NYC Language Mapping Project: Visualizing Levenshtein Distance in a Multilingual City

Cherie Chu (cc4203) | Jennifer Lee (jgl2153) | Anavi Lohia (al3750)

1. **Goal**

Our goals for this project were twofold: (1) to identify relationships between high-resource and low-resource languages that allow existing technologies for high-resource languages to be better leveraged for use with low-resource languages, and (2) to create a way for laypeople to see, access, and interact with this data, sparking newfound interest in marginalized languages.

Even as technology has created new barriers to digital ascent for languages struggling to survive in the 21st century, it has also created new avenues by which to leverage emerging tools to better understand under-resourced languages. For our project, we were curious if it might be possible to look at the relations between languages—some more privileged in the contemporary world than others—and see if by better understanding their relations, we might enable technology to be better leveraged to create space for, rather than shut doors to, the ascent of marginalized and endangered languages. Technology enables us to study relations between languages in new ways, and we wondered if by analyzing new relations between these languages, we might be able to identify new relations that could be potentially useful for both computational researchers and the everyday person.

1. **Relation to Language Justice**

The primary motivation for this project was out of consideration of the unequal ways in which languages are prioritized in their transition to the digital sphere. In “Digital Language Death,” Kornai writes that 95% of today’s languages are unlikely to make it to digitization—a generous estimate (Kornai 2). And the 5% of languages that do stand a chance of digital ascent are not a random subset of all existing languages; rather, this prioritization of languages is shaped by powerful hierarchies passed down to us from centuries of imperialist and colonial legacies that existed long before the digital era. As Lydia Liu writes in “Scripts in Motion: Writing as Imperial Technology,” the technology of writing itself is one that has always been accessible only to a minority of the world’s languages, and which has perpetuated the privileging of certain cultures, communities, and modes of memories of others. In the digital age, this practice has only been exacerbated, as “digital media are transforming the world by turning one of the oldest technologies of world civilization—alphabetic writing—into a new imperial coding system” (Liu 382). The primary goal for our project then was to create a means by which the inequality of which languages can gain access to digital tools and support could be mitigated.

However, in consideration of what actually drives change in digital support for a variety of languages, we believed it was imperative that our project go beyond simply creating a “means” by which these inequalities could be mitigated. After all, I is not that the “means” of fixing digital inequalities do not exist or are impossible to imagine—rather, as we have seen again and again through discussion of different stakeholders who control how technological access is allowed and disallowed to various communities, it has become clear that the issue goes beyond mere technological capability and is about ongoing value systems that reify and are reified by technological inequalities. In fact, Martin Benjamin states in “Digital Language Diversity: Seeking the Value Proposition” that although “DLD [Digital Language Diversity] will not come about of its own accord,” it is also the case that “no great technological leaps are required to create a full panoply of resources for any given language. The heavy lifting … that is undertaken for languages at the top of the scale can be applied to other languages at relatively high speed and low cost” (Benjamin 56). Yet despite the fact that such expansion of technologies is not insurmountably difficult, it has not yet happened—thus, we believed that it would not be enough to simply create a computational component that might be of interest to programmers or computer scientists already interested in developing multilingual tools.

As a result, while the first part of our project is trying to identify where such “high speed and low cost” expansion of existing technology might be possible, the second part aims to try to bridge the gap between the existing possibilities and people who might potentially become interested in helping develop or advocate for such technologies. We believed this meant engaging laypeople who would otherwise have no reason to encounter such literature on computational approaches to language diversity, either because (a) they have little computational background and/or (b) they have no preexisting interest in linguistic diversity or language justice. We believed it was important to engage both groups because in treading, it became clear that while there are huge structural and technological inequities which imperil the survival of the vast majority of languages, that technologists are not the only ones capable of creating the tools for change. In the article “How to Prevent Language Extinction” Jose Mira is quoted saying that the “exogenous injection of just a few speakers into one group or another can determine whether a language lives or dies”—that is, that even just a handful of people can have a huge effect on the fate of a small language, and so for our map, we really wanted to believe that every individual interaction that might spark an interest and commitment to an endangered language is valuable. Thus, the second goal of our project is inspired by this desire to create a visual front-end component to our findings that enables everyday people to interact with our data.

1. **Existing Research**

The article “Machine learning has been used to automatically translate long-lost languages,” describes using ML to decipher extinct languages that humans had been unable to decipher through comparing them to their known, modern forms (arXiv “Machine learning”). This article was a starting point for us wondering if by identifying new relationships between languages, we might be able to identify low-resource languages that are computationally similar to high-resource languages, thus suggesting that technologies developed for those high-resource languages could be adapted to be used for those low-resource languages as well.

In “Generalized Data Augmentation for Low-Resource Translation,” Mengzhou Xia et al. discuss methods for using a modified machine learning translation framework to inject low-resource words into high-resource sentences to better convert high-resource data to low-resource languages, another method where being able to better identify languages to use in relation to each other would be helpful, and which is different work than the initial impetus for our project, where we saw only that tools designed for high resource languages might be directly re-applied to low-resource languages. This idea of creating new tools using injection from a LRL to HRL suggests to us that it could be especially useful to identify not only pairs or triangulations of relations between languages but rather map relations between multiple languages whose tools could be used in conjunction with each other, and our project would help with this kind of development.

Furthermore, in “On the Importance of Pivot Language Selection for Statistical Machine Translation,” Paul et al. describe the importance of choosing an effective intermediary language for translating between low-resource pairs. English is often used as the pivot language, but English is not necessarily the most useful pivot language when working with low-resource languages that are very different from English, making it imperative that there be means of finding more effective pivot languages or language pairs. In “Improving Candidate Generation for Low-resource Cross-lingual Entity Linking,” Shuyan Zhou et al. emphasize the importance of choosing an effective pivot language, as a poorly chosen one can decrease translation effectiveness by up to 20%, and finds that choosing the HRL that shares the largest number of character n-grams with the LRL is very effective—moreso than manually selecting for languages with linguistic similarities. This paper highlighted for us the importance and value of developing a system to help identify the most potentially effective pivot languages, as well as attuned us to the idea that more complicated is not always better: as has often been demonstrated in computational analysis of texts, often basic methods yield the most effective and timely results. It also suggested to us that there is real value in looking at computationally derived similarity rather than similarity based only on manually-identified similarities, since often the results may be surprising and counterintuitive, but yield results that are more useful for computational analysis.

This idea is similar to what initially drove the creation of LangRank, which is what we initially planned to use as a primary tool. LangRank is a Python tool based on the 2019 article “Choosing Transfer Languages for Cross-Lingual Learning” by Yu-Hsiang Lin, et al.. Here, the paper explains that NLP performance on low-resourcing languages can be improved through better identification of language pairs for “cross-lingual transfer, where a high-resource transfer language is used to improve the accuracy of a low-resource task language,” and approaches this as a ranking problem based on a number of dataset-dependent and dataset-independent features including dataset size, type-token ratio, word and subword overlap, geographic distance, genetic distance, inventory distance, and so on (Lin 3125, 7, 8). Given that LangRank was specifically developed for this purpose of ranking what language pairs might work especially well together, we had thought it fit well with our project goals as well.

In addition, there were two papers we found which attempt to computationally suggest language groupings or relations and compare that to existing classifications. In 2019, Georgi et al. published a paper “Comparing Language Similarity across Genetic and Typologically-Based Groupings” where the authors compared the World Atlas of Language Structures (WALS) family classifications to language groupings formed by clustering models using topological features, and found that the induced clusters had higher typological similarity than genetic grouping, and that the two groupings led to different results. In 2007, in “Indo-European languages tree by Levenshtein distance,” Maurizio Serva and Filippo Petroni calculated something called the Levenshtein Distance between two languages to identify their similarities. The method essentially looks at how many letters between two words are different. In the paper, the method was scaled up to 200-word corpuses for 50 different Indo-European languages, and the results were then used to construct a genealogy tree for those languages.

Although for our project we had initially intended to use LangRank because it was designed specifically and intentionally developed for the purpose of identifying languages that can serve well as a transfer and task language pair and so we believed we could demonstrate that usefulness and visualize such relations for our project, because we ultimately could not get LangRank to work, we decided to focus on reproducing this Levenshtein Distance methodology.

Finally, as discussed in the language justice goals of our project, we wanted a significant front-end component to our project. To do this, we looked for examples of interactive maps. The most closely related one to our project we found was of endangered languages is of Queens—the “Garden of Forked Tongues” map created by Miriam Ghani. In this map, regions were color-coded by language family of the language chosen to be featured in that region, but depicting or ranking actual closeness between languages is not something that was attempted. This map is the closest interactive tool to what we are planning to do that we came across, and in our own work we hoped to be able to implement the successes of that project—especially the interactive aspect and the visual appeal—while representing a wider range of things (all of NYC instead of all of Queens, the computationally determined closeness between languages rather than simply family trees). The one piece of information we chose not to reproduce visually that was available in the Endangered Language Alliance map was relative number of speakers. On the ELA map, more speakers meant a bigger label; for us, we chose to keep all labels equal inside, while population was noted in numbers when a language was selected by the user.

Given that Martin Benjamin’s article “Digital Language Diversity: Seeking the Value Proposition” discusses how the expectations of different language users that their language either will or will not be valued is what ultimately has a bearing on the language’s ultimate flourishing, it was important for us to create a map where all languages are featured as holding equal weight or value, such that it might normalize the idea that all these languages should be supported and able to thrive in all spaces, including digital ones.

1. **Results**

**Quantiative**

For the language analysis, we downloaded corpora for the languages we chose to work with from the *An Crúbadán* website (Scannell). The files containing word and frequency count pairs for each language were used, and any words that appeared three or less times were removed, to reduce processing time for each language. Then language corpora were processed in pairs, through a Python script calculating the Levenshtein distance of each word in the first language with each word in the second language. These values were used to calculate an aggregate similarity index for the two languages, visible below:

| **RAW RESULTS** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **English (en-US)** | **Quechua**  **(qu)** | **Spanish (es-PR)** | **Tagalog**  **(tl)** | **Cebuano**  **(ceb)** | **Uzbek**  **(uz-Latn)** | **Indonesian (id)** | **Brazilian Portuguese (pt-BR )** |
| **English (en-US)** | 5800057454 | 1358782478 | 1308306430 | 4634571488 | 1747033910 | 6395876358 | 8033721888 | 1521680730 |
| **Quechua (qu)** |  | 263082526 | 300107560 | 973597411 | 376504770 | 1362198719 | 1774238557 | 355008851 |
| **Spanish (es-PR)** |  |  | 287418368 | 1027990188 | 387665304 | 1420574153 | 1785240968 | 335972233 |
| **Tagalog (tl)** |  |  |  | 3306331788 | 1267959059 | 4753585015 | 6097091272 | 1206702113 |
| **Cebuano (ceb)** |  |  |  |  | 477430366 | 1828456776 | 2305310535 | 449317394 |
| **Uzbek (uz-Latn)** |  |  |  |  |  | 6364980082 | 8542227564 | 1678199223 |
| **Indonesian (id)** |  |  |  |  |  |  | 10618939790 | 2071489698 |
| **Brazilian Portuguese ( pt-BR )** |  |  |  |  |  |  |  | 384199592 |

| **SIMILARITY INDEX** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **English**  **(en-US)** | **Quechua (qu)** | **Spanish**  **(es-PR)** | **Tagalog**  **(tl)** | **Cebuano (ceb)** | **Uzbek**  **(uz-Latn)** | **Indonesian (id)** | **Brazilian Portuguese ( pt-BR )** |
| **English**  **(en-US)** | 1 | 4.268569508 | 4.433256094 | 1.251476532 | 3.319945549 | 0.906843274 | 0.721963933 | 3.811612607 |
| **Quechua**  **(qu)** |  | 1 | 0.876627453 | 0.270216953 | 0.698749517 | 0.193130798 | 0.148279117 | 0.741059062 |
| **Spanish**  **(es-PR)** |  |  | 1 | 0.279592521 | 0.741408542 | 0.202325495 | 0.16099696 | 0.855482506 |
| **Tagalog (tl)** |  |  |  | 1 | 2.607601377 | 0.695544895 | 0.542280186 | 2.73997348 |
| **Cebuano (ceb)** |  |  |  |  | 1 | 0.261111103 | 0.20710024 | 1.062568181 |
| **Uzbek**  **(uz-Latn)** |  |  |  |  |  | 1 | 0.745119471 | 3.792744029 |
| **Indonesian (id)** |  |  |  |  |  |  | 1 | 5.126233454 |
| **Brazilian Portuguese ( pt-BR )** |  |  |  |  |  |  |  | 1 |

There were many surprising results—for example Quechua scoring as more similar to English than Spanish, or the highest degree of similarity we got being between Brazilian Portuguese and Indonesian.

**Qualitative**

For the map portion, a demo can be viewed here:

<https://www.youtube.com/watch?v=LDCni7vcOmE>

We built a visualizer that maps languages on the Endangered Language Alliance map of New York City languages. It creates an interactive display combining the information on the original map of languages in the city with the information we found on computational similarities for particular languages.

The map works as follows: When no language is selected, all languages are displayed alongside each other on the map, specifically placed in areas of the city where that language has a high concentration of speakers. For languages that are used extensively in multiple areas of the city, the language appears multiple times. The labels consist of a colored dot indicating the geographic origin of the language along with the name (exonym) of the language. When a language is clicked, the user can see details of the borough, region of origin, size of the language community, language family, and similarity ranking to other languages based on our Levenshtein distance calculation (in order from most to least similar). When a language is selected, the circles remain colored for all instances of that language while the other dots turn gray. The size of the other languages’ dots also change—larger for languages with a low Levenshtein distance (high similarity) and smaller for languages with less similarity.

1. **Challenges**

The primary computational challenge for our work was that there are very few linguistic tools readily available for checking the linguistic similarity of corpora. Given that this is something very difficult even for trained human linguists to do, it is unsurprising that it is difficult for computers to classify this as well. What few tools that are available are poorly documented; e.g. while we could get LangRank working as-is, we could not follow the documentation to get run it to on our own corpora of languages. Additionally, for less common languages, getting parallel datasets was difficult. Finally, running queries on larger corpora requires a lot of processing time and power, which limited the computational work we could do. At the same time, using smaller datasets meant potentially less accurate results.

For the front-end portion of our work, the first challenge was in finding an efficient way to map the languages. Since we had no precise location data for the languages, our method was to look at the map published by the Endangered Language Alliance and then use Google maps to approximate the latitude and longitude. For languages with multiple locations, it meant that although we tried to identify and map all occurrences of the language, it is very possible that we missed some. This work was not very efficient and maybe not as accurate as we would have liked—for an expanded version of the project, it would be helpful to discuss with ELA how they determined the precise location within a larger neighborhood to place the language label, and if they have some kind of record of the latitude-longitude of each of the language points.

Finally, another challenge was in visualizing the data we produced as well. We wanted a way to view each language’s similarity index to other languages utilizing both the color depicting what region of the world it originates from (as was done in the ELA map and in our demo) and size to show similarity, but it ended up being confusing to try to discern the selected color versus all the other colors of the map. To mitigate that confusion and make the map more interpretable, when the user clicks on a language all the other languages turn gray, so they don’t distract and the color of the selected language stands out. The hope was that the text labels would still make it easy to identify other languages regardless of color.

1. **Post-project assessment of goals met**

We believe we met our overall goals well in that we were able to demonstrate the possibility of using Levenshtein distance to calculate language similarity and to extend that beyond an Indo-European family tree and devise a visualizer prototype to engage users in the possibility of easily seeing not only what languages are related to each other, but how close in proximity they are to each other New York City. We hope this kind of project might make somebody want to go out and meet someone a subway ride away who speaks another language, either because the language is surprisingly similar to or delightfully unrelated to languages they themselves know.

At the same time, we believe this work exposes in a self-explanatory way some of the many limitations to and lack of scalability of certain methods. Computationally, it seems from our results that while Levenshtein distance is an interesting and thought-provoking method, by itself it is not enough to identify effective high-resource language and low-resource languages for computational work. The results themselves were unconvincing in that certain language pairs returned a similarity index above 1, which should not happen—a language should be most similar to itself, with a similarity index of 1. In addition, the inability to scale it to non-Latin scripts is also problematic for making the project truly able to accommodate the widest possible breadth of languages. Similarly, the idea of calculating Levenshtein distance is somewhat reliant on the idea of a language having a standardized system of writing, as it is not very forgiving of variance in spelling—again, leaving out many less-standardized languages. There are also shortcomings in the scalability of our map, the most primary one being that we had to manually identify the longitude and latitude coordinates of languages on the ELA map to adapt that for our own work. Ideally, for a larger-scale project, we would have a database of coordinates to work with.

1. **Conclusions**

Based on our results, one conclusion we reached was that languages from different language trees will often still result in computationally calculated linguistic similarities—such as for example in our case Indonesian (Ethnologue classification: Austronesian, Malayo-Polynesian, Malayo-Chamic, Malayic, Malay) and Brazilian Portuguese (classification: Indo-European, Italic, Romance, Italo-Western, Western, Gallo-Iberian, Ibero-Romance, West Iberian, Portuguese-Galician), which had the highest similarity index of all languages we tested, despite belonging to entirely different language families (Austronesian versus Indo-European).

At the same time, another conclusion we reached was that Levenshtein distance, while interesting, is itself not nearly enough to identify promising high-resource language and low-resource pairs. However, we do believe we demonstrated the potential of Levenshtein distance to be a factor in calculating computational similarity between languages, as well as the value in a project that calculates a similarity index between different languages to discern data that can then be used by others attempting to determine low-cost, high-impact ways of scaling existing technologies to a greater number of languages.

Finally, a more methodological conclusion we reached was the importance of developing code documentation that is easy to follow and implement on another machine, as a large bulk of the difficulty in our project was in getting LangRank to work: first because the version available on Github had problems, but then once we got the working version through our Professor, while we could get it to work on the pretrained languages, we could not follow the documentation to get it to work on our own set of corpora. This issue comes back to the goals of our own project—since our desire is to inspire people to develop their own technologies based on our work, it is imperative that we be able to make our work legible and understandable to others.

1. **Future Work**

For future work on our project, the primary computational improvement we would like to make is in evaluating similarity through heuristics that go beyond Levenshtein distance. As useful as it is as a rough way of suggesting similarity between languages, it would be nice to both be able to calculate similarity using a multitude of metrics as LangRank does, as well as to be able to compare the rankings derived through a multiple of different processes (eg. N-gram matching) to see how they compare to something like Levenshtein distance. It would also be helpful to check whether this calculation is actually giving us a way of ranking languages that is in fact computationally useful, by for example taking a high-resource and low-resource languages pair that showed high similarity in our calculations and seeing how well the technology tailored to the HRL can be applied to the LRL, for example by trying to run a machine translation program developed for the HRL on the LRL.

To further explore Levenshtein distance as a metric itself, it would be useful to have larger, parallel corporas, as well as to expand the set of languages. Being limited by the Latin alphabet was for us a function of how we calculated the Levenshtein distance, but for a future project, being able to calculate language similarity for languages that do now use the Latin alphabet would be important as well. It may also be interesting to see if a language genealogy could be constructed through Levenshtein for a broader range of languages than was done by Maurizio Serva and Filippo Petroni. Additionally, it would also be interesting as an extension of this work to see to what extent the size of the corpora affects our results, as there are huge differences in computational time as we scale up the size of the corpora, and knowing what an optimal corpora size is and where the trade-off of time over accuracy stops being worth it is. Finally, just as there are a multitude of ways of analyzing language similarity and as different technologies focus on different aspects of a language to work, it could be useful if instead of simply trying to identify high-resource and low-resource language pairs to be used for any computational task, we could pull apart the idea of “computational tasks” further to identify language pairs that would work well for specific use-cases (eg. When developing a machine translation technology, does it make sense to choose a different language pair than for sentence completion suggestion? Do different methods of machine translation require different language pairs? Etc.).

For future work on the interactive map portion, it would be great if we could find more ways to display relationships between languages, such as different types of relationships between languages (ie. Not only Levenshtein distance but other heuristics, as well as WALS classifications) and unexpected patterns or dissimilarities between languages. Another feature we could add is a way to filter what gets displayed, so users can choose to look only at languages of a certain region or size, as well as if users could choose to look only at different dialects of a single language. To move beyond this one-way interaction with data, some exciting ways to develop this technology for future use would be to create a way for users to self-report qualitative experiences of similarities between languages—for example, cognates with languages or some other linguistic similarity that they may have noticed in personal use. These kinds of personal anecdotes that users could add are the kind of work that we hope goes beyond the computational aspect of the work and really engages the public with these questions of multilingual living in the city. It is also something that could be especially useful in considering the most recent, oral changes in languages that might not be reflected in corpora of texts scraped off the internet, and could be a way to track how distinct languages that are used in close proximity of each other in a place as diverse and densely packed as the city may lead to transformations in those languages, as they mix and mesh within the same physical sphere.

Works Cited

arXiv, Emerging Technology from the. “How to Prevent Language Extinction.” *MIT Technology Review*, MIT Technology Review, 16 June 2010, https://www.technologyreview.com/s/419394/how-to-prevent-language-extinction/. Accessed December 6, 2019.

arXiv, Emerging Technology from the. “Machine Learning Has Been Used to Automatically Translate Long-Lost Languages.” *MIT Technology Review*, MIT Technology Review, 1 July 2019, https://www.technologyreview.com/s/613899/machine-learning-has-been-used-to-automatically-translate-long-lost-languages/. Accessed March 3, 2020.

Benjamin, Martin. “Digital Language Diversity: Seeking the Value Proposition.” *Proceedings of the LREC 2016 Workshop “CCURL 2016 – Towards an Alliance for Digital Language Diversity*,” 23 May 2016, pp. 52-58.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2019. *Ethnologue: Languages of the World*. Twenty-second edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com. Accessed 5 May 2020.

Georgi, Ryan, et al. “Comparing Language Similarity across Genetic and Typologically-Based Groupings.” *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010),* Aug 2010, pp. 385–93, www.aclweb.org/anthology/C10-1044.pdf.

Ghani, Miriam. “The Garden of Forked Tongues.” *Forked Tongue Map*,  ww.mariamghani.com/forkedtongues/tonguemap/.

Kornai, András. “Digital Language Death.” *PLoS ONE*, vol. 8, no. 10, 2013, doi:10.1371/journal.pone.0077056.

Lin, Yu-Hsiang, et al. “Choosing Transfer Languages for Cross-Lingual Learning.” *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3125–35, doi:10.18653/v1/p19-1301.

Liu, Lydia H. “Scripts in Motion: Writing as Imperial Technology, Past and Present.” *Pmla*, vol. 130, no. 2, 2015, pp. 375–383., doi:10.1632/pmla.2015.130.2.375.

Paul, Michael, et al. “On the Importance of Pivot Language Selection for Statistical Machine Translation.” *Association for Computational Linguistics, vol. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, Jun 2009, pp. 221–24, www.aclweb.org/anthology/N09-2056.pdf.

Scannell, Kevin P. “An Crúbadán -.” *An Crúbadán - Corpus Building for Minority Languages*, crubadan.org. Accessed 1 May 2020.

Serva, M., and F. Petroni. “Indo-European Languages Tree by Levenshtein Distance.” *EPL (Europhysics Letters)*, vol. 81, no. 6, 2008, p. 68005., doi:10.1209/0295-5075/81/68005.

Xia, Mengzhou, et al. “Generalized Data Augmentation for Low-Resource Translation.” *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, doi:10.18653/v1/p19-1579.

Zhou, Shuyan, et al. “Improving Candidate Generation for Low-Resource Cross-Lingual Entity Linking.” *Transactions of the Association for Computational Linguistics*, vol. 8, 2020, pp. 109–24, doi:10.1162/tacl\_a\_00303.