Teaching Endangered World Languages with a Conversational Artificial Intelligence Application

Analyzing the Network Linking Minority Languages, Technology and Government

A Thesis Project Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science
University of Virginia - Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree

Bachelor of Science in Computer Science, School of Engineering

Charles Edward West Beall
Fall, 2023

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Travis Elliott, Department of Engineering and Society

Briana Morrison, Department of Computer Science

Introduction

Language is one of the most important aspects of human culture and society. It fosters cultural identity, enables communication between people, and serves as a vehicle for thought. Nearly everything that anyone does is within the context of a language, thus language is a lens through which we comprehend the world around us. Part of a large network of actors, language has very close ties to technology and government policy, both of which shape how it is used by people all over the world. No language has the luxury of being as ubiquitous as English, and many are losing speakers as they are not being passed on from one generation to the next, despite efforts and methods employed to preserve them. This leads to loss of culture, identity, and history for many communities that speak endangered languages. Such loss is the impetus for the technical focus of this project, which is the proposal of a conversational artificial intelligence program that can be used to teach languages with low numbers of speakers. The development of this product would make it a part of the network of world languages, government policy and technology, so the STS research portion of this project will seek to analyze the interactions of actors in said network, especially the way in which social media and the internet interact with language in communities.

Technical Topic: Revitalizing Endangered World Languages with Conversational Artificial Intelligence

The goal of the technical portion of this project is to propose a conversational artificial intelligence program that can be used to teach languages with low amounts of speakers in order to bolster communities' efforts to revitalize endangered languages. Governments around the world have ongoing efforts to preserve languages spoken in communities where the language

was once prominent, but is being passed onto future generations at rates that put the language at risk of extinction. Some examples of ways in which local governments promote the use of minority or endangered languages are through bilingual or immersion schools (ENTR en, 2022), or promoting use and exposure to languages in quotidian life through things like road signs, government documents, or political operations (Wilson & Cook, 2023). Most of these actions are supplemented by use of technology for language teaching and translation, which is a role that this proposal seeks to fill.

One example of a language learning platform that this project has taken inspiration from is Duolingo. This widely known and used application has over the years developed language courses for at-risk languages such as Irish, Hawaiian, Welsh, Navajo, and Yiddish (Griffiths, 2019). The benefits of this development are that it raises awareness about these languages, and more importantly supports a learning community for the language, all without any required cost to users. Another program that has influenced the idea behind this project proposal is ChatGPT, a text-based natural language processing and generation program that can be used by language learners to practice conversation, learn grammar and vocabulary, or even develop study plans ("Learning a New Language with ChatGPT", 2023). This proposal intends to use the mentioned aspects of both of these platforms to become a flexible language learning application powered by artificial intelligence that can be used to teach endangered languages.

This product would take the form of a language learning application that can be accessed on smartphones or in web browsers on computers. In order to develop this application, it would need to harness an artificial intelligence program, and train it to be able to produce and understand languages that have less media and fewer people as training resources. Training on less data would pose a challenge, especially because successful artificial intelligence programs

are trained on vast datasets. This program would risk being biased towards specific regional accents or dialects of a language, or even certain ideologies. Additional challenges for using small training datasets exist in making sure that the program's language use is accurate, which is important for correctly teaching its users a language. There are measures to improve training outcomes in this situation. These include transfer learning from other language models (Orynycz, 2022), and pre-training on monolingual datasets (Zheng et al., 2021). In addition to text processing and generation, the program would have speech and listening capabilities to be able to interact with users in all use cases for language practice. The main part of the application would be a chat interface, where users have the option of selecting a language, and can choose to communicate with the program via text or speech. Due to the flexibility that the program would provide, the user would be able to practice conversation with it, and could even ask it to generate personalized lessons or exercises.

If this proposal were implemented, it would join the network of actors associated with language use in various communities. Local governments might decide to use it in schools or other language teaching programs. It could even fill part of the role of a shortage of teachers in languages with few speakers. In those spaces, students would interact with this program on smartphones or computers, and the program would learn from those interactions. Students' language abilities would improve, and they would be able to use their new skills to communicate with each other, and other members of their community who speak the language.

STS Research Topic: Analyzing the Network Linking Languages, Technology and Government Policy

Introduction to Frameworks

Language is a key feature of humanity, being used by all for communication, thinking, and processing the surrounding world. Communication occurs between people (or organized groups of people) in many forms, be it face-to-face or with technology as an intermediary. These communications form links, meaning language is a core part of an interconnected network of people, technology, and organizations, which can be described using Actor-network theory (ANT). In ANT, "An *actor* is, everything that in some causal way affects the production of scientific statements and theories," which must "be able to perform actions as a kind of behavior describable under some intention" and "a *network* is a set of actors such that there are relations and translations between the actors that are stable, in this way determining the place and functions of the actors within the network" (Detel, 2001). The purpose of employing ANT is to parse the nature of the relationships between different entities in the network as a whole. Using this framework will bring forth a comprehensive understanding of how the network linking language to technology and government works.

Language Networks, Technology, and Policy

Languages naturally change over time, as do the ways that people use them. Important parts of the force driving the change in language are technology and policy, which themselves are shaped by language. This interplay between these actors can be seen in various contexts throughout the globe.

One example of this type of network is with the Uyghur community in the Xinjiang province of China. The Uyghurs are an ethnic and cultural group in northwestern China that have been subject to a concerted effort by the government to culturally and linguistically integrate the population into the rest of the country. The Chinese government employs methods to persecute

and repress Uyghur language and culture such as internet censorship, re-education detention centers, and arrests of dissenters. The Uyghur language serves as an important aspect of the identity of Uyghur community members in China (Chen, 2010). In the face of these actions taken by the Chinese government, Uyghur community members have previously been able to leverage social spaces on the internet in order to preserve their linguistic community and culture, as well as promote hidden messages of resistance (Clothey & Koku, 2017). While the government continued to act against the Uyghur community on the internet, users adapted, finding ways around censorship to spread messages (Borak, 2022). Strict government policy in China has even led Chinese tech companies to preemptively restrict the use of minority languages on their social platforms (The China Team, 2021). It is quite clear from this example that the Chinese government is an actor creating policies that attempt to dictate use of language in Uyghur communities, who in turn act by creating content and talking in their language on the internet. The government responds to internet content through censorship and policy changes, while users adjust their behavior accordingly. Internet social platforms can also be identified as actors because they change their use policies to adapt to the government's laws, and affect how users can express themselves and their ideas. There is a reactionary cycle between the different actors of this network, in which the actions of one member influence the actions of the others.

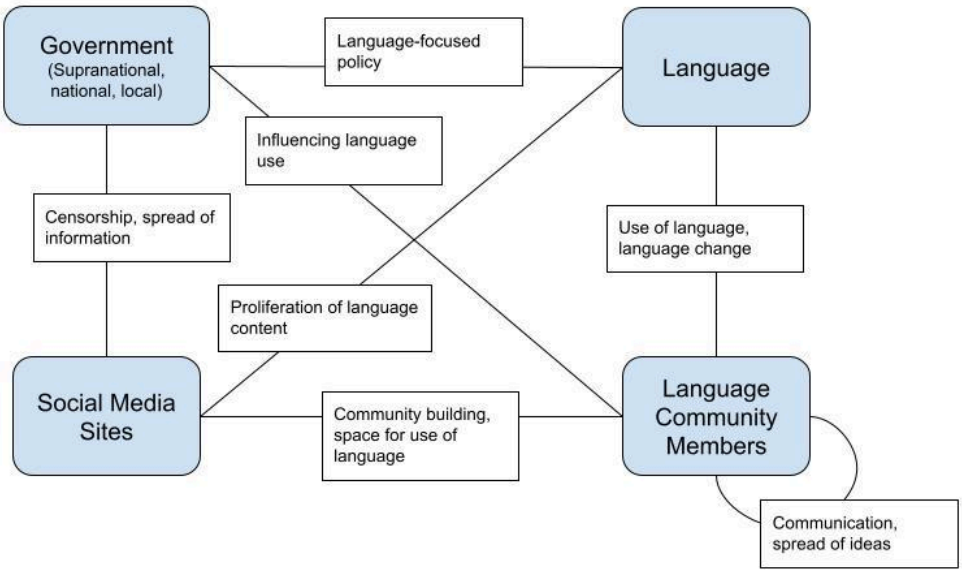
Another example of a network linking government, technology and a minority language is that of the Irish government, social media, and the Irish language. Nowadays, Ireland is known as an English speaking country due to a long history of the slow death of the Irish language. A combination of factors in the 1800s including national education in English, the political union of Ireland with its neighbor, and famine are part of what ultimately led to the dominance of English language on the island (Hindley, 1990). Even though the Irish language has long been

out of mainstream use in Ireland, there are ongoing efforts to revive the language to its past prominence. These efforts are spearheaded by the Irish government, whose plans to promote the language have the aim of increasing use and knowledge of the Irish language throughout the country, as it is an important part of Irish heritage (Government of Ireland, 2010). Part of this includes providing public resources (such as digital documents) for its citizens in both Irish and English (Government of Ireland, 2010). Public efforts and enthusiasm surrounding the revitalization of the Irish language cultivates an environment in which content creators focusing on teaching the language are able to without opposition create their own spaces on the internet (Fox, 2022), and Irish language users have taken to expressing themselves in the language on social media platforms (Caulfield, 2013). In this example, the Irish government, an actor, creates policies that are favorable for use of the Irish language. Other actors, namely people that speak Irish or are trying to learn it are able to use social spaces on the internet to connect with the community of speakers and practice their language skills. Social media sites hold increasing amounts of Irish-language content as their digital communities grow. The Irish language itself has already seen changes as it is used more frequently online (TedX Talks, 2016). As a result, the government assesses fluctuation in Irish language use among the population in order to decide future policies related to language promotion.

Analyzing these examples enables us to identify common actors in a network that is framed by language, technology, and government, which are the governing body that makes policies relating to minority language use, technology in the form of social media and underlying algorithms, speakers and learners of a language, and the language itself. They can be represented by the following actor-network diagram:

Figure 1

Actor-Network Diagram for Language, Internet, and Government



Using ANT to examine the components of this network diagram reveals that all of the actors interact with each other in some way. National governments enact policy that influences public perception of a language and ability to use it in public spaces and on the internet through social media. In those spaces, language users create and consume content, defining how language is used within the limits imposed by social platforms and the government. Language use online transmits ideas from one person to the other, as well as to governments and tech companies. The government implements new policy in a reactionary manner to the nature of language use to either support or restrict it, while tech companies can choose to accommodate the language users and/or government policy regarding the language in question. Governments also enforce their policy relating to the language through technology, which could include censorship or uploading helpful language resources to the internet.

Research Plan

These networks will be examined in further detail by consulting sources that describe the actions of their actors, such as further examples of government policy or action in response to or with the purpose of influencing language use on the internet, and how language communities have manifested themselves on internet and social platforms in the cases of the Uyghur and Irish languages. The information gathered will be used to construct a full understanding of the inner workings of languages' actor-networks.

Conclusion

Language is an important tool for communication and a significant part of cultural identity. Most world languages are losing more speakers than they are gaining, risking the disappearance of communities' heritage. The technical portion of this paper describes the conceptual development of a conversational artificial intelligence program that focuses on flexible language-learning programs for low-resource languages. The purpose of such a program would be to support efforts to teach endangered languages and ultimately preserve culture. This product would fit the role of social interaction based platforms on the internet, and would be subject to the actions of government policy and language communities. By generating language, the program would also take action that influences the other members of the network. Actor-Network Theory is used in this paper to examine the network linking language to government and technology, with examples from the Uyghur and Irish languages. Understanding such networks is key to understanding how actions taken by technological actors have an influence on languages, and how society may influence technological actors in the network. Language's role as a tool for communication and a hub of culture highlight its importance to society, thus understanding the factors that help shape it in the technological age is invaluable.

References

Abu-Irmies, A., & Al-Khanji, R. R. (2019). The Role of Social Media in Maintaining Minority Languages: A Case Study of Chechen Language in Jordan. *International Journal of Linguistics*, *11*(1), 62-75.

Belmar, G., & Glass, M. (2019). Virtual communities as breathing spaces for minority languages: Re-framing minority language use in social media. *Adeptus*, (14).

Bertemes, J.-P. (2014, November 27). *Une évolution importante à l'ère des médias sociaux*. science.lu | Wessenschaft fir jiddereen. https://www.science.lu/fr/limportance-du-luxembourgeois/une-evolution-importante-lere-des-medias-sociaux

Borak, M. (2022, November 2). *The strange death of the uyghur internet*. Wired. https://www.wired.com/story/uyghur-internet-erased-china/

Caulfield, J. (2013). *A social network analysis of Irish language use in social media* (Doctoral dissertation, Cardiff University).

Chen, Y. (2010). Boarding school for Uyghur students: Speaking Uyghur as a bonding social capital. *Diaspora, Indigenous, and Minority Education*, *4*(1), 4-16.

Clothey, R. A., Koku, E. F., Erkin, E., & Emat, H. (2016). A voice for the voiceless: online social activism in Uyghur language blogs and state control of the Internet in China. *Information, Communication & Society*, *19*(6), 858-874.

Clothey, R. A., & Koku, E. F. (2017). Oppositional consciousness, cultural preservation, and everyday resistance on the Uyghur Internet. *Asian Ethnicity*, *18*(3), 351-370.

Clothey, R., & Meloche, A. (2022). Don't lose your moustache: community and cultural identity on the Uyghur internet in China. *Identities*, *29*(3), 375-394.

Detel, W. (2001). Social Constructivism. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 14264–14267). Pergamon. https://doi.org/10.1016/B0-08-043076-7/01086-X

Dewey, C. (2021, December 1). *How the internet is killing the world's languages*. The Washington Post. https://www.washingtonpost.com/news/worldviews/wp/2013/12/04/how-the-internet-is-killing-the-worlds-languages/#

ENTR en. (2022, December 5) *The Breton language is in danger, and it's France's fault | Raising voices* [Video]. YouTube. URL https://www.youtube.com/watch?v=cUtxMV7O00A

Ferré-Pavia, C., Zabaleta, I., Gutierrez, A., Fernandez-Astobiza, I., & Xamardo, N. (2018). Internet and social media in European minority languages: analysis of the digitalization process. *International Journal of Communication*, *12*, 22.

Fox, K. (2022, August 21). *Day in the life: Digital creator using social media to promote the Irish language*. BreakingNews.ie. https://www.breakingnews.ie/day-in-the-life/day-in-the-life-digital-creator-using-social-media-to-promote-the-irish-language-1319040.html

France24 English. (2021, June 3) *Linguistic treasures: Exploring France's regional languages* [Video]. YouTube. URL https://www.youtube.com/watch?v=tq7QsB_v8dc

Government of Ireland. (2010). 20-Year Strategy for the Irish Language 2010–2030.

Griffiths, J. (2019, October 4). *The internet threatened to speed up the death of endangered languages. could it save them instead? | CNN business*. CNN. https://www.cnn.com/2019/10/04/tech/duolingo-endangered-languages-intl-hnk/index.html

Hermes, M., Bang, M., & Marin, A. (2012). Designing indigenous language revitalization. *Harvard Educational Review*, *82*(3), 381–402. https://doi.org/10.17763/haer.82.3.q8117w861241871j

Hindley, R. (1990). *The death of the Irish language: A qualified obituary*. Taylor & Francis.

Internet and social media – blessing or curse for linguistic minorities?. Terminology Coordination European Parliament. (2015, December 14). https://termcoord.eu/2015/12/internet-and-social-media-blessing-or-curse-for-linguistic-minorities/

Learning a New Language with ChatGPT. Microsoft. (2023, May 5). https://www.microsoft.com/en-us/microsoft-365-life-hacks/writing/using-chatgpt-for-foreign-language-learning

Library of Congress. (n.d.). *Research guides: Reading in French: A student's Guide to Francophone Literature & Language Learning: Regional & Minority languages in France*.

Orynycz, P. (2022, May). Say It Right: AI Neural Machine Translation Empowers New Speakers to Revitalize Lemko. In *International Conference on Human-Computer Interaction* (pp. 567-580). Cham: Springer International Publishing.

Rehg, K. L., & Campbell, L. (2018). Abstract. In *The Oxford Handbook of Endangered Languages*. essay, Oxford University Press.

Simon, E. (2023, July 22). *6 ways to use CHATGPT to learn a foreign language: ICLS: International Center for Language Studies: Washington D.C.* ICLS. https://www.icls.edu/6-ways-to-use-chatgpt-to-learn-a-foreign-language/

TedX Talks. (2016, February 6). *How social media breathes life into the Irish language | Teresa Lynn | TEDxFulbrightDublin* [Video]. YouTube. URL https://www.youtube.com/watch?v=LM3ISST2eg8

The China Team. (2021, November 8). *Chinese tech companies appear to censor Uyghur and Tibetan*. Protocol. https://www.protocol.com/china/bilibili-talkmate-uyghur-tibetan-tech

UNESCO WAL. (n.d.). https://en.wal.unesco.org/

Voice Bots and conversational AI. Voice bots and Conversational AI | Microsoft Power Virtual Agents. (2023). https://powervirtualagents.microsoft.com/en-us/voicebots-conversational-ai/

Wilson, J., & Cook, L. (2023, September 19). *Spain allows Catalan, Basque and Galician languages in Parliament. EU ponders use in Brussels*. AP News. https://apnews.com/article/spain-catalan-basque-galician-languages-parliament-3209def249eabbb3a446a9de55cb3479

Woodbury, A. C. (n.d.). *What is an endangered language? - Linguistic Society of America*. Linguistic Society of America. https://www.linguisticsociety.org/sites/default/files/Endangered_Languages_0.pdf

Zheng, F., Reid, M., Marrese-Taylor, E., & Matsuo, Y. (2021, June). Low-resource machine translation using cross-lingual language model pretraining. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas* (pp. 234-240).