The Biological Lumbering of Linguistic Trees

A Comparison of Traditional and "New School" Approaches to Phylogenetic Reconstruction in Historical Linguistics

Jens Fleischhauer, Johann-Mattis List
Heinrich Heine University Düsseldorf

June 30, 2010

1 Introduction

"Old" and "New School" Approaches

- Historical processes are inferred from synchronic structures.
- This is a general inference pattern, not unique to linguistics but also used in e.g. biology and geology (cf. Christy 1983).
- Intra- and interdisciplinarily different methods for such an inference exist.
- (Marris 2008) claims that the methods of historical linguistics are "Old School" while the biological methods (applied to language data) are "New School".
- **Leading question of the talk:** What are the differences between "Old" and "New School"?

2 Proof of Genetic Relationship

2.1 Lexical Similarities

When considering lexical resemblences which can be observed between languages, one can roughly distinguish three different reasons for these resemblences to occur. They can be

1. coincidental (they are simply due to chance),
2. natural (they are due to general patterns of denotation observable in all human languages),
3. historical (they are due to a shared history of the languages under observation).

The latter kind of resemblences can be further divided in genealogical and non-genealogical resemblences, i.e. resemblences due to a shared genealogical history and resemblences due to a shared history of contact.

2.2 Sound Correspondences as the Primary Notion of Similarity

1. In genetically related languages we can find a certain amount of semantically similar words which are structurally similar in so far as certain sounds, which do not necessarily have to be identical, correspond to each other in so far as they regularly occur in the same position of the words and hence carry the same distinctive function.
2. As opposed to other kinds of language change, sound change happens to be quite *regular*. This can be seen when comparing older written sources of languages and their modern descendants. The regularity of sound change is reflected in the fact that, once a certain sound changes, this change does not only affect a couple of words but may affect *large parts of the lexicon* of a language.

Given the observation *2* one can conclude that the functional similarity of certain sounds in certain languages is due to genealogical relatedness of these languages: The result of sound changes in languages which diverged from each other shows up in their lexicon as sounds which correspond to each other functionally in words inherited from the ancestor language.

3 Methods for Phylogenetic Reconstruction

3.1 The Comparative Method

Key assumptions of the comparative method:

- Innovations languages which are not reflected in other genetically related languages indicate that the respective languages have evolved separately.
- Shared innovations allow to reconstruct a phylogenetic tree which depicts the process of how an ancestor language split into several descendents.

Working procedure:

1. **Proof of Genetic Relationship:** Carry out a comparative analysis of languages previously assumed to be genetically related. Search the languages for possible cognates, thereby identifying regular sound correspondences and carrying out putative reconstructions.
2. **Shared Innovations:** Search the languages for shared innovations.
3. **Phylogenetic Reconstruction:** Reconstruct a language tree which explains the identified shared innovations of the different subgroups in a most parsimonious way.

Phylogenetic Reconstruction:

No.	Innovation	Languages	Example	Reference
1	1st Germ. Cons. Shift	Germanic Languages	PIE *p > PGM *f	Fox 1995
2	2nd Germ. Cons. Shift	High German	PGM *p > GER pf	Trask 2000
3	Vowel Syncopation	Western Romance	LAT cĭnĕrĕm > FRE cendre	Geisler 2008

3.2 Lexicostatistics

Key assumptions (cf. Geisler & List 2009, Dyen *et al.* 1992):

- The lexicon of every human language contains words which are relatively resistant to borrowing and relatively stable over time due to the meaning they express: these words constitute the *basic vocabulary* of languages.
- The process of replacement of words belonging to the realm of *basic vocabular* is reflected in the amount of shared cognates in the *basic vocabulary* of languages.
- Shared cognates in the *basic vocabulary* of genetically related languages reflect their degree of genetic closeness and allow to reconstruct their phylogeny.

Working Procedure:

1. **Swadesh-List Compilation:** Compile a list of basic vocabulary items (a Swadesh list).
2. **Swadesh-List Translation:** Translate the items into the languages that shall be investigated.
3. **Cognate Judments:** Search the language entries for cognates.
4. **Cognate Percentages:** Compute percentages of shared cognates for every language pair.

5. **Subgrouping:** Construct a graphical representation out of the information on percentages of shared cognates (this is usually, but not necessarily, a genealogical tree).

Phylogenetic reconstruction[1]**:**

	German	English	Italian	French
German	100	82	40	38
English	82	100	40	36
Italian	40	40	100	94
French	38	36	94	100

3.3 The Method of the "New School"

Key assumptions[2]**:**

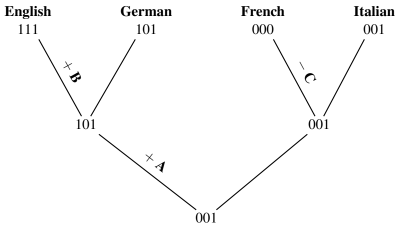
- Borrowing is rare in basic vocabulary.
- There are significant similarities between linguistic and biological evolution which allow the same methods to be applied for phylogenetic reconstruction.
- The distribution of cognate sets over a sample of languages can be used to model linguistic evolution within a phylogenetic tree with the help of phylogenetic software packages used in biology.

Working procedure:

1. **Swadesh-List Compilation:** Compile a list of basic vocabulary items (a Swadesh list).
2. **Swadesh-List Translation:** Translate the items into the languages that shall be investigated.
3. **Cognate Judgments:** Search the language entries for cognates.
4. **Binarization of Cognate Information:** Convert the data into a binary matrix reflecting for each cognate set its presence (1) or absence (0) in the respective language.
5. **Subgrouping:** Use phylogenetic software to construct a phylogenetic tree which explains the distribution of cognate-sets best.

Phylogenetic reconstruction:

Char.	English	German	French	Italian
A	1	1	0	0
B	1	0	0	0
C	1	1	0	1



4 Comparison of the Methods

4.1 Similarities

- inference of language phylogenies from synchronic data
- reconstruction of evolutionary trees depicting the historical processes
- inference of language splits

[1] Percentages were computed with STARLING (cf. Starostin no date).
[2] Cf. (Gray & Atkinson 2003), (Atkinson & Gray 2006).

4.2 Differences

- qualitative vs. quantitative data
- innovations vs. character distributions

Lexicostatistics vs. the “New School” approach:

- replacement vs. gain/loss

5 Conclusion

Different Theories – Same Problems

- Conflicting data (as a result of undetected borrowing events or linguistic convergence) are problematic for all methods.
- The evolutionary process is supposed to be tree-like and reticulate evolution is rejected.

The Need for a New “New School” Approach

- Contradicting Marris (2008), the “New School” approach is not superior to the “Old School” approaches.
- A real “New School” account should allow to solve the well known problems and abandon the *a priori* assumption of tree-likeness.

References

Atkinson, Quentin D., & Russell D. Gray. 2006. How old is the Indo-European language family?Illumination or more moths to the flame? In *Phylogenetic methods and the prehistory of languages*, ed. by Peter Forster & Colin Renfrew, McDonald Institute monographs, 91–109, Cambridge UK , Oxford UK , Oakville CT USA ,. McDonald Institute for Archaeological Research; Distributed by Orbow Books.

Christy, Craig. 1983. *Uniformitarianism in linguistics*, volume 31 of *Studies in the history of linguistics*. Amsterdan and Philadelphia: John Benjamins.

Dyen, Isidore, Joseph B. Kruskal, & Paul Black. 1992. An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society* 82.iii–132.

Fox, Anthony. 1995. *Linguistic reconstruction: An introduction to theory and method*. Oxford University Press.

Geisler, Hans. 2008. Konvergenz und divergenzphänomene in der romania: Lautentwicklung. In *HSK Romanische Sprachgeschichte, Bd. 3*, ed. by G. Ernst, M. Glessgen, C. Schmitt, & W. Schweickard. de Gruyter.

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.

Marris, Emma. 2008. The language barrier. *Nature* 453.446–448.

Starostin, Sergej Anatol’evič, no date. The STARLING database program. Online available under http://starling.rinet.ru.

Trask, Robert L. (ed.) 2000. *The dictionary of historical and comparative linguistics*. Edinburgh: Edinburgh Univ. Press.