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

Introduction

The Yucatec Maya language 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.

In this context, creating a linguistic corpus of the Maya language is of utmost importance. The cornerstone of our linguistic corpus 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.

`http://taantsil.com.mx/`

The system incorporates multiple search indices, facilitating information retrieval through two distinct approaches: the Vector Space Model (VSM) and, more recently, the Mayasoundex algorithm.

VSM presupposes user familiarity with Maya script, a proficiency not widespread among speakers. 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.

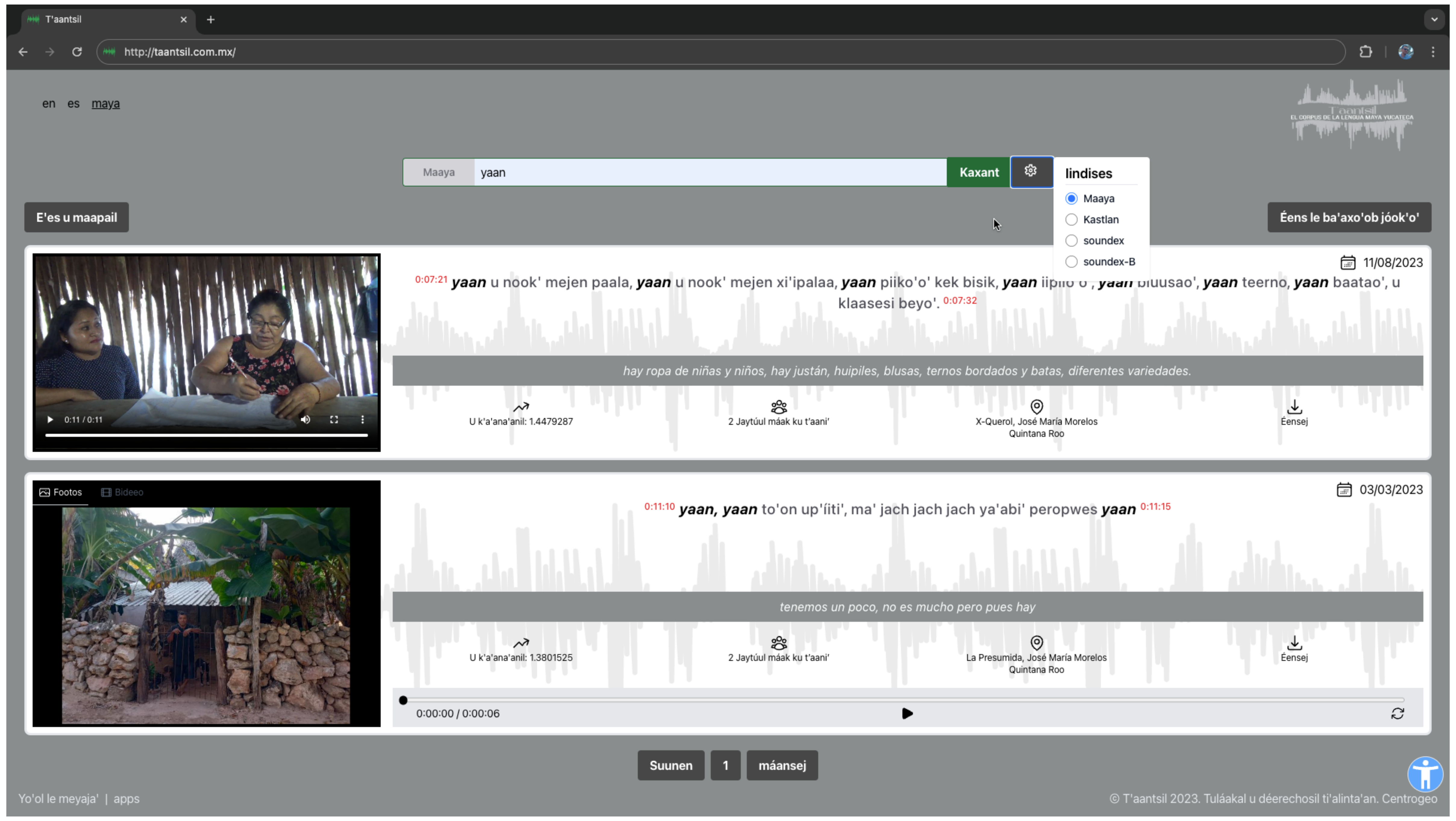
This limitations prompted the implementation of a Soundex-based variant for searches.

Mayasoundex

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 Martín Sobrino Gomez 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.

Development of Information Technologies for the Linguistic Corpus of the Maya Language


T'aantsil
http://taantsil.com.mx/
en es maya
T'aantsil
EL CORPUS DE LA LENGUA MAYA YUCATECA
Maaya
yaan
Kaxant
Iindises
● Maaya
○ Kastlan
○ soundex
○ soundex-B
E'es u maapail
Éens le ba'axo'ob jóok'o'
11/08/2023
0:11 / 0:11
U k'a'ana'anil: 1.4479287
2 Jaytúul máak ku t'aani'
X-Querol, José María Morelos Quintana Roo
Éensej
Footos
Bideeo
03/03/2023
U k'a'ana'anil: 1.3801525
2 Jaytúul máak ku t'aani'
La Presumida, José María Morelos Quintana Roo
Éensej
0:00:00 / 0:00:06
Suunen
1
máansej
Yo'ol le meyaja' | apps
© T'aantsil 2023. Tuláakal u déerechosil ti'alinta'an. Centrogeo

0:07:21 ***yaan*** u nook' mejen paala, ***yaan*** u nook' mejen xi'ipalaa, ***yaan*** piiko'o' kek bisik, ***yaan*** iipilo'o', ***yaan*** biiusao', ***yaan*** teerno, ***yaan*** baatao', u klaasesi beyo'. 0:07:32

hay ropa de niñas y niños, hay justán, huipiles, blusas, ternos bordados y batas, diferentes variedades.

0:11:10 ***yaan, yaan*** to'on up'íiti', ma' jach jach jach ya'abi' peropwes ***yaan*** 0:11:15

tenemos un poco, no es mucho pero pues hay

Figure: T'aantsil is the first linguistic corpus platform in Maya and is searchable through an information retrieval system.

• Convert Letters to Digits: Replace letters with digits using the rules in ad-hoc dictionaries. 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. • Remove Consecutive Duplicates: If two or more consecutive letters are converted to the same digit, retain only the first digit. • Ensure Length of Characters: Add padding if needed. • Retain a First Letter: Keep the class representative first letter of the word as the first character.

	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** (u)			**y** (i)		
Lateral		**l** (r)				

Table: Consonantal System of Yucatec Maya by Martin Sobrino-Gomez.

Example of word "ts’uulo’ob" (Gentlemen) transformation steps of Mayasoundex algorithm:
» start: ts’uulo’ob
» after priorclass3: suulob
» after priorclass2: sulob
» after priorclass1: 74601
» final code with padding: S4601********.

Results

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.

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

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

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

The complete code can be accessed from a Colab notebook:

bitly

`https://bit.ly/MayaSoundex`