Intersection of Linguistics, Language, and Culture NSF REU Site
High School Essay Contest, June 2019
First-Place Winner: Pravan Chakravarthy
https://undergrad-language-research.org/

White Rhinos, Endangered Languages, and the Power of Computational Linguistics

Sometimes, I think of myself as a white rhinoceros. By this I mean that I am one of the last of my kind. At home, I speak a dialect of Tamil unique to the Iyengar culture found in the South Indian state of Karnataka. General Tamil is spoken by about 82 million people and is in little danger of dying out in the next few generations, but Iyengar Tamil has very few speakers left (the exact number is unknown because official tallies are hard to come by). The language is becoming endangered because its youngest speakers (i.e., my generation) are more likely to speak the wider dialects of Tamil and Kannada common around the region, or one of the national languages of India. This is why it has become important to me to continue to uphold both the Iyengar Tamil language and the cultures it represents.

This is happening all around the world with all kinds of languages. There is no easy way to combat the global rise in endangered languages, but one solution that is becoming more and more common (and yielding better and better results) is using computational strategies to digitally document and enhance acquisition of rare languages.

Iyengar Tamil is just one endangered language that can be impacted by the "linguistics breakthrough" described in the article I read (dated May 9, 2017 in the MIT Technology Review). It describes the work of two scientists in Munich, Ehsaneddin Asgari and Hinrich Schütze, who have endeavored to preserve some of the rarest languages on earth with a surprisingly low amount of information.

There are two main differences between their approach and others. First, they started by examining tenses, which in most languages is regular enough to be used to map out the general structure of the language. Second, instead of starting with English or another well-known language, they began with some lesser-known creoles, which have not been corrupted by time and a large number of speakers. These different approaches, and their alleged success, might signal a new way to use machine learning to aid the preservation of endangered languages. As the article notes, machine learning in its current state is only adding to the problem of language endangerment, because it enhances the popularity of languages that are already spoken by hundreds of millions of people.

The article brings up a good point about preserving rare languages in the modern world. In order for these languages to remain intact, we have to be able to either harness the technology we have in order to help us in this goal, or create new technologies entirely. The former has been and is being done in many ways, such as digital categorization being done by the Endangered Language Alliance and Wikitongues (the subject of another of my favorite articles I perused).

However, we have also used technology to help us by creating computer programs that not only digitally document the language, but give users an interactive element which (in some cases) helps them acquire the language quite easily. I have some experience with these kinds of

programs; recently I helped Canadian researchers create and modify a tutorial for a conjugator of the endangered Mohawk language Kanyen'kéha. The tutorial involves the utilization of finite-state transducers (FSTs) for computational morphology, a lot of which made no sense to me until I started working on it. But as I continued to work on the project, I learned about the different ways in which linguistics and computer science intersect, and especially how this helps with real-world linguistic issues like the fact that in the next century more than a third of the world's languages are in danger of disappearing. In the conjugator I worked on, the FSTs were another uniquely efficient way to make the most out of small data sets; some of them have to be tweaked manually (unlike in the article) but are very reliable when it comes to taking smaller, more endangered languages and bringing them to an organized and effective program.

The biggest limitation that I see with the methods described in the article is that Asgari's and Schütze's database is simply too small of a data set to be viable in translating and documenting the more complex endangered languages. I have taken the NACLO contest for the past two years, and I know that it's very difficult to solve some problems with only the given information; this is why a simple false conclusion with just one word can lead to an entirely incorrect perception of the whole language and grammatical structure.

Despite these limitations, the article is still a big leap forward in the realm of computationally documenting endangered languages. It highlights the important work to be done with rare languages but also proposes a solution to these problems. For example, one of the main reasons Iyengar Tamil and other regional languages are dying out is that modernization is occurring at an unprecedented rate. With new advances in technology and infrastructure in the areas where these languages are spoken, it becomes very easy to lose cultural traditions that have been around for millennia; often the language is the first of these traditions to disappear. What Asgari and Schütze are doing, along with many other scientists, is using this modernization to their advantage by pioneering the technological ways in which these languages can be preserved.

I hope to help further with these developments by going to India this summer to catalog and digitize Sanskrit texts (although Sanskrit is already extinct, it can teach us something about preserving other less widespread languages that are near that fate). While I'm there, I hope to find a white rhino, although my chances of that are less optimistic.