Visualizing Etymology: The Evolution of Language

Shreya Patel, Helen Lei
Rensselaer Polytechnic Institute

Abstract—Throughout history, more recent languages have evolved from historical scripts. Words are often originated from previous languages and vocabulary. Their meanings will evolve and change over time, carrying different nuances and weight, which is important for the study of Linguistics. In this paper we aimed to create a tool to visualize etymology that would be intuitive enough for beginners to use, yet also informative enough for someone more experienced to find useful.

1 Introduction

Etymology and Linguistics is a critical and crucial part of our society and intellectual development. Understanding not only the meaning, but the nuances that words implicate reinforce the importance of languages and their derivations. In fact, there have often been theories and studies tied to knowledge and abstract concepts which heavily tie to language [1].

We wanted to aim at creating a tool for visualizing etymologies, or tracing the origins of a word that could be neat and intuitive to use and view for perhaps a hobbyist, but still useful for someone more experienced in the field, such as a linguistics student.

We centered our drive around testing and finding the best visualization to express etymologies in a new and engaging way. Thus, we asked the question: “How can we create a new, interesting and engaging way to visualize etymologies that would be intuitive enough for someone who is not an expert to comprehend how to use, and yet still be useful enough for someone who has studied linguistics?”

2 Related Work Summary/Background

WordNet

A lot of prior work was involved in collection of etymological data and the historical roots of words. One of these works included Wordnet: Tracing the History of Words, by De Melo, G [4]. The work has used the Wiktionary dataset for tracing the relationship between words. The information is mined using pattern matching techniques. They had then cleaned the data to deprive it of any external links, excessive whitespace, and encoded characters. As the API is intuitive to use and highly extensive, we decided to utilize WordNet as our main data source.

2.1 Other Etymological Visualizations

Another paper we came across was Visualizing Etymology: A Radial Graph Displaying Derivations and Origins, by Dixit and Karrfelt [2]. This work built an actual visualization for etymologies. Based on prior work seen, they aimed to reduce clutter and also make the end result interactive.

Figure 2.1 The main interface implemented by the paper through which a user may input a word to plot and how many levels of children nodes (etymological roots) to find.

The researchers had provided input boxes to allow users to specify a word they would like to map and how much depth they wanted (how many level of roots to find), as can be seen in Figure 2.1.

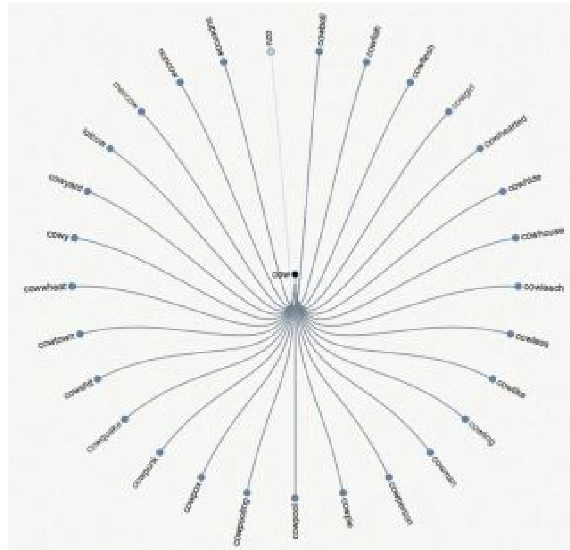


Figure 2.2. This an example of the radial structure portrayed for the etymological relationships between words.

The visualization then portrayed a radial graph with the parent word being in the middle and roots displayed in the perimeter, as seen in Figure 2.2. We were inspired by the freedom of letting a user choose a word(s) to map and also letting him/her highlight certain relationships by hover/click. This paper used WordNet as a data source as well. Similar features were included in our final design.

3 Conceptualization

We started our base design at a tree for starters, as this was the most standard format to visualize etymologies. From there we asked ourselves, “what are new ways to demonstrate etymologies that don’t obviously follow a tree structure?” We devised a number of rough drafts, that will be described below.

The first to be described will be the circular formation, similar in structure to the radial graph described in the related works section. The basic structure would work in a similar radial structure, however the more modern words would be centered around the middle and the roots would outline points in the outer ring (Figure 3.1 and Figure 3.1.2). We quickly scrapped this idea though, as the visualization gets bigger and conglomerates, it becomes less and less intuitive to comprehend, and no user would be able to easily pick out words and trace them with ease.

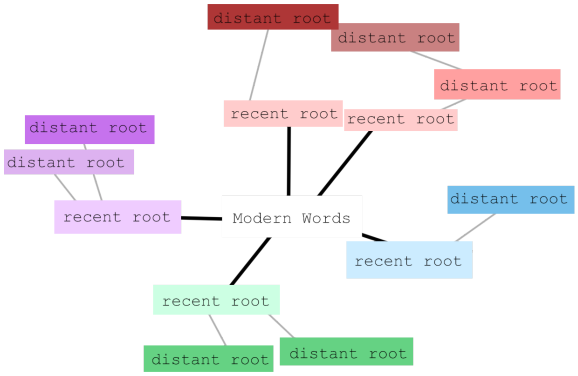


Figure 3.1. A more polished rough draft of our visualization pitch.

Figure 3.1.2. The structure from [3] that we tried to emulate with this visualization.

Another visualization we pitched was a streamgraph-like tree mutation, Figure 3.2) which was indeed a very interesting concept, but would have been too complex to implement reasonably within the time limitations we had.

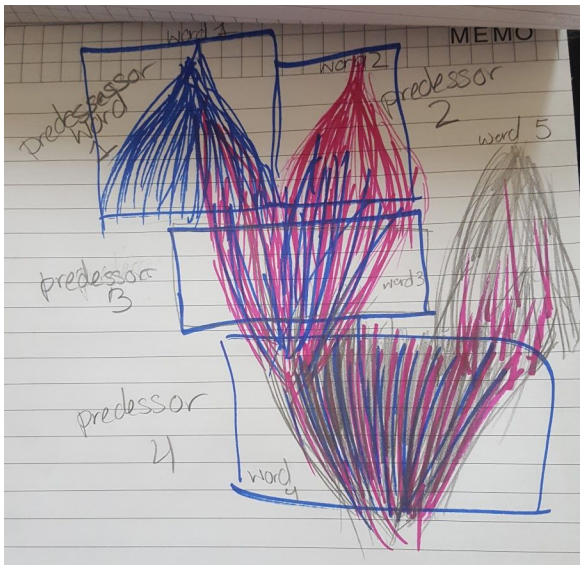


Figure 3.2 A very rough streamgraph structure arranged in a hierarchical branching structure.

As an attempt to dumb down the idea in (Figure 3.2) a little bit, we tried to go for a more standard stream graph, snapped to the x axis which represents the passage of time, (Figure 3.3) The heights of each section would represent the frequency of the usage of a word (or languages, we've considered) over time. Again, we quickly scrapped this idea, this time due to its simplicity and it's lack of engagement for a user. Additionally, the datasets we found and used would not have provided us with the information to create this graph.

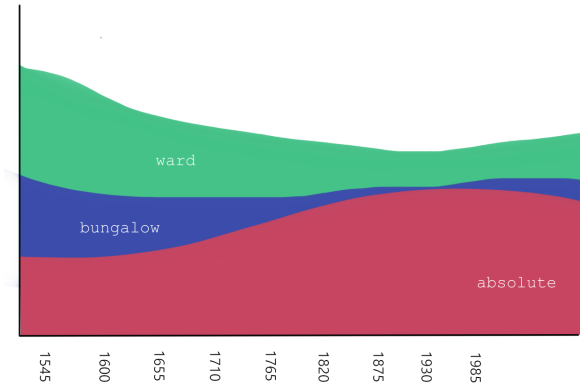


Figure 3.3 A streamgraph-like stacked graph that represents the usage of words over time.

3.1 Peer Review and Feedback

By the time the peer review came around, we had fallen back upon the tree structures due to their intuition and simplicity (Figure 3.4 and Figure 3.4.2), one using the d3 tree implementation, and the other using a force directed graph. We also entertained the idea of mapping physical locations to words, but that'll come later.

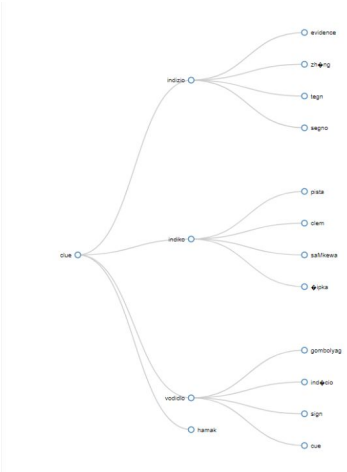


Figure 3.4

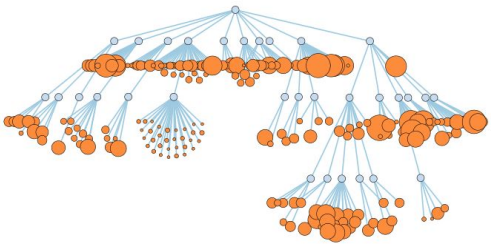


Figure 3.4.2

Helen focused on the representation of the tree and the density of information to be represented. From the discussion, we decided that people tended to prefer a manageable structure, with small amounts of information represented at a time, to avoid clutter and give people the opportunity to explore and expand trees at their own pace, instead of giving them "everything" and having no place to start. Tooltips was also a suggestion from peers to explore words more in depth.

Shreya focused on finding new ways to represent the data, once again trying to pry away from showing the bland standard tree model. She found that users preferred having guided navigation of the data set, ie. reinforcing the use of a small amount of data initially and expanding the more they explore, and from this there was a suggestion to superimpose the tree upon a map. Using a map to give a physical sense of an origin of the word gave our dataset the extra dimensionality we needed and this suggestion led us to our final representation that we settled upon.

3.2 Final Concept

Our final result rode on the suggestions to improve our design, and thus we created a geographic map representation with a tree superimposed upon the map's locations. With this visualization we could express the "origin" of a word, or rather the word that a root and word have been used frequently in and tie it to a physical and recognizable location. This map gave our visualization the extra dimension we needed to create an interesting and new interactive visualization.

To use it you insert your preferred word and the language you wish to search the word from in the search bar in the top left corner of the visualization (Figure 3.5.2). Mapping it draws a single node where the word originates from, and by clicking on the node it draws and maps the roots of the words and the countries. By clicking on the successive nodes, you can continue expanding the tree, allowing users to explore the map at their own pace.

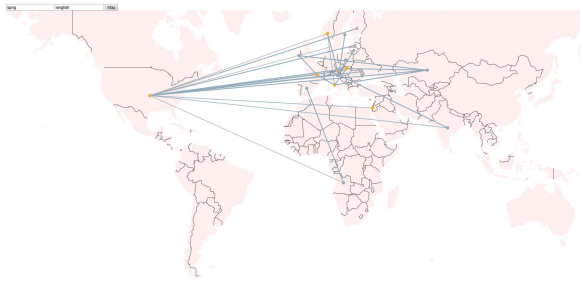


Figure 3.5 The full map

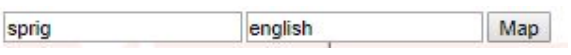


Figure 3.5.2 The Search bar

Additionally, we added tooltipping, which allows users to hover over a node they are interested in and see information about the node, such as the word itself, and the language tied to it. In future work we can expand the tooltipping further and give it far more powerful capabilities, but we stuck to basics for now due to time limitations (Figure 3.5.3).

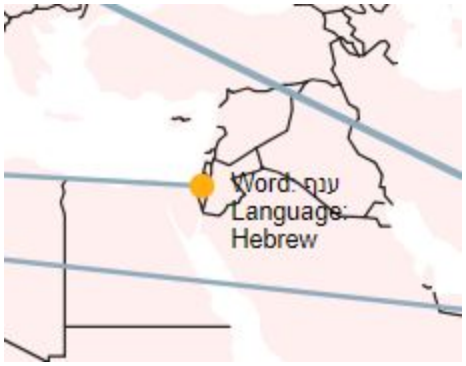


Figure 3.5.3 The tooltipping shown by hovering over a node/circle element.

4 Methods

The Java API from WordNet was used to obtain words and their etymological roots. The problem with this tool was that not all etymologies are actual historical roots, but contain some relationship [4].

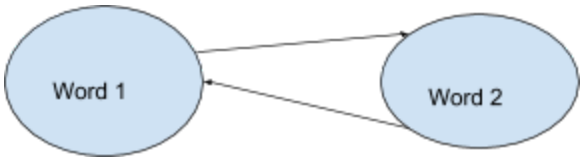


Figure 4.1 Words would often point to each other in terms of their etymological roots.

Figure 4.1 shows a common problem we were having using the Java API. Words would often contain each other in their etymological roots section of Wiktionary, and therefore portray that they originated from each other. As a result, the initial

visualization debugging graphs generated with the help of GraphViz turned out to be cyclic.

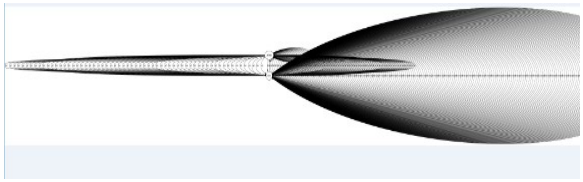


Figure 4.1.2 The diagram above shows an example output of a word and its roots, demonstrating the multiple calls back to the original parent node from its' children roots. Due to this, we used a HashMap structure which stored a tuple of a word and its' language as keys. Each words' children (only first level kid nodes) were kept as a list under the value for each key. With the help of this structure, we eliminated/skipped through any words already utilized.

Figure 4.1.3 This shows a revised structure with the repetitive relationships filtered out.

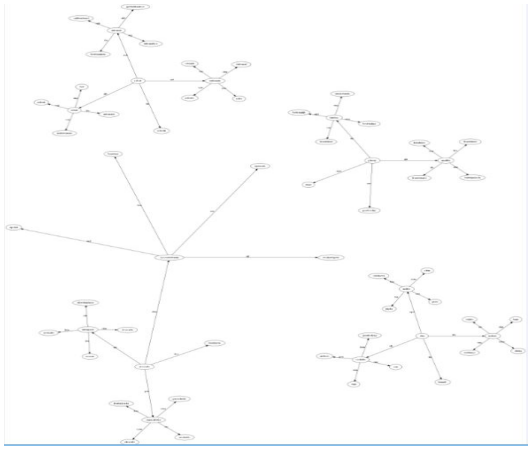


Figure 4.2 This shows roots of more than one word with a depth of 2 levels.

As there were multiple definitions of words, each vocab term had a code and weight associated with it, such as the following:

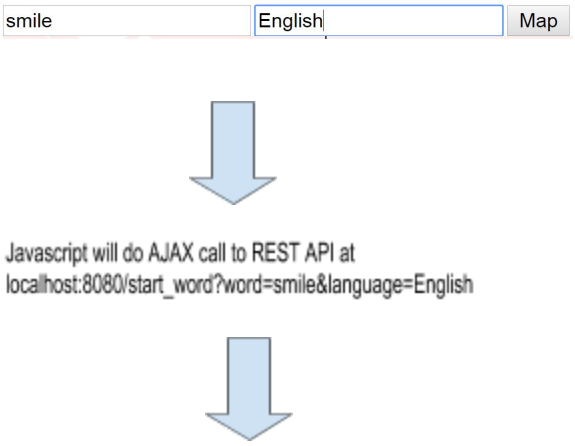
ISO-36546 (.56)
ISO-297983 (.89)
ISO-876349 (.05)
ISO-120978 (1.0)

The codes depict different meanings of a given word and their weights state a value between 0-1. A value closer to 0 means the given definition for the word is rarely used, while a value closer to 1 means the given definition for the word is used more often. For example, from the figure above, we can see the word definition associated with the code ISO-120978 is used more often than ISO-876349. For the purpose of our visualization, we ended up taking the code of a word which had the maximum value, or the most utilized definition.

As shown in Figure 3.5, our final visualization ended up being a map with user capabilities to include a word of their choosing. Of course, since WordNet mostly parsed English Vocabulary for their dataset, searching terms from other languages will not be feasible. A certain word from another language will only be included in the library if that word is a child root or has an etymological relationship to an English term.

In order to allow quick interactivity for the user, we decided to build a REST API for mapping words and their roots. The Spring MVC framework was used to take in

a word a user inputs within the text field or clicks over and output its roots.



Rest API made through Spring MVC will then send root words back to Javascript and it will then be used to map the point on the map.

Points are plotted with the word in the country where the associated language is used most. The country/language and related coordinate information was taken from CLDR (http://cldr.unicode.org/translation/country-names). This information was obtained with the help of a Java library called JSOUP. We found the first country and the most common language used within the country, and then parsed the coordinates of the country to be loaded to another HashMap. This parsing is done before the start of the program/Rest API controller. After the mapping of the initial start word, the user can click on the actual point where the node is mapped to then visualize its roots, and so on to get more levels/children nodes.

Figure 4.3 This diagram depicts a user interacting with the map. These are some iterations of mapping the origin of the word "come". Each orange node demonstrates a new unexplored child node of the parent word. After clicking on the orange root nodes, they will turn to blue, demonstrating that its' roots have already been mapped. If a node already exists in a location where another root is to be mapped, the root word replaces the node currently at the location. As there are multiple roots to a given word, we decided to limit this number to eight so as not to entirely clutter the map.

Currently the edges mapped for the relationships to depict etymologies are straight lines and therefore the map can become easily cluttered. In the future we

hope to draw edges such that the minimum amount of overlapping happens. In addition, we want the user to delete edges which are no longer useful to him/her anymore. In addition, we hope to not only show from which country a word may have originated, but also go into further detail and show which city it was derived from.

5 In Summary

While still a little finicky, we believe that the general user would be able to pick up our tool intuitively and that it can intrigue them to explore more of the tool. We also successfully managed to subvert the previous etymology visualizations by introducing a new and refreshing dimension to visualize words with.

The main data source for this visualization was the WordNet API. WordNet has parsed Wiktionary to obtain etymological roots of words. As the tool was intuitive, and extremely extensive, the information may not be completely reliable. Wiktionary is, of course, a biased source of data and all the etymological roots are not entirely where a specific root may have originated from. In the future, we hope to use a reliable source like the Oxford etymological dictionary and web parse it ourselves.

5.1 Division of Labor

As for the division of labor for this project:

Helen focused primarily on conceptualization, and rough drafting with input from Shreya, as well as aided in debugging later bits in the code.

Shreya primarily dug her hands into accessing, parsing and finding ways to utilize the WordNet library and packages associated with it. She also worked on the initial visualization/interface design.

Both Shreya and Helen worked together on the presentation and paper.

6 References

[1] Newton, A.M., Villiers, J.G. (2007). Thinking While Talking: Adults Fail Nonverbal False-Belief Reasoning. *Psychological Science,* 18(7), 574-579. DOI: 10.1111/j.1467-9280.2007.01942.x

[2] Dixit, C., Karrfelt, F. Visualizing Etymology: A Radial Graph Displaying Derivations and Origins. https://web.stanford.edu/class/archive/cs/cs448b/cs448b.1166/cgi-bin/wiki/images/7/7a/Cs448b_chinmayi_filippa_final_paper.pdf

[3] Jia, Y., Garland, M., Hart, J.C. (2011). Social Network Clustering and Visualization using Heirarchical Edge Bundles. *Computer Graphics Forum Vol 30. Number 8, pp. 2314-2327.* (2011). DOI: 10.1111/j.1467-8659.2011.02037.x

[4] G. Melo. "Etymological Wordnet: Tracing the History of Words." *LREC 2014*. ELRA. http://www1.icsi.berkeley.edu/~demelo/etymwn/