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 phylogenic tree of languages that closely resembles the generally accepted phylogeny; the languages all fell neatly into their respective language subfamilies.

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 ([ðoː] and [toː]).

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 phylogenic 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.

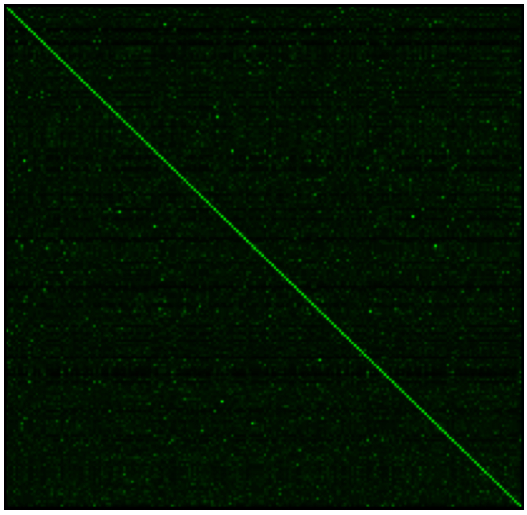

Figure 1: A visualization of the lexical metric for French.

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 Phylogenic 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 Phylodendron tool from the University of Indiana to visualize this information as a proper tree (Gilbert, 1999).

This process was repeated separately for both the KL distance matrix and the Rao distance matrix.

Results

The results from using Kullback-Liebler distance are found in Figure 2, and the results from using Rao distance are found in Figure 3.

In our results, it can be seen that the algorithm does a good job of grouping similar languages together in the tree. Well-accepted subfamilies of languages can be seen together. The Rao distance tree seems to do a better job of this grouping, seeing more definable groups with greater consistency.

Some outliers are evident in this tree. Danish, Norwegian, and Swedish find themselves far separated from their closely related Scandinavian languages (Faroese, Icelandic, etc.), and other Germanic languages (German, English, Dutch) are not grouped well with them at all. French and Catalan are also similarly separated from their other Romance counterparts (Latin, Spanish, Italian).

Outside of these outliers, the groupings seems quite sensible. Indo-Iranian languages (Urdo, Ossetic, etc.), Slavic languages (Russian, Czech, Polish, etc.), and most of the Romance languages (Spanish, Italian, etc.) are seen grouped together. However, the combining of these smaller groupings into larger groupings seems to not make as much sense.

Discussion

Overall, our results seem to reflect positively on the validity of the methodology. It appears to group similar languages together by an objective measure of the language into generally well-accepted families. Our results, however, seem weaker than those of Kirby and Ellison.

There are several routes that could be explored for improving the results. We could have experimented more with how to handle meanings with multiple words, perhaps by taking the minimum confusion probability value or average of the values. This likely would only have a minor influence on the results, but perhaps noticeable.

The data may also have been lacking, given that so few of the word meanings were found in the data set for all languages. A larger and more consistent spread of word meanings in the languages may have resulted in more representative lexical metrics and better comparisons between them for measuring distance.

Another avenue for improvement may lie in that our means of calculating edit distances between phonemic transcriptions was not as effective as using orthography, as such transcriptions are not consistent and can be difficult to get an accurate distance measurement out of. Dealing with how these transcriptions were encoded and comparing them is also challenging, and we may have had better results using third-party libraries that handled this better. It may even be the case that orthography does a better job of comparing the relatedness of words than phonemic transcriptions, as the orthography could better capture the ancestry of the word.

Conclusion

This paper presented a reimplementation of an automatic language phylogeny constructing algorithm described by Kirby and Ellison (2006), which uses inter-language comparisons of words to build a lexical network that is compared against other languages, measuring their similarity. This was conducted on a new data set, using phonemic transcriptions instead of orthography for comparing individual words. We obtained encouraging results that seemed to group languages well, but the results were likely less strong than those from Kirby and Ellison. Some possible explanations for this were identified, and opened paths for further exploration.

References

Felsenstein, J. (1989). *PHYLIP (Phylogeny Inference Package) version 3.6.* Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Gilbert, D. (1999). *Phylodendron.* 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).

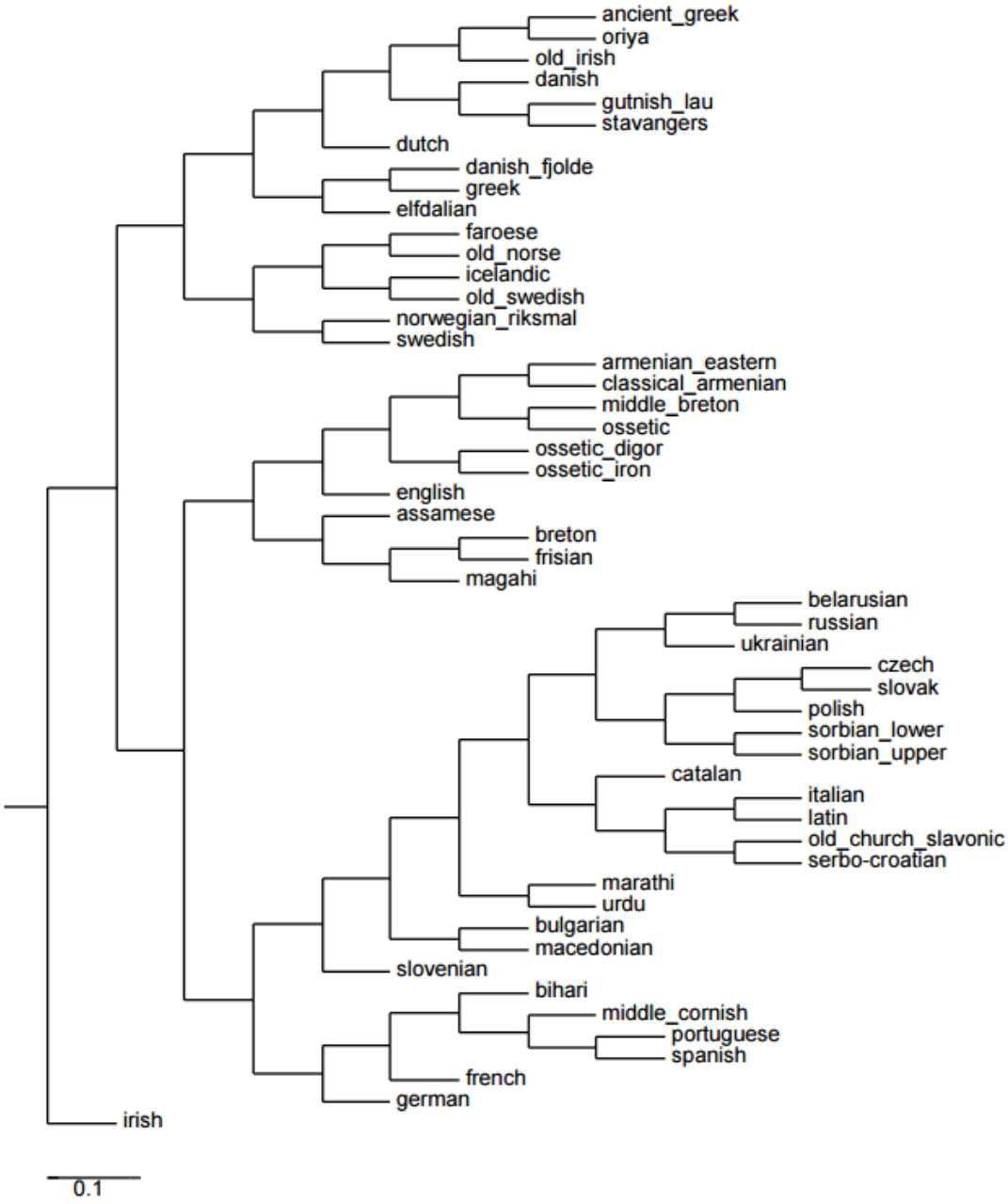

Figure 2: Tree visualization of KL-distance measurements

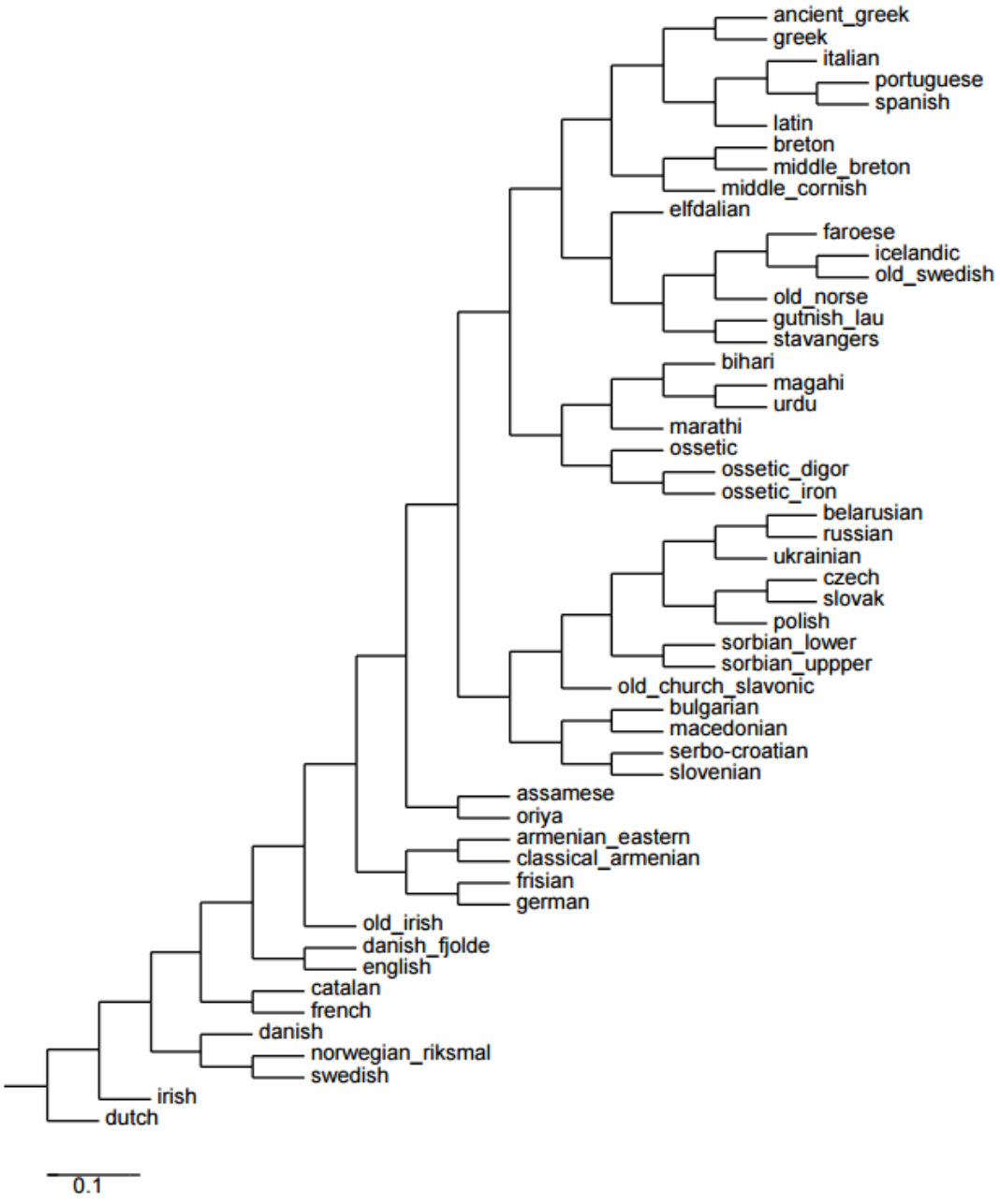

Figure 3: Tree visualization of Rao-distance measurements