Reimplementation of an Intra-Lexical Technique for Determining Language Similarity

Christopher Dilley
christopher.dilley@stu~.de

Erik Schill
erik.schill@stu~.de

Inna Pirina
inna.pirina@stu~.de

Abstract

Determining the ancestry and relationship of languages is one of the primary challenges that historical linguists face. The task is daunting and tedious, and is prone to the subjectivity of the researcher. For these reasons, it has been of great interest to develop means of automating the process. In this paper, we discuss a reimplementation of one such process, which builds a representation of the internal structure of a language to compare against those of other languages and establish a measure of their relatedness.

Background

This paper aims 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.

When linguists perform the task of building phylogenic tree, they typically do so by comparing cognates between languages, and reconstructing their forms in a shared proto-language. Some attempts at automation build upon this idea, comparing words of the same meaning across languages and judging their similarity and potential cognacy. The approach used here instead compares words within a language in order to avoid issues of languages using different orthography or having different phonetic spaces.

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; nearly all of the languages 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.

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

In order to compare all languages to one another, an objective, numerical representation of the language needed to be created. This representation is referred to as the language's **lexical metric**.

The measurement used by Kirby and Ellison (2006) to create the lexical metric was the **confusion probability** for the language's word pairs, which represents the likelihood that a particular word could be produced or understood instead of an intended word. More similar words are more likely to be confused, and thus have a higher confusion probability. Confusion probabilities can be obtained experimentally by testing these word pairs and recording whether they were confused or not, but collecting enough such data for large variety languages would be challenging. Instead, an estimation must be made based on the surface-level similarity of the words.

The simplest way to measure the similarity of two words would be to use the **edit distance** between the two words, which is a measure of the minimum number of insertions, deletions, and substitutions of characters that are required to transform one word into another. In order to adjust for high edit distances that result from large disparities in the size of words, this value is also normalized by the words' combined length.

Kirby and Ellison used the orthography of words to obtain their edit distance measurements. This has the problem that some words that are similar in pronunciation are quite different in their spelling. To use the example from earlier, the words *though* and *toe* have an orthographic edit distance of 4, despite only having a change in one phoneme. We instead

opted to use the phonetic transcriptions of the words, with the hypothesis that this would better capture the similarity of words and produce a more accurate lexical metric for the language. The words *though* and *toe* would be transcribed as [ðoː] and [toː], respectively, and these transcriptions have an edit distance of just 1.

Once having found the edit distance for all words, these values are converted into probabilities, and assembled into a matrix. The rows and columns of the matrix represent each of the words in the language for which there is data, and the values in each element represent the probability that the row's word may be confused for the column's word. This matrix is then normalized so that all of the values add up to 1, serving as the final lexical metric for the language.

Figure 1 serves as a visualization of the lexical metric for French. Each row and column of pixels represents rows and columns in the matrix, and the brightness of each pixel corresponds to the magnitude of the probability value in that position.

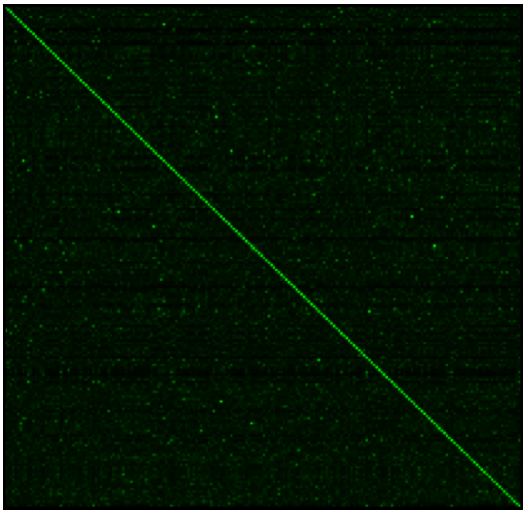

Figure 1: A visualization of the lexical metric for French.

Calculating Language Distance

Once we have constructed the lexical metrics for each of the 52 languages in our corpus, we could apply the methods described in the Section 3 of the Kirby and Ellison paper.

Before computing the distances between the languages, however, we had to ensure that all of the language matrices were consistent. That implied that in each of the languages matrices, the rows and columns should correspond to the same meanings; this proved to be slightly problematic.

Firstly, we found, that in the corpus we were working with, there were very few meanings that were shared by all of the languages. That led us to the decision of constructing large matrices (for all 208 meanings), but completely ignoring the values of the meanings that were not present in a language.

Secondly, some of the languages contained multiple forms for one meaning, while others did not. That raised a question of which of the several present forms we should choose to include to the languages matrix. After some discussion, our final solution was to only take the form with the greatest sum of probabilities. We argue that this makes the most sense, as word with higher confusion probabilities means that it is more similar to other words in the language, and may better represent the lexical structure of the language. However, this decision was arbitrary, and other methods could instead be used, such as using the minimum or the average.

The authors argued that in order to compare two languages, one should compute the distance between them. Since all of the languages were already represented as probability distributions (matrices), the distance between two matrices could be calculated. There were two types of distance calculations used in the paper, computed using the language matrices: Kullback-Liebler (KL) distance and Rao distance. Both distances are commonly used by statisticians to measure the divergence of two probability distributions. The formulas and their application to the data is explained in greater detail by Kirby and Ellison (2006).

Calculating the KL distance and Rao distance for a pair of languages involved comparing all of the elements of both matrices for which both languages had word meaning data. The script for implementing the formulas was written in Python.

Having compared the matrices of two languages using both KL and Rao distances, we obtained a numerical distance value for each of these methods. We then repeated the process for each pair of the languages in the data set and got the distance measures for each pair. These values were then assembled into another matrix, where the rows and columns represent each language, and the values represent the distance between the row's and the column's languages. In the end, we had two of such matrices: one containing the KL distances for all language pairs, and another containing the Rao distances.

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 a noticeable impact.

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 more effectively. 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 construction 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).

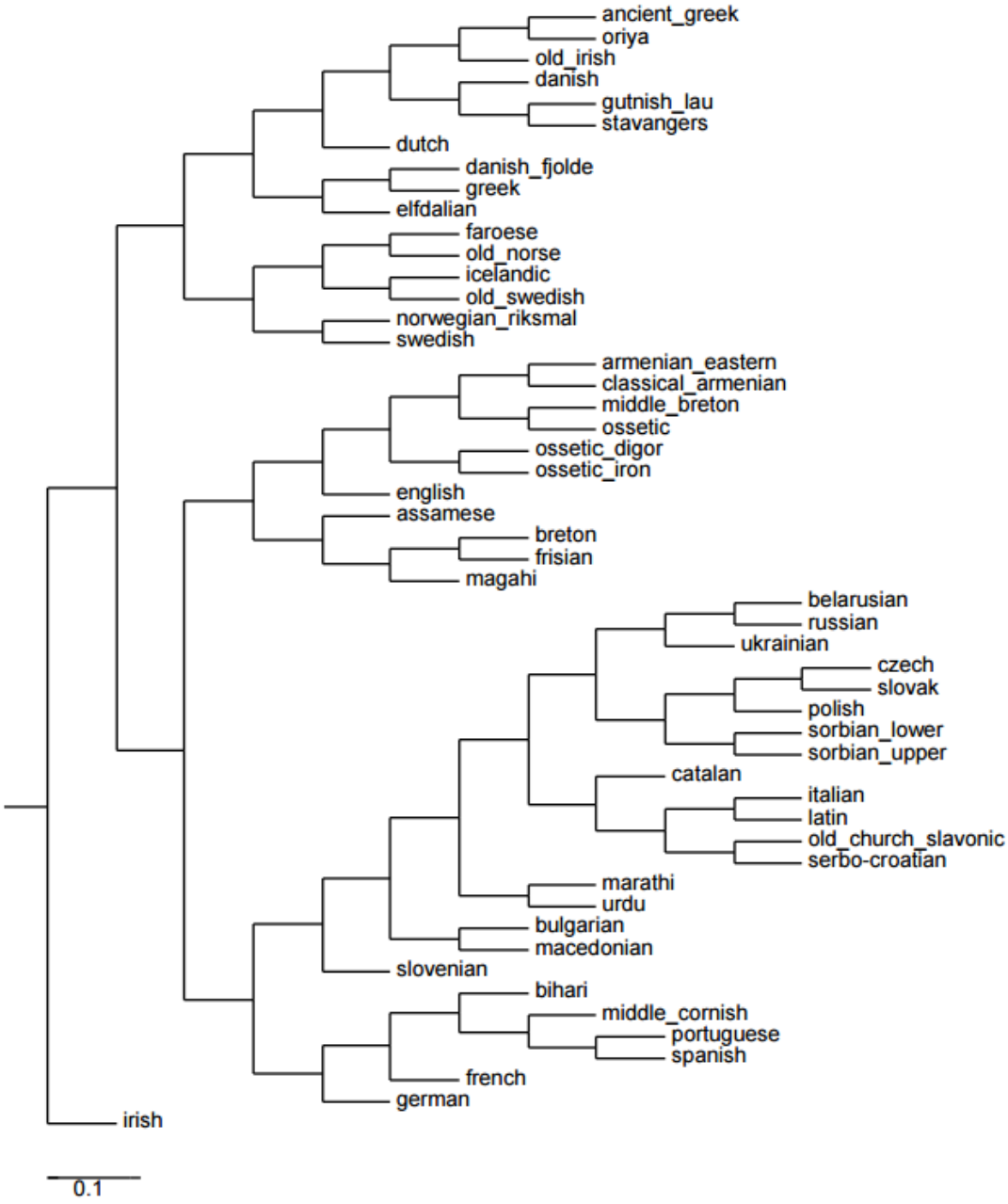

Figure 2: Tree visualization of KL-distance measurements

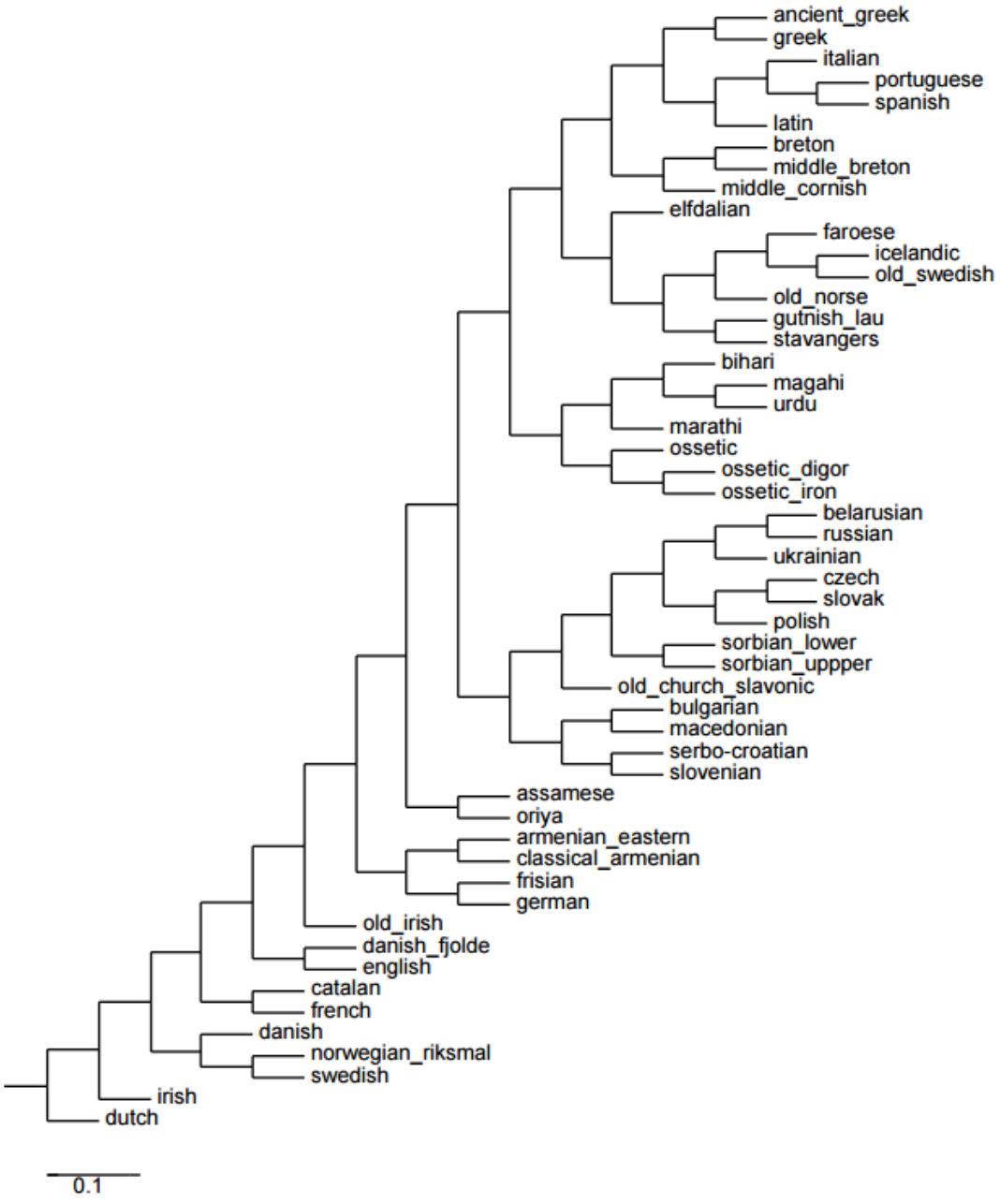

Figure 3: Tree visualization of Rao-distance measurements