Syntactic Trees and Parsimony Analysis

Andrea Ceolin (andceo88@gmail.com) – Università degli Studi di Trieste

Summary

In the last years linguists have been applying to their data computational methods developed by biologists to generate phylogenetic trees (e.g. Dyen et al. 1992, Ringe et al. 2002). Recently, such methods have been successfully applied to syntactic parameters (**Parametric Comparison Method**, **PCM**, Longobardi and Guardiano 2009).

This poster compares two trees, one built through KITSCH and the other through PAUP*. The comparison addresses the differences between distance-based and character-based methods (here, **maximum parsimony**, **MP**, Edwards and Cavalli-Sforza 1963) when applied to syntactic parameters.

The results show that, while neither tree displays borrowing effects, homoplasy impacts negatively on MP.

Introduction

PCM used distances between languages for classification purposes. In particular, the **Jaccard-Tanimoto** distance has been used, to deal with cases of non-informative parameters (i.e. **implicated parameters**, Longobardi & Guardiano 2009).

The programs which produced the best results are those relying on the molecular-clock-hypothesis (i.e. KITSCH and UPGMA).

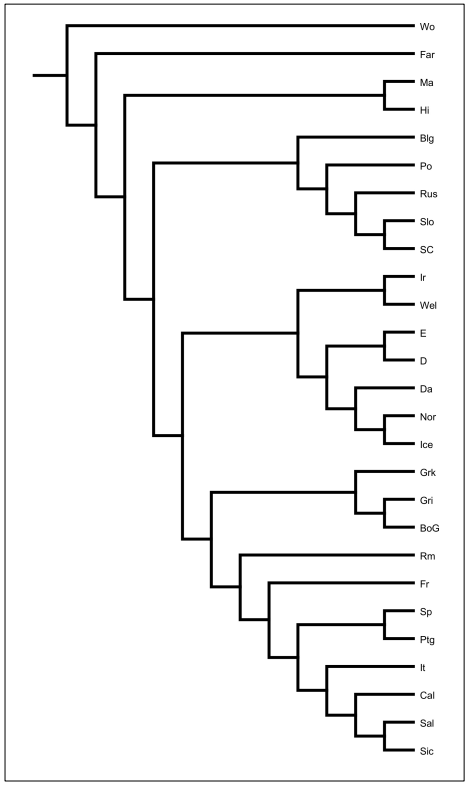
There are at least two reasons to expect that parsimony analyses could work as well for syntactic parameters:

1) Parameters seem to follow the **Inertial Principle** (Longobardi 2001, Keenan 2002), and in particular syntactic evolution revealed to be slower than lexical one (cf. Longobardi & Guardiano 2009). This is coherent with a parsimony approach.

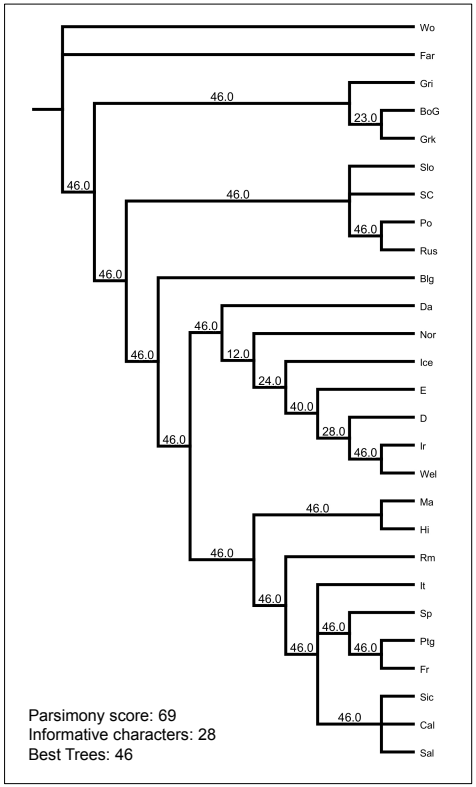
2) Experimental work is showing that maximum parsimony should correctly deal with the implicational structure held by parameters (Ceolin *in prep.*).

A parsimony analysis has been performed through PAUP* on the 56-parameters full database of Longobardi et al. (2013), and the resulting tree has been compared with the KITSCH tree from the same dataset.

Distance Tree from KITSCH



Parsimony Tree from PAUP*



Results and Discussion

PAUP* performs worse than KITSCH in at least three points, where parameters displaying **parallel development** are clearly detectable. There could be at least three reasons for this:

1) Parameters are binary. It means that parallel development is more frequent here than in lexical domains.

2) Data are insufficient for a parsimony analysis.

3) Distance-based methods take advantage of parameters displaying a single value versus all the languages, while these parameters are considered uninformative by PAUP*. In general, parameters isolating certain groups (e.g. Indic) are more informative in KITSCH.

Conversely, there are at least two cases where one should expect PAUP* to perform better than KITSCH. Suppose that a language is exposed to massive external borrowing or isolation and evolves differently from its sisters. KITSCH would detach it from its family while PAUP* would keep the family together following the **shared innovations** of its languages. This is something expected here for Farsi with respect to the Indic languages, but it must be the case that homoplastic effects are preventing a correct classification.

Notice that syntactic borrowing in the Balcanic area and in Southern Italy does not affect the trees.

Conclusions

Distance-based methods proved to be more suitable for syntactic parameters than those character-based, at least in dealing with homoplasy (especially parallel development). There are some cases where parsimony should work better, but at the moment none is attested. Regarding future research, there are at least three ways to improve these analyses:

1) More data need to be added to the database to reach reliable results.

2) Horizontal effects must be detected and singled out at least in the parsimony search.

3) Parameters must be investigated in order to get information about weights, ancestral states and directionality of change.

Methods

The low number of data allows running a **branch-and-bound** search, so that the results are expected to be correct. The trees represent the following IE languages (plus Wolof, **Wo**, used as outlier):

Sicilian: **Sic;** Northern Calabrese: **Cal**; Italian: **It**; Salentino: **Sal**; Spanish: **Sp**; French: **Fr**; Portuguese: **Ptg**; Rumanian: **Rm**; Bovese Greek of Southern Calabria: **BoG**; Grico, i.e. Greek of Salento: **Gri**; standard Greek: **Grk**; English: **E**; German: **D**; Danish: **Da**; Icelandic: **Ice**; Norwegian: **Nor**; Bulgarian: **Blg**; Serbo-Croat: **SC**; Slovenian: **Slo**; Polish: **Po**; Russian: **Rus**; Irish: **Ir**; Welsh: **Wel**; Farsi: **Far**; Marathi: **Ma**; Hindi: **Hi**.

Comments

This tree is consistent to both the classical ones and those resulted from modern lexicostatistical analyses.

There are two points of major differences:

1) Farsi, **Hindi** and **Marathi** are close to each other but not clustered together, thus the Indo-Aryan cluster is not recognized. The acknowledge of the macro-family has also been a problem for Dyen et al. (1992) and its robustness has been occasionaly debated (Lazzeroni 1968).

2) Bulgarian is the outlier of the Slavic family, while one should expect it to be clustered with Slovenian and Serbo-Croat.

Comments

PAUP* tree differs from the previous one in the following points:

1) The Indo-Aryan puzzle is even more problematic, with Hindi and Marathi attracted by Western IE languages (especially Romance). Farsi seems to be attracted by Wolof.

2) Bulgarian is attracted by Western IE to the point that it goes out its main family, though being displaced close to it.

3) Celtic and Germanic, close to each other in the previous tree, collapse together.

Romance and Greek languages are correctly detected.

References

Dyen, I., Kruskal, J., Black, P., 1992. An Indoeuropean classification: a lexicostatistical experiment. *Transactions of the American Philosophical Society* 82 (5).

Edwards, A.W.F., Cavalli-Sforza L.L., 1963. The reconstruction of evolution. *Annals of Human Genetics* 27, 106-106.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Lazzeroni, R. 1968. Per una definizione dell'unita indo-iranica. *Studi e Saggi Linguistici*. Supplement to *L'Italia Dialettale* 8, 131-159.

Longobardi, G., Guardiano, C., 2009. Evidence for syntax as a signal of historical relatedness. *Lingua* 119 (11), 1679-1706.

Longobardi, G., Guardiano, C., Silvestri, G., Boattini, A., Ceolin, A., 2013. Toward a syntactic phylogeny of modern Indo-European languages. *Special Issue of Journal of Historical Linguistics 3:1*, 122-152.

Ringe, D., Warnow, T., Taylor, A., 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100 (1), 59-129.

Swofford, D. L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.Sinauer Associates, Sunderland, Massachusetts.