Inteligencia Artificial, 28(76) (2025), 283-300
doi: 10.4114/intartif.vol28iss76pp283-300

INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

Generating a Culturally and Linguistically Adapted Word Similarity Benchmark for Yucatec Maya

Alejandro Molina-Villegas[1,*], Joel Suro-Villalobos[2], Jorge Reyes-Magaña[3], Silvia Fernandez-Sabido[1]

[1] SECIHTI - Centro de Investigación en Ciencias de Información Geoespacial
Scientific and Technological Park of Yucatan, Merida, 97302, Yucatan, Mexico
[*] amolina@centrogeo.edu.mx

[2] ShogunOS `shogunos.com`

[3] Facultad de Matemáticas- Universidad Autónoma de Yucatán, Mexico

Abstract In the field of AI, word embedding models have proven to be one of the most effective methods for capturing semantic and syntactic relationships between words, enabling significant advancements in natural language processing. However, producing word embeddings for low-resource indigenous languages—such as Yucatec Maya—often suffers from poor reliability due to limited data availability and unsuitable evaluation benchmarks. In this work, we propose a novel methodology for constructing reliable word embeddings by adapting the Swadesh List for semantic similarity evaluation. Our approach involves translating the Swadesh List from a high-resource pivot language into the target language, applying linguistic and cultural filtering, and correlating similarity scores between pivot-language embeddings from large language models and target-language embeddings. Our results demonstrate that this method produces reliable and interpretable embeddings for Yucatec Maya. Furthermore, our analysis provides compelling evidence that the choice of evaluation benchmark has a far greater impact on reported performance than hyperparameter optimization. This approach establishes a robust new framework with the potential to be adapted for improving word embedding generation in other low-resource languages.

Keywords: Yucatec Maya, Low-resource NLP, Word embeddings for Indigenous languages, Swadesh list, Culturally grounded NLP.
Palabras Clave: Maya yucateco, PLN en lenguas subrepresentadas, Incrustaciones de palabras para lenguas indígenas, Lista de Swadesh, PLN con base cultural.

1 Introduction

In Natural Language Processing (NLP), word embedding models play a crucial role in capturing semantic and syntactic relationships between words and phrases, enabling advancements in areas such as machine translation, information retrieval, and large language models (LLMs). However, these breakthroughs have primarily benefited high-resource languages such as English, Spanish, and Chinese, while low-resource languages—particularly indigenous languages—remain significantly underrepresented.

Among these underrepresented languages is Yucatec Maya, a Mayan language spoken primarily in the Yucatán Peninsula, covering regions of Mexico (Yucatán, Campeche, and Quintana Roo), as well as parts of Belize and northern Guatemala. According to the 2020 Mexican census, over 770,000 people speak Yucatec Maya, making it one of the most widely spoken languages in the country. It holds official status

ISSN: 1137-3601 (print), 1988-3064 (on-line)

under Mexico's General Law of Linguistic Rights of Indigenous Peoples, which grants it the same legal standing as Spanish in its territory of use.

Despite this recognition, digital resources for Yucatec Maya remain scarce. The most recent orthographic standard, established by the Instituto Nacional de Lenguas Indígenas (INALI) in 2014, provides a foundation for linguistic and educational materials but has not yet been widely adopted in computational contexts. The lack of large corpora, annotated datasets, digital dictionaries, and pre-trained models continues to hinder the development of NLP applications for Yucatec Maya.

This technological gap not only limits linguistic and computational research but also contributes to the digital marginalization of its speaker community. Addressing this challenge requires innovative strategies that maximize the value of limited resources while ensuring cultural and linguistic relevance.

In this work, we propose a novel methodology for constructing reliable word embeddings for Yucatec Maya by adapting the Swadesh List for semantic similarity evaluation. Originally developed by linguist Morris Swadesh in the 1950s, the Swadesh List identifies core vocabulary items—such as body parts, natural elements, and basic actions—that tend to be stable across time and languages. Leveraging this culturally grounded lexical resource, we construct a benchmark tailored to the linguistic and cultural context of Yucatec Maya.

Our methodology involves translating the Swadesh List from a high-resource pivot language into Yucatec Maya, applying linguistic and cultural filtering, and then correlating similarity scores between embeddings derived from large language models in the pivot language and those trained in the target language. The resulting set of word pairs serves as a benchmark for evaluating and optimizing word embeddings trained with the Skip-gram with Negative Sampling (SGNS) algorithm. By systematically tuning hyperparameters, we aim to ensure that the embeddings capture meaningful semantic structures—even under conditions of limited data.

While this study focuses on Yucatec Maya, the proposed methodology offers a potentially adaptable framework for other low-resource languages. More broadly, our work highlights the need to tailor NLP evaluation techniques to the cultural and linguistic characteristics of the languages involved, promoting the inclusion of indigenous languages in the digital era.

2 Related Work

Developing word embeddings for low-resource languages presents unique challenges due to the scarcity of large-scale corpora, limited orthographic standardization, and the lack of appropriate evaluation benchmarks. Various approaches have sought to adapt classic techniques such as SGNS to these constrained environments.

In [13], authors trained SGNS word embeddings for Guarani using a corpus assembled from news articles, tweets, Wikipedia, religious texts, and other online content. To evaluate embedding quality, they translated the MC-30 benchmark into Guarani and reported Spearman correlations between 0.403 and 0.569, demonstrating the feasibility of adapting standard evaluation resources to indigenous contexts.

Similarly, in [2], authors presented intrinsic evaluations for Yorùbá and Twi by translating the WordSim-353 benchmark [12]. Despite preserving the original similarity scores unless discrepancies were culturally significant, they reported relatively low correlations (0.391 and 0.437, respectively), highlighting the difficulties of semantic transfer across languages. Their work utilized fastText and character-level embeddings [5, 8].

Beyond corpus-based approaches, researchers have explored algorithmic innovations tailored for data scarcity. In [17], authors introduced a PU-learning method for learning embeddings with extremely sparse corpora, while in [4], authors derived semantic representations from bilingual dictionaries for languages such as Wolastoqey and Mi'kmaq. These studies reflect a growing interest in embedding models that accommodate linguistic and resource diversity.

However using translated evaluation benchmarks for low-resource languages raises questions of cultural and semantic pertinence. While translating MC-30 into indigenous languages is a viable strategy, it risks erasing culturally specific meanings or introducing translation artifacts. Note that, the English word *shore* denotes land beside a body of water, while its Spanish equivalent *orilla* may refer to any kind of edge. Such nuances may misalign perceived similarity across languages.

The process reported by [15] to build the MC-30 cross-lingual benchmarks is that first highly proficient English speakers who were native in Spanish, Romanian, and Arabic were tasked with translating word pairs from two data sets. They translated each pair while considering the relationship between the words to help disambiguate terms with multiple possible translations. Annotators were instructed to avoid multi-word expressions and could replace slang or culturally specific terms when necessary.

To assess the accuracy of the bilingual judges under these conditions, an experiment was conducted. Five judges provided Spanish translations, which were consolidated by a sixth judge who resolved any disagreements. Next, five experts evaluated the translations using the same scale applied in the original English data set. The resulting correlation of 0.86 indicated that the translations accurately preserved the relatedness of the original words.

In over 74% of cases, at least three judges agreed on the same translation, and when disagreements occurred, they typically involved synonyms. This high agreement demonstrated that the annotation approach effectively identified correct translations, even for ambiguous terms. Encouraged by the successful validation of this process for Spanish, only one annotator was used to translate the data sets for Arabic and Romanian.

Adapting a similarity benchmark to a language other than English involves several complex aspects. It requires bilinguals whose mother tongue is the target language but who are also proficient in English, which is extremely difficult to obtain for indigenous languages.

Table 1 presents columns with the translation in Spanish of MC-30 in addition to the original English version. [21] used the translation in Spanish of [15] listed in Table 1 to test word embeddings that were trained using Norm Association words in a graph structure resulting a Spearman correlation of 0.53 between their against MC-30.

Intrinsic evaluation remains a foundational method for assessing embedding quality by comparing model-generated cosine similarities with human judgments. Key benchmarks include:

- *MC-30* [19], a refined version of [22], containing 30 carefully curated word pairs with relatedness scores.
- *WordSim-353* [12], which includes 353 word pairs but conflates relatedness and similarity.
- *SimLex-999* [16], which explicitly distinguishes similarity from relatedness across 999 pairs.
- *MEN-3k* [6], crowdsourced scores for 3,000 word pairs rated for relatedness.
- *MTURK-771* [14], a 771-pair benchmark obtained via Amazon Mechanical Turk.
- *RG-65* [22], an early benchmark focused on synonymy over 65 non-technical word pairs.

Despite their continued relevance, these benchmarks reflect Anglo-Saxon cultural assumptions. Their direct application to other languages—particularly indigenous ones—requires careful cultural and linguistic adaptation to avoid semantic misalignments and biased evaluation.

This need motivates our proposal of a culturally grounded alternative: a word similarity benchmark based on the Swadesh list, curated and filtered to reflect the linguistic structure and cultural salience of Yucatec Maya. While the Swadesh list has been widely used in historical linguistics and language classification, its application in NLP remains limited. To the best of our knowledge, the Swadesh list—or any subset of it—has not previously been employed as a benchmark for evaluating word embeddings.

3 Materials and Methods

3.1 Methodological Overview

We identified two key limitations that make MC-30, or any other larger pair of standard words, unsuitable for intrinsic evaluation of Maya.

First, some word pairs are not culturally relevant in low-resourced languages like Maya. For instance, the concept of *asylum* —as a designated retreat for the elderly— is a Western construct. In Maya-speaking villages of the Yucatan Peninsula, elders are deeply respected for their wisdom and typically

ENG 1	ENG 2	SPA 1	SPA 2	Similarity
car	automobile	coche	automóvil	3.92
gem	jewel	gema	joya	3.84
journey	voyage	viaje	pasaje	3.84
boy	lad	chico	muchacho	3.76
coast	shore	costa	orilla	3.70
asylum	madhouse	asilo	manicomio	3.61
magician	wizard	mago	hechicero	3.50
midday	noon	mediodía	mediodía	3.42
furnace	stove	horno	estufa	3.11
food	fruit	comida	fruta	3.08
bird	cock	pájaro	gallo	3.05
bird	crane	pájaro	grulla	2.97
tool	implement	herramienta	implementar	2.95
brother	monk	hermano	monje	2.82
lad	brother	muchacho	hermano	1.66
crane	implement	grúa	implemento	1.68
journey	car	viaje	coche	1.16
monk	oracle	monje	oráculo	1.10
cemetery	woodland	cementerio	bosque	0.95
food	rooster	comida	gallo	0.89
coast	hill	costa	colina	0.87
forest	graveyard	forestales	cementerio	0.84
shore	woodland	orilla	bosque	0.63
monk	slave	monje	esclavo	0.55
coast	forest	costa	bosque	0.42
lad	wizard	muchacho	hechicero	0.42
chord	smile	acorde	sonrisa	0.13
glass	magician	vidrio	mago	0.11
rooster	voyage	gallo	viaje	0.08
noon	string	mediodía	cuerda	0.08

Table 1: MC-30 word pairs annotated with scores based on perceived semantic relatedness commonly used as Benchmark for intrinsic evaluation in NLP models.

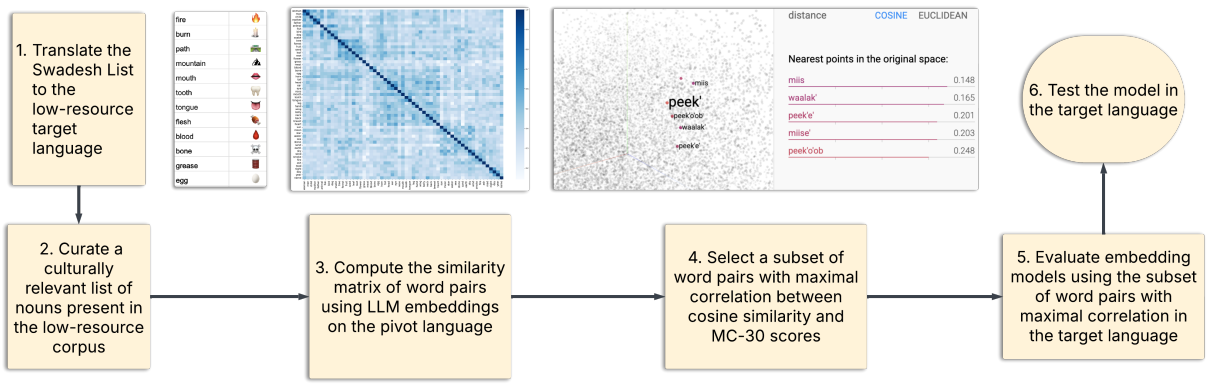

Figure 1: Graphical abstract of the main stages to generate a Culturally and Linguistically Adapted Word Similarity Benchmark for Maya.

remain with their families until the end of their lives, rather than being placed in institutionalized care. Similarly, words like *furnace* and *stove* are problematic as well, as these objects are largely absent from Maya-speaking communities due to the region's intense heat, where traditional cooking methods differ significantly from those in colder climates.

Second, the availability of quality textual data in Maya is severely limited. As we describe in Section 4.1, the frecuency distribution for the Maya corpus exhibits a premature decline, rapidly tapering off into the long tail of low-frequency words. Many words from MC-30 does not appear in the corpus or they appear with extremely low frequencies, rendering them unreliable for high quality word embeddings models.

To address these challenges, we adopt a lexicon of simple words, based on the Swadesh List, as a foundation for constructing a culturally and linguistically adapted word similarity benchmark for low-resource language models evaluation.

The Figure 1 is a graphical abstract of the main stages in the proposed methodology. In our approach, we select a high-resource pivot language, such as English or Spanish, to ensure an initial lexicon with varied degrees of semantic relatedness derived from the Swadesh List. The lexicon is then translated into the target language (Maya), creating a baseline that aligns the semantic relationships with the pivot language (stage 1). We apply a series of filtering steps until we obtain a curated list of nouns which are culturally relevant and present in the corpus in Maya (stage 2). With the curated list, we compute the cosine similarity of word pairs using embeddings of the pivot language derived from a LLM (stage 3). We identify a subset of word pairs whose cosine similarity scores exhibit maximal correlation with the similarity scores series from the MC-30 dataset (stage 4). The refined set is then used to guide the training of Maya word embeddings models, employing the SGNS algorithm by systematically adjusting the main model parameters (stage 5). As a result, we obtain embeddings models that preserve the semantic relationships observed in the pivot language while adapting to the linguistic constraints of the Maya language Data described in section 3.2 (stage 6).

We employed the SGNS to train word embeddings guided both by empirical considerations and by the desire to enable comparison with prior work in a similar low-resource setting—specifically, the Guaraní–Spanish study by [13], which also used SGNS. Given the comparable constraints in terms of corpus size and language resources, aligning our methodology allowed us to contrast results under similar conditions, particularly in terms of correlation.

Besides,Skip-gram generally performs better than Continuous Bag of Words (CBOW) on sparse data by more effectively capturing semantic relationships between infrequent words, which is crucial for morphologically rich and under-resourced languages like Yucatec Maya.

Subword tokenization was not used in this study since our primary goal was to evaluate the feasibility of culturally adapted similarity benchmarks using word-level embeddings.

All embeddings used in this study are static, meaning each word is represented by a single vector irrespective of context. This choice favors interpretability and aligns with our focus on establishing a

linguistically grounded, low-resource benchmark rather than on high-capacity, contextual models, which require substantially larger corpora and computational resources.

3.2 Data in Maya

For the experiments, we used a corpus derived from three different sources in Maya. Each of the three corpora described below comes from different contexts and projects, but all have been curated and managed by our research team to ensure quality and consistency.

The first part of our corpus comes from K'iintsil section by La Jornada Maya. La Jornada is one of Mexico's major newspapers since 1984. In the Yucatan Peninsula, La Jornada Maya publishes a dedicated section in Maya called K'iintsil. Through a collaboration between our research team and the editorial board of La Jornada Maya, we gained access to a sample of the K'iintsil. The topics in K'iintsil include world and national news. Practically all kind of topics are covered: politics, arts, sports, sciences, among others. This part of the corpus is valuable since it is full of neologisms given the modern topics covered by La Jornada Maya. The raw original data contains 505,760 words (2,486 texts). Heavy revision and pre-processing was necessary, as many records included HTML code and a mixture of Spanish and Maya. In some cases, the title was in Maya, but the body of the text was in Spanish.

The second corpus is an extended version of the Maya dataset described in [9], publicly available for the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. The corpus is part of a collaborative effort between the Geospatial Information Sciences Research Center (CENTROGEO, Mexico) and the Secretariat of Culture and the Arts of Yucatan (SEDECULTA, Mexico), as referenced in Agreement SEDECULTA-DASJ-149-04-2024. This corpus part consists of 14,438 simple sentences in Maya (75, 767 words), closely related to the Maya culture. The data is divided into 35 topics reflecting aspects of daily life in Maya-speaking communities and was originally created for educational purposes. The topics covered in the corpus include phrases related to cornfields, family, work, town, location, daily life, greetings, farewells, parks, markets, school, weather, courtesy, shopping, travel, pets, birds, insects, among others. This corpus part is particularly valuable as it provides short sentences in culturally relevant contexts, ensuring linguistic authenticity and alignment with the daily experiences of Maya-speaking communities.

The third corpus used in our study is part of the T'aantsil corpus project, a significant resource for the Maya language. T'aantsil platform is a search engine dedicated to explore oral corpus data designed to promote and preserve Maya by providing a wealth of authentic linguistic data [20]. It features a diverse range of subject from narratives recorded in the maya speaking comunities all aroud the Yucatan Peninsula ensuring broad coverage of modern language uses. Beyond its linguistic relevance, the corpus holds deep cultural significance, offering insights into the traditions, beliefs, and daily life of Maya communities. We used 10,478 fragments (166, 216 words) recovered from the transcriptions of interviews within the T'aantsil platform. All the retrieved texts are exclusively in Maya, ensuring linguistic purity while also reflecting the evolving nature of the language. By capturing culturally embedded discourse, this corpus provides a robust foundation for language processing tasks, supporting initiatives in documentation, education, and computational modeling. This resource is instrumental in developing robust word embedding models that accurately reflect the linguistic nuances and cultural context of the language.

4 Empirical Results

4.1 Corpus Exploration with MC-30 baseline

In NLP studies, it is widely assumed that word frequency distributions follow well-established statistical regularities. Zipf's Law, for instance, posits that in any corpus, the frequency of a word is inversely proportional to its rank, leading to a predictable long-tailed distribution. Such regularities have been extensively documented in high-resource languages, where large-scale corpora provide sufficient data to observe these patterns consistently.

Yet, a major finding in our research suggests that this presumption manifests differently concerning Maya. Owing to the absence of uniform writing standards, our analysis of the Maya corpus illustrates a notable extension in the frequency distribution's long tail.

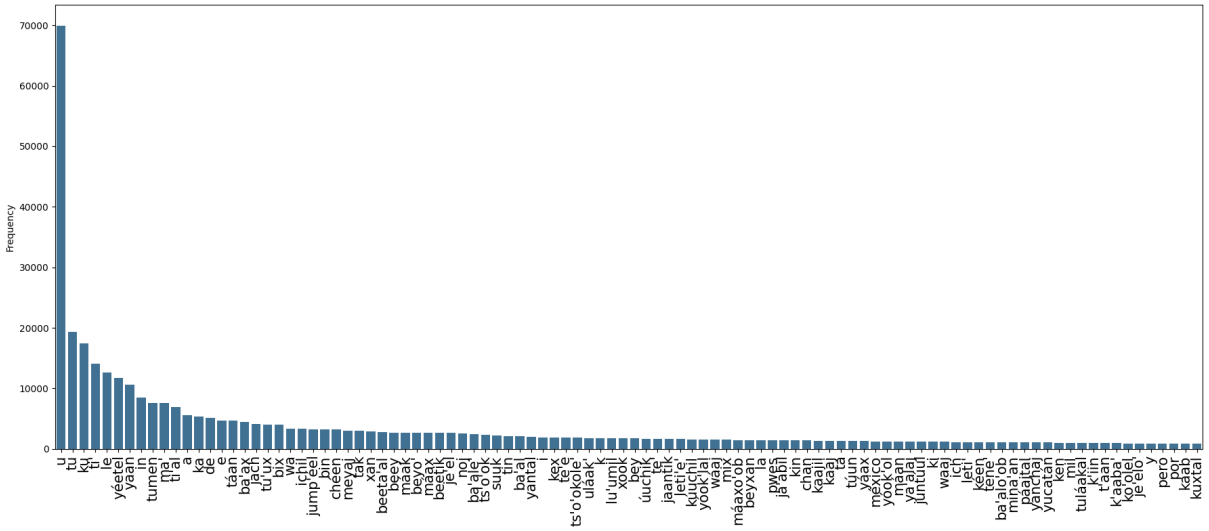

Figure 2: First 100 words in the frecuency distribution for the Maya corpus.

Figure 2 illustrates the word frequency distribution in our corpus for the top 200 terms. Unlike what is typically observed in high-resource languages, the frequency distribution in Maya exhibits a premature decline, rapidly tapering off into the long tail of low-frequency words. This irregular behavior suggests sparsity and a lack of lexical homogenity and highlights the challenges posed by data sparsity in low-resource languages.

As a result, we found that intrinsic evaluation of word embedding models using complete standard word similarity datasets, such as the well-known MC-30 [19], was unfeasible for our study. Originally developed within an Anglo-Saxon linguistic and cultural context, MC-30 lacks adaptation to Maya, as it includes terms with no direct cultural equivalent.

Table 2 presents the only word pairs from MC-30 that were found in the Maya corpus. In this table, we retained only the words that are both present in the corpus and culturally relevant, ensuring that the evaluation remains pertinent within the linguistic and cultural context of Maya. In the rest of the experiments, we used this reduced word pair list version of MC-30 as a baseline to compare against our approach.

MAYA 1	**MAYA 2**	**ENG 1**	**ENG 2**	f**(MAYA 1)**	f**(MAYA 2)**	**Similarity**
ch'íich'	t'eel	bird	cock	59	13	3.05
ch'íich'	garza	bird	crane	59	7	2.97
táankelem	láak'	lad	brother	60	28	1.66
garza	nu'ukul	crane	implement	7	107	1.68
xíimbal	kooche	journey	car	357	6	1.16
janal	t'eel	food	rooster	461	13	0.89
táankelem	meen	lad	wizard	60	411	0.42
paax	che'ej	chord	smile	165	100	0.13

Table 2: Word pair frequencies and similarity scores for some MC-30 pairs in Maya and English.

4.2 The Swadesh List as Word Pairs Benchmark Generator

The first version of the Swadesh list [23] was developed by linguist Morris Swadesh to study the evolution and relationship between languages through the analysis of their vocabularies [19]. Morris Swadesh adopted the idea that by comparing basic words across different languages, it is possible to estimate how closely related the languages are and when their common ancestors diverged. The Swadesh lists contain

a few hundred words from core areas of life such as: family, body parts, foods, basic actions, natural elements and some abstract concepts. The inclusion or exclusion of terms has been the subject of debate among linguists, so various versions of the lists exist, and some authors may refer to them as "Swadesh lists". Nevertheless, any version Swadesh list provides a culturally neutral starting point. Therefore, words on the Swadesh list are more likely to be represented in low-resource indigenous corpora due to their universality as a proxy for the natural world and the life of people.

For this study, we started with a Swadesh list of 207 English terms translated into various Maya languages, part of Wiktionary collaborative multilingual dictionary project [24]. With the intention of preserving the MC-30 criteria, we removed all non-noun words from the English Swadesh list (noun criterion). Lets denote $\mathbb{S}_{eng}$ to the set of the resulting original English nouns used in the rest of the experiments.

To adapt the Swadesh list from English (the pivot language) to Maya (the target language), we manually translated each word into $\mathbb{S}_{eng}$. Since the Wiktionary Maya version of Swadesh lacked 31-term translations, we completed the translations using two maya dictionaries: Diccionario Maya Popular [1] and the Diccionario Maya Cordemex [3]. Were multiple translations were available they were disambiguated by maya linguists collaborators. Words without direct translations were discarded. For instance *nu'ukulil meyaj* is a composition that expresses the idea of a tool but does not have a direct translation since the expression is closer to *working equipment*. We received aid from native speakers and linguistic experts to determine which words should be removed (cultural criterion).

Maximizing frequencies is the last criterion for filtering the words from the Maya Swadesh list. The words from the Swadesh list were translated into Maya, then filtered according to the cultural criterion and finally sorted by frequency in descending order. Words with frequency lower than the set threshold (threshold=10) were discarded. The determination of the threshold was established during the data exploration step, where it was observed that generating models from embeddings with frequencies lower than this value generated unexpected results (frequency criterion).

Let's denote $\mathbb{S}_{yua}$ to the set of the resulting 57 Yucatec Maya words that fullfilled the above mentioned criteria used the rest of the experiments. Table 3 lists the words fullfilling the criterion.

4.3 Computing Word Similarities

From the candidate word list, $\mathbb{S}_{yua}$, and its equivalents in the pivot language, $\mathbb{S}_{eng}$, we computed semantic similarity for each word pair using word embeddings from the pivot language. The embeddings were generated with a GPT model (LLM), specifically GPT-3.5 model, accessed via the OpenAI API with the implementation `text-embedding-3-large`. A vector dimensionality of 512 was set as a parameter for the embeddings to ensure the same dimensions as the Maya generated models.

To ensure consistency, embeddings were generated without any contextual information —that is, the LLM processed each word in isolation, without surrounding text. Word similarities were then computed using cosine similarity, defined as the normalized dot product between the embedding vectors of each word pair in $\mathbb{S}_{eng} \times \mathbb{S}_{eng}$ using the scikit-learn library (version 1.6).

Figure 3 presents the resulting similarity matrix for the word pairs in $\mathbb{S}_{eng} \times \mathbb{S}_{eng}$, derived from the LLM-generated embeddings. Note that Figure 3 displays a wide range of similarity values, spanning from very low to very high scores. Additionally, due to word ordering, distinct "similarity zones" emerge, aligning with intuitive groupings. For instance, one such region clusters together semantically related words like animal, fish, bird, dog, and snake, showing consistently high similarity scores. In contrast, another region groups body parts such as tail, head, ear, eye, nose, mouth, tooth, tongue, leg, and hand, forming a distinct similarity pattern separate from other semantic groups. This property is crucial for identifying word pairs with both high and low similarity in the target language to create a final benchmark set, as outlined in Section 4.4.

4.4 Selecting the Final Subset of Word Pairs

At this stage, we identify a subset of word pairs whose cosine similarity scores exhibit the highest correlation with the similarity scores from the MC-30 dataset. Prior studies have translated MC-30 word pairs into a target language and evaluated the correlation between the original similarity scores and the cosine similarity scores obtained from word embeddings in the new language. However, the specific words used

$\mathbb{S}_{eng}$	$\mathbb{S}_{yua}$	$f(\mathbb{S}_{yua})$	$\mathbb{S}_{eng}$	$\mathbb{S}_{yua}$	$f(\mathbb{S}_{yua})$
man	máak	2693	bird	ch'íich'	59
fruit	ich	1141	stone	tuunich	59
eye	ich	1141	leaf	le'	57
sun	k'iin	963	sky	ka'an	57
name	k'aaba'	932	meat	bak'	56
woman	ko'olel	925	grass	xíiw	42
water	ja'	561	moon	uj	34
year	ja'ab	522	smoke	buuts'	34
forest	k'áax	444	root	moots	33
head	pool	416	ear	xikin	32
earth	lu'um	400	neck	kaal	30
night	áak'ab	203	snake	kaan	27
road	bej	198	ash	ta'an	27
back	paach	195	nose	ni'	25
father	yuum	185	star	eek'	24
egg	je'	178	tooth	koj	22
fire	k'áak'	167	tongue	aak'	22
tree	che'	163	wing	xiik'	20
mouth	chi'	161	blood	k'i'ik'	18
child	paal	146	breast	iim	18
mother	na'	136	belly	nak'	17
dog	peek'	136	flower	nikte'	15
hand	k'ab	133	tail	nej	15
sea	k'áak'náab	119	bone	baak	13
animal	ba'alche'	117	horn	baak	13
leg	ook	105	seed	neek'	12
fish	kay	94	sand	sus	12
wind	iik'	86	day	k'iin	11
heart	puksi'ik'al	69			

Table 3: Filtered Swadesh List words in English and their corresponding translations in Yucatec Maya. These words were selected following three key criteria: (1) the *noun criterion*, ensuring only nouns were retained to align with MC-30 standards; (2) the *cultural criterion*, where words without direct or culturally relevant translations were removed; and (3) the *frequency criterion*, where words appearing fewer than 10 times in the corpus were excluded. The resulting set, $\mathbb{S}_{yua}$, consists of 57 culturally relevant and frequently occurring nouns in Yucatec Maya, forming the basis for the subsequent similarity evaluations.

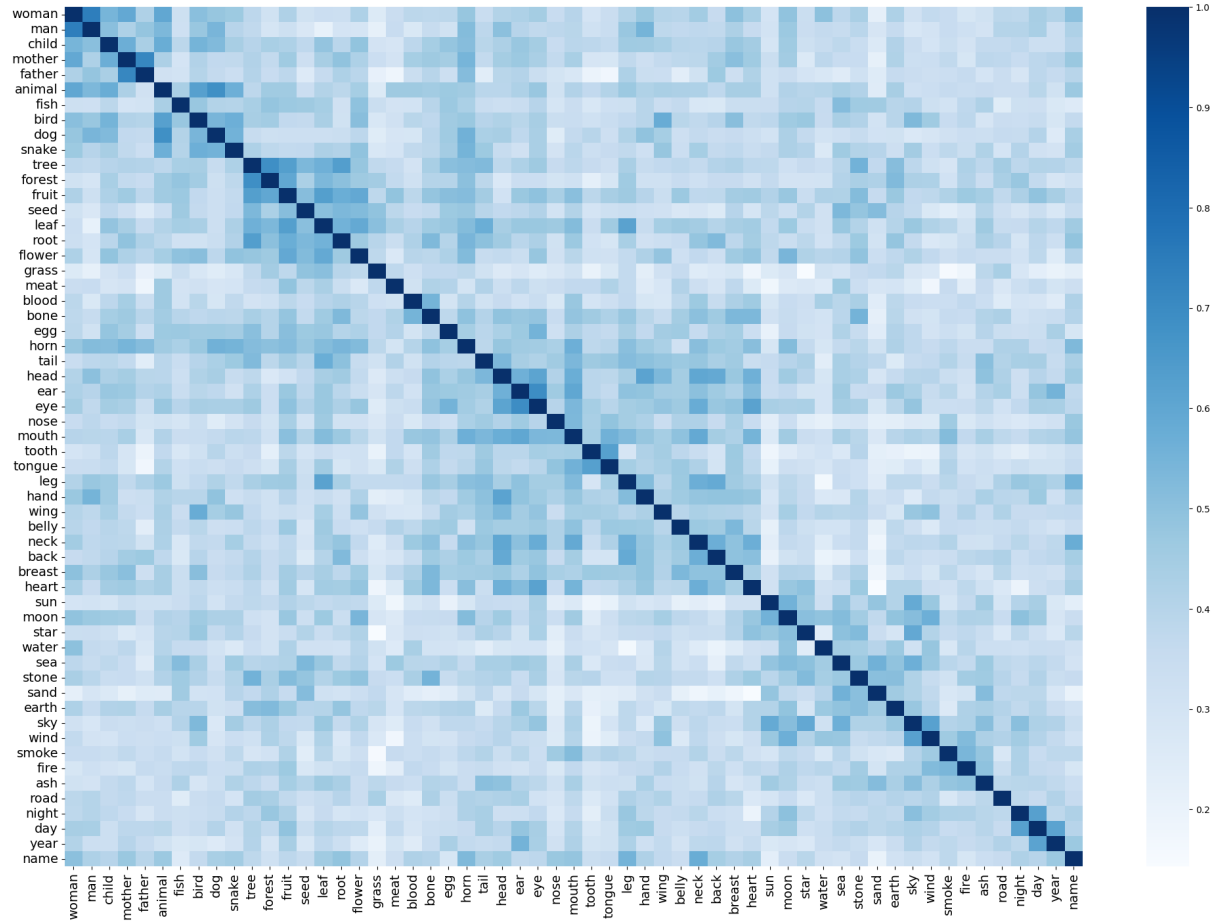

Figure 3: Cosine Similarity Pivot Language (English).

in translation are not inherently important—what matters is achieving a similarity ranking that closely aligns with the MC-30 benchmark.

In fact, different word pairs may be used as long as they yield a high correlation with the original MC-30 similarity rankings. Following this principle, we select word pairs from $\mathbb{S}_{eng} \times \mathbb{S}_{eng}$ that best reproduce the numerical similarity distribution of MC-30, as shown in the final column of Table 1. By constructing this adapted benchmark based on similarity scores, we develop a culturally and linguistically tailored evaluation set derived from the Swadesh List, specifically designed for an Indigenous language.

To enable numerical comparison between the similarity scores from MC-30 ($MC30_{sim}$) and the cosine similarity scores from the pivot language, we apply Min-Max scaling. The original $MC30_{sim}$ series range from a minimum of $min = 0.0$ to a maximum of $max = 4.0$. The normalized similarity scores are then computed as in Equation 1.

$$norm(MC30_{sim}) = \frac{MC30_{sim} - min}{max - min} \tag{1}$$

We then select 30 word pairs whose cosine similarity values in the pivot language most closely match the scaled MC-30 similarity scores. The resulting subset is listed in Table 4.

The Spearman rank correlation coefficient between$norm(MC30_{sim})$ and the cosine similarity scores from the pivot language is $\rho = 0.96439$ with a p-value of 1.0355×10^{-17}. Similarly, the correlation between the original $MC30_{sim}$ and the cosine similarity scores remains at $\rho = 0.96439$ with a p-value of 1.03554×10^{-17}, demonstrating a strong alignment between the adapted benchmark and the original MC-30 similarity distribution.

Eng 1	**Eng 2**	$cosine(LLM)$	$norm(MC30_{sim})$
woman	man	0.74397	0.9800
mother	father	0.71223	0.9600
ear	eye	0.69491	0.9600
tree	forest	0.68541	0.9400
animal	dog	0.68501	0.9250
tree	root	0.63488	0.9025
tree	fruit	0.63168	0.8750
fruit	leaf	0.63075	0.8550
tooth	tongue	0.62587	0.7775
animal	bird	0.62310	0.7700
night	day	0.62260	0.7625
sky	wind	0.61923	0.7425
leaf	leg	0.61567	0.7375
eye	heart	0.61543	0.7050
wing	road	0.41529	0.4150
night	name	0.41985	0.4200
grass	sky	0.29006	0.2900
mouth	sand	0.27511	0.2750
seed	smoke	0.23731	0.2375
back	sea	0.22215	0.2225
belly	sand	0.21729	0.2175
back	sun	0.20979	0.2100
grass	smoke	0.15623	0.1575
heart	sand	0.14313	0.1375
grass	star	0.14503	0.1050
leg	water	0.16450	0.1050
meat	wind	0.17448	0.0325
father	tongue	0.17641	0.0275
meat	sun	0.17688	0.0200
father	meat	0.18043	0.0020

Table 4: Word pairs with associated scores ($\rho = 0.96439$, $p - value = 1.0355 \times 10^{-17}$).

4.5 Training Word Embeddings Models for Maya

To generate the word embeddings model for Maya, we used an implementation of the SGNS algorithm [18]. The SGNS algorithm basically optimizes word embeddings models by maximizing the dot product between a target word w and its context words c, i.e $w \cdot c$; while simultaneously minimizing the same function for negative examples, i.e $w \cdot n$; where n is a word sampled from the corpus but does not necessarily co-occur with w. This negative sampling strategy enhances the model's ability to distinguish between meaningful word associations and noise.

For each target word w, k negative samples are drawn from the corpus using the smoothed frequency distribution following Equation 2, where $f(c)$ represents the frequency of word c and the parameter $nsamp \in [-1.0, +1.0]$ controls the smoothing effect.

$$P(c) = \frac{f(c)^{nsamp}}{\sum_{i=1}^{k} f(c_i)^{nsamp}}. \tag{2}$$

Referencing Equation 2, it is evident that the parameter $nsamp$ decisively affects the sampling probability. With $nsamp = 1$, the sampling aligns with frequency, whereas $nsamp = 0$ results in uniform sampling. Negative $nsamp$ values enhance the probability of choosing infrequent words over common ones.

While many empirical studies have often set $nsamp = 0.75$ as a standard choice, prior research has highlighted the benefits of using negative values in cases where the corpus exhibits irregular frequency distributions [7]. Given the unique characteristics of our corpus in Maya, we explored several values of $nsamp$ to determine the most suitable configuration for learning robust word embeddings.

To optimize the final model, we employed a grid search method, systematically exploring a predefined subset of the hyperparameter space. The key hyperparameters considered were embedding dimension (dim), number of training epochs (epochs), window size (win), and negative sampling rate (nsampling).

The selected ranges for each were as follows:

- Embedding dimension (dim): 128, 256, 512;
- Epochs: 30, 60, 90;
- Window size (win): 5, 7, 9, 11;
- Negative sampling rate (nsamp): -1.0, -0.75, -0.5, -0.25, 0.0, 0.25, 0.5, 0.75, 1.0.

Model performance was assessed by computing the Spearman rank correlation coefficient (ρ) between the cosine similarity scores derived from the trained embeddings and those obtained from the pivot language using the LLM-based embeddings. During the grid search process, we recorded the correlation ρ values and $p - value$ for each trained embedding model.

To systematically evaluate our approach, we conducted two separate grid search runs: one using our proposed word pair selection method based on the Swadesh list and another using the eight word pairs from MC-30 listed in Table 2.

Figure 4 presents the distribution of Spearman correlation values obtained for both benchmarks, comparing our Swadesh-based approach with the limited MC-30 pairs. The Swadesh-trained models demonstrated substantially higher correlation coefficients ($median\ \rho = 0.57, interquartile\ range = 0.50 - 0.63$) compared to MC-30-trained models ($median\ \rho = 0.20, interquartile\ range = 0.12 - 0.32$). This represents an approximately threefold improvement in median performance, with the lowest-performing quartile of Swadesh models still outperforming the highest quartile of MC-30 models. Notably, Figure 4 prove that our Swadesh-based approach consistently achieves in general higher correlation values, indicating a stronger alignment with the semantic relationships observed in the pivot language.

In contrast, the distribution of correlation values obtained from the MC-30 pairs suggests a significant degree of randomness, reinforcing the notion that these pairs are less reliable for training word embeddings in a low-resource language setting. The markedly lower p-values in our Swadesh-based benchmark confirm the validity and effectiveness of our word pair selection strategy.

The final best-performing model achieved a Spearman correlation of $\rho = 0.7331$ with a statistical significance of $pval = 3.07 \times 10^{-5}$. The optimal hyperparameter setting for SGNS embeddings model in

Maya was found to be: $dim = 256$, $epochs = 60$, $win = 7$, $nsamp = -0.25$. This configuration resulted in embeddings that best aligned with the semantic relationships observed in the pivot language while adapting to the linguistic characteristics of Yucatec Maya.

Two selected examples of the best output model performance are the following similar words in the constructed space: The most similar words to *peek'* (dog) are yaalak' (domestic) with sim =0.9654 and miis (cat) with sim=0.9434. The most similar words to *k'áax* (forest) are *ja'* (water) with sim=0.7921 and k'a'amkach (strong, referring to rain) with sim=0.7653.

5 Results Analysis

5.1 Grid Search Analysis

We analyzed in depth the grid search results of our approach using the Mann-Whitney U test, a nonparametric alternative to the independent samples t-test that evaluates whether two independent samples are drawn from the same distribution. The Mann-Whitney U makes no assumptions about the normality of the underlying distributions, which is particularly important when analyzing correlation coefficients that are bounded between -1 and 1 and often exhibit skewness [10]. In addition, the test evaluates differences across the entire distribution rather than solely focusing on measures of central tendency, allowing detection of stochastic dominance relationships that might be missed by parametric alternatives.

Our implementation follows the standard Mann-Whitney procedure where we calculate the U statistic defined in Equation 3.

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \tag{3}$$

Where n_1 and n_2 are the sample sizes of grid search 1 and grid search 2, and R_1 is the sum of ranks for grid search 1. For reporting purposes, we converted the U statistic to a standardized Z-score, allowing the calculation of effect size $r = Z/\sqrt{N}$, which quantifies the magnitude of difference between distributions independently of sample size.

This methodological choice aligns with best practices in computational linguistics evaluation, where non-parametric tests are increasingly favored for comparing model performances due to the inherent variability in NLP tasks [11]

The Mann-Whitney U test comparing correlation coefficients between models trained on Swadesh and MC-30 datasets revealed stark differences in performance. As shown in Figure 4, the test yielded an extremely significant result ($U = 2134, p = 4.20 \times 10^{-105}$), indicating that the difference between ranking correlation distributions is not attributable to chance. The effect size $r = 0.86$ far exceeds Cohen's threshold of 0.5 for a large effect, suggesting minimal overlap between distributions.

5.2 Statistical Interpretation of SGNS Parameters Effects

Figure 5 provides a detailed breakdown of parameter-specific analysis, revealing varying degrees of influence on model performance. The negative sampling parameter (*nsamp*) demonstrated the largest impact with an average range of 0.236 in ρ values across different settings, as evident in the bottom panel. Window size ranked second in importance (impact = 0.039), followed by training epochs (impact = 0.018), with dimensionality showing minimal effect (impact = 0.009). Both datasets exhibited parameter-specific response patterns that, while following similar trends, revealed key differences in optimal configurations.

The dimensionality parameter, top panel in Figure 5, shows remarkably stable performance across all settings if dim parameter (128, 256, 512) for both datasets. The Swadesh dataset maintains consistent ρ values around 0.57-0.58, while MC-30 remains steady around 0.20. This stability suggests that even the smallest dimension size adequately captures the semantic relationships, with minimal additional benefit from increased dimensionality.

For training epochs parameter, second panel in Figure 5, we observe slightly different patterns between datasets. The Swadesh dataset shows optimal performance at 60 epochs ($\rho \approx 0.58$), with a slight decrease at 90 epochs. Conversely, MC-30 shows a minimal but consistent downward trend as epochs increase, suggesting potential overfitting with extended training.

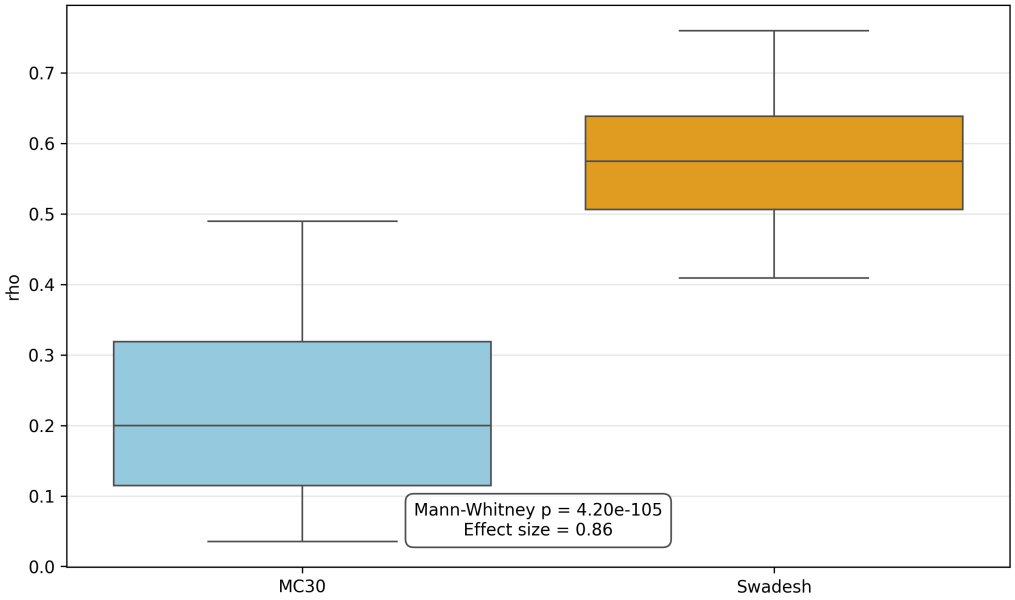

Figure 4: Boxplot comparison of Swadesh and MC-30-trained models.

The window size parameter , third panel in Figure 5, reveals a more pronounced impact on MC-30 compared to Swadesh. While Swadesh maintains relatively consistent performance across window sizes ($\rho \approx 0.55 - 0.58$), MC-30 shows a clear upward trend as window size increases, nearly doubling from window=5 ($\rho \approx 0.16$) to window=11 ($\rho \approx 0.28$). This indicates that larger context windows significantly benefit the MC-30 dataset, though not enough to approach Swadesh performance levels.

The negative sampling parameter in the bottom panel Figure 5, displays the most dramatic impact on both datasets, but with notably different patterns. Swadesh exhibits a clear bell curve distribution, with peak performance at $nsamp = -0.25$ ($\rho \approx 0.69$) and decreasing performance toward both extremes. This symmetrical pattern suggests robust performance regardless of whether sampling focuses on the right or left tail of the corpus distribution. In contrast, MC-30 shows an almost monotonically increasing trend from negative to positive nsamp values, with optimal performance at $nsamp = 0.5 - 0.75$ ($\rho \approx 0.35$). This striking difference in patterns provides valuable insight into how each dataset interacts with the embedding algorithm's negative sampling strategy.

The negative sampling parameter (nsamp) emerged as by far the most influential factor for both datasets, but with strikingly different response patterns. Swadesh-based models exhibit a bell curve distribution of ρ values ranging from -1 to 1, with optimal performance occurring around nsamp = -0.25. This symmetrical pattern suggests that negative sampling for both the most and least represented words in the corpus produces similarly robust results, regardless of whether sampling focuses on the right or left tail of the corpus distribution.

6 Conclusions

We have described and tested a new alternative methodology for generating culturally and more linguistically adapted word similarity benchmarks for intrinsic Evaluation Benchmarks in low resource languages. The fundamental idea of the methodology is based on considering simple, more generalizable word sets when dealing with low-resource language corpora. In particular, we base our approach on the Swadesh list to construct an evaluation benchmark for Maya.

The Swadesh list proved superior performance across all parameters in a grid search analysis, but its mayor advantage regarding negative sampling resistance clearly demonstrates its robustness for Maya. This is particularly relevant given the resource limitations faced when working with such languages, including challenges in acquiring clean, well-structured corpora and the lack of standardized orthography

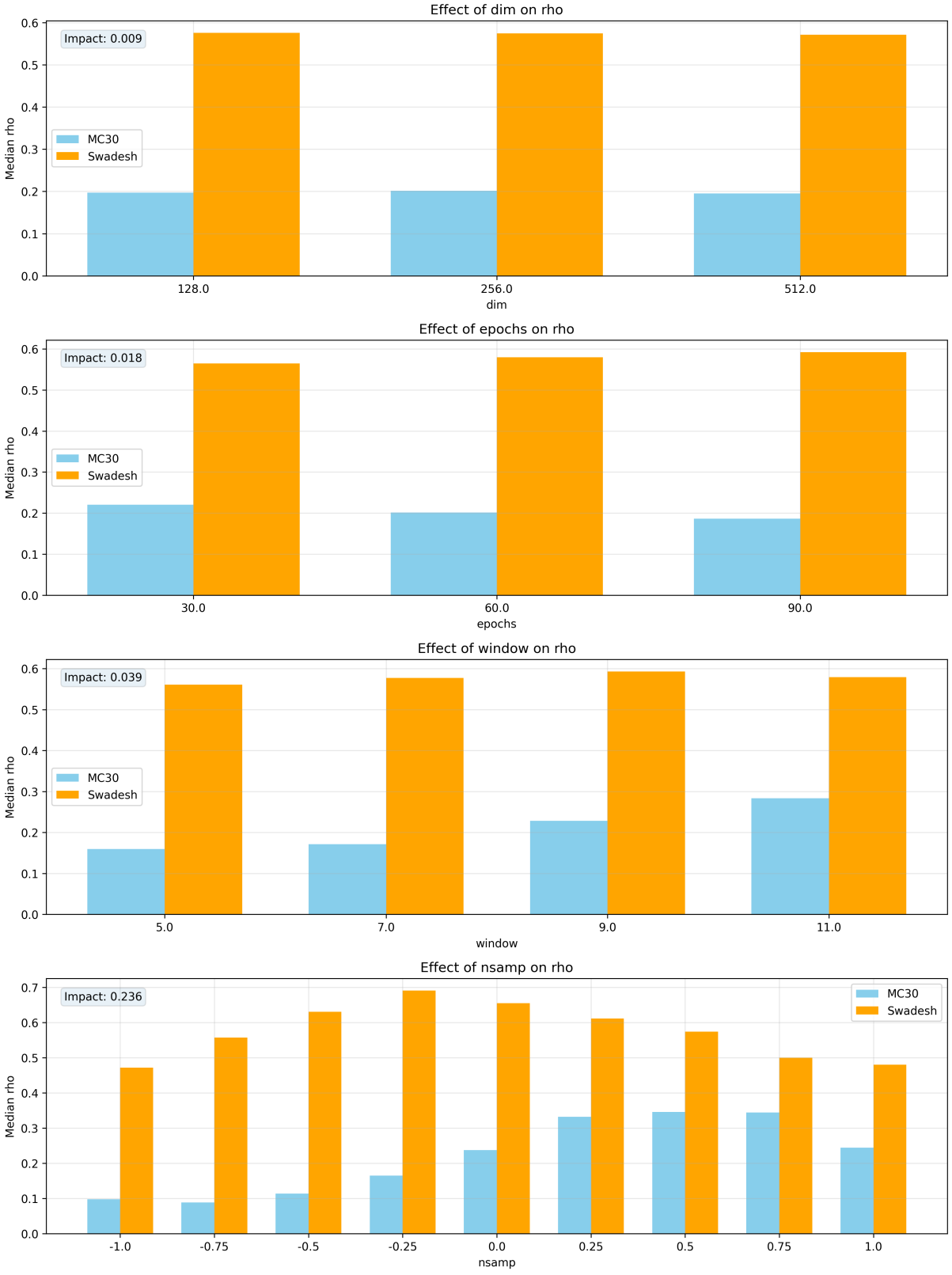

Figure 5: Parameter-specific analysis of word embedding models.

that varies significantly between communities and even individuals. The consistency of Swadesh-based embeddings across varying hyperparameter settings suggests that models trained on this dataset might be more generalizable across dialectal variations, an essential consideration for languages with limited institutional standardization.

The statistical significance and large effect size observed in our analysis provide compelling evidence that the Swadesh dataset is fundamentally more suitable for training word embedding models for our target language. This difference persists across all hyperparameter configurations, suggesting an intrinsic compatibility between Swadesh word pairs and the semantic structures present in our indigenous language corpus. The analysis provides compelling evidence that the choice of evaluation approach has a far greater impact on reported performance than any hyperparameter optimization, with the Swadesh based benchmark consistently outperforming MC-30 across all parameter settings.

These findings have profound implications for computational linguistics approaches to low-resource languages, suggesting that benchmark sets developed specifically for such contexts can dramatically improve model performance and evaluation validity compared to sets originally designed for high-resource languages.

Acknowledgments

This research was supported by the W.K. Kellogg Foundation under the grant *Development of Information Technologies for the Linguistic Corpus of the Maya Language of Yucatán* (Grant No. P-6005156-2021). We also acknowledge the institutional collaboration of *La Jornada Maya*, as well as the support provided through the cooperation agreement between *CentroGeo* and *SEDECULTA* (Agreement No. SEDECULTA-DASJ-149-04-2024). Our gratitude extends to the maya speakers, linguists, developers, and community contributors involved in the T'aantsil project, whose efforts were fundamental to the creation of the Yucatec Maya language resources used in this study. The resulting Yucatec Maya word embedding model is publicly available at: `https://huggingface.co/alemol/maya2vec`.

References

[1] Academia de la Lengua Maya de Yucatán, AC. *Diccionario Maya Popular: Maya-Español, Español-Maya*. Instituto de Cultura de Yucatán, Mérida, Yucatán, México, 2003.

[2] Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina Espana-Bonet. Massive vs. curated embeddings for low-resourced languages: the case of yorùbá and twi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2754–2762, 2020.

[3] Alfredo [dir] Barrera Vásquez. *Diccionario Maya Cordemex: Maya-Español, Español-Maya*. Ediciones Cordemex, Mérida, Yucatán, México, 1980.

[4] Diego Bear and Paul Cook. Fine-tuning sentence-roberta to construct word embeddings for low-resource languages from bilingual dictionaries. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 47–57, 2023.

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[6] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics, 2012.

[7] Hugo Caselles-Dupré, Florian Lesaint, and Jimena Royo-Letelier. Word2vec applied to recommendation: Hyperparameters matter. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 352–356, 2018.

[8] Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huan-Bo Luan. Joint learning of character and word embeddings. In *IJCAI*, pages 1236–1242. Citeseer, 2015.

[9] Luis Chiruzzo, Pavel Denisov, Alejandro Molina-Villegas, Silvia Fernandez-Sabido, Rolando Coto-Solano, Marvin Agüero-Torales, Aldo Alvarez, Samuel Canul-Yah, Lorena Hau-Ucán, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense, and Manuel Mager. Findings of the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. In Manuel Mager, Abteen Ebrahimi, Shruti Rijhwani, Arturo Oncevay, Luis Chiruzzo, Robert Pugh, and Katharina von der Wense, editors, *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 224–235, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[10] Marie Delacre, Christophe Leys, Youri L Mora, and Daniël Lakens. Taking parametric assumptions seriously: Arguments for the use of welch's f-test instead of the classical f-test in one-way anova. *International Review of Social Psychology*, 32(1):13, 2019.

[11] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, 2018.

[12] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131, 2002.

[13] Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. Can we use word embeddings for enhancing Guarani-Spanish machine translation? In Sarah Moeller, Antonios Anastasopoulos, Antti Arppe, Aditi Chaudhary, Atticus Harrigan, Josh Holden, Jordan Lachler, Alexis Palmer, Shruti Rijhwani, and Lane Schwartz, editors, *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 127–132, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[14] Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1406–1414, New York, NY, USA, 2012. ACM.

[15] Samer Hassan and Rada Mihalcea. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1201, Singapore, August 2009. Association for Computational Linguistics.

[16] Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, 2015.

[17] Chao Jiang, Hsiang-Fu Yu, Cho-Jui Hsieh, and Kai-Wei Chang. Learningword embeddings for low-resource languages by pu learning. *arXiv preprint arXiv:1805.03366*, 2018.

[18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[19] George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

[20] Alejandro Molina-Villegas. Mayasoundex: A phonetically grounded algorithm for information retrieval in the maya language. In *Proceedings of the LatinX in AI (LXAI) Research Workshop 2024 [Poster Presentation]*, pages 1–4, Mexico City, 2024. North American Chapter of the Association for Computational Linguistics.

[21] Jorge Reyes-Magaña, Helena Gómez Adorno, Gemma Bel-Enguix, and Gerardo Sierra. Representaciones vectoriales de palabras de un corpus de normas de asociación. *Research in Computing Science*, 147:109–118, 2018.

[22] Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

[23] Morris Swadesh. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21:121 – 137, 1955.

[24] Wiktionary. Appendix:mayan swadesh lists, 2025.