Do Roots Really Grow Trees?

Quantitative Root-Based Approaches in Historical Linguistics

Hans Geisler, Johann-Mattis List *
Institute for Romance Languages and Literature
Heinrich Heine University Düsseldorf

SLE – 43rd Annual Meeting, Vilnius, 2 – 5 September 2010

1 Introduction

1.1 Comparison and Reconstruction

Goal of Comparison

- Comparative linguistics tries to reconstruct how genetically related languages evolved from a common ancestor language.

Comparison and Reconstruction

- Up to now, it is not clear, whether the comparison of languages should be based on phonetical, morpho-syntactical or lexical features, or a mixture of all.
- Phonetical and morpho-syntactical criteria prevailed in 19th century linguistics.
- From the 1950s onwards, there was an ever-growing tendency to use lexical comparison as the basis of phylogenetic reconstruction.
- The dominant model was lexicostatistics (Swadesh 1950, 1952 & 1955).
- Lexicostatistics has some severe methodological and practical drawbacks (cf. Geisler & List 2009) and we should try to improve it by a root-based approach.

1.2 Lexicostatistics vs. Root-Based Approaches

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

Table 1: Root-Based Methods vs. Lexicostatistics

Apparent Advantages of Foot-Based Approaches over Lexicostatistics

- Root-based approaches do not depend on the basic vocabulary assumption.
- Use of roots, i.e. to account for regular formal and semantic correspondences, gives a more fine-graded analysis of phylogenetic relationships.

*The research leading to this talk was carried out in the research project "Evolution and Classification in Biology, Linguistics and the History of Science (EvoClass: http://www.evoclass.de)" funded by the German Federal Ministry of Education and Research (BMBF). If You have any further questions, feel free to contact us under: listm@phil.uni-duesseldorf.de

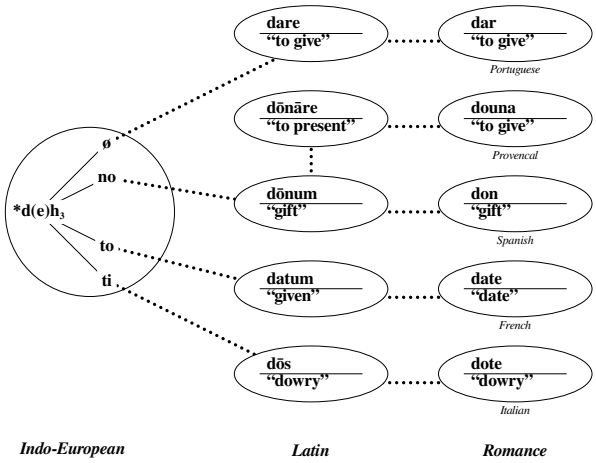

Figure 1: The Root-Concept in Historical Linguistic

Concept	Latin	Italian	Romanian	Spanish	French
BIRD	avis "bird"	ucello "bird"	pasăre "bird"	pájaro "bird"	oiseau "bird"
	1	1	2	2	1

Table 2: Lexicostatistical Analysis of Cognacy in Romance Languages for the Concept BIRD

Latin	Meaning	Italian	Romanian	Spanish	French
passer	"sparrow"	passero "sparrow"	pasăre "bird"	pájaro "bird"	passereau "little bird"
		1	1	1	1
avis	"bird"	ucello "bird"	-	ave "big bird"	oiseau "bird"
		1	0	1	1

Table 3: Root-Based Analysis of Descendent Words of Latin *passer* "sparrow" and *avis* "bird" in Romance Languages

2 Two Different Root-Based Approaches

2.1 The Separation Base Method (Holm 2000 & 2008)

The Evolutionary Model of the Separation Base Method

- The theoretical basis of Holm's (2000) method is a model of language change where language evolution is described as a process of random root loss in descendent languages after the split of the ancestor languages.
- The result is a distribution of roots which all were present in the ancestor language within the descendant languages.

Datasets for the Separation Base Method

Language	Value	Coding
Proto	*h_2ent-	1
Hittite	hant-	1
Old Indian	ánti	1
Avestan	-	0
Armenian	-	0
Greek	antí	1
Slavic	-	0
Baltic	ānt-i	1
Germanic	*anθ-ia	1
Latin	ante	1
Celtic	*antono	1
Albanian	-	0
Tokharian	ānt	1

Table 4: Coding of data according to the *Separation Base Method*

2.2 Etymostatistics (Starostin 2000[1989])

Evolutionary Model of Etymostatistics

- In contrary to the loss-only model proposed by Holm (2000 & 2008), the model proposed by Starostin (2000) incorporates innovations.
- Hence, language evolution is described as a process of root loss and root gain.
- Starostin further assumes that the process of root loss and gain is not driven by random but by regular forces.

Datasets for Etymostatistics

1. Start with a set of words (e.g. a list of translated basic concepts) of a given language where all borrowings are excluded.
2. Extract the roots from the words with help of etymological dictionaries of the given language.
3. Take this set of roots and look, with help of etymological dictionaries, for each root, whether it has a descendent word in other genetically related languages that shall be compared.
4. Repeat the procedure for the other languages that shall be investigated by changing the basic-language. (cf. Starostin 2000)

Word	Meaning	Root	English	Swedish	Dutch
groß	"big"	PGM *grauta- "groat; big"	great	gröt "pulp"	groot "big"
			1	1	1
Vogel	"bird"	PGM *fugla- "bird"	fowl	fågel "bird"	vogel "bird"
			1	1	1
schwarz	"black"	PGM *swarta- "black"	-	svart	zwart
			0	1	1
Feuer	"fire"	WGM *fewur- "fire"	fire	-	vuur
			1	0	1
viel	"much"	PGM *felu- "much"	-	-	veel
			0	0	1

Table 5: Exemplary Etymostatistical Analysis for Four Germanic Languages with German as Basic Language

2.3 Comparison of the Approaches

	Separation Base Method	Etymostatistics
Evolutionary Model	Root loss	Root loss and gain
Data Basis	Complete etymological dictionaries listing all reconstructable roots of a proto-language.	Random samples of roots extracted from texts or word-lists, analyzed with help of different etymological dictionaries.
Method of Reconstruction Proposed by the Author	Pairwise quasi-distances of the languages (based on the assumption that the root reflexes in the descendant languages are hypergeometrically distributed) are analysed with help of a specific clustering algorithm.	Uncorrected distances (Percentages of common character states) are clustered with a cluster method assuming an evolutionary clock (e.g. UPGMA).

Table 6: Comparison of the Two Approaches

3 Testing the Different Root-Based Approaches

3.1 Testing the Separation Base Method

Data Set and Analyses

- **Dataset:** Stefenelli's (1992) collection of the 1000 most frequent Latin words and their reflexes in nine Romance languages (Romanian, Sardinian, Portuguese, Spanish, French, Occitan, Catalan, Rhaeto-Romance, Italian).
- **Analyses:** Cluster analyses (Neighbor-Joining, cf. Saitou & Nei 1987) based on different distance measures (Cosine distance, Holm's N-values converted to distances), Bayesian analysis using the MrBayes software package (Ronquist & Huelsenbeck 2003).

Results of the Analysis

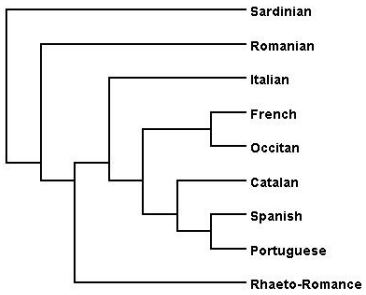

Figure 2: Bayesian Analysis of Stefenelli (1992)

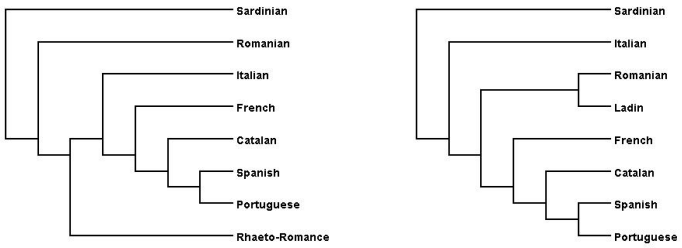

Figure 3: Separation Base Method (Left) vs. Lexicostatistics (Right)

Comparison with the Traditional View on Romance Language's Phylogeny

- Comparing the results for lexicostatistics (Gray & Atkinson 2003) and our analysis of Stefenelli (1992), we clearly see that the method reproduces the traditional view of Romance linguistics much better than lexicostatistics.
- The grouping of Catalan and Occitan in different branches (Gallo-Romance vs. Ibero-Romance), however, is not in concordance with the view of many scholars in Romance linguistics who tend to group both languages together.

3.2 Testing Etymostatistics

Dataset and Analyses

- **Dataset:** Etymostatistical analysis of 7 Romance languages (Sardinian, Romanian, Portuguese, Catalan, Spanish, Italian, French) based on basic vocabulary lists of 110 items translated into the respective languages (Starostin 2008).
- **Analyses:** Cluster analyses (Neighbor-Joinging) based on different distance measures (Cosine distance, uncorrected distances), Bayesian analysis using the MrBayes software package (Ronquist & Huelsenbeck 2003).

Results of the Analysis

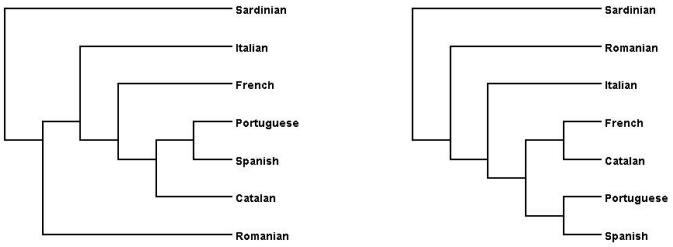

Figure 4: Distance- (Left) and Bayesian (Right) Analysis of the Data

Comparison of the Results with the Traditional View on Romance Language's Phylogeny

- Distance- and Bayesian analyses of the data result in nearly equivalent tree-topologies.
- The results for the analysis come much closer to the traditional view on the phylogeny of the Romance languages.
- The different grouping of Catalan in the Neighbor-Joining and the Bayesian analysis reflects the differing opinions regarding the position of the language within Romance linguistics. Our analysis does not leave a conclusive result here.

4 Conclusion

Do Roots Grow Trees?

- Root-based approaches applied to Romance language data show a clear improvement over lexicostatistical approaches.
- Nevertheless, root-based approaches are no miracle cure against the well-known and longstanding problems of historical linguistics.
- As in all approaches which are based on the assumption that language evolution can be simply characterized by a process of split and divergence, there remains a considerable amount of uncertainty and variation within the reconstructed phylogenies.

Models and Reality

- Both models assume that languages split (dichotomously) into daughter languages – language contact and language mixing is neglected.
- The transfer of phylogenetic methods with tree-like genetic models as background supports this 19th century approach to language evolution.
- These assumptions do not fit real language evolution: only in distantly related standardized written languages we have clear-cut divergences between languages. In all other cases we have continuity between language varieties in space.
- This complex linguistic reality cannot be captured by tree-like structures, the intricate relationships between linguistic varieties seem to be better described by networks.

$\Rightarrow$ Instead of sticking to trees as the only way of representing language history, we need new models which reflect the vertical as well as the horizontal aspects of language evolution.

References

Geisler, Hans, & Johann-Mattis List, 2009. Beautiful trees on unstable ground. Notes on the data problem in lexicostatistics. Presentation held at the Arbeitstagung der Indogermanischen Gesellschaft 2009: Die Ausbreitung des Indogermanischen. Thesen aus Sprachwissenschaft, Archäologie und Genetik. Würzburg. 24. - 26. September 2009. Handout available under: www.evoclass.de.

Gray, Russell D., & Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426.435–439.

Holm, Hans J. 2000. Genealogy of the main indo-european branches applying the separation base method. *Journal of Quantitative Linguistics* 7.73–95.

Holm, Hans J. 2008. The distribution of data in word lists and its impact on the subgrouping of languages. In *Data Analysis, Machine Learning, and Applications. Proceedings of the 31th Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, März 7-9, 2007*, ed. by C. Preisach, H. Burkhard, L. Schmidt-Thieme, & R. Decker, 629–636, Heidelberg; Berlin.

Ronquist, Frederik, & J. P. Huelsenbeck. 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19.1572–1574.

Saitou, N, & M Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4.406–425.

Starostin, George, 2008. Tower of babel. an etymological database project. http://starling.rinet.ru/main.html.

Starostin, Sergej Anatol'evič. 2000. Comparative-historical linguistics and lexicostatistics. In *Time depth in historical linguistics*, ed. by Colin Renfrew, April McMahon, & Larry Trask, Papers in the prehistory of languages, 223–265. Cambridge: The McDonald Institute for Archaeological Research. Translation: Peiros, Ilia. Originally published in 1989. Online available under: http://starling.rinet.ru/Texts/method.pdf.

Stefenelli, Arnulf. 1992. *Das Schicksal des lateinischen Wortschatzes in den romanischen Sprachen*. Passau: Rothe.

Swadesh, Morris. 1950. Salish internal relationships. *International Journal of American Linguistics* 16.157–167.

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* 96.452–463.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21.121–137.