What is it?
Towards a Generalizable Native American Language Identification System

Ivory Yang **Weicheng Ma**
Carlos Guerrero Alvarez **William Dinauer** **Soroush Vosoughi**
Department of Computer Science, Dartmouth College
{Ivory.Yang.GR, Soroush.Vosoughi}@dartmouth.edu

Abstract

This paper presents a research thesis proposal to develop a generalizable Native American language identification system. Despite their cultural and historical significance, Native American languages remain entirely unsupported by major commercial language identification systems. This omission not only underscores the systemic neglect of endangered languages in technological development, but also highlights the urgent need for dedicated, community-driven solutions. We propose a two-pronged approach: (1) systematically curating linguistic resources across all Native American languages for robust training, and (2) tailored data augmentation to generate synthetic yet linguistically coherent training samples. As proof of concept, we extend an existing rudimentary Athabaskan language classifier by integrating Plains Apache, an extinct Southern Athabaskan language, as an additional language class. We also adapt a data generation framework for low-resource languages to create synthetic Plains Apache data, highlighting the potential of data augmentation. This proposal advocates for a community-driven, technological approach to supporting Native American languages.

1 Introduction

Language is more than a means of communication; it is a vessel of culture, history, and identity (Miller and Hoogstra, 1992; Bucholtz and Hall, 2004; Sirbu, 2015). For many Indigenous communities, the loss of a language represents not just linguistic erosion but the disappearance of traditions, worldviews, and ways of knowing (Grenoble and Whaley, 1998; Khawaja, 2021). Despite increasing efforts in computational linguistics to support low-resource languages (Ranathunga et al., 2023; Singh et al., 2024), the landscape remains starkly imbalanced. Google's LangID (Caswell et al., 2020), one of the most commercialized language identification systems, covers over 200 languages, but overlooks almost all North American Native languages. **This exclusion is an alarming reflection of how centralized language technologies systematically marginalize Indigenous voices** (Khubchandani, 2016; Yim, 2024).

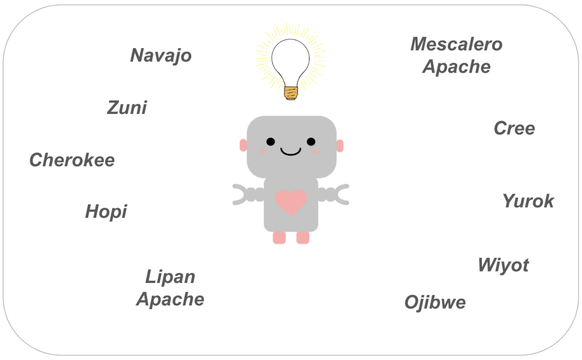


Figure 1: A simplified, stylized rendition of the proposed generalizable Native American Language identification system.

The state of New Mexico (NM) stands as a crucial focal point in this discussion. Home to eight Native American languages[1] (New Mexico Secretary of State, 2025), the state exemplifies both the resilience and fragility of Indigenous linguistic heritage. While computational linguistics has explored the most-widely spoken Navajo to some extent (Liu et al., 2021; Yang et al., 2025b), progress remains constrained by the scarcity of accessible linguistic data (Meek, 2012; Goswami et al., 2024). To address the current gap in commercialized language technologies, we propose a research agenda to build a generalizable Native American language identification system, the first of its kind, as exemplified in Figure 1.

Our approach consists of two key initiatives: (1) Data Resource Aggregation: A comprehensive, systematic effort to manually collect and curate linguistic datasets across all available Native Amer-

[1] The eight languages are Tiwa, Tewa, Keres, Towa, Zuni, Navajo, Mescalero Apache and Jicarilla Apache. There are eleven New Mexico counties with Native American lands.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 105–111
April 30 - May 1, 2025 ©2025 Association for Computational Linguistics

ican languages, ensuring high-quality, representative training data. (2) Synthetic Data Generation: Applying an established data augmentation framework for endangered languages to expand existing data, particularly for languages with few or no remaining fluent speakers. For proof of concept, we manually curated a small dataset of 25 Plains Apache sentences, an extinct Southern Athabaskan language, and successfully integrated it into an existing rudimentary Athabaskan language classifier (Yang et al., 2025b). We then adapted a data generation framework for low-resource languages (Yang et al., 2025a) to create 5 syntactically-coherent new Plains Apache sentences, displaying the promise of our approach. **This paper serves as both a research thesis proposal and a call to action, working towards a future where Native American languages are not only included but actively supported by commercialized language technologies**.

2 Related Work

Efforts to develop Natural Language Processing (NLP) technologies for endangered languages are hindered by scarce datasets (Maimaiti et al., 2022) and non-specialized model architectures (Lin et al., 2018). This section reviews emergent research in two key areas: Native American language classification, and synthetic data generation for endangered languages.

2.1 Native American Language Classification

Yang et al. (2025b) exposed the shortcomings of centralized NLP systems in handling Native American languages. Google's LangID system (Caswell et al., 2020), despite covering over 100 languages, failed to include any Native American languages, even the most widely spoken Navajo (Palakurthy, 2022). To address this gap, they developed a Random Forest classifier (Ho, 1995) trained on Navajo and 20 of its most frequently confused languages, achieving a near-perfect accuracy (97-100%). Further experiments revealed that the classifier generalized well to other Athabaskan languages[2] under the same family tree, suggesting potential scalability across related language families. However, while this work introduced a novel approach to Native American language identification, its scope was limited, covering only five languages. Expanding its applicability requires broader generalization across diverse linguistic groups.

2.2 Synthetic Data Generation for Endangered Languages

Data scarcity is a persistent challenge in low-resource NLP (Ghafoor et al., 2021; Adimulam et al., 2022), particularly for languages with few or no fluent speakers (Bansal et al., 2021). Yang et al. (2025a) demonstrated the effectiveness of synthetic data augmentation for endangered languages on Nüshu, a near-extinct ancient Chinese script (Congrong, 2024). Using a language-specific data generation framework, they produced a novel dataset of 98 linguistically coherent synthetic sentences in Nüshu, demonstrating a viable approach to language revitalization.

Applying this approach to Native American languages presents both opportunities and challenges. Unlike Nüshu's text-to-text structure (Di, 2024), many Indigenous languages require careful handling of phonetic, morphological, and orthographic variation (Link et al., 2021). Still, a synthetic data pipeline remains a promising strategy for expanding training resources, especially for those on the verge of extinction.

2.3 Towards a Unified Approach

Building on prior work, this paper proposes a hybrid approach that combines language classification and synthetic data generation to create a scalable Native American language identification system. Unlike previous efforts that addressed classification or data expansion in isolation, we argue that both are essential for developing a truly generalizable, resource-efficient, and community-driven model. By integrating rigorous classifier development with targeted augmentation, we aim to surpass existing limitations and advance linguistic inclusivity in commercialized language technologies.

3 Native American Language Landscape

Native American languages form a vast and diverse linguistic ecosystem (Oberg and Olsen-Harbich, 2022), reflecting centuries of cultural, historical, and geographical significance (Clements, 2021). While many of these languages once flourished across North America, colonization (Huang, 2024), forced assimilation policies (Ellinghaus, 2022), and systemic marginalization (Sear and Turin, 2021)

[2] The languages tested with the Navajo classifier were Western Apache, Mescalero Apache, Jicarilla Apache and Lipan Apache, which are all Southern Athabaskan languages.

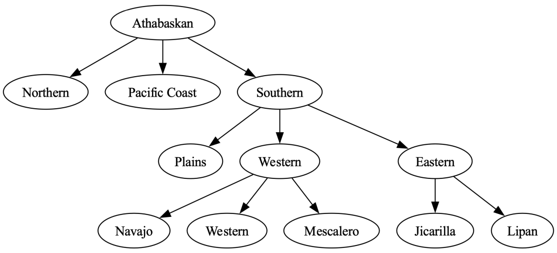


Figure 2: Family Tree for Athabaskan Languages

have led to widespread language loss. Today, their survival depends on urgent and deliberate revitalization efforts (De Costa, 2021), including the development of computational tools for language preservation and accessibility.

3.1 Statistics

At the time of European contact, over 300 Native American languages were spoken across North America (Williams, 2022), belonging to numerous distinct language families (Sutton, 2021). These languages exhibited immense structural diversity, with some featuring polysynthetic morphology (e.g., Mohawk, Inuktitut) (Arkadiev, 2023), complex tone systems (e.g., Athabaskan languages) (Uchihara, 2023), or elaborate evidential marking (e.g., Quechua) (Kalt, 2021). In present-day United States, about 175 Native American languages are still spoken (Antoine, 2021). While some languages like Cherokee and Navajo are better documented, with existing text corpora (Zhang et al., 2021; Goldhahn et al., 2012), many others have little to no surviving linguistic records (Leonard, 2023).

3.2 Endangered Status and Language Loss

The vast majority of Native American languages in the United States are either moribund (Dorzheeva et al., 2021), where they are spoken only by the elderly, or critically endangered (Estrada et al., 2022), where fewer than 100 speakers remain. The statistics are stark: Only about 20 Native American languages are being acquired by children as a first language (Clifton, 2021), and by 2050, at least 90% of Native American languages are predicted to become extinct (Yerian and Halima, 2024).

These figures highlight an accelerating crisis - one driven not only by natural language shift but by centuries of forced assimilation policies, including residential schools that punished Indigenous children for speaking their native tongues (Lomawaima and McCarty, 2025). Even today, Native American communities face systemic barriers to language transmission, from limited access to bilingual education (McCarty and Brayboy, 2021) to the shortage of digital language tools that support continued learning and usage (Meighan, 2021). Without deliberate investment in technological solutions tailored to Indigenous languages, these languages risk further exclusion from digital spaces, thus accelerating their decline.

4 Language Detection Experiments

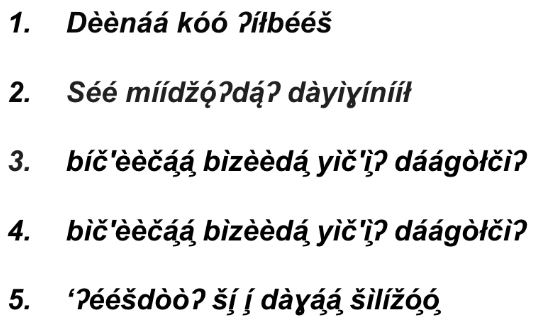
To demonstrate the feasibility of our proposed approach mentioned in Section 2.3, we conduct a small-scale experiment using Plains Apache (Saxon, 2023), an extinct member of the Southern Athabaskan language family. This proof of concept serves as as a preliminary step in our broader effort to build a generalizable Native American language identification system.

4.1 Why Plains Apache?

Plains Apache presents a unique case study for two key reasons. Firstly, the Athabaskan language classifier proposed by Yang et al. (2025b) covered nearly all Southern Athabaskan languages except Plains Apache. Given its linguistic proximity to Navajo and other Apache languages, as shown in Figure 2, incorporating it into the classifier offers a straightforward and scalable expansion. Secondly, unlike Navajo, which still has thousands of speakers, Plains Apache is extinct (Tellmann, 2021), with no known fluent speakers. This makes it an ideal candidate for synthetic data augmentation using the text generation framework for endangered languages proposed by Yang et al. (2025a). If successful, this experiment could serve as a blueprint for generating linguistically sound training data for other highly endangered or extinct Native American languages. By implementing the classifier expansion and synthetic data pipeline with the Plains Apache language, we aim to evaluate the feasibility of our broader research approach on a small scale before scaling to a multi-language setting.

4.2 Manually Gathering Plains Apache Data

Due to the absence of publicly available digital corpora for Plains Apache, we manually scraped and transcribed sentences from various linguistic sources (Wikipedia, 2025; Morgan, 2012). As an



1. Dèènáá kóó ʔíłbééš
2. Séé míídžǫ́ʔdą́ʔ dàyìɣíníił
3. bíč'èèčą́ą́ bìzèèdą́ yìč'į̀ʔ dáágòłčìʔ
4. bìč'èèčą́ą́ bìzèèdą́ yìč'į̀ʔ dáágòłčìʔ
5. ‘ʔééšdòòʔ šį́ į́ dàɣą́ą́ šìlížǫ́ǫ́

Figure 3: Sample sentences of manually curated Plains Apache text

initial effort, we curated 25 Plains Apache sentences in CSV format, with a small sample shown in Figure 3. This manually curated dataset underscores the challenges of working with endangered and extinct Indigenous languages, highlighting the urgent need for automated, scalable solutions such as data augmentation.

4.3 Integration into Athabaskan Classifier

Integrating Plains Apache as an additional language class into the Random Forest classifier yielded interesting results. With all other training weights of the original classifier unchanged, Plains Apache sentences were classified as Navajo with 100% likelihood, as shown in Table 1. In the original experiments conducted by Yang et al. (2025b), Western Apache and Mescalero Apache had the highest classification rates as Navajo at 96.00% and 100%, respectively, while Jicarilla Apache and Lipan Apache performed lower at 92.32% and 62.16%. This disparity was previously attributed to subgroup distinctions, as Jicarilla and Lipan Apache belong to the Eastern branch of Southern Athabaskan, whereas Navajo, Western Apache, and Mescalero Apache fall under the Western subgroup, as illustrated in Figure 2. However, Plains Apache, despite being its own distinct subgroup, exhibited classification behavior identical to Mescalero Apache. This raises new questions about the lexical and syntactic relationships among the Southern Athabaskan subgroups, warranting further analysis.

Language	Classified as Navajo	Total Sentences
Western Apache	96.00%	25
Mescalero Apache	100.00%	32
Jicarilla Apache	92.31%	13
Lipan Apache	62.16%	37
Plains Apache	100.00%	25

Table 1: Classification Results for Apache Languages: Percentage of sentences classified as Navajo and total number of sentences examined for each Apache language, with the addition of Plains Apache, highlighted in pink.

4.4 Synthetic Data Generation for Plains Apache

We applied the framework introduced by Yang et al. (2025a) to expand our Plains Apache text. Originally developed for the endangered Nüshu language, this approach combines few-shot prompting with language-specific tailoring to generate new synthetic data. Using the GPT-4o model, we provided a dataset of 25 Plains Apache sentences and prompted the model to generate 5 new artificial sentences, which it successfully produced[3]. While this represents a small-scale test, it highlights the potential of synthetic augmentation for highly endangered Indigenous languages, even in cases of extreme data scarcity. Moving forward, this methodology could be extended to other extinct or moribund Native American languages, significantly increasing the amount of available data for classification, modeling, and revitalization efforts.

5 Conclusion

This paper presents a long-term research vision for developing a generalizable Native American language identification system, addressing the critical absence of Indigenous languages in commercial language technologies. By building on existing work in Native American language classification and synthetic data generation, we propose a unified approach that leverages both to bridge this gap. Our small-scale experiments integrating Plains Apache demonstrate the promise and feasibility of this method. Beyond its technical contributions, this work serves as a call to action for the broader NLP community to invest in decentralized, community-driven language technologies that prioritize linguistic diversity. Through collective efforts, we can ensure that these languages are not only preserved, but actively recognized and used in the digital age.

[3] These generated sentences have not yet been rigorously validated beyond a visual review; we propose this as a viable method for data augmentation rather than asserting complete accuracy.

Limitations

While this study lays the groundwork for Native American language identification, limitations remain. The Plains Apache experiment, though informative, is constrained by scarce natural data, and while synthetic augmentation mitigates this, it cannot fully replicate the depth of naturally spoken language. Our focus on Athabaskan languages also raises questions about the broader applicability of this approach to other linguistic families. Additionally, reliance on synthetic data poses risks of capturing artifacts rather than true linguistic features. Beyond identification, future work must explore applications like translation and speech recognition for meaningful impact. Expanding datasets, refining augmentation techniques, and engaging Indigenous communities will be essential to ensuring these technologies support both linguistic and cultural preservation.

Ethics Statement

Ethical considerations are important when developing technologies for Native American languages, which have a big role in cultural, spiritual, and historical settings. This study recognizes that these languages are not only tools for communication but also symbols of culture and heritage. Thus, the development of language technologies for Native American languages should happen in close collaboration with community members and leaders to ensure language preservation rather than cultural homogenization and appropriation. We are actively engaging with the Native American and Indigenous Languages department at our institution to ensure this project is conducted in a thoughtful, respectful, and community-centered manner.

References

Thejaswi Adimulam, Swetha Chinta, and Suprit Kumar Pattanayak. 2022. Transfer learning in natural language processing: Overcoming low-resource challenges. *International Journal of Enhanced Research In Science Technology & Engineering*, 11:65–79.

Jurgita Antoine. 2021. New grant to support aihec's native languages program. *Tribal College*, 33(2):1–2.

Peter Arkadiev. 2023. Polysynthesis: lessons from northwest caucasian languages. In *Mediterranean Morphology Meetings*, volume 13, pages 1–26.

Rachit Bansal, Himanshu Choudhary, Ravneet Punia, Niko Schenk, Émilie Pagé-Perron, and Jacob Dahl. 2021. How low is too low? a computational perspective on extremely low-resource languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 44–59.

Mary Bucholtz and Kira Hall. 2004. Language and identity. *A companion to linguistic anthropology*, 1:369–394.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language id in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608.

William M Clements. 2021. *Native American verbal art: Texts and contexts*. University of Arizona Press.

PYE Clifton. 2021. Documenting the acquisition of indigenous languages. *Journal of Child Language*, 48(3):454–479.

Li Congrong. 2024. History, characteristics, and modern vitality of nüshu: A cultural anthropology perspective. *Anthropological Explorations of Gender, Identity, and Economics*, 85.

Peter I De Costa. 2021. Indigenous language revitalization: how education can help reclaim "sleeping" languages. *Journal of Language, Identity & Education*, 20(5):355–361.

Ming Di. 2024. The other mother tongues and minority writing in china. *Mother Tongues and Other Tongues: Creating and Translating Sinophone Poetry*, 53:224.

Victoria Vladimirovna Dorzheeva et al. 2021. Preservation of indigenous languages in the united states. legislation and challenges. *European Proceedings of Social and Behavioural Sciences*.

Katherine Ellinghaus. 2022. *Blood will tell: Native Americans and assimilation policy*. U of Nebraska Press.

Alejandro Estrada, Paul A Garber, Sidney Gouveia, Álvaro Fernández-Llamazares, Fernando Ascensão, Agustin Fuentes, Stephen T Garnett, Christopher Shaffer, Júlio Bicca-Marques, Julia E Fa, et al. 2022. Global importance of indigenous peoples, their lands, and knowledge systems for saving the world's primates from extinction. *Science advances*, 8(31):eabn2927.

Abdul Ghafoor, Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, Rakhi Batra, Mudasir Ahmad Wani, et al. 2021. The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing. *IEEE Access*, 9:124478–124490.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*.

Dhiman Goswami, Sharanya Thilagan, Kai North, Shervin Malmasi, and Marcos Zampieri. 2024. Native language identification in texts: A survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3149–3160.

Lenore A Grenoble and Lindsay J Whaley. 1998. *Endangered languages: Language loss and community response*. Cambridge University Press.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*.

Yi-Wen Huang. 2024. Language loss and translingual identities near the navajo land. *International Journal of Language Studies*, 18(2).

Susan E Kalt. 2021. Acquisition, loss and innovation in chuquisaca quechua—what happened to evidential marking? *Languages*, 6(2):76.

Masud Khawaja. 2021. Consequences and remedies of indigenous language loss in canada. *Societies*, 11(3):89.

Lachman Mulchand Khubchandani. 2016. The relationship between language and culture is interwoven in a unique manner in different traditions. one of the major consequences of technology-driven globalization has been the increasing marginalization of less-populated language communities and the intimidating hegemony of larger socio-economic networks. this phenomenon acquires more vis. *The Language Loss of the Indigenous*, page 183.

Wesley Y Leonard. 2023. Challenging "extinction" through modern miami language practices. In *Global Language Justice*, pages 126–165. Columbia University Press.

Ying Lin, Shengqi Yang, Veselin Stoyanov, and Heng Ji. 2018. A multi-lingual multi-task architecture for low-resource sequence labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 799–809.

Adrianna Link, Abigail Shelton, and Patrick Spero. 2021. *Indigenous languages and the promise of archives*. U of Nebraska Press.

Ling Liu, Zach Ryan, and Mans Hulden. 2021. The usefulness of bibles in low-resource machine translation. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 44–50.

K Tsianina Lomawaima and Teresa L McCarty. 2025. *" To remain an Indian": Lessons in democracy from a century of Native American education*. Teachers College Press.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2022. Data augmentation for low-resource languages nmt guided by constrained sampling. *International Journal of Intelligent Systems*, 37(1):30–51.

Teresa L McCarty and Bryan McKinley Jones Brayboy. 2021. Culturally responsive, sustaining, and revitalizing pedagogies: Perspectives from native american education. In *The Educational Forum*, volume 85, pages 429–443. Taylor & Francis.

Barbra A Meek. 2012. *We are our language: An ethnography of language revitalization in a Northern Athabaskan community*. University of Arizona Press.

Paul J Meighan. 2021. Decolonizing the digital landscape: The role of technology in indigenous language revitalization. *AlterNative: An International Journal of Indigenous Peoples*, 17(3):397–405.

Peggy J Miller and Lisa Hoogstra. 1992. Language as tool in the socialization and apprehension of cultural meanings. *New directions in psychological anthropology*, 3:83–101.

Juliet Liane Morgan. 2012. *Classificatory Verbs in Plains Apache*. Ph.D. thesis, University of Oklahoma.

New Mexico Secretary of State. 2025. Native american languages in new mexico.

Michael Leroy Oberg and Peter Jakob Olsen-Harbich. 2022. *Native America: a history*. John Wiley & Sons.

Kayla Palakurthy. 2022. New speakers and language change in diné bizaad (navajo). *International Journal of Bilingualism*, 26(5):601–619.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Leslie Saxon. 2023. 39 dene–athabaskan. *The Languages and Linguistics of Indigenous North America: A Comprehensive Guide, Vol. 2*, 13:875.

Victoria Sear and Mark Turin. 2021. Locating criticality in the lexicography of historically marginalized languages. *history of humanities*, 6(1):237–259.

Gurinder Singh, Astha Gupta, Pranay Verma, Naina Chaudhary, Rajneesh Kler, and Ayush Thakur. 2024. Catalyzing multilingual nlp: New methods for low-resource language support. In *2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pages 67–75. IEEE.

Anca Sirbu. 2015. The significance of language as a tool of communication. *Scientific Bulletin" Mircea cel Batran" Naval Academy*, 18(2):405.

Mark Q Sutton. 2021. *An introduction to native North America*. Routledge.

Bryce D Tellmann. 2021. *The Great Plains and the Available Means of Regionalism*. The Pennsylvania State University.

Hiroto Uchihara. 2023. 3 tone. *The Languages and Linguistics of Indigenous North America: A Comprehensive Guide, Vol 1*, 13:63.

Wikipedia. 2025. Plains apache language.

Roger Williams. 2022. *A Key Into the Language of America: The Help to the Language of the Natives in That Part of America Called New-England*. DigiCat.

Ivory Yang, Weicheng Ma, and Soroush Vosoughi. 2025a. Nüshurescue: Reviving the endangered nüshu language with ai. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7020–7034.

Ivory Yang, Weicheng Ma, Chunhui Zhang, and Soroush Vosoughi. 2025b. Is it navajo? accurate language detection in endangered athabaskan languages. *arXiv preprint arXiv:2501.15773*.

Keli Yerian and Bibi Halima. 2024. Language endangerment and revitalization. *Learning How to Learn Languages*.

Thomas Yim. 2024. Technology's dual role in language marginalization and revitalization. *GRACE: Global Review of AI Community Ethics*, 2(1).

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2021. Chrentranslate: Cherokee-english machine translation demo with quality estimation and corrective feedback. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 272–279.

A Exploration with Support Vector Machines

While not discussed in detail in this paper, we also explored Support Vector Machines (SVM) as a potential alternative or complement to the proposed Random Forest classifier. We initialized an SVM classifier with a linear kernel and probability outputs, using `GridSearchCV` with cross-validation on the F1 score for hyperparameter tuning. Due to the computational demands of SVM training, we leveraged research computing resources, setting `n_jobs` to 32 for parallel processing. Initial results were largely coherent, though further investigation is needed to assess its comparative effectiveness.