

Thesis Proposal:

Computational Evolutionary Linguistics

Tracy van Cort

Readers/Advisors: Prof. “Z” Sweedyk (HMC CS),
Prof. Kossuth (Pomona Linguistics), Prof. Levin (HMC Math)

19 September 2000

Introduction

As a math/linguistics dual thesis project, I would like to investigate the application of mathematical methods developed for evolutionary biology to problems in historical linguistics. This is an idea that intrigues me enough that I’m determined to work on it eventually, so why not now, when I’ve got access to lots of great resources, including really smart people who can give me feedback.

Background

Phylogenetic trees are used to describe the evolutionary history of a group of related species. Cladistics attempts to construct such trees by comparing homologous characters (traits with a common evolutionary origin). When two species (or groups of species) diverge on a trait, the tree branches. It is assumed that reversals (when a species’ trait changes away from an ancestor’s and then back to the ancestral state) are extremely rare, so the fewer reversals in a tree, the better. A perfect phylogeny is one in which no reversals occur.

It seems to me that methods used to construct trees for species of organisms should work for describing language change as well, but I would like to better understand the mathematics involved (more on this in the next paragraph). I also want to try applying cladistics to a group of languages whose history is already fairly well-known (the emergence of the Romance languages from Latin or the origins of English and others from ancient Germanic languages) and see whether the result matches the standard historical linguistic analysis.

Research Plan

“I would be very surprised if something like this hasn’t been done before,” I wrote back in April when I first came up with this project idea. As it turns out, a friend of Prof. Z’s has: Tandy Warnow, then at the University of Pennsylvania, now at the University of Texas, Austin. Warnow was a co-author on a paper that proved the perfect phylogeny problem (a special case of the problem driving cladistic methods of drawing phylogenetic trees) is NP-Complete (Bodlaender et al 1992). What little background reading I’ve done on the subject so far reveals that I have a *lot* of graph theory to relearn, but there are special cases of the perfect phylogeny problem that can be solved in polynomial-time, as

well as heuristics for tackling NP-Complete problems to the best of our capabilities. So there's plenty of math for me to learn about.

On the linguistic side of things, I have to learn a lot more about language change. For example, I was worried that reversal might be more common among languages than it is among species, but it turns out that sound changes, to give an example of an entire class of traits that's been pretty extensively documented, are pretty strictly one-way. So phonetic features will be useful data, and I'll be looking into syntactic and lexical changes as well (especially cognates: words that have the same linguistic origin, but may or may not have different meanings). One tricky problem that I don't yet know how to reconcile is the fact that these three kinds of changes tend to occur at very different rates (lexical changes can happen almost overnight, but changes to the grammar or phonetics of a language are much more gradual). Hopefully there's a nice clean answer to that (maybe even some cool math to back it up!)

Intended Reading

Right now I have a pretty extensive list of linguistic resources to investigate: Merritt Ruhlen and Derek Bickerton's work on the origins and evolution of the languages of the world, as well as Luca Cavalli-Sforza's *Genes, Peoples, and Languages* and *The History and Geography of Human Genes*, which trace historical patterns in linguistic and genetic traits. Based on these, I should be able to choose a good set of languages to compare and contrast. My mathematical resources are a bit more limited: right now what I really need is a better graph theory book than my Discrete Math notes from frosh year. From there I should be better able to work on the computational aspects of this problem (especially as we get to the NP-Completeness section of TheoComp).