Mayasoundex: A Phonetically Grounded Algorithm for Information Retrieval in the Maya Language

Alejandro Molina-Villegas
Conahcyt - Center for Research in
Geospatial Information Sciences
amolina@centrogeo.edu.mx

Abstract

This paper introduces Mayasoundex, a phonetically grounded algorithm tailored for information retrieval in the Maya language. Mayasoundex utilizes phonetic principles to generate consistent codes for words with similar sounds, promoting phonetic similarity in information retrieval tasks. Drawing upon the distinctive phonological characteristics of the Maya language, the algorithm offers a robust approach to indexing and searching linguistic data. The proposed method addresses challenges posed by the oral tradition and the recent adoption of Latin characters in Maya writing, providing a versatile solution for preserving and promoting the Maya language through advanced information retrieval technologies. The Mayasoundex algorithm is made publicly accessible through a Colab Notebook, facilitating broader utilization and fostering future developments in this field.

1 Introduction

The Yucatec Maya language, also known as Maaya T'aan, belongs to the broader Maya language family, encompassing approximately 30 distinct languages distributed across Guatemala, Belize, and Mexico. Within Mexican territory, Yucatec Maya stands as the second most spoken indigenous language, with over 500 thousand speakers residing in the states of Yucatán, Quintana Roo, Campeche, and parts of Belize, as reported by The National Institute of Statistics and Geography [1].

In this context, creating a linguistic corpus of the Maya language is of utmost importance. On one hand, it compiles the communicative practices of native speakers in audio and video files. At the same time, these files document and preserve the knowledge expressed in the oral language, ensuring that its legacy is safeguarded for current and future generations. Thanks to this project, various technological and educational resources for the Maya language are planed to be developed (Can-Canul et al., 2023).

2 Deploying a Search Engine for Maya

The cornerstone of our project is the T'aantsil platform, an information retrieval system designed specifically to help people learn and understand the Maya language. Its development presented significant challenges, initially stemming from the lack of requisite data to construct deep learning-based models. To address this constraint, we adopted a phonetic-grounded algorithm, as elaborated upon in subsequent sections.

T'aantsil enables searches in Spanish and Maya through a dedicated search bar interface, depicted in Figure 1. The system incorporates multiple search indices, facilitating information retrieval through two distinct approaches: the Vector Space Model (VSM) (Schütze et al., 2008) and, more recently, the Mayasoundex algorithm. In VSM-operated indices, documents and queries are represented as weighted vectors within a multidimensional space, with each unique index term serving as a dimension and weights assigned based on tf-idf values. The documents within the system consist of transcriptions derived from audiovisual segments obtained during the project's linguistic documentation fieldwork.

However, while VSM enables retrieval of relevant information from queries, it presupposes user familiarity with Maya script, a proficiency not widespread among speakers. This limitation prompted the implementation of a Soundex-based variant for searches. Besides, we sought to encourage Maya speakers' utilization of T'aantsil by enabling queries beyond standard writing forms and the spellings prescribed in the 1984 agreements. As highlighted by Sobrino de Gómez and Paz Avila

[1] XII general population and housing census of the INEGI: https://www.cuentame.inegi.org.mx/monografias/informacion/yuc/poblacion/diversidad.aspx?tema=me&e=31

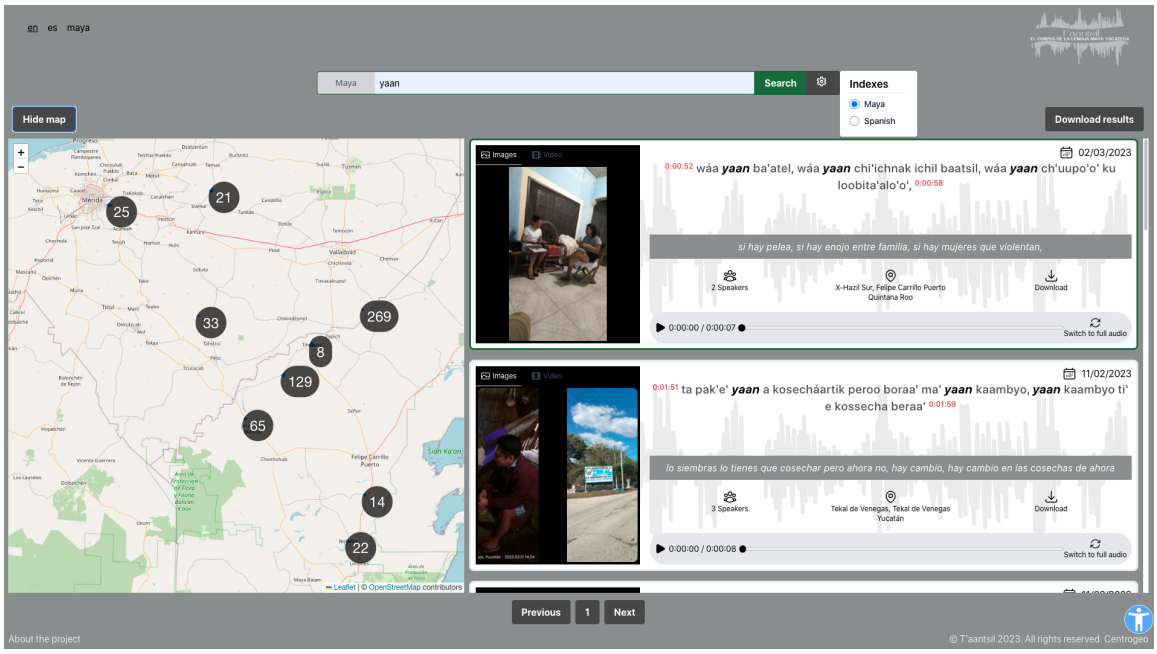

Figure 1: T'aantsil, the information retrieval system in Maya, allows searches in Maya through the Mayasoundex algorithm.

(2011), "*The central problem facing the Maya ethnicity (...) is the limited number of people who can write in their own language.*" Indeed, the adoption of Latin characters for Maya writing is relatively recent and heavily influenced by Spanish. Given the predominantly oral nature of the Maya language and its distinct consonantal system, we devised a phonetic based algorithm as a tailored solution.

3 Mayasoundex

We adapted the original Soundex algorithm (Russell and Margaret, U.S. Patent 1435663, issued 1922-11-14) to develop the Mayasoundex algorithm, tailored specifically for the Maya language.

The Mayasoundex algorithm generates a code based on the phonetics of Maya from a given word. The basic idea is for the algorithm to generate identical codes for two words that sound identical or almost identical. For the design of Mayasoundex, we considered the Mayan phonological inventory. More precisely, the consonantal system proposed by Gómez (2012) and reproduced in Table 1.

The system indicates that sounds sharing linguistic features exhibit similar sounds, as evident from the proximity of items in the table. For instance, the **m**, **n** group, both nasal sounds, are commonly confused by speakers and thus receive the same code in Mayasoundex. Similarly, **b**, **p**, both bilabial plosives, share a code in Mayasoundex. In this way, we proceeded with the complete consonantal system assisted by experts bilingual maya linguists until we mapped all this rules in dictionaries.

In the case of the vowel system, the soft vowels share a common code: zero (as in the original Soundex). However, in the Mayan language, some vowels, such as **u** and **i**, become strong in specific combinations; thus, they get their own code in those situations. Short vowels and long vowels are treated as equivalents in the algorithm.

It is worth noticing that certain sounds are represented in writing by three characters, some by two and some by one. This is the reason we defined three dictionaries to guide the algorithm. The highest priority corresponds to the exchange of phonetic patterns represented in 3 characters (priorclass3), the next priority corresponds to phonetic patterns represented in 2 characters (priorclass2), and finally, patterns of a single character (priorclass1). After sucesive replacements, the algorithm proceeds similarly to the original Soundex: by keeping the first letter (or a one-character equivalent), compressing repeated vowels, and adding a padding character to obtain codes of the same length.

For example, the word "ts'uulo'ob" (Gentlemen) will be transformed in these steps: » start: ts'uulo'ob » after priorclass3: suulob » after priorclass2: sulob » after priorclass1: 74601 » final code with padding: S4601********.

The complete code can be accessed from a Co-

	Bilabial	Alveolar	Postalveolar	Palatal	Velar	Glotal
Nasal	**m**	**n**				
Plosive	**p b**	**t**			**k**	**’**
Glottalized Plosive	**p’**	**t’**			**k’**	
Fricative		**s**	**x**			**ʃ**
Affricate		**ts**	**ch**			
Glottalized Affricate		**ts’**	**ch’**			
Approximant	**w**			**y**		
Lateral		**l**				

Table 1: Consonantal System of Yucatec Maya by Gómez (2012).

lab notebook [2], which includes the Mayasoundex algorithm.

4 Evaluation

To evaluate Mayasoundex, we curated one hundred instances of misspelled words entered by real users in our system. For each misspelled we manually write the correct version that was used as the groundtruth in the evaluation.

Each word in the dataset was encoded using the Mayasoundex algorithm and subsequently corrected by a baseline spell checker. To establish the baseline, we employed a digital Mayan-Spanish dictionary containing 3500 words to develop a basic spell-checker system.

A successful match was recorded each time the code generated for the misspelled word by Mayasoundex corresponded to the code for the correctly spelled word. Similarly, if the spell checker correctly identified the misspelled word and provided the correct suggestion, it was deemed a success for the baseline system.

For example, consider the word "ts’uulo’ob" with the correct spelling. The user-entered version "sulob" was identified as misspelled. While the spell checker suggested "ts’úumul," penalizing the baseline, both the correct version and the user-entered version yielded the code "S4601********" using Mayasoundex, resulting in a correct match for our proposal.

5 Results and Discussion

Table 2 illustrates the results of our evaluation, demonstrating that the proposed Mayasoundex approach outperformed largely the spelling checker baseline. Notably, in the case of the spelling checker, only words found in the Mayan-Spanish dictionary were considered for the evaluation, omitting out-of-vocabulary (OOV) words (marked as UNK). It’s crucial to emphasize that our proposal exhibits robustness to OOV words, as it doesn’t rely on a lexicon for encoding any given string, thus rendering it advantageous for real-world deployment.

	Correct	**Incorrect**	**Support**
Speller	15.59%	84.41%	77
Mayasoundex	85.00%	15.00%	100

Table 2: Comparison evaluation results of the Mayasoundex algorithm versus spelling correction for an information retrieval system in the Maya Language.

In contrast, the spelling checker approach encountered difficulties providing the exact correct suggestion for more than 80% of the total support, indicative of its limitations since only canonical words appear in the dictionary. Another noteworthy observation is the robustness of our algorithm to borrowings from Spanish, a frequent phenomenon among Mayan speakers due to the interaction between Mexican Spanish and Mayan (commonly referred to as code-switching). Unlike the spelling checker approach, which would necessitate consulting separate dictionaries for each language and detecting the appropriate one before making a suggestion, our method remains effective in handling such scenarios.

6 Conclusion

In conclusion, we have presented a novel algorithm grounded in the cultural and phonetic characteristics of the Maya language, facilitating the generation of consistent codes for words with similar sounds and promoting phonetic similarity in information retrieval tasks. We believe that this contribution will prove invaluable to researchers and technicians working on related problems, offering a robust foundation upon which to build further advancements in this field.

[2] https://colab.research.google.com/drive/1oWhAvd30vmWXMXFEIY4wS91tmAqeqe73?usp=sharing

7 Acknowledgements

We extend our special acknowledgment to the project titled "Development of Information Technologies for Maya" (W.K. Kellogg Foundation, Grant Number: P-6005156-2021), which is dedicated to the revitalization of the Maya language.

References

César Can-Canul, Igor Vinogradov, Samuel Yah, and Alejandro Molina-Villegas. 2023. *Corpus lingüístico de la lengua maya de Yucatán: una propuesta metodológica*. IHCC Publicaciones, Ensenada, Mexico.

Martín Sobrino Gómez. 2012. Breve historia de los modelos del sistema vocálico del maya yucateco y su discusión actual. *Temas antropológicos: Revista científica de investigaciones regionales*, 34(1):159–175.

Robert Russell and King Margaret. U.S. Patent 1435663, issued 1922-11-14. Soundex.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Carlos Sobrino de Gómez and Lillian Paz Avila. 2011. La condicón actual de la lengua maya en yucatán. *Archipielago. Revista cultural de nuestra América*, 16(60).

Apendix

Mayasoundex Codes	
M09********** man	M09********** maan
B08********** péek	B08********** pe'ek
B0601******** paalo'ob	B0601******** palob
B060********* paalo'o'	B060********* paaló
T01********** ta'ab	T01********** ta'ap'
X0901******** xanab	X090********* xana'
B02********** boox	B02********** bosh
U5206******** wishar	U5206******** wichar
X0909******** shaman	X0909******** xaman
T09********** den	T09********** ten
T09********** taan	T09********** t'aan
T506********* ti'al	T506********* tya'al
A642********* arux	A642********* alux
S46********** ts'uul	B010746****** papadzul
K098********* kang	K098********* kank
U9106******** jump'éel	U9106******** ump'éel
U0906******** hanal	U0906******** janal
A404********* ahaw	A404********* ajau

Table 3: "Example of Maya words with similar phonetics and their corresponding Mayasoundex generated codes.