

# Influences of Uncertainty Optimization on the Development of Language

**Christopher Dilley**

christopher.dilley@student.uni-tuebingen.de

**Erik Schill**

email@email.com

**Inna Pirina**

email@email.com

## Abstract

This is where an abstract would go.

## Background

In this paper, we aim to determine the linguistic relatedness of a number of languages in the same manner as described in Kirby and Ellison (2006), using an improved dataset.

Kirby and Ellison describe a process whereby a language is quantified as a matrix comparing the probability of confusing one word for another for every word pair in the language, and these matrices are compared and clustered based on their relatedness. By doing so, they achieved very encouraging results, automatically generating a phylogenetic tree of languages that closely resembles the generally accepted phylogeny; the languages all fell neatly into their respective language sub-families.

In employing a new set of data, we hoped to obtain equal or better results. This new data set contains more data than that used by the authors in the paper, and uses each word's phonemic transcription rather than its orthography to calculate confusion probabilities. We hypothesized that the language's phonemics would expose more meaningful relationships between words than orthography, as orthography can vary widely. For example, the English words 'though' and 'toe' have significantly different orthography, while being pronounced very similarly ([ðo:] and [to:]).

However, it is possible that this may give worse results as well. Orthography may better represent the ancestry and relatedness of individual words, even when their phonetics converge. The fact that 'though' and 'toe' are pronounced similarly may simply be due to random sound shifts causing phonetic convergence by chance, and this phonetic similarity may lead measures of word interrelatedness astray, negatively affecting intra-language comparisons.

This study seeks to identify which of these hypotheses is more likely.

**- PHONETICS vs. PHONEMICS vs. PHONOLOGY... WHICH IS CORRECT USAGE HERE?**

**- THIS SECTION COULD PROBABLY BE EXPANDED**

## Implementation

Our methodology follows much of the same procedures as Kirby and Ellison (2006). We developed our own tools for generating the lexical metrics of all languages and for comparing them as well, while using an external tool to render the phylogenetic trees from the resulting data.

As our input, we obtained data from the Indo-European Lexical Cognacy Database. This data set spans 208 word meanings across 52 languages (not all languages had data for all word meanings), and includes phonemic transcriptions for all words. We used these transcriptions instead of the word orthography for the inter-lexical comparisons, expecting to better detect similarity between words.

## Calculating Lexical Metrics

(Talk about this process of comparing all words inside each language to generate the matrices)

Hey, look at the lexical metric in Figure 1.

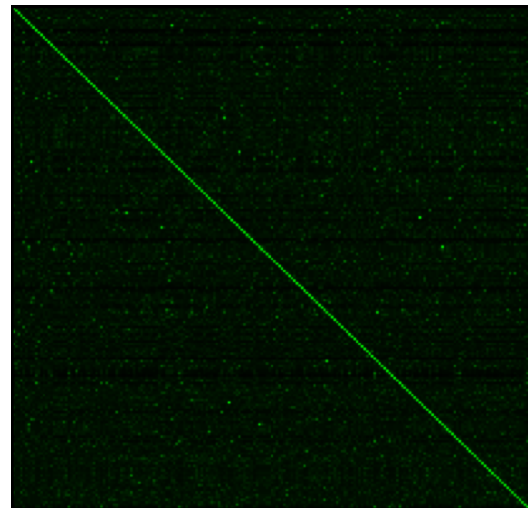


Figure 1: A visualization of the lexical metric for French.

## Calculating Language Distance

(Talk about the process of comparing these matrices, and how we got around the various difficulties of languages not having certain words or having multiple words for a single meaning)

## Constructing Phylogenetic Trees

Once having constructed a matrix that represents the distance between all language pairs in the collection, these results need to be converted into an analyzable form. Specifically, we want to visualize how close certain language pairs are related, and how they cluster together into larger groups. From this, we hope to see some of the same taxonomic relations that are generally accepted as language families and subfamilies.

In order to accomplish this, we began by using the *NEIGHBOR* program from the *PHYLIP* package (Felsenstein, 1989), as was done in the paper. This program takes a matrix input, and constructs a tree that represents the relatedness of all of the languages based on their distance measurement. This tree

is output in a text form (in the standardized 'Newick' format) that describes the nodes on the tree and the distance to their child nodes. Once having generated this, we used the Phylo-dendron tool from the University of Indiana to visualize this information as a proper tree (Gilbert, 1999).

This process is repeated separately for each of the distance measurement matrices (the KL distance matrix and the Rao distance matrix).

## **Results**

## **Discussion and Conclusion**

## **References**

- Felsenstein, J. (1989). *PHYLP (Phylogeny Inference Package) version 3.6*. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
- Gilbert, D. (1999). *Phylo-dendron*. Department of Biology, University of Indiana. Retrieved from <http://iubio.bio.indiana.edu/treeapp>
- Kirby, S., & Ellison, T. M. (2006). Measuring language divergence by intra-lexical comparison. In *Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics* (pp. 273–280).