Is It Navajo? Accurate Language Detection for Endangered Athabaskan Languages

Ivory Yang   Weicheng Ma   Chunhui Zhang   Soroush Vosoughi
Department of Computer Science, Dartmouth College
{Ivory.Yang.GR, Weicheng.Ma, Chunhui.Zhang.GR, Soroush.Vosoughi}@dartmouth.edu

Abstract

Endangered languages, such as Navajo—the most widely spoken Native American language—are significantly underrepresented in contemporary language technologies, exacerbating the challenges of their preservation and revitalization. This study evaluates Google's Language Identification (LangID) tool, which does not currently support any Native American languages. To address this, we introduce a random forest classifier trained on Navajo and twenty erroneously suggested languages by LangID. Despite its simplicity, the classifier achieves near-perfect accuracy (97-100%). Additionally, the model demonstrates robustness across other Athabaskan languages—a family of Native American languages spoken primarily in Alaska, the Pacific Northwest, and parts of the Southwestern United States—suggesting its potential for broader application. Our findings underscore the pressing need for NLP systems that prioritize linguistic diversity and adaptability over centralized, one-size-fits-all solutions, especially in supporting underrepresented languages in a multicultural world. This work directly contributes to ongoing efforts to address cultural biases in language models and advocates for the development of culturally localized NLP tools that serve diverse linguistic communities.

1 Introduction

The urgency of preserving endangered languages is not merely a linguistic issue but one deeply connected to the preservation of cultural, historical, and ecological knowledge (Tulloch, 2006; Zariquiey et al., 2022; Zhang et al., 2022; Cusenza and Çöltekin, 2024; Yang et al., 2025). These languages reflect the intellectual heritage of diverse communities, playing a critical role in maintaining global cultural diversity. Yet, despite this significance, the development of language technologies has been disproportionately focused on languages with large speaker bases and economic clout, leaving languages with smaller populations—such as Native American languages—largely unsupported.

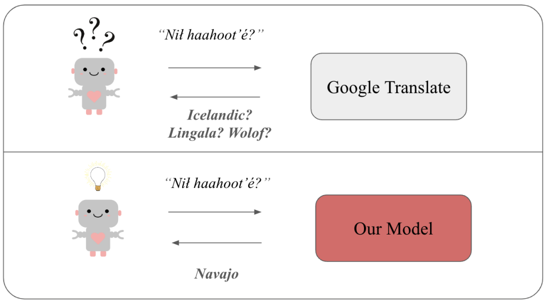


Figure 1: Google's LangID does not currently support any Native American languages, including Navajo and other Athabaskan languages. Our model addresses these challenges effectively.

This study centers on Navajo, the most widely spoken Native American language (Dietrich et al., 2022), which remains unsupported by commercial language technologies like Google's Language Identification (LangID) tool (Caswell et al., 2020). The lack of comprehensive linguistic datasets and dedicated tools impedes both language preservation and learning efforts (Shamsfard, 2019). This technological gap is even more pronounced for other Native American languages, many of which are on the verge of extinction due to minimal technological integration and educational resources (Meredith, 2013; Flavelle and Lachler, 2023).

Google LangID's performance on the *Navajo 10k* dataset (Goldhahn et al., 2012) revealed complete misidentification of Navajo sentences as unrelated languages, an expected outcome given that LangID does not currently support any Native American language. In response, we developed a language identification model tailored to accurately distinguish Navajo from languages erroneously suggested by LangID, achieving near-perfect accuracy. This success illustrates that low-resource languages, often overlooked by major technological platforms, can be effectively supported with targeted approaches and resources. Beyond Navajo, we extended our model to other languages in the

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 277–284
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

Athabaskan family—including *Western Apache, Mescalero Apache, Jicarilla Apache, and Lipan Apache* (George and Lopraisová). Our model's robustness across these related languages underscores its potential applicability across broader linguistic groups (see Figure 1). This suggests a viable path for NLP technologies to not only support individual endangered languages but to facilitate revitalization efforts across entire language families.

We highlight how centralization in language technology disproportionately benefits global languages, often sidelining underrepresented languages and thus contributing to the erosion of linguistic diversity (Schneider, 2022). Our findings underscore the feasibility of creating decentralized, robust language identification tools, which, by focusing on the unique needs of specific languages, can play a significant role in preserving endangered languages. Furthermore, it offers promising pathways for leveraging NLP tools across culturally and linguistically related groups, enriching both academic research and community-driven language revitalization by fostering tools that are responsive to the specific needs of these communities. This aligns with the broader goal of developing NLP technologies that not only accommodate but also actively support the linguistic and cultural diversity of our vibrant multicultural world.

2 Background

While there exist studies on endangered languages (Zariquiey et al., 2022; Zhang et al., 2022; Cusenza and Çöltekin, 2024), their integration into business technologies remains insufficient. For example, although Google's LangID supports over a hundred languages, it fails to include any Native American language, and so provides completely inaccurate suggestions when encountering Navajo. Similarly, advanced NLP models, such as ChatGPT, struggle with Navajo due to a lack of training data, which is predominantly derived from more widely spoken languages (Hangya et al., 2022).

The scarcity of digital resources for Navajo further compounds these challenges, as it lacks sufficient digital presence needed for effective NLP tool development (Magueresse et al., 2020). This scarcity not only limits the use of standard NLP methodologies but also hampers preservation and revitalization efforts. These issues reflect broader market-driven priorities in language technology, which overlook less commercially viable languages (Costa-jussà et al., 2022), creating significant barriers to preserving cultural heritage and emphasizing the need for more inclusive technological support for endangered languages (Todacheeny, 2014).

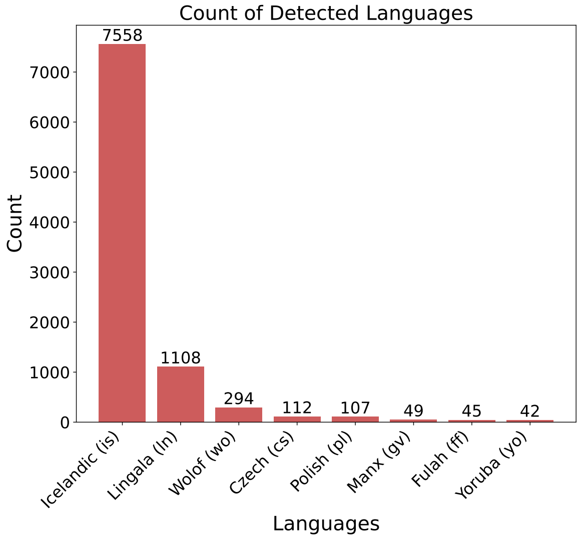


Figure 2: Visual representation of some erroneously suggested languages by Google LangID, along with their frequency counts.

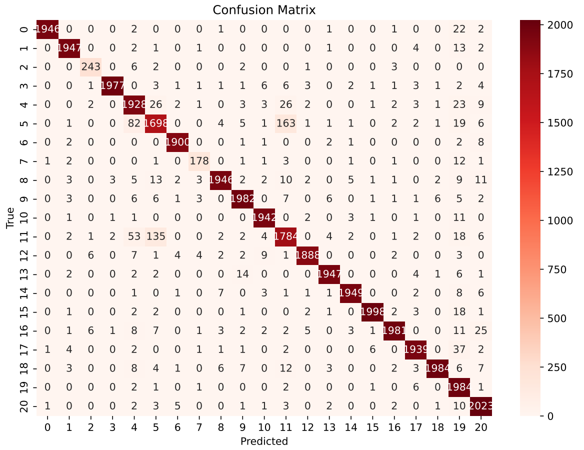
3 Identification for NatAm Languages

The benefits of Native American language identification are twofold: to bolster the development of linguistic tools tailored to these languages, and to aid in their preservation and revitalization (Mohanty et al., 2023). Effective identification is foundational for creating technologies that understand and process these languages, addressing the significant digital divide in language technology support (Mohanty et al., 2024). Our evaluation aims at a detailed assessment of the models' capability to accurately recognize and differentiate between Native American languages and others. By understanding the strengths and limitations of our models, we can refine our techniques to better serve the needs of Native American language communities.

3.1 Benchmark Construction

To construct our dataset for evaluating language identification models, we used two distinct approaches to account for diversity and specificity. The first dataset was formed based on twenty languages[1] that Google's LangID misidentifies as when presented with Navajo sentences, with their distribution shown in Figure 2. Each entry consists

[1] The languages are Icelandic, Lingala, Wolof, Czech, Polish, Manx, Fulah, Yoruba, Portuguese, Somali, Slovak, Tsonga, Spanish, Oromo, Indonesian, Igbo, Northern Sami, Irish, Arabic and English.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	1946	0	0	0	2	0	0	0	1	0	0	0	0	1	0	0	1	0	0	22	2
1	0	1947	0	0	2	1	0	1	0	0	0	0	0	1	0	0	0	4	0	13	2
2	0	0	243	0	6	2	0	0	0	2	0	0	1	0	0	0	3	0	0	0	0
3	0	0	1	1977	0	3	1	1	1	1	6	6	3	0	2	1	1	3	1	2	4
4	0	0	2	0	1928	26	2	1	0	3	3	26	2	0	0	1	2	3	1	23	9
5	0	1	0	0	82	1698	0	0	4	5	1	163	1	1	1	0	2	2	1	19	6
6	0	2	0	0	0	0	1900	0	0	1	1	0	0	2	1	0	0	0	0	2	8
7	1	2	0	0	0	1	0	178	0	1	1	3	0	0	1	0	1	0	0	12	1
8	0	3	0	3	5	13	2	3	1946	2	2	10	2	0	5	1	1	0	2	9	11
9	0	3	0	0	6	6	1	3	0	1982	0	7	0	6	0	1	1	1	6	5	2
10	0	1	0	1	1	0	0	0	0	0	1942	0	2	0	3	1	0	1	0	11	0
11	0	2	1	2	53	135	0	0	2	2	4	1784	0	4	2	0	1	2	0	18	6
12	0	0	6	0	7	1	4	4	2	2	9	1	1888	0	0	0	2	0	0	3	0
13	0	2	0	0	2	2	0	0	0	14	0	0	0	1947	0	0	0	4	1	6	1
14	0	0	0	0	1	0	1	0	7	0	3	1	1	1	1949	0	0	2	0	8	6
15	0	1	0	0	2	2	0	0	0	1	0	0	2	1	0	1998	2	3	0	18	1
16	0	1	6	1	8	7	0	1	3	2	2	2	5	0	3	1	1981	0	0	11	25
17	1	4	0	0	2	0	0	1	1	1	0	2	0	0	0	6	0	1939	0	37	2
18	0	3	0	0	8	4	1	0	6	7	0	12	0	3	0	0	2	3	1984	6	7
19	0	0	0	0	2	1	0	1	0	0	0	2	0	0	0	1	0	6	0	1984	1
20	1	0	0	0	2	3	5	0	0	1	1	3	0	2	0	0	2	0	1	10	2023

Figure 3: Classification results for Navajo and 20 other languages, presented as a confusion matrix. The intensity of the color indicates the frequency of classifications. Labels on x and y axes correspond to the languages listed in Table 1.

Language	Class	Precision	Recall	F1-score	Support
Navajo	0	1.00	0.99	0.99	1975
Icelandic	1	0.99	0.99	0.99	1971
Lingala	2	0.94	0.95	0.94	257
Wolof	3	1.00	0.98	0.99	2014
Polish	4	0.91	0.95	0.93	2032
Czech	5	0.89	0.85	0.87	1987
Manx	6	0.99	0.99	0.99	1917
Fulah	7	0.92	0.88	0.90	202
Yoruba	8	0.99	0.96	0.97	2020
Portuguese	9	0.98	0.98	0.98	2030
Somali	10	0.98	0.99	0.99	1963
Slovak	11	0.88	0.88	0.88	2018
Tsonga	12	0.99	0.98	0.98	1929
Spanish	13	0.99	0.98	0.99	1979
Oromo	14	0.99	0.98	0.99	1980
Indonesian	15	0.99	0.98	0.99	2031
Igbo	16	0.99	0.96	0.98	2059
Northern Sami	17	0.98	0.97	0.98	1996
Irish	18	0.99	0.97	0.98	2046
Arabic	19	0.89	0.99	0.94	1998
English	20	0.96	0.98	0.97	2054

Table 1: Detailed classification results for each language: precision, recall, f1-score, and support counts.

of 1k or 10k sentences extracted from the Leipzig corpora (Goldhahn et al., 2012), ensuring standardization against the Navajo 10k dataset in terms of linguistic features and contexts.

The second dataset focuses on languages from the Athabaskan family, which are linguistically related to Navajo, including Western Apache, Mescalero Apache, Jicarilla Apache, and Lipan Apache (Saxon, 2023). A sample of aligned words across these languages are included in Appendix A. To compile this dataset, we curated a limited sample of texts contributed by native speakers and language preservation organizations. Specifically, we collected texts in Western Apache (Glosbe, 2024), Mescalero Apache (Library, 2024), Jicarilla Apache (Wikipedia, 2024), and Lipan Apache (Indians.org, 2024). These curated texts not only enrich the model's exposure to authentic linguistic scenarios but also enable the evaluation of its ability to discern subtle linguistic nuances among closely related languages, thereby enhancing the precision and applicability of our language identification models.

3.2 Model Specifics

For the language identification task, we used a Random Forest classifier (Ho, 1995) with 100 trees for its robustness, interpretability, and effectiveness in handling complex classification tasks. The model builds multiple decision trees and predicts by aggregating their results, which reduces overfitting and enhances generalization. This model was chosen for its high interpretability, crucial when working with lesser-known languages, as it helps identify key features that distinguish languages, supporting linguistic analysis and improving endangered language identification. Additionally, its simplicity and widespread use make it ideal for replication and comparison with similar research.

3.3 Experimental Results

Our experimental evaluation of the Random Forest classifier is conducted using a dataset that included Navajo, labeled as '0', alongside a selection of languages that Google's LangID erroneously identifies as similar to Navajo. The dataset comprised 153,832 training samples and 38,458 test samples, each vectorized into 5,000 features to capture a wide range of linguistic attributes. The classifier achieved an overall accuracy of 97%, demonstrating its strong capability to distinguish Navajo from the misidentified languages.

The performance metrics show high precision, recall, and f1-scores across all tested languages, with Navajo achieving a precision of 1.00, a recall of 0.99, and an f1-score of 0.99, as shown in Table 1. These results highlight the classifier's effectiveness in accurately differentiating Navajo from other unrelated languages, which are erroneously suggested by Google's LangID. The confusion matrix in Figure 3 highlights the classifier's performance, particularly for Navajo. The model achieved 1,946 true positives for Navajo, with only a few false negatives. These results demonstrate the model's robustness in identifying Navajo and underscore its potential for supporting language identification

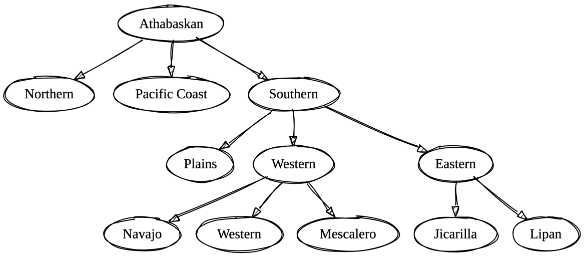


Figure 4: Family Tree for Athabaskan Languages

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

Table 2: Classification Results for Apache Languages: Percentage of sentences classified as Navajo and total number of sentences examined for each type of Apache language (out of 107 sentences).

in underrepresented Native American languages. Nevertheless, addressing the few misclassifications through enhanced training could further improve accuracy and generalization.

The stable performance of our classifier serves as a counterresponse to Google's LangID, and its lack of support for Native American languages, including Navajo. This omission directly results in the erroneous suggestions of linguistically unrelated languages, highlighting a critical need to include these languages in global technology platforms to better respect and reflect linguistic diversity.

3.4 Model Generalizability

Following the successful differentiation of Navajo from languages erroneously suggested by Google's LangID, we further tested the classifier's capability with our second dataset of curated Apache languages. Upon running this subset through the classifier, initially trained to distinguish Navajo from other languages, we observed that the classifier often identified these Apache languages as Navajo. This result is particularly significant given the linguistic similarities within the Athabaskan language family, to which both Navajo and the Apache languages belong. The classifier's performance here underscores its ability not only to identify Navajo with high accuracy but also to generalize across related languages within the same family. This generalizability is indicative of the model's potential utility in broader linguistic applications, especially in creating tools that support multiple but related Native American languages.

These findings also raise interesting questions about the classifier's sensitivity to the nuances between closely related languages and its potential role in developing more sophisticated NLP tools that can accurately differentiate between languages with subtle linguistic differences. The detection for Navajo performed best for Western Apache and Mescalero Apache, as shown in Table 2. Both these languages fall under the Western Apachean subgroup along with Navajo, as shown in Figure 4. On the other hand, Jicarilla Apache and in particular, Lipan Apache, performed less well in Navajo detection, which could be because they fall under the Eastern Apachean subgroup. This observation could be pivotal for linguistic preservation, allowing for the development of specialized educational and communicational tools tailored to each language's unique characteristics.

4 Conclusion and Future Work

This study demonstrates the effectiveness of our Random Forest classifier in accurately distinguishing Navajo from languages erroneously suggested by Google's LangID, as well as effectively recognizing related Athabaskan languages. These results emphasize the potential for broader applications in language identification, particularly for underrepresented languages. Our findings highlight a significant gap in support for Native American languages in current digital platforms, and urge the need for refined, inclusive language models.

Future work can focus on expanding the classifier's training to include additional Native American languages, improving its adaptability, and extending its utility to different language groups. The development of tools capable of distinguishing closely related languages is crucial for supporting educational and communication needs within Native communities[2]. We also advocate the decentralization of NLP research efforts, emphasizing the need for targeted investment in endangered languages. Such initiatives are essential to ensure that advances in language technology promote linguistic equity, thereby preserving cultural diversity and heritage in the digital age.

[2] This study represents a preliminary exploration, and we acknowledge the importance of direct collaboration with Native American communities. Moving forward, we plan to engage with community members and linguistic experts to ensure our work aligns with their perspectives, priorities, and cultural considerations.

Limitations

While the study successfully demonstrates the Random Forest classifier's efficacy in distinguishing Navajo from languages commonly misidentified by Google Translate and identifying related Athabaskan languages, it does have limitations that impact its broader applicability. Firstly, the language variety included in the study is limited; the classifier was tested primarily against a small set of languages suggested by Google's LangID and a few Athabaskan languages. This narrow scope might not capture the classifier's effectiveness across a broader range of Native American languages, potentially limiting its utility for other endangered language families. Secondly, the experimental design assumes a binary distinction between Navajo and other languages without considering intra-group variations and dialectical differences within the Athabaskan language family, which could affect accuracy in real-world applications. Lastly, reliance on vectorized features of 5,000 dimensions may overlook some finer linguistic nuances, which are crucial for distinguishing between closely related languages. Addressing these limitations in future work will be essential for developing more robust and applicable language identification systems.

Ethics

Ethical considerations are paramount in the development of language technology, especially for Native American languages, which are deeply intertwined with cultural identity and heritage. This study emphasizes the importance of respectful engagement with these communities, recognizing the cultural, spiritual, and historical significance of their languages. Technology development involving Native American languages should proceed with close collaboration with native speakers and community leaders to ensure that these tools support and reinforce language preservation rather than contributing to cultural homogenization or appropriation. Additionally, data privacy and consent are critical, as much of the linguistic data involves sensitive cultural content. Ensuring that communities retain control over how their linguistic resources are used is essential for maintaining trust and upholding ethical standards in research. Moreover, to ensure transparency and foster research, we have made our code and datasets used publicly available at https://github.com/ivoryayang/Isitnavajo.

Acknowledgment

This work was partially funded by a Google Research Award, and in part by a Dartmouth Alumni Research Award. We extend our gratitude to the Dartmouth graduate alumni for their generous support and commitment to fostering academic inquiry.

References

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.

Chiricahua Apache Mimbreno Nde Nation. 2024. Chiricahua apache mimbreno nde nation.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Giulio Cusenza and Çağrı Çöltekin. 2024. Nlp for arbëresh: How an endangered language learns to write in the 21st century. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*.

Sandy Dietrich, Erik Hernandez, et al. 2022. Language use in the united states: 2019. *American community survey reports*.

Darren Flavelle and Jordan Lachler. 2023. Strengthening relationships between indigenous communities, documentary linguists, and computational linguists in the era of nlp-assisted language revitalization. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP*.

Michael George and Jana Lopraisová. The cultural differences between the tribes of na-dené linguistic family.

Glosbe. 2024. Western apache texts.

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*.

Viktor Hangya, Hossain Shaikh Saadi, and Alexander Fraser. 2022. Improving low-resource languages in pre-trained multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

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

Indians.org. 2024. Lipan apache songs.

UVA Library. 2024. Mescalero apache texts.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

America Meredith. 2013. Racing against extinction: Saving native languages.

Sushree Sangita Mohanty, Satya Ranjan Dash, and Shantipriya Parida. 2024. *Applying AI-based Tools and Technologies Towards Revitalization of Indigenous and Endangered Languages*. Springer.

Sushree Sangita Mohanty, Shantipriya Parida, and Satya Ranjan Dash. 2023. Role of nlp for corpus development of endangered languages. *Grenze International Journal of Engineering and Technology. Jan Issue. Grenze ID*.

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

Britta Schneider. 2022. Multilingualism and ai: The regimentation of language in the age of digital capitalism. *Signs and Society*.

Mehrnoush Shamsfard. 2019. Challenges and opportunities in processing low resource languages: A study on persian. In *International conference language technologies for all*.

Frank Todacheeny. 2014. *Navajo Nation in crisis: Analysis on the extreme loss of Navajo language use amongst youth*. Arizona State University.

Shelley Tulloch. 2006. Preserving dialects of an endangered language. *Current Issues in Language Planning*.

Wikipedia. 2024. Jicarilla language.

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

Roberto Zariquiey, Arturo Oncevay, and Javier Vera. 2022. CLD^2 language documentation meets natural language processing for revitalising endangered languages. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP help revitalize endangered languages? a case study and roadmap for the Cherokee language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

A Aligned Words Across Native American Languages

Figure 5 lists 20 aligned words in four Native American languages, together with their English and French translations.

English (Français)	**Navajo**	**Jicarilla Apache**	**Mescalero Apache**	**Western Apache**
One (Un)	Táátłáí	Dałaa	Dáłe’é	Dałaa
Two (Deux)	Naaki	Naakii	Naaki	Naki
Three (Trois)	Táá	Kái’ii	Tai’	Táági
Four (Quatre)	Dįį	Dįį’ii	Dįį’	Dįį’i
Five (Cinq)	Ashdla	Ashdle’	Ąąshdlai’	Ashdla’i
Man (Homme)	Hastįį	Haskįįyįį	Haastį’	Ndeeń
Woman (Femme)	Asdząą	Izdzánįí	Izdząą	Izdzán
Dog (Chien)	Łééchąąí	Chíníí	Chúúné	Góshé or Łichánee
Sun (Soleil)	Shá	Ya’íí	Sháa	Ya’áí
Moon (Lune)	Tłéé	Tł’é’na’áí	Tł’é’na’áí	Tł’é’gona’áí
Water (Eau)	Tó	Kóh	Tú	Tú
White (Blanc)	Łigai	Łigai	Łiga	Łigai
Yellow (Jaune)	Łitso	Łitso	Łitsu	Łitsog
Red (Rouge)	Łichíí	Łichíí	Łitu	Łichíí
Black (Noir)	Łizhíní	Łizhįį	Łizhį	Diłhił
Eat (Manger)	Yiyąąsh	Iyą	Iyą	Yiyąą
See (Voir)	Nááłníí	Yaa’įį	Yiiłtse	Yo’įį
Hear (Entendre)	Yóónįįh	Yidiits’e	Yidiits’e	Yidits’įh
Sing (Chanter)	Hataał	Ha’dishéí	Haadi’a	Ha’do’aał
Leave (Partir)	Deesháál	Deeyá	Idzúút’i	Deyaa

Figure 5: Aligned Words from Chiricahua Apache Mimbreno Nde Nation (2024).