der]	= das				
Verbraucher	"consumer"	Brauch	*brūkanan	to brook	1
weiter	"further"	weit	*wīđaz	wide	1
aufgehellt	"brighten"	"hell"	OHG *hellan*	-	0

[1] Translation: "The strong economic growth has further brightened the mood of the customers."

HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

Phylogenetic Reconstruction

Distance-Based Methods

Convert the binary data into distances, and analyze it with help of common cluster algorithms (e.g. Neighbor-Joining, cf. Saitou & Nei 1987; UPGMA, cf. Sokal & Michener 1958).

Character-Based Methods

Take the binary form of the data, and analyze it with help of specific algorithms which explain the distribution of characters according to certain evolutionary models (e.g. probabilistic models, cf. Ronquist 2003; parsimony models, cf. Camin & Sokal 1965).

Comparison of the Models

	Separation Base Method	Etymostatistics
Evolutionary Model	Root loss	Root loss and gain
Data	Complete etymological dictionaries listing all reconstructable roots of a proto-language	Random samples of roots extracted from texts or word-lists
Reconstruction	Quasi-distances based on the assumption that the root reflexes in the descendant languages are hypergeometrically distributed	Uncorrected distances (Percentages of common character states)

Testing the Methods

- Simulations of the Evolutionary Models
- Testing the Models on Real Data

Simulations of the Evolutionary Models

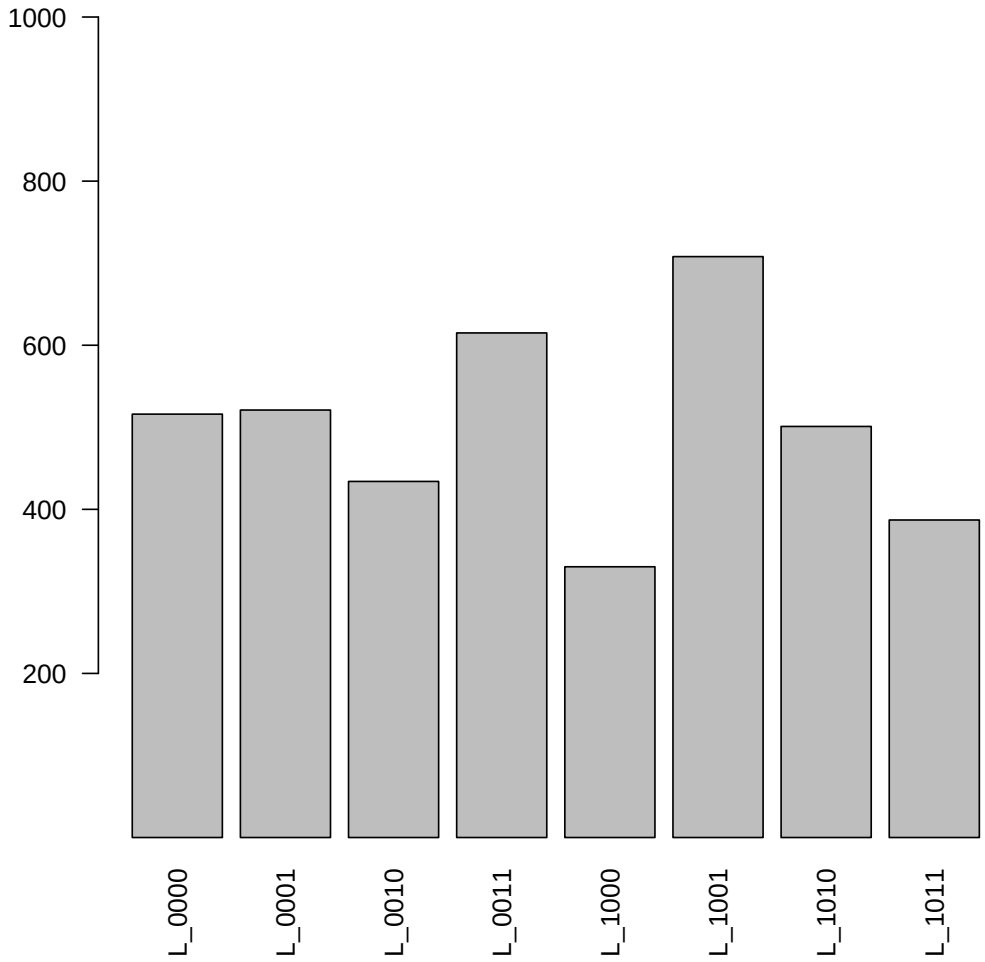
+++ short description of the programs +++

Simulations of the Evolutionary Models

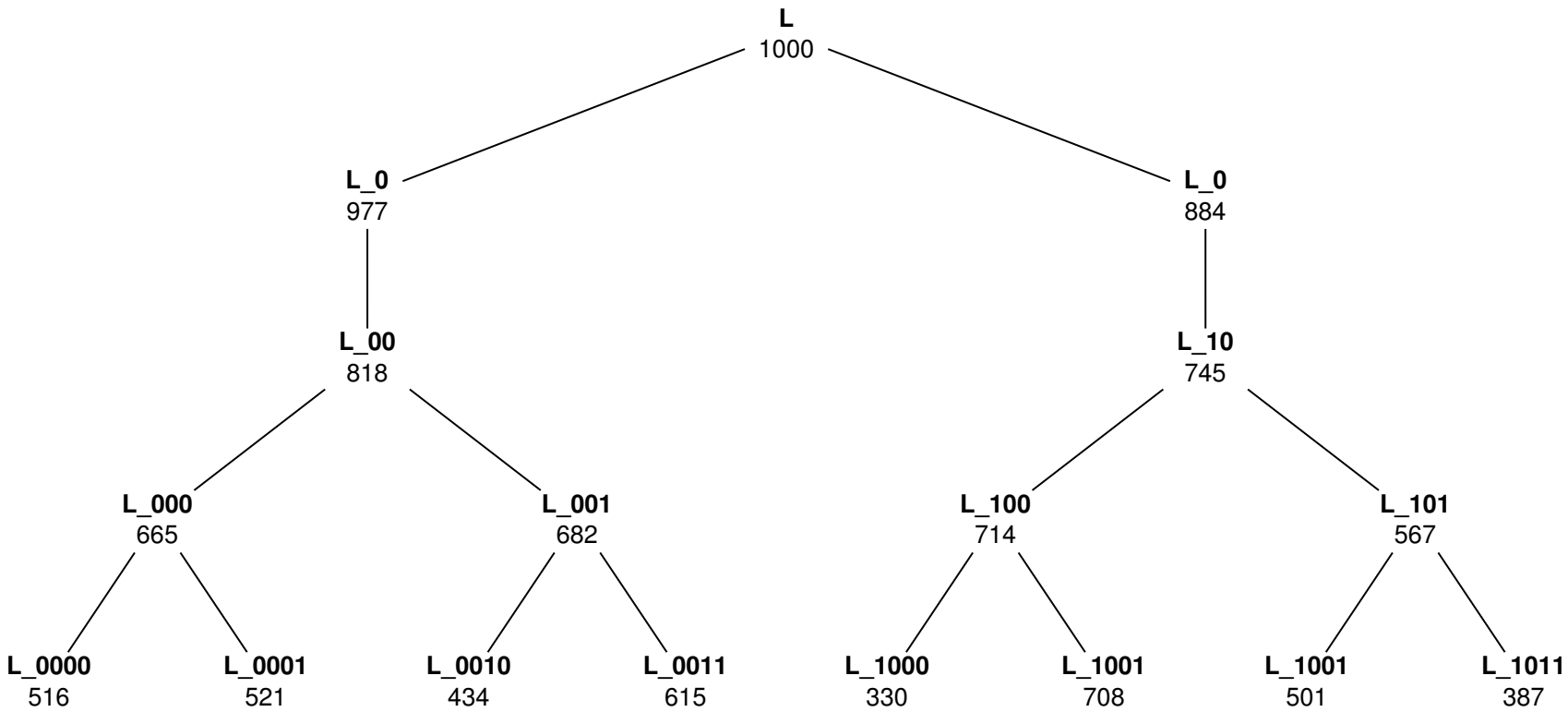
Python Program for the Simulation of the Models

- Program starts with one language *L*.
- Language goes through different *generations of change*.
- A *generation of change* is characterized by a possible split of the language into two descendant languages and a random amount of root-loss (Separation Base Method) or root-loss and root-gain (Etymostatistics).
- The result is a certain amount of descendant languages in the last *generation of change* and a specific distribution of roots among these languages.

Simulations of the Evolutionary Models



Simulations of the Evolutionary Models



Simulations of the Evolutionary Models

Testing the Separation Base Method

+++ description of the test+++

Testing the Separation Base Method

+++ graphic/tree +++

Testing the Separation Base Method

+++ graphic/lexstat/stefenelli+++

Testing the Separation Base Method

+++ zusammenfassen der Resultate+++

Testing Etymostatistics

+++ description of the test+++

Testing Etymostatistics

+++ graphic/results+++

Testing Etymostatistics

+++ zusammenfassen der resultate+++

Conclusion

- Model-Internal Problems
- Models and Reality

Model-Internal Problems

+++ Information loss in the models +++ +++ more rigid testing of the appropriate method for reconstruction +++

Models and Reality

+++ split as the key assumption
+++ evolution is not always tree-like
+++ datasets are problematic