Advancing Uto-Aztecan Language Technologies: A Case Study on the Endangered Comanche Language

Jesus Alvarez C, Daua D. Karajeanes, Ashley Celeste Prado, John Ruttan, Ivory Yang, Sean O'Brien, Vasu Sharma, Kevin Zhu
Algoverse AI Research
ivory.yang.gr@dartmouth.edu, kevin@algoverse.us

Abstract

The digital exclusion of endangered languages remains a critical challenge in NLP, limiting both linguistic research and revitalization efforts. This study introduces the first computational investigation of Comanche, an Uto-Aztecan language on the verge of extinction, demonstrating how minimal-cost, community-informed NLP interventions can support language preservation. We present a manually curated dataset of 412 phrases, a synthetic data generation pipeline, and an empirical evaluation of GPT-4o and GPT-4o-mini for language identification. Our experiments reveal that while LLMs struggle with Comanche in zero-shot settings, few-shot prompting significantly improves performance, achieving near-perfect accuracy with just five examples. Our findings highlight the potential of targeted NLP methodologies in low-resource contexts and emphasize that visibility is the first step toward inclusion. By establishing a foundation for Comanche in NLP, we advocate for computational approaches that prioritize accessibility, cultural sensitivity, and community engagement.

1 Introduction

The decline of endangered languages represents not only a linguistic loss (Low et al., 2022) but also the erosion of invaluable cultural, historical, and ecological knowledge (Tulloch, 2006; Cámara-Leret and Bascompte, 2021; Sallabank and Austin, 2023). Despite growing advancements in language technology, computational efforts overwhelmingly favor widely spoken languages, leaving endangered languages largely unsupported (Meighan, 2021; Yang et al., 2025a; Jerpelea et al., 2025). Over 88% of the world's languages have minimal to no representation in mainstream language technologies, exacerbating their digital marginalization (Rangel, 2019). This exclusion hinders linguistic research and deepens the digital divide (Valijärvi and Kahn, 2023; Yang et al., 2025b), complicating preservation and revitalization efforts. Among these, Comanche, an Uto-Aztecan language, faces imminent extinction, with fewer than 50 fluent speakers remaining (Chaika et al., 2024).

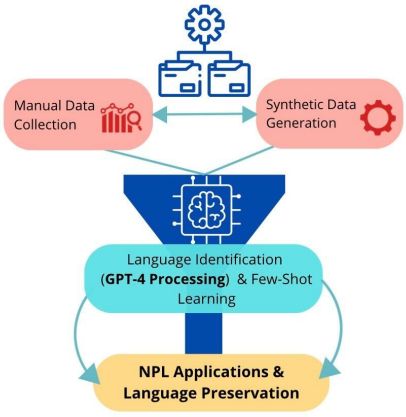

Figure 1: Stylized overview of our exploration of NLP applications for the endangered Comanche language.

We present a case study on the Comanche language, demonstrating that with minimal cost and computational resources, it is possible to achieve what large corporations and academic institutions have largely neglected. As shown in Figure 1, we contribute (1) a manually curated dataset, (2) synthetic data generation pipeline for resource expansion, and (3) an empirical evaluation of GPT-4o and GPT-4o-mini in zero-shot and few-shot language identification. Our findings highlight the potential of large language models (LLMs) in low-resource settings, offering insights into their applicability for endangered language preservation. **This work marks the first-ever introduction of Comanche into the NLP domain, laying groundwork for future research and linguistic equity**.

Contact other authors at: jalvarezc@my.canyons.edu, d.d.karajeanes@student.tue.nl, aprad054@fiu.edu, jruttan3@uwo.ca, seobrien@ucsd.edu, vasus@andrew.cmu.edu

Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 27–37

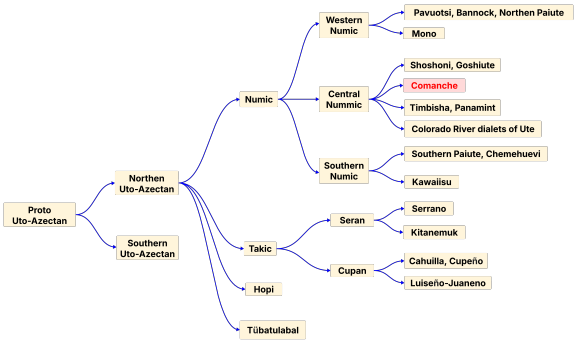

Figure 2: Family tree for Uto-Aztecan Languages, with Comanche highlighted.

2 Related Work

Efforts to preserve endangered languages, particularly Native American languages, date back to at least the early 20th century (Charney, 1993), with early approaches relying heavily on linguistic documentation and literary preservation (Schwartz and Dobrin, 2016). While these foundational efforts paved the way, they were hindered by the scarcity of available datasets and standardized benchmarks, leading researchers to explore alternative strategies for text processing (Lorenzo et al., 2024; Spencer and Kongborrirak, 2025). Despite these advancements, modern computational linguistics continues to face significant challenges when working with polysynthetic languages such as Comanche and Apache, due to their intricate orthographic and morphological structures (Kelly, 2020). In recent years there have been promising community-led revitalization initiatives, including immersive education programs, digital archives, and collaborations with computational linguists (of Indian Affairs, 2023; Schwartz et al., 2021).

Data scarcity remains a fundamental challenge in NLP (Glaser et al., 2021). Unlike widely spoken languages with abundant corpora, low-resource languages lack annotated datasets, limiting the effectiveness of LLMs for preservation (Zhong et al., 2024; Dinh et al., 2024). Few-shot prompting has emerged as a promising solution, allowing LLMs to generate synthetic data from minimal examples (Zhang et al., 2021), though its success hinges on data quality. Transfer learning (Adimulam et al., 2022) has also been explored to improve low-resource NLP, but without robust evaluation frameworks tailored for Indigenous languages (Shu et al., 2024), achieving meaningful generalization remains a challenge (Mager et al., 2023).

3 Native American Language Landscape

The linguistic diversity of Native American languages is vast, spanning multiple language families with distinct phonetic, morphological, and syntactic properties. Despite this richness, many of these languages are critically endangered, with fluency declining due to historical policies of forced assimilation, boarding schools, and sociopolitical marginalization (Krauss, 1992). Language documentation efforts have attempted to counteract this loss, but computational resources remain scarce, and mainstream NLP models are ill-equipped to process these languages effectively (Blasi et al., 2022). The lack of digital resources further exacerbates the challenge, preventing these languages from benefiting from advances in language technologies (U.S. Department of the Interior, 2022).

Comanche belongs to the Uto-Aztecan language family, one of the largest language families in the Americas, encompassing over 60 languages spoken across the western United States, Mexico, and Central America (Opler, 1943). While some Uto-Aztecan languages, such as Nahuatl (Andrews, 2003), have relatively larger speaker populations and a degree of digital presence, others, including Comanche, face imminent extinction. As shown in Figure 2, Comanche developed as a distinct language after diverging from Shoshone in the 18th century, evolving unique phonological and lexical features (Casagrande, 1955). Today, with fewer than 50 fluent speakers, Comanche lacks sufficient linguistic resources for computational modeling.

4 Data

4.1 Manual Data Collection

To construct a foundational dataset for Comanche, we conducted a systematic review of linguistic resources, including academic literature, digital archives, and historical records. Given the scarcity of publicly available corpora, we aggregated and curated data from 15 distinct domains (Appendix B), ensuring consistency through transcription and standardization. To enhance data reliability, we cross-referenced linguistic materials with community-driven documentation efforts, validating authenticity and linguistic accuracy. This structured dataset of 412 Comanche phrases, the first digitalized dataset of its kind, serves as a crucial resource for both language preservation and computational linguistic research in Comanche.

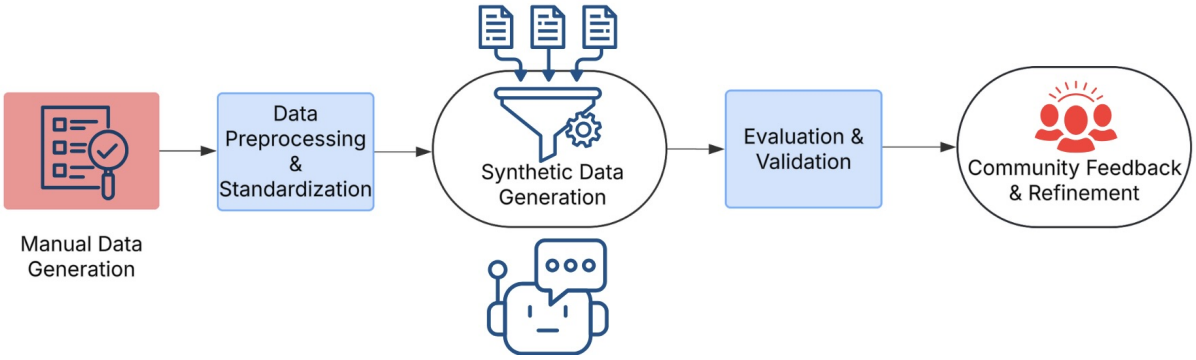

Figure 3: Data pipeline.

4.2 Synthetic Data Generation

Given the extreme scarcity of parallel Comanche–English text, we leveraged few-shot prompting with GPT-4o to generate synthetic translations. Using a manually curated dataset of 100 Comanche–English sentence pairs, we split the data into an 80% training set and a 20% test set. During training, GPT-4o was provided examples from the training subset and then prompted to generate translations for the test set (Appendix C). The generated outputs were evaluated using normalized Levenshtein similarity, ensuring a minimum quality threshold of 0.1[1] before incorporation into the dataset. This controlled expansion strategy maintained linguistic integrity while demonstrating that even minimal data can be effectively leveraged to create valuable resources for endangered language NLP. While the pipeline shown in Figure 3 is in early stage, it underscores the potential of leveraging NLP for endangered language documentation and expansion. As data scarcity persists, synthetic augmentation offers a scalable approach to bridge resource gaps and support revitalization efforts.

5 Language Identification

While data collection and synthetic expansion are crucial aspects of language preservation, identification is equally essential. Despite supporting over 200 languages, Google's LangID system (Caswell et al., 2020) does not include a single Native American language, including Comanche, highlighting the systematic exclusion of these languages from mainstream computational resources. This absence not only limits automatic language identification

Figure 4: GPT-4o achieves a remarkable improvement in language identification performance, with the help of few-shot examples.

capabilities, but also further marginalizes endangered languages in digital spaces, making their preservation even more challenging.

While LLMs have demonstrated remarkable proficiency in high-resource language tasks, their ability to identify low-resource languages remains a critical challenge. In our zero-shot prompting experiments using a dataset of 412 Comanche entries, GPT-4o achieved only 13.5% accuracy, correctly identifying 56 instances. These results highlight the broader issue that without explicit guidance, even state-of-the-art models struggle to recognize endangered languages. To address this limitation, we introduced few-shot prompting with both Comanche and English samples, training the model to actively identify features of the Comanche language. For this experiment, we used a sample of 100 randomly

[1]Given that Comanche has never been explored in NLP, we set a baseline threshold of 0.1 due to the difficulty of the task. As the pipeline matures, we will refine our evaluation criteria and increase the required similarity score.

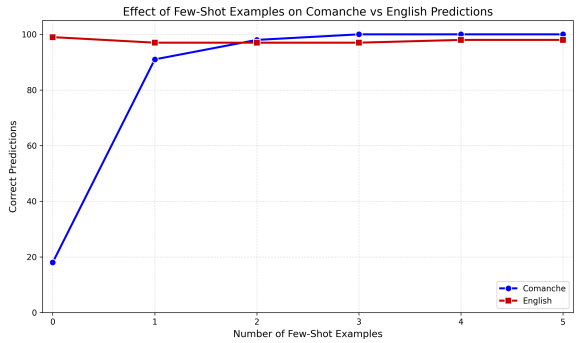

Figure 5: Effect of Few-Shot Examples on Comanche Prediction Accuracy.

selected entries from our original dataset. Each few-shot pair included one Comanche phrase and a randomized English entry from the dataset, as shown in Figure 4. With just one Comanche example, GPT-4o achieved 91% accuracy in identification of Comanche. Extending to a three-shot strategy consistently yielded 100% accuracy, as shown in Figure 5. Notably, English identification accuracy remained consistently high (97-100%) across all experimental conditions. These findings underscore the limitations of default language identification systems and demonstrate that even minimal targeted prompting can significantly enhance recognition capabilities. The stark performance gap between Comanche and English underscores the model's inherent bias toward high-resource languages when tasked with identification. Our results provide a scalable, low-resource approach for integrating endangered languages into NLP systems, offering a pathway toward more inclusive computational language technologies.

6 Community Feedback

To ensure that our approach to NLP-driven language preservation is both transparent and respectful, we engaged with a community member of Comanche and Rarámuri heritage through a semi-structured interview. The interview provided insights into the lived experiences of individuals connected to endangered languages, highlighting both the cultural significance of linguistic preservation and the challenges posed by data scarcity.

The interviewee shared that although the Comanche and Rarámuri languages were not passed down to him, he maintains a profound connection to his Native American heritage. He recounted a childhood experience in which he struggled to communicate with members of a Rarámuri community in Chihuahua, Mexico, due to language barriers[2]. His reflections highlight the critical role that digital resources and computational methods can play in language preservation. While exposure to artificial intelligence and NLP technologies remains limited in many Indigenous communities, the potential for these tools to support language revitalization is immense. Our study emphasizes that responsible NLP research must engage directly with affected communities, ensuring that technological interventions align with cultural needs and ethical considerations.

7 Future Work

Future efforts will focus on expanding the manually curated Comanche dataset, refining the synthetic data generation pipeline, and developing a real-time language identification demo. Given the largely oral nature of Comanche, we will also investigate audio-based approaches to support speech recognition and transcription, as well as exploring learning (Wang and Guo, 2019; Mangar et al., 2025) and reading comprehension tasks (Zhang et al., 2024). Additionally, we will actively engage with more Comanche community members to ensure our work remains aligned with their needs and perspectives. We hope to eventually secure the resources to conduct a deeper analysis of Comanche and other indigenous languages—work that has largely been limited to high-resource languages—examining dimensions such as linguistic features (Lee et al., 2024), implicit versus explicit expression (Wang et al., 2025), persuasive strategies (Wang et al., 2024; Yang et al., 2024) and intellectual humility (Guo et al., 2024).

8 Conclusion

This study represents the first computational effort to integrate Comanche into the NLP landscape, addressing critical gaps in language documentation and technological accessibility. Through manual data collection, synthetic data expansion, and empirical evaluations of LLM-based language identification, we demonstrate that even minimal resources can yield meaningful improvements in language modeling for endangered languages. While this work marks an initial step, continued collaboration

[2]The interviewee recalled attempting to explain to local Rarámuri residents that his disposable camera differed from a Polaroid and would not produce an immediate photograph. This miscommunication left a lasting impression on him, reinforcing the importance of language technologies.

with Comanche speakers, expansion into audio-based methods, and refinement of evaluation metrics will be essential to advancing these efforts. We advocate for a NLP research paradigm that actively includes Indigenous and low-resource languages, ensuring that they are not only preserved but empowered through computational advancements.

Limitations

Despite the contributions of this study, several limitations must be acknowledged. Firstly, the manually curated Comanche dataset remains small, constraining both model performance and generalizability. Future work must expand this dataset to improve model robustness and alignment (Zeng et al., 2025), as well as to prevent biases (Guan et al., 2025). In addition, while synthetic data augmentation offers a promising avenue for resource expansion, the quality of generated translations is inherently dependent on the prompting strategy (Jian et al., 2022) and the capabilities of the underlying language model. Further refinements to the pipeline and more rigorous evaluation methodologies are necessary to ensure linguistic accuracy. Moreover, our experiments focus primarily on text-based language identification, overlooking the oral tradition of Comanche. Future research should incorporate audio-based approaches, such as automatic speech recognition, to better align with the language's natural form. Lastly, our engagement with community members, while valuable, represents only an initial step. Sustained collaboration with Comanche speakers and language advocates will be essential to ensuring that computational interventions align with community priorities and ethical considerations.

Ethics Statement

Our research adheres to ethical principles that prioritize Indigenous data sovereignty, cultural sensitivity, and responsible engagement. We collected Comanche words, affixes, and phrases exclusively from publicly available sources, ensuring transparency in our data practices and proper attribution of all resources. All relevant citations for the manual dataset can be found in Appendix D. Consistent with the principles outlined by Schwartz (2022), we acknowledge that Indigenous languages are deeply tied to cultural identity, historical continuity, and community sovereignty. We explicitly recognize the Comanche Nation as the rightful stewards of their language and are committed to ensuring that our work aligns with their goals of preservation and revitalization. Our research seeks not only to document but to actively contribute to the accessibility and visibility of Comanche within the computational linguistics community. We emphasize the importance of relational engagement with Indigenous communities, acknowledging that linguistic data is not merely an artifact for academic study but also a living expression of cultural heritage (Appendix A).

Finally, we uphold ethical obligations of cognizance, beneficence, accountability, and non-maleficence. We remain committed to avoiding harm, ensuring that our findings and datasets serve as tools for language empowerment rather than extraction. Future work will continue to involve direct engagement with Comanche speakers, fostering a collaborative research framework that respects community agency and cultural priorities. In the spirit of transparent and ethical research, our full dataset and code have been made available at (`https://github.com/comanchegenerate/ComancheSynthetic`).

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.

James Richard Andrews. 2003. *Introduction to classical Nahuatl*, volume 1. University of Oklahoma Press.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Rodrigo Cámara-Leret and Jordi Bascompte. 2021. Language extinction triggers the loss of unique medicinal knowledge. *Proceedings of the National Academy of Sciences*, 118(24):e2103683118.

Joseph B Casagrande. 1955. Comanche linguistic acculturation iii. *International Journal of American Linguistics*, 21(1):8–25.

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.

O. Chaika, N. Sharmanova, and O. Makaruk. 2024. Revitalising endangered languages: Challenges, successes, and cultural implications. *Futurity of Social Sciences*, 2(2):38–61.

Jean Ormsbee Charney. 1993. *A Grammer of Comanche*. University of Nebraska Press.

Nguyen Dinh, Thanh Dang, Luan Thanh Nguyen, and Kiet Nguyen. 2024. Multi-dialect vietnamese: Task, dataset, baseline models and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7498.

Ingo Glaser, Shabnam Sadegharmaki, Basil Komboz, and Florian Matthes. 2021. Data scarcity: Methods to improve the quality of text classification. In *ICPRAM*, pages 556–564.

Xin Guan, Nate Demchak, Saloni Gupta, Ze Wang, Ediz Ertekin Jr, Adriano Koshiyama, Emre Kazim, and Zekun Wu. 2025. Saged: A holistic bias-benchmarking pipeline for language models with customisable fairness calibration. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3002–3026.

Xiaobo Guo, Neil Potnis, Melody Yu, Nabeel Gillani, and Soroush Vosoughi. 2024. The computational anatomy of humility: Modeling intellectual humility in online public discourse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5701–5723.

Alexandru-Iulius Jerpelea, Alina Radoi, and Sergiu Nisioi. 2025. Dialectal and low resource machine translation for aromanian. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7209–7228.

Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2022. Contrastive learning for prompt-based few-shot language learners. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5577–5587.

Kevin Kelly. 2020. *An Evaluation of Parallel Text Extraction and Sentence Alignment for Low-Resource Polysynthetic Languages*. Ph.D. thesis, University of Groningen.

Michael Krauss. 1992. The world's languages in crisis. *Language*, 68(1):4–10.

Dongryeol Lee, Yerin Hwang, Yongil Kim, Joonsuk Park, and Kyomin Jung. 2024. Are llm-judges robust to expressions of uncertainty? investigating the effect of epistemic markers on llm-based evaluation. *arXiv preprint arXiv:2410.20774*.

Abelardo Carlos Martínez Lorenzo, Pere-Luís Huguet Cabot, Karim Ghonim, Lu Xu, Hee-Soo Choi, Alberte Fernández Castro, and Roberto Navigli. 2024. Mitigating data scarcity in semantic parsing across languages: the multilingual semantic layer and its dataset. In *The 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Dylan Scott Low, Isaac Mcneill, and Michael Day. 2022. Endangered languages: A sociocognitive approach to language death, identity loss, and preservation in the age of artificial intelligence. *Sustainable Multilingualism*, 21(1):1–25.

Manuel Mager, Arturo Oncevay, Annette Rios, Jamshidbek Mirzakhalov, and Katharina Kann. 2023. The role of computational linguistics in indigenous language revitalization: Challenges and opportunities. In *Proceedings of the 1st Workshop on Computation for Indigenous Languages (C3NLP)*, pages 23–31.

Ravindra Mangar, Cesar Arguello, David Inyangson, Tina Pavlovich, Karen Gareis, and Tushar M Jois. 2025. Engaging students from under-represented groups to pursue graduate school in computer science and engineering. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, pages 742–748.

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.

Bureau of Indian Affairs. 2023. Native language revitalization literature review.

Marvin K Opler. 1943. The origins of comanche and ute. *American Anthropologist*, 45(1):155–158.

Jhonnatan Rangel. 2019. Challenges for language technologies in critically endangered languages. In *UNESCO International Conference Language Technologies for All (LT4All)*, Paris, France. ⟨hal-02917830⟩.

Julia Sallabank and Peter K Austin. 2023. Endangered languages. In *The Routledge handbook of applied linguistics*, pages 362–373. Routledge.

Lane Schwartz. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 724–731, Dublin, Ireland. Association for Computational Linguistics.

Lane Schwartz, Emily Chen, Hyunji Hayley Park, Edward Jahn, and Sylvia L.R. Schreiner. 2021. A digital corpus of st. lawrence island yupik. *arXiv preprint*.

Saul Schwartz and Lise M Dobrin. 2016. The cultures of native north american language documentation and revitalization. *Reviews in Anthropology*, 45(2):88–123.

Peng Shu, Junhao Chen, Zhengliang Liu, Hui Wang, Zihao Wu, Tianyang Zhong, Yiwei Li, Huaqin Zhao, Hanqi Jiang, Yi Pan, et al. 2024. Transcending language boundaries: Harnessing llms for low-resource language translation. *arXiv preprint arXiv:2411.11295*.

Piyapath T Spencer and Nanthipat Kongborrirak. 2025. Can llms help create grammar?: Automating grammar creation for endangered languages with in-context learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10214–10227.

Shelley Tulloch. 2006. Preserving dialects of an endangered language. *Current Issues in Language Planning*, 7(2-3):269–286.

U.S. Department of the Interior. 2022. Federal Indian Boarding School Initiative Investigative Report. Accessed: 2025-03-04.

Riitta-Liisa Valijärvi and Lily Kahn. 2023. The role of new media in minority-and endangered-language communities. *Endangered Languages in the 21st Century. Abingdon, Oxon, England: Routledge*, pages 139–157.

Chixiang Wang and Junqi Guo. 2019. A data-driven framework for learners' cognitive load detection using ecg-ppg physiological feature fusion and xgboost classification. *Procedia computer science*, 147:338–348.

Yuxin Wang, Ivory Yang, Saeed Hassanpour, and Soroush Vosoughi. 2024. Mentalmanip: A dataset for fine-grained analysis of mental manipulation in conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3747–3764.

Yuxin Wang, Xiaomeng Zhu, Weimin Lyu, Saeed Hassanpour, and Soroush Vosoughi. 2025. Impscore: A learnable metric for quantifying the implicitness level of sentences. In *The Thirteenth International Conference on Learning Representations*.

Ivory Yang, Xiaobo Guo, Sean Xie, and Soroush Vosoughi. 2024. Enhanced detection of conversational mental manipulation through advanced prompting techniques. In *Eighth Widening NLP Workshop (WiNLP 2024) Phase II*.

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*.

Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. 2025. Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10602–10617.

Ningyu Zhang, Luoqiu Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, and Huajun Chen. 2021. Differentiable prompt makes pre-trained language models better few-shot learners. *arXiv preprint arXiv:2108.13161*.

Yuhong Zhang, Shilai Yang, Gert Cauwenberghs, and Tzyy-Ping Jung. 2024. From word embedding to reading embedding using large language model, eeg and eye-tracking. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4. IEEE.

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, et al. 2024. Opportunities and challenges of large language models for low-resource languages in humanities research. *arXiv preprint arXiv:2412.04497*.

A Appendix A

Figure 6: Map showing approximate locations of Indigenous peoples of the Great Plains prior, to displacement in the 19th century. Comanche territory is depicted in the bottom-left region. Source:https://www.britannica.com/place/Great-Plains#/media/1/243562/330.

Figure 7: Comanche cultural artifacts.

Figure 8: Lloyd Heminokeky, Jr., Language Consultant for the Comanche Nation Language Department, hosts an event honoring Comanche Code Talkers, including his grandfather, Technician Fifth Grade Wellington Mihecoby, whose distinguished service is highlighted in the portrait beside him. Source: https://youtu.be/M_JO8C63Ins?si=uc8JzAiAX7sCrF9A.

B Appendix B

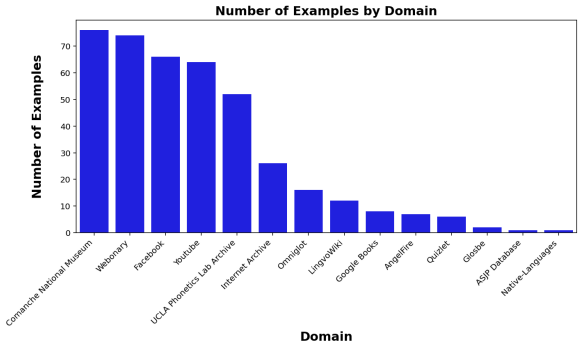

Figure 9: Distribution of manually collected Comanche-English phrases across 15 sources. The Comanche National Museum, Webonary, and Facebook (via the Comanche Nation Language Department) contributed the highest number of examples. This distribution underscores the variability in available linguistic resources for Comanche.

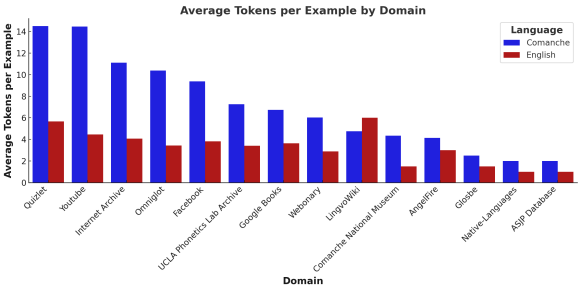

Figure 10: The average token length per example differs notably between Comanche (blue) and English (red). Some sources, such as Quizlet and Youtube, contain significantly longer Comanche phrases, while others, such as LingvoWiki, show an inverse pattern due to the presence of affixes and bound morphemes. These variations highlight source-specific differences, particularly in how morphology and translation conventions impact token length.

C Appendix C

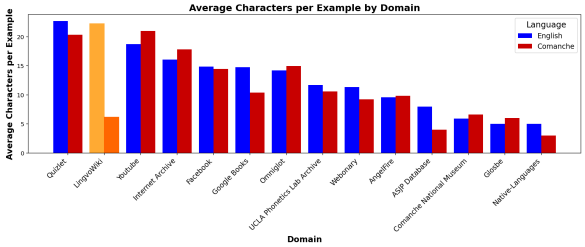

Figure 11: Comparison of average characters per example across various sources of English and Comanche. Notably, the Comanche data from LingvoWiki appears unusually short due to the presence of affixes in the collected samples, which artificially lowers the character count for that source.

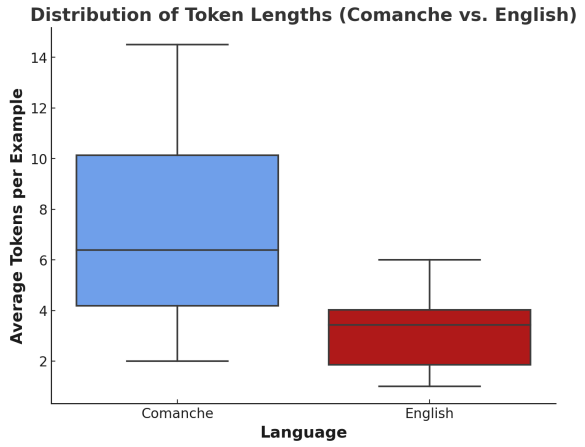

Figure 12: Box plot distribution of token lengths in Comanche and English phrases. Comanche phrases exhibit a wider range of token lengths, with a median of around 6 tokens per example, and an extended upper quartile value, reflecting its polysynthetic structure. English translations, by contrast are more compact with less variability.

English	Real Comanche	GPT Comanche	Levenshtein Similarity
Everyday speak Comanche.	Oyo?tabeni nɨmɨtekwa	Oyotabeni nɨmɨ tekwapɨ	69.23%
I see a bird	huutsu?a nɨ puni	huutsuu nɨ puni	72.22%
I lost my money that I had put in a can	nɨ buhiwi hta nɨ' narohtɨma kɨ nɨ rɨki 'ih a watsi kɨ'	Nɨ wana nɨ ɨnɨmɨ ɨ hɨpi nɨ puni	31.48%
I don't understand you	ke nɨ ɨ nakɨsupana?itɨ	Keta nɨ nakɨsupana?itɨ nɨ	64.29%
We dance for our People	Nɨnɨse taanahnɨmɨnɨ?a nɨhkakɨrɨ	Taanɨmɨnɨ?a?nanɨsutaikɨrɨ	36.36%

Figure 13: Comparison of English sentences with their corresponding Real Comanche translations and GPT-generated Comanche translations.

D Appendix D

Table 1: Online sources referenced to construct the Comanche-English dataset.

Cite Key	Author (Year)	Title	URL
omniglotwriting	Omniglot (n.d.)	Comanche language, alphabet and pronunciation	https://www.omniglot.com/writing/comanche.htm
omniglotphrases	Omniglot (n.d.)	Comanche phrases	https://www.omniglot.com/language/phrases/comanche.htm
ucla1992	UCLA Phonetics Lab Archive (1992)	Comanche word lists (1992)	https://archive.phonetics.ucla.edu/Language/COM/
angelfire	Angelfire (n.d.)	Comanche language page	https://www.angelfire.com/creep2/fracod/comanche.html
rosettaproject	Internet Archive (n.d.)	rosettaproject_com_morsyn-1	https://archive.org/details/rosettaproject_com_morsyn-1/page/n3/mode/2up
cnlanguagefb	Comanche Nation Language Dept. (n.d.)	Facebook videos	https://www.facebook.com/CNLanguage/videos/
comanchemuseum	Comanche National Museum and Cultural Center (n.d.)	Comanche dictionary	https://www.comanchemuseum.com/dictionary.html
nativelang	Native Languages of the Americas (n.d.)	Comanche language: Word sets	https://www.native-languages.org/comanche_words.htm
glosbecomanche	Glosbe (n.d.)	English–Comanche dictionary	https://glosbe.com/en/com
asjpcomanche	ASJP (n.d.)	COMANCHE	https://asjp.clld.org/languages/COMANCHE
webonarycomanche	Comanche Dictionary Project (n.d.)	Comanche webonary	https://www.webonary.org/comanche/
lingvoforum	LingvoForum (n.d.)	Comanche dictionary (LingvoForum Wiki)	https://wiki.lingvoforum.net/wiki/Comanche_dictionary
quizletcomanche	Quizlet (n.d.)	Comanche phrases flashcards	https://quizlet.com/718424723/comanche-phrases-flash-cards/
youtubecn	Comanche Nation Language Dept. (n.d.)	CNLanguage YouTube channel	https://www.youtube.com/@CNLanguage/videos