# Corpora duplication for NLP in low-resource languages: A case study of Nahuatl

**Juan-José Guzmán-Landa[1], Juan-Manuel Torres-Moreno[1], Luis-Gil Moreno-Jiménez[3], Elvys Linhares Pontes[4], Miguel Figueroa-Saavedra[2], Graham Ranger[1], Martha-Lorena Avendaño-Garrido[2]**

[1]Université d'Avignon (France), [2]Universidad Veracruzana (Mexico), [3]Independent Researcher (France), [4]Trading Central Labs

**Correspondence:** juan-manuel.torres@univ-avignon.fr

## Abstract

In this paper, we aim to answer the following question: could corpus duplication be useful in Natural Language Processing (NLP) for low-resource languages? In these languages (or $\pi$-languages), corpora available for training Large Language Models are virtually non-existent. Specifically, we study the impact of corpus expansion in Nahuatl, an agglutinative and polysynthetic Amerindian $\pi$-language characterised by extensive dialectal variation. Our goal is to increase the size of Nahuatl corpora, which currently consist of a limited number of tokens, through controlled duplication techniques. Our experimental setup employs incremental duplication alongside appropriate corpus balancing, with the objective of training embeddings optimised for downstream NLP tasks. Consequently, static embeddings were trained and evaluated on a sentence-level semantic similarity task. Our results show a significant improvement in performance when incremental duplication is applied, compared to results obtained without corpus expansion. To our knowledge, this technique has not yet been explored in this field.

## 1 Introduction

It is well established that Large Language Models (LLMs) require training corpora comprising substantial volumes of textual data in order to acquire a deep contextual understanding of linguistic structures and usage. These amounts often run into the hundreds of millions or even billions of words. Furthermore, it has been found that performance increases logarithmically with corpus size (Kaplan et al., 2020). This massive data requirement implies a major problem for the development of LLMs trained on languages with few computational resources ($\pi$-languages), as opposed to $\tau$-languages or languages with abundant resources (Berment, 2004; Abdillahi et al., 2006). Indeed, $\pi$-languages suffer from a severe lack of representative, large-scale textual corpora, making it impossible to train LLMs adequately. Consequently, these languages remain under-represented in Natural Language Processing (NLP), perpetuating a linguistic bias that limits their usefulness for the communities that speak them. One example of the Americas' $\pi$-languages is Nawatl (also known as Nahuatl), one of Mexico's indigenous national languages. In this country, Nawatl has been recognised as the second national language, after Spanish, with approximately 1.65 million Nawatl speakers (INEGI, 2020).

Nawatl has a significant number of dialectal varieties, with 29 recognised varieties spread across four major regions in Mexico: Western, Central, Eastern, and Huasteca[1]. This diversity represents an enormous linguistic challenge for the development of NLP tools, as it involves correctly handling significant variations in spelling and lexical choices (Zimmermann, 2019; Olko and Sullivan, 2016; Hansen, 2024). To solve this, a symbolic unifier for Nawatl spellings has recently been proposed (Guzman-Landa et al., 2025). Although the publication of digital content in Nawatl is constantly increasing, the dispersion and considerable dialectal diversity of this content mean that it cannot easily be included within the few available corpora.

The availability of digital Nawatl documents and their written application are nonetheless essential for the ongoing revitalisation of the language (Pugh et al., 2025). Our approach to addressing the scarcity of corpora involves the controlled duplication of available textual data. Combined with other techniques, this strategy could serve as a basis — in the case of $\pi$-languages — for expanding corpora on a larger scale. These corpora, in turn, could be used to train static word embeddings (Tunstall et al., 2022; Goyal et al., 2018). More specifically,

[1]See Ethnologue, 2025: https://www.ethnologue.com and (Lastra de Suárez, 1986).



*Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 115–127

our objective is to expand the Nawatl $\pi$-YALLI corpus[2] sufficiently to generate a positive impact on training models that produce static embeddings.

The structure of the paper is as follows: Section 2 provides a review of corpus expansion techniques in languages with few resources. Section 3 introduces the Nawatl language and the $\pi$-YALLI corpus. Section 4 introduces the strategy used in balancing the corpora, and Section 5 the duplication technique. Section 6 presents our experimental setup on a semantic similarity task. Section 7 describes the results. Finally, Section 8 concludes the paper and suggests avenues for future research.

## 2 Previous works

In the existing literature, research on Nawatl encompasses multiple levels of linguistic analysis. At the morphological level, a finite-state transducer has been employed to model the language's inflectional and derivational processes (Pugh et al., 2021). At the level of dialectal similarity, character-based representations have been explored using an LSTM architecture (Pugh and Tyers, 2021). Furthermore, at the syntactic level, both textual and audio modalities have been integrated to investigate syntactic structure (Pugh et al., 2024). Other models and resources for translation tasks and syntactic and speech analysis in dialects other than Nawatl have been published in (Shi et al., 2021; Gutierrez-Vasques et al., 2025)

Data duplication is found in literature primarily as a problem in $\tau$-languages, rather than as a technique for corpora expansion. Indeed, the massive amount of data on the internet leads to significant redundancy in collected corpora, which negatively impacts model training (Lee et al., 2022; Penedo et al., 2024). Consequently, most research focuses on the detection and removal of duplicates (or deduplication), particularly in the context of corpora intended for LLM training.

This problem is particularly pronounced in $\tau$-languages, where the volume of available text data is substantial but also highly redundant. For this reason, most research aims to produce large-scale training corpora whilst minimizing duplication as much as possible. Thus, FineWeb (Penedo et al., 2024) and CCNet (Wenzek et al., 2020) show filtering and deduplication techniques to produce high-quality, non-redundant corpora.

However, the situation is different for $\pi$-languages, where the problem is not an excess of data, but a scarcity of it. In this context, data augmentation (DA) could be an interesting strategy for expanding currently available corpora to compensate for the lack of resources (Feng et al., 2021; Chen et al., 2023). There are two main approaches proposed for DA techniques, in the literature: at the lexical level and at the syntactic level.

### 2.1 Lexical level DA

The EDA (Easy Data Augmentation) method (Wei and Zou, 2019) performs simple operations such as synonym substitution, as well as the insertion, deletion or random replacement of words. It has been applied exclusively in the context of text classification, and the results show performance ranging from 87.8% to 88.6%, representing improvements of less than 1%. These techniques use dictionaries to deal with synonyms.

### 2.2 Syntactic level DA

For languages lacking dictionaries, there are some techniques such as EDDA (Easy Distributional Data Augmentation) and TSSR (Type Specific Similar word Replacement) (Mahamud et al., 2023), which utilise distributional context and morphosyntactic labels to address this shortcoming. TSSR requires the data to be annotated with POS[3] tags. EDDA relies on the latent space generated by Word2Vec rather than a dictionary. These techniques have previously been applied to Swedish corpora.

In this article, we propose an approach to corpus expansion that utilises techniques requiring no lexical resources, such as part-of-speech taggers or dictionaries. Indeed, for Nawatl, these resources (where they exist at all) are difficult to apply directly due to the language's high degree of agglutination and polysynthesis. Furthermore, the limited dictionaries available do not cover all dialectal varieties.

Finally, we contend that EDA-type techniques and their stochastic mechanisms can introduce syntactic and semantic biases. Consequently, we aim to avoid such methods in our proposal.

[2]This corpus is available at: https://demo-lia.univ-avignon.fr/pi-yalli

[3]*Part-Of-Speech.*

## 3 The Nawatl language and the $\pi$-YALLI corpus

The Nawatl language is an Amerindian polysynthetic and agglutinative language. In other terms, verbal or nominal root morphemes and a range of inflexional morphemes combine productively to form new "words". At the syntactic level, Nawatl sentences follow a basic **verb–subject–object** (**VSO**) word structures, although this can be flexible. Thus, there are VO, VS, VOS and, less frequently, SV, SVO and SOV word orders (de Durand-Forest et al., 1995; Guzmán-Landa et al., 2026), depending on speakers' needs. Furthermore, the syntactic and semantic relationships between words and clauses are established through the valency of the verb and the use of conjunctive particles. These particles may also function as markers and discourse connectors.

Another distinctive feature of Nawatl is that words can be written as complete sentences, and this is particularly true of predicative words featuring verbs or verbal derivations. We therefore refer to them as phrase-words or "single-word phrases", as their morphology includes the subject and predicate, as well as information on the actants, and modal, directional and relational elements (Launey, 1978; Charles, 2016; Flores Nájera, 2019; Sasaki, 2022). Given its oral nature, there are very few written resources available for this language. Combined with the lack of standardised writing systems, this makes automated processing extremely difficult (Guzmán-Landa et al., 2025).

### 3.1 Some available Nawatl resources

There are very few tools and textual resources available for the Nawatl language. To our knowledge, only one machine translation tool is available for the Western Huasteca variety[4] since 2024.

In 2017, the *Instituto de Ingeniería* at the *Universidad Nacional Autónoma de México* (UNAM) published *Axolotl*, a bilingual Spanish/Nawatl corpus[5]. Furthermore, a Nawatl spelling unifier (Guzman-Landa et al., 2025) and the new $\pi$-YALLI corpus has recently been introduced. However, many dialectal varieties and texts remain inaccessible. This has a detrimental impact on the development of machine learning-based tools, thereby hindering their widespread use and adoption by Nahua-speaking communities.

### 3.2 The Nawatl $\pi$-YALLI text corpus

The $\pi$-YALLI corpus (Guzmán-Landa et al., 2025) is a Nawatl text resource available for machine learning and NLP algorithms. It is a heterogeneous corpus covering 16 topics and 26 dialectal varieties of Nawatl, spoken mainly in Mexico and El Salvador. It contains a limited number of words (around 6.6 million) and sentences, but it has been used successfully in various NLP tasks (Guzmán-Landa et al., 2025; Guzmán-Landa et al., 2026).

Despite its limited size, $\pi$-YALLI is, however, useful for training vector models: TF-IDF (Manning and Schütze, 1999), BM25 (Robertson et al., 2004), TF-PDF (Bun and Ishizuka, 2002) or static embedding models such as Word2Vec (Mikolov et al., 2013b), FastText (Bojanowski et al., 2017) or GloVe (Pennington et al., 2014), but clearly unsuitable for training contextualised vector models using BERT-style transformers (Devlin et al., 2019).

The acronyms used in this paper concerning the 26 dialectal varieties and the 16 topics, are listed in the Appendix A.1.

## 4 Statistical corpora balance

It is accepted by the scientific community that corpora must be balanced in order to avoid any bias (Arbach and Ali, 2013). However, the current $\pi$-YALLI corpus is not balanced at all (see Figure 1) either topically or dialectally.

We decided to evaluate the impact of balanced corpora on NLP tasks. For this reason, we will use two types of corpora: (i) unbalanced corpora and (ii) statistically balanced corpora (in our case, balanced only in terms of topics and dialectal varieties). Subsequently, both categories will be incrementally duplicated in order to find an optimal duplication ratio $\rho$ that maximises the performance of the models on an NLP task.

Firstly, we proceeded to establish a statistical balancing. In this regard, we introduced two types of corpora balancing: uniform balancing and positional balancing.Both techniques can be applied on topics, dialectal varieties or others corpus categories. Corpus balancing begins by sorting the $N$ categories in descending order, classifying them according to their number of tokens $T_i, i = 1, 2, ..., N$.

[4]See Google Translate https://translate.google.com.mx/?hl=es&sl=nhe&tl=es&op=translate

[5]The *Axolotl* corpus is also available at the following address: http://www.corpus.unam.mx/axolotl

### 4.1 Uniform balancing

This strategy involves balancing the varying token counts across the $N$ categories (topical or dialectal) relative to the initial value $T_1$, which contains the largest number of tokens. This enables the $T_i; i = 2, 3, 4, \ldots, N$ tokens within each category to be balanced until they all match the token count of $T_1$.

Once the process of uniform distribution to the $\pi$-YALLI corpus is applied, in the case of the $N = 16$ topics, each one will account for 3.2 million tokens. The resulting corpus will therefore be uniformly balanced, increasing from 6.6 million to **47.9** million tokens. However, we observed that some topics—such as literature (LIT) and history (HIS)— are duplicated fewer than 5 times, while others—such as politics (POL) and music (MUS)—are duplicated more than 1,000 times. This represents a significant increase that could have a major impact on model training. In the case of $N = 26$ dialectal varieties, the new balanced corpus increases from 6.6 million to **31.3** million tokens. The number of tokens for each dialectal variety increases at a different rate, as there are fewer available topics than there are dialectal varieties.

### 4.2 Positional balancing

In this statistical balancing strategy, we multiply the number of tokens $T_i$ of each topic (or dialectal variety) by their position $i = 1, 2, 3, ..., N$ in the ranking. This allows the $N$ topics (or varieties) to be balanced positionally.

Positional balancing aims to correct the excessive uniform duplication of certain topics (or varieties). Furthermore, in $N = 16$ topic positional balancing, the corpus increases from 6.6 million to **13.9** million tokens. As there are 16 topics, no single topic will be duplicated more than 16 times. For $N = 26$ dialectal varieties, the number of tokens increases from 6.6 million to **28.5** million. A larger number of tokens is obtained because there are more dialectal varieties represented (see Figure 1).

## 5 Incremental corpora duplication

It has been reported that LLMs require between 10 and 100 million tokens to obtain stable embeddings (Micheli et al., 2020). We therefore decided to expand the $\pi$-YALLI corpus using balancing strategies combined with an incremental duplication technique. This study aims to assess the impact of both factors on static embedding training algorithms.

At first glance, such a strategy might appear to have no positive impact on embedding learning. Indeed, it has been found that corpus deduplication is a crucial step in achieving successful embedding learning (Lee et al., 2022). In the case of $\tau$-languages, certain sentences are repeated 60,000 times or more; this poses a significant challenge for dense word representations, as such redundancy often leads to the overfitting of neural models.

Concerning $\pi$-languages, in addition to the lack of resources, it should be borne in mind that Nawatl is an agglutinative and polysynthetic language; consequently, the frequent use of compound words reduces the number of what we normally understand as "words", compared with other types of languages. Put differently, what would take five or six words in non-agglutinative languages is expressed in Nawatl with just one. This is very obvious in translation. Ultimately, all of this has an impact on the number of words (tokens) available in the corpora.

However, our hypothesis is that a *controlled* and *moderate* increase in the number of occurrences could facilitate the learning of textual representations in the case of $\pi$-languages, and in particular in Nawatl. We decided to empirically test our hypothesis regarding the impact of corpus expansion on learning algorithms. The aim, of course, is to seek a positive impact on the quality of static word embeddings.

## 6 Experimental setup

The protocol on a semantic task and the experiments concerning the incremental duplication strategy will be detailed in this section.

### 6.1 Similarity Semantic Task using static embeddings

Semantic similarity, a classic NLP task, involves evaluating various models (statistical, neural networks, etc.) using appropriate evaluation protocols (Francis-Landau et al., 2016). In our study, the aim is to calculate the semantic similarity between the reference sentences and the sets of candidate sentences, which may be semantically close to or distant from the references. This results in rankings of the candidate sentences, which will be compared to rankings produced by 5 native Nawatl speakers, via a statistical estimator. This is the same evaluation

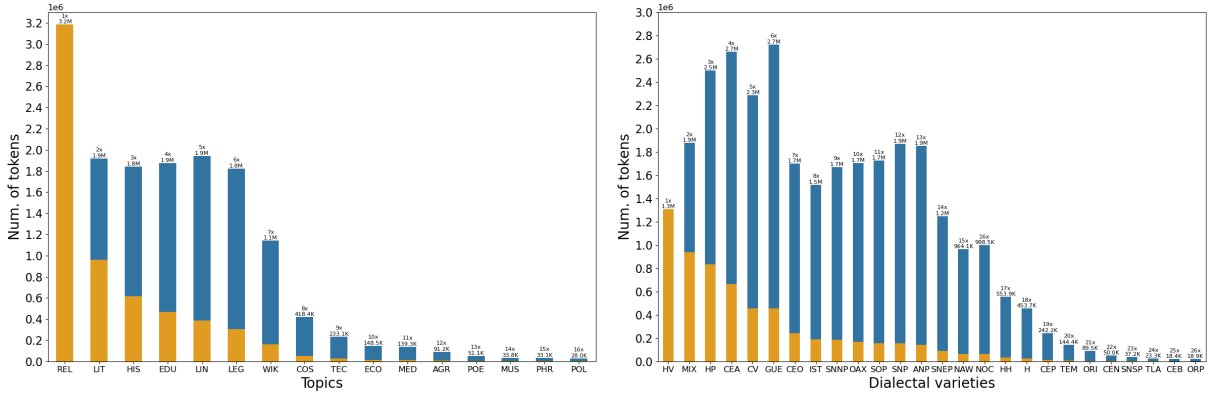


Figure 1: Corpora distribution by topics (at left) and dialectal varieties (at right). Orange bars: original unbalanced corpus. Blue bars: positional balanced corpus (numbers are in millions of tokens; see other details on the Appendix A.5).

protocol found in the recent literature (Guzmán-Landa et al., 2025): 30 reference sentences and 5 candidate sentences per reference. The final ranking of candidates allows us to estimate the impact of incremental duplication on embedding learning and also to measure their quality on a semantic proximity task. An example of sentences used in this semantic task is shown in the Appendix A.2.

Static embeddings have been widely used in NLP tasks (classification, analogies, semantic similarity, etc.), but they have been largely abandoned in favour of transformer-based contextual embedding models (such as the BERT model), whose popularity is due to their excellent performance (Devlin et al., 2019). Although transformers have shown their superiority in NLP tasks, this has only been possible in $\tau$-languages. Indeed, this type of model requires large amounts of textual data to learn effectively. The situation changes completely when it comes to processing $\pi$-languages. In this context, non-contextual embeddings are competitive, as they can be generated from scratch, are quick to train and, most importantly, non-contextual models require small corpora to achieve meaningful learning.

Among the popular static embedding training algorithms are Word2Vec (Mikolov et al., 2013a), FastText (Bojanowski et al., 2017), and GloVe (Pennington et al., 2014). Consequently, we employed these algorithms to train word embeddings on our extended (duplicated) corpora. To operate at the vector-sentence level, we compute the average of the word vectors in the all sentences. In the Appendix A.3 we show the training hyper-parameters used for the models. The quality of the embeddings was subsequently evaluated using the aforementioned sentence semantic similarity task, establishing a ranking (Guzman-Landa et al., 2025).

The cosine similarity between each candidate phrase vector $\vec{C}_{i,j}$ and the reference phrase vector $\vec{R}_j$ for a block $j$, where $i = 1, 2, ..., 5$ and $j = 1, 2, ..., 30$, allows us to calculate a ranking $O_j$ of the candidate phrases. This ranking is compared with the ranking $O_j^*$ produced by human annotators. The correlation between the obtained $O_j$ and ground-truth $O^*$ rankings is evaluated by Kendall's rank correlation coefficient $\tau$.

Kendall's coefficient $\tau$ is a non-parametric measure of correlation that assesses the ordinal association between two variables, i.e. the degree of agreement between two rankings (Kendall, 1938).

### 6.2 Corpus balancing and incremental duplication

In order to expand the original corpus, we incrementally duplicate $\rho$ times the $\pi$-YALLI corpus– whether balanced or unbalanced– where $\rho = [1, 2, 4, ..., 28, 30]$ times its original size. The aim is to determine the optimal value of $\rho^\star$ that maximises the efficiency of the models.

On the one hand, in the case of unbalanced corpora, we have incrementally generated corpora ranging from: 6.6 million ($\rho = 1$), 13.2 million ($\rho = 2$), 19.8 million ($\rho = 3$), ..., to approximately 198 million words (duplicated $\rho = 30\times$ times), without regard to topical or dialectal varieties.

On the other hand, using statistically balanced corpora, we have corpora generated by progressive incrementation. Table 1 shows the corpus size in millions of words from starting point (without duplication) where $\rho = 1$, to our final experimental

point with $\rho = 30$ duplications.

| Balancing | Topical | | Dialectal | |
|---|---|---|---|---|
| | $\rho$=1 | $\rho$=30 | $\rho$=1 | $\rho$=30 |
| **Uniform** | 47.9 | 1437.7 | 31.3 | 938.2 |
| **Positional** | 13.9 | 417.5 | 28.5 | 856.3 |
| | **Original corpus** | | | |
| **Unbalanced** | $\rho$=1 | $\rho$=30 | | |
| | 6.6 | 198 | | |

Table 1: Balancing and duplication impact on corpora's size in millions of words (tokens).

## 7 Results and Discussion

The corpus was pre-processed using a spelling Nawatl unifier (Guzman-Landa et al., 2025), followed by processes of cleaning, segmentation into paragraphs and sentences, and the removal of some stop-words (Guzmán-Landa et al., 2026).

Figure 1 presents the new statistical distributions (blue bars) obtained by applying the positional balancing techniques (orange bars) to the topical and dialectal varieties. This addresses the bias introduced by the variability and productivity of the texts selected for the corpus. This variability affects not only the quantity but also the quality of future text processing and generation.

These transformations involve considerable volumes of text, thereby significantly increasing the size of the $\pi$-YALLI corpus and reducing the underrepresentation of topic-based and speech communities.

### 7.1 Unbalanced corpora

Figure 2 shows the results using incremental duplication of unbalanced corpora, with an increment ratio $\rho = [1, 2, ..., 30]$. For each ratio $\rho$, we show the average Kendall's $\langle\tau\rangle$ over five runs and the respective standard deviation (shown as a coloured band). The FastText algorithm in skip-gram mode achieves the best $\langle\tau\rangle$ performance across most ratios $\rho$. However, it should be noted that Word2Vec, also in Skip-gram mode, benefits most from the unbalanced incremental duplication technique, with consistent improvements between $1\times \leq \rho \leq 16\times$. In contrast, the GloVe algorithm shows diminishing results as the number of duplications increases.

Table 2 shows further details concerning these results: the maximum average values of $\langle\tau\rangle$, the gain and the training time. Except GloVe, the unbalanced duplication yields moderate or significant benefits. We have observed that Word2Vec achieves better results when the duplication rate $\rho = [20, 22]$. On the other hand, FastText reaches its maximum values with $\rho = [8, 10]$. This shows that FastText, with a lower duplication rate, generates higher-quality static embeddings.

In order to compare our results, we have used as our baseline three well known pre-trained embedding models. These models uses the same learning algorithms but they were trained on three commonly available corpora–without duplication or balance–: (i) FastText trained on Common Crawl[6]; (ii) FastText trained on Nawatl Wikipedia[7]; and (iii) Word2Vec trained on Nawatl Common Crawl[8]. Our results outperform the three baselines, as expected.

### 7.2 Balanced corpora

Table 3 presents a comparison between models trained on the original unbalanced $\pi$-YALLI corpus and those trained on the newly proposed uniform and positionally balanced corpora. This study was conducted using only the Word2Vec and FastText models employing Skip-gram architectures, which yielded the best results when applying unbalanced incremental duplication to the corpora (see Figure 2). As shown in Table 3, topic positional balancing ($T_{pos}$) enables FastText to achieve a Kendall's $\langle\tau\rangle = 0.477$, the highest recorded across all balancing methods. Meanwhile, with uniform topic balancing ($T_{+}$), Word2Vec shows a **19.9%** improvement, increasing from a Kendall's $\langle\tau\rangle$ of 0.357 to 0.428. This constitutes the largest percentage gain among all balancing techniques.

We found that positional (pos) and uniform (+) balancing, when applied to topics, yield the highest scores. However, in practice, both types of balancing applied to the topic (T) or dialectal varieties (D) yield an improvement in the base Kendall's $\tau$ for Word2Vec and FastText. For this reason, we decided to apply incremental duplication to all cases: T+, $T_{pos}$, D+, and $D_{pos}$.

Figure 3 shows that topic positional balancing $T_{pos}$ stands out significantly compared to the other cases. FastText once again achieves the highest mean Kendall's value of $\langle\tau\rangle = 0.515$ for $\rho = 12$. The percentage increase relative to the baseline $\tau$

[6] https://commoncrawl.org

[7] FastText has been trained on 157 languages: https://fasttext.cc/docs/en/crawl-vectors.html

[8] https://sparknlp.org/2022/03/16/w2v_cc_300d_nah_3_0.html

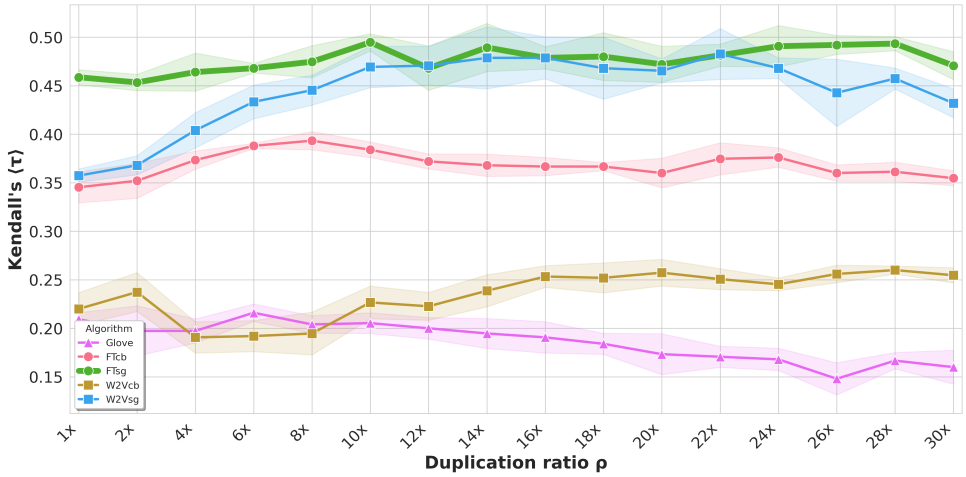


Figure 2: Unbalanced $\pi$-YALLI corpus. Kendall's coefficient $\langle\tau\rangle$ on sentence semantic similarity task. Learning on incrementally duplicated unbalanced corpus $\pi$-YALLI, using the static models GloVe, FastText (CBOW=FTcb, Skip-gram=FTsg) and Word2vec (CBOW=W2Vcb, Skip-gram=W2Vsg). There are 5 runs per duplication point $\rho$.

— which rises from 0.477 to 0.515 — is 8%. FastText also achieves a maximum absolute[9] value of $\tau = \mathbf{0.547}$. Word2Vec achieves its highest average $\langle\tau\rangle = 0.481$ at $\rho = 12$, and a maximum absolute $\tau = 0.500$. However, this is not the highest $\tau$ for Word2Vec, as it achieves a maximum absolute[10] $\tau = \mathbf{0.527}$ at $\rho = 18$. Therefore, $\rho^{\star} = \mathbf{12}$ can be established as a reasonably optimal duplication ratio for this NLP task.

Given the upward trend in Kendall's $\tau$ for the topic positional balancing ($T_{pos}$) in both Word2Vec and FastText, we decided, with purely exploratory intent, to investigate whether this improvement would persist in these algorithms. Figure 4 (see Appendix A.4) shows the performance using topic positionally balancing for values of $\rho$ up to 50. A limit to the improvement from incremental duplication can be observed; the average $\langle\tau\rangle$ no longer exceeds the results obtained at $\rho = 12$. Indeed, for $\rho = 50$ in Word2Vec, the $\langle\tau\rangle$ obtained is lower than the initial value. In FastText, the final four average $\langle\tau\rangle$ values remain very close to the initial baseline.

Finally, a significant difference was observed between the CBOW and Skip-gram architectures of the learning algorithms. CBOW is an architecture that focuses on predicting an unknown word $X$ based on its context — the set $C(X)$ of words surrounding $X$ (Mikolov et al., 2013c). $X$ is then predicted based on the information provided by its context, $C$. In contrast, Skip-gram predicts the words surrounding $X$. Word2Vec generates a single vector (embedding) for each word in the vocabulary, whereas FastText generates an embedding for each Nawatl character $n$-gram. This allows for the construction of vectors containing more information, by virtue of these $n$-grams. We confirmed experimentally that FastText's Skip-gram architecture significantly outperforms the other algorithms.

## 8 Conclusions and Future work

The results obtained highlight the effectiveness of the balancing and duplication techniques proposed for the corpora. Indeed, when these techniques are applied to agglutinative and polysynthetic languages having limited computational resources, they seem to facilitate the training of models that produce static representations. In this way, static representations trained with these techniques capture the language's structure more effectively.

Specifically, in the case of FastText, the topical positional balancing, combined with the $\rho = 12$ replicas of the corpus improved the Kendall coefficient $\langle\tau\rangle$ from 0.459 (using the original, unbalanced corpus) to $\langle\tau\rangle = 0.515$, representing a significant gain of **12.2%**. Although Word2Vec does not achieve the highest mean value of Kendall's $\langle\tau\rangle$, it is the model that obtained the most representative gain, illustrating the advantages of the approach presented. Their average Kendall's coefficient at the starting point (without balancing or duplication) is $\langle\tau\rangle = 0.357$, whilst topic positional balancing, combined with incremental duplication, achieves a $\langle\tau\rangle = 0.481$, i.e. a **34.7%** improvement.

[9] Not visible in the Fig. 3, at left.
[10] Not visible in the Fig. 3, at right.

| Model | $\langle\tau\rangle$ | $\max\langle\tau\rangle_{\rho\times}$ | $\rho$ | Gain % | Time (min) |
|---|---|---|---|---|---|
| FastText Skip-gram | 0.459 | **0.495** | 10 | 7.8 | 46.6 |
| Word2Vec Skip-gram | 0.357 | 0,483 | 22 | 35.3 | 39.3 |
| FastText CBOW | 0.345 | 0.393 | 8 | 13.9 | 43.7 |
| Word2Vec CBOW | 0.220 | 0.257 | 20 | 16.8 | 14.9 |
| GloVe | 0.209 | 0.216 | 6 | 3.4 | 6.5 |
| **Baselines** | $\langle\tau\rangle$ | | | | |
| FastText/Wikipedia | 0.242 | - | - | - | - |
| FastText/Common Crawl | 0.240 | - | - | - | - |
| Word2Vec/Wikipedia | 0.240 | - | - | - | - |

Table 2: Unbalanced $\pi$-YALLI corpus. Kendall's $\langle\tau\rangle$ over five runs of the models without duplication, and $\max\langle\tau\rangle_{\rho\times}$: the maximum $\tau$ obtained with $\rho\times$ duplications. %: percentage of $\langle\tau\rangle$ improvement. The learning time is approximate for each single $\rho$ run, executed on a cluster with a requirement of [8, 12] cores and [16, 64] GB of RAM, running under GNU/Linux in SLURM (*Simple Linux Utility for Resource Management*) mode. The baselines use pre-training models.

| | **Unbalanced corpus** | **Balanced corpora** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Topical** | | | | **Dialectal** | | | |
| **Skip-gram Model** | $\langle\tau\rangle$ | $\langle\tau\rangle$ | | **Gain%** | | $\langle\tau\rangle$ | | **Gain%** | |
| | | $T_+$ | $T_{pos}$ | $T_+$ | $T_{pos}$ | $D_+$ | $D_{pos}$ | $D_+$ | $D_{pos}$ |
| FastText | *0.459* | 0.467 | **0.477** | 1.7 | **3.9** | 0.465 | 0.468 | 1.3 | 1.9 |
| Word2Vec | *0.357* | **0.428** | 0.425 | **19.9** | 19.0 | 0.413 | 0.381 | 15.7 | 6.7 |

Table 3: Starting point (without duplication) of unbalanced and balanced corpora. Kendall's $\langle\tau\rangle$ on 5 runs of Skip-gram models. +: Uniform balancing, pos: Positional balancing, T: Topical, D: Dialectal.

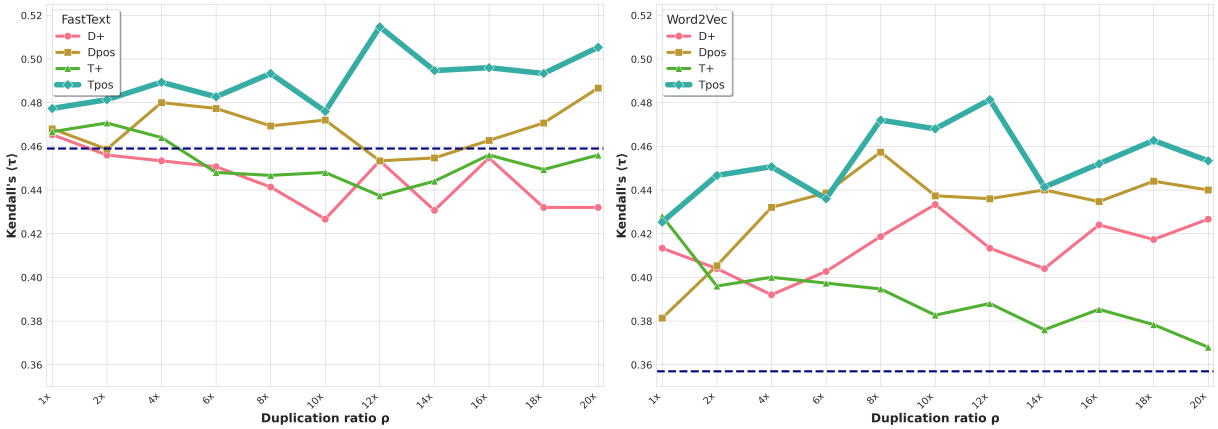


Figure 3: Balanced and incrementally duplicated corpus $\pi$-YALLI. **FastText** (left) vs. **Word2Vec** (right) in Skip-gram mode. Kendall's coefficient $\langle\tau\rangle$ of the sentence semantic similarity task. D+: Uniform balancing by dialectal variety, T+: Topic uniform balancing, Dpos: Positional balancing by dialectal variety, Tpos: Topic positional balancing. The dashed lines represent the $\langle\tau\rangle$ obtained using the original corpus (without balancing or duplication). There are 5 runs per duplication point $\rho$ (the scale is the same on both graphics).

We found that positional balancing outperforms uniform balancing, particularly when applied to topics. We have reached the conjecture that uniform balancing tends to favour data redundancy to a greater extent. Furthermore, this is exacerbated when incremental duplication is applied; consequently, in several cases starting from $\rho = 2$, a deterioration in Kendall's $\tau$ can be observed. This suggests that while maintaining a proper balancing between classes within a corpus is crucial, it is equally important to prevent them from becoming excessively redundant.

Our research highlights the positive impact on model learning of applying a statistically balanced,

positional strategy to heavily imbalanced corpora. Furthermore, these benefits are amplified when combined with an incremental corpus-duplication technique. Even without any balancing applied, the duplication strategy alone shows positive effects on the model training. Duplication is thus an efficient and comprehensive alternative for corpora expansion, particularly when dealing with NLP of $\pi$-languages.

In future work, we will explore other ways of balancing corpora and their impact on the incremental duplication technique. Similarly, we intend to evaluate the contribution of expanded corpora to the tasks of Automatic Text Summarisation, Sentiment analysis and Named Entity recognition—such as toponyms detection—in Nawatl.

## Limitations

Our results indicate that the proposed corpus balancing and duplication methods yield better results than using the original corpora.

Although these results are very promising, we recognise that further experiments are needed with other types of balancing and duplication, particularly using other NLP tasks to assess the limitations in greater detail. This is especially true for $\pi$-languages, which have very few computational resources.

## Ethics Statement

We are mindful of the potential risks associated with data duplication, including the possibility of encouraging redundancy and bias, and the risk of minimizing or excluding some lects and scripts of Nawatl speech.

We therefore strongly advocate that this technology be used exclusively to promote the appreciation of Nawatl and to support the development of digital resources that facilitate its study and dissemination.

## Acknowledgments

This research work has been financed by the Agorantic NAWA project and the Intermedius PhD Grant, and supported by the Laboratoire Informatique d'Avignon, from Avignon Université (France).

## References

Nimaan Abdillahi, Pascal Nocera, and Juan Manuel Torres. 2006. Boites a outils TAL pour les langues peu informatisées : Le cas du Somali. In *Journées d'Analyses des Données Textuelles*, Besançon, France.

Najib Arbach and Saandia Ali. 2013. Aspects théoriques et méthodologiques de la représentativité des corpus. *Corela [En ligne], HS-13*.

Vincent Berment. 2004. *Méthodes pour informatiser les langues et les groupes de langues "peu dotées"*. Ph.D. thesis, Université Joseph-Fourier - Grenoble I.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

Khoo Khyou Bun and Mitsuru Ishizuka. 2002. Topic extraction from news archive using tf* pdf algorithm. In *3rd International Conference on Web Information Systems Engineering (WISE'02)*, pages 73–82. IEEE.

Wright-Carr. David Charles. 2016. *Lectura del náhuatl*. Instituto Nacional de Lenguas Indígenas.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in NLP. *Transactions of the ACL*, 11:191–211.

Jaqueline de Durand-Forest, Danièle Dehouve, and Eric Roulet. 1995. *Parlons Nahuatl. La langue des Aztèques*. L'Harmattan.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the ACL: Human Language Technologies, Vol 1*, pages 4171–4186, Minneapolis, Minnesota. ACL.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the ACL: ACL-IJCNLP 2021*, pages 968–988, Online. ACL.

Lucero Flores Nájera. 2019. *La gramática de la clausula simple en el náhuatl de Tlaxcala*. Ph.D. thesis, CIESAS.

Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing semantic similarity for entity linking with convolutional neural networks. In *NAACL: Human Language Technologies*, pages 1256–1261, San Diego, California. ACL.

Palash Goyal, Sumit Pandey, and Karan Jain. 2018. *Deep Learning for Natural Language Processing*. Springer.

Ximena Gutierrez-Vasques, Robert Pugh, Victor Mijangos, Diego Barriga Martínez, Paul Aguilar, Mikel Segura, Paola Innes, Javier Santillan, Cynthia Montaño, and Francis Tyers. 2025. Py-elotl: A python NLP

package for the languages of Mexico. In *5th Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 38–47, Albuquerque, New Mexico. ACL.

Juan-José Guzmán-Landa, Juan-Manuel Torres-Moreno, Martha-Lorena Avendaño-Garrido, Miguel Figueroa-Saavedra, Ligia Quintana-Torres, Graham Ranger, Carlos-Emiliano González-Gallardo, Elvys Linhares-Pontes, Patricia Velázquez-Morales, and Luis-Gil Moreno-Jiménez. 2025. $\pi$-YALLI : un nouveau corpus pour des modèles de langue nahuatl / Yankuik nawatlahtolkorpus pampa tlahtolmachiotl. In *TALN, vol 1*, pages 802–816, Marseille, France. ATALA.

Juan-José Guzman-Landa, Jesús Vázquez-Osorio, Juan-Manuel Torres-Moreno, Ligia Quintana-Torres, Miguel Figueroa-Saavedra, Martha-Lorena Avendaño Garrido, Graham Ranger, Patricia Velázquez-Morales, and Gerardo Sierra-Martínez. 2025. A symbolic algorithm for the unification of nawatl word spellings. In *Advances in Soft Computing: 24th MICAI'25, Guanajuato, Mexico, 2025, Part I*, page 141–154, Berlin, Heidelberg. Springer-Verlag.

Juan-José Guzmán-Landa, Juan-Manuel Torres-Moreno, Graham Ranger, Miguel Figueroa-Saavedra, Ligia Quintana-Torres, Carlos-Emiliano González-Gallardo, Luis-Gil Moreno-Jiménez, and Martha-Lorena Avendaño-Garrido. 2026. Nawatl context-free grammars for Natural Language Processing. In *15th Language Resources and Evaluation Conference (LREC)*, pages 3333–3342, Palma, Spain. ELRA.

Juan-José Guzmán-Landa, Juan-Manuel Torres-Moreno, Miguel Figueroa-Saavedra, Carlos-Emiliano González-Gallardo, Graham Ranger, and Martha Lorena-Avendaño-Garrido. 2026. Classifying several dialectal nawatl varieties. *Preprint*, arXiv:2601.02303.

Magnus Pharao Hansen. 2024. *Nahuatl Nations: Language Revitalization and Semiotic Sovereignty in Indigenous Mexico*. Oxford University Press.

INEGI. 2020. Censo de población y vivienda 2020. In *CENSO 2020*. `https://www.inegi.org.mx/rnm/index.php/catalog/632/study-description`.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Yolanda Lastra de Suárez. 1986. *Las áreas dialectales del náhuatl moderno*. UNAM, Instituto de Investigaciones Antropológicas, Mexico.

Michel Launey. 1978. *Introduction à la langue et à la littérature aztèques*, volume 1. L'Harmattan, Paris.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *60th Annual Meeting of the ACL (V1)*, pages 8424–8445, Dublin, Ireland. ACL.

Mosleh Mahamud, Zed Lee, and Isak Samsten. 2023. Distributional data augmentation methods for low resource language. *Preprint*, arXiv:2309.04862.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. 2020. On the importance of pre-training data volume for compact language models. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858, Online. ACL.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *NIPS - Vol 2*, NIPS, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL – HLT 2013)*, pages 746–751, Atlanta, GA, USA. ACL.

Justyna Olko and John Sullivan. 2016. Bridging gaps and empowering speakers: An inclusive, partnership-based approach to nahuatl research and revitalization. *Integral strategies for language revitalization*, pages 347–386.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Preprint*, arXiv:2406.17557.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. ACL.

Robert Pugh, Varun Sreedhar, and Francis Tyers. 2024. Wav2pos: Exploring syntactic analysis from audio for Highland Puebla Nahuatl. In *4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 121–126, Mexico City, Mexico. ACL.

Robert Pugh and Francis Tyers. 2021. Investigating variation in written forms of Nahuatl using character-based language models. In *1st Workshop on Natural*

*Language Processing for Indigenous Languages of the Americas*, pages 21–27. ACL.

Robert Pugh, Francis Tyers, and Marivel Huerta Mendez. 2021. Towards an open source finite-state morphological analyzer for zacatlán-ahuacatlán-tepetzintla Nahuatl. In *4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Vol 1*, pages 80–85. ACL.

Robert Pugh, Cheyenne Wing, María Ximena Juárez Huerta, Ángeles Márquez Hernandez, and Francis Tyers. 2025. Ihquin tlahtouah in tetelahtzin-cocah: An annotated, multi-purpose audio and text corpus of western sierra Puebla Nahuatl. In *Conference of the Nations of the Americas Chapter of the ACL: Human Language Technologies (Vol 1)*, pages 3549–3562, Albuquerque, New Mexico. ACL.

Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple BM25 extension to multiple weighted fields. In *Thirteenth ACM international conference on Information and knowledge management*, pages 42–49.

Mitsuya Sasaki. 2022. Divide y entenderás: El papel de la polarización sintáctica en el náhuatl moderno y colonial. In *Coloquio de Investigación Lingüística, Universidad de Sonora (Mexico)*.

Jiatong Shi, Jonathan D. Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan, and Shinji Watanabe. 2021. Highland Puebla Nahuatl speech translation corpus for endangered language documentation. In *1st Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 53–63. ACL.

Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2022. *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. O'Reilly Media.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP-IJCNLP*, pages 6382–6388, Hong Kong, China. ACL.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *20th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. ELRA.

Klaus Zimmermann. 2019. Estandarización y revitalización de lenguas amerindias: funciones comunicativas e ideológicas, expectativas ilusorias y condiciones de la aceptación. *Revista de Llengua i Dret, Journal of Language and Law*, 71:111–122.

# A Appendix

In this Appendix, we present: (A.1) the list of acronyms used (both dialectal and topical); (A.2) an example of Nawatl semantic similarity between a reference sentence and its candidates; (A.3) the learning hyper-parameters for the models; (A.4) the comparison of the two best-performing learning algorithms, and (A.5) the statistical distribution of topics and dialectal varieties.

## A.1 Dialectal and Topical acronyms

The following acronyms will be used to designate the 26 dialectal varieties available of the $\pi$-YALLI corpus:

***CV**: Veracruz Central Nawatl, **H**: Huasteca, **HH**: Hidalgo's Huasteca, **HV**: Veracruz Huasteca, **HP**: Potosí Huasteca, **CEA**: Upper Central Region Mexican, **CEO**: Central-Western Mexican, **CEN**: Central Mexican, **CEB**: Central Low Mexican, **ORI**: Eastern Mexican, **GUE**: Guerrero's Mexican, **NOC**: Central Northwest, **IST**: Isthmus Nawatl, **NAW**: Nawat, **OAX**: Oaxaca's Nawatl, **CEP**: Puebla Center, **ANP**: Puebla's Northern Highlands, **ORP**: Eastern Puebla's Mexican, **SNP**: Puebla's Sierra Negra, **SNNP**: Puebla's Northern Sierra Negra, **SNSP**: Puebla's Southern Sierra Negra, **SNEP**: Puebla's Northeast Sierra, **SOP**: Puebla's Western Sierra, **TEM**: Temixco's Mexican, **TLA**: Tlaxcala's Mexican, and **MIX**: Mixture of dialectal varieties*[11].

The acronyms for the 16 topics are as follows[12]: ***REL**: Religion, **LIT**: Literature, **HIS**: History, **EDU**: Education, **LIN**: Linguistics, **LEG**: Legislation, **WIK**: Wikipedia, **COS**: Cosmovision, **TEC**: Technology, **ECO**: Economics, **MED**: Medicine, **AGR**: Agriculture, **POE**: Poetry, **MUS**: Music, **PHR**: Sentences without context, and **POL**: Politics.*

[11]The original INALI Spanish names for the dialectal varieties are as follows: **CV**: Nawatl Central de Veracruz, **H**: Huasteca, **HH**: Huasteca Hidalguense, **HV**: Huasteca Veracruzana, **HP**: Huasteca Potosina, **CEA**: Mexicano del Centro Alto, **CEO**: Mexicano Central de Occidente, **CEN**: Mexicano del Centro, **CEB**: Mexicano Central Bajo, **ORI**: Mexicano del Oriente, **GUE**: Mexicano de Guerrero, **NOC**: Noroeste Central, **IST:** Nawatl del Istmo, **NAW**: Nawat, **OAX**: Nawatl de Oaxaca, **CEP**: Centro de Puebla, **ANP**: Alto del Norte de Puebla, **ORP**: Mexicano del Oriente de Puebla, **SNP**: Sierra Negra de Puebla, **SNNP**: Sierra Negra Norte Puebla, **SNSP**: Sierra Negra Sur de Puebla, **SNEP**: Sierra Noreste de Puebla, **SOP**: Sierra Oeste de Puebla, **TEM**: Mexicano de Temixco, **TLA**: Mexicano de Tlaxcala, **MIX**: Mezcla de variedades.

[12]Classification of the Atlas of Languages INALI: https://atlas.inali.gob.mx/agrupaciones/info/0211

### A.2 Example of a reference-candidates block 10, for the semantic similarity task

**REFERENCE SENTENCE (10):**
Yewehkatlahtolli momachtia ken okatka tlakayotl /
*History studies the past of Humanity.*

**RANKED CANDIDATE SENTENCES:**

1. Tikmatih tlen opanok, ken okatka tlakayotl ika yewehkatlahtolli.
   *We know about humanity's past thanks to history.*
2. Tlen ye wehkah otlamochih momachtia ipan weyi tlamachtiloyan.
   *Historical events are studied at university.*
3. Momachtilistli itechpa wehkawitl techpalewia pampa tikachtopaittaskeh yakapankawitl.
   *Studying the past helps us to predict the future.*
4. In wehkawitl ye wehka opanok
   *The past is just that: the past.*
5. Nonemilis nesi ihkin inemilis notahtzin: ohwi.
   *My personal story is much like my father's: complicated.*

### A.3 Training Hyper-parameters used for the models

The hyper-parameters used for all models are as follows:

Number of epochs: **20**; Context window size: **5 tokens**; Embeddings' dimension: **300**; and only for GloVe algorithm: Cutoff = **100** and $\alpha$ = **3/4**.

### A.4 Comparing mean $\langle\tau\rangle$ of FastText vs. Word2Vec with $1 \leq \rho \leq 50$

In this comparison, both algorithms use the skip-gram architecture. Kendall's $\langle\tau\rangle$ coefficient for the sentence semantic similarity task is shown in the Figure 4. The results reflect the algorithms' performance on incrementally duplicated and topic-positionally balanced corpora.

The dotted lines in the figure indicate the $\tau$ values for the trained models (FastText in green, Word2Vec in blue) using the original $\pi$-YALLI corpus (without duplication or balancing).

In this experiment there are 10 runs per duplication ratio $1 \leq \rho \leq 50$, where the standard deviation is shown as a coloured band.

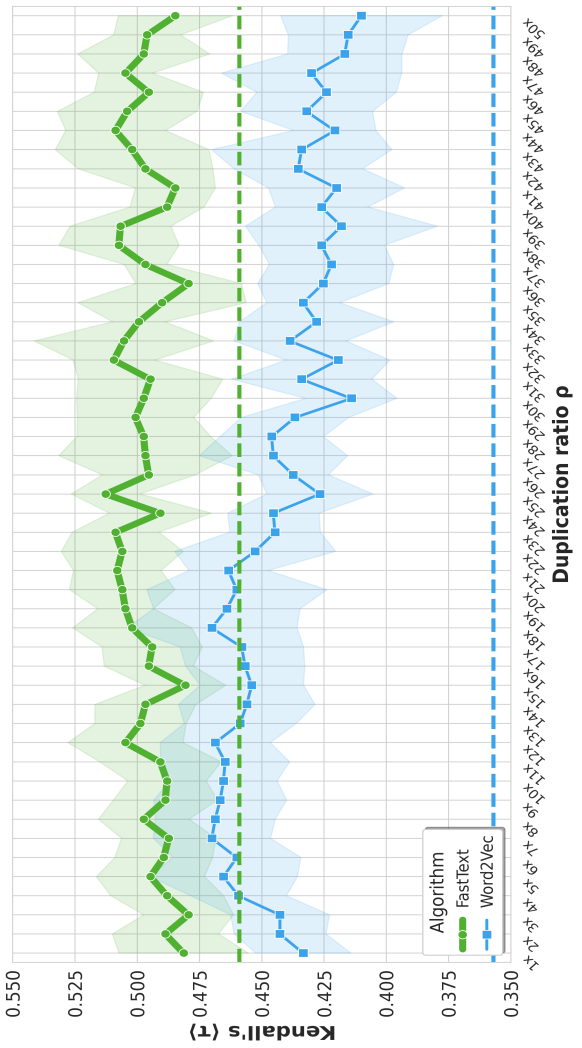


Figure 4: **FastText** (green) vs. **Word2Vec** (blue) using topic positionally balanced corpora. The dotted lines show Kendall's $\tau$ for the trained models on the original $\pi$-YALLI corpus.

### A.5 Statistical distribution of tokens of the nawatl corpus

In Tables 4 and 5, we show the statistical distribution of words (tokens) by topics and dialectal varieties within the nawatl $\pi$-YALLI corpus.

| Topics | | |
|---|---|---|
| **Acronym** | **#** | **Num.** |
| REL | 3 183,638 | 1 |
| LIT | 958,679 | 2 |
| HIS | 613,682 | 3 |
| EDU | 468,585 | 4 |
| LIN | 388,546 | 5 |
| LEG | 303,748 | 6 |
| WIK | 162,994 | 7 |
| COS | 52,304 | 8 |
| TEC | 25,899 | 9 |
| ECO | 14,854 | 10 |
| MED | 12,668 | 11 |
| AGR | 7,604 | 12 |
| POE | 3,934 | 13 |
| MUS | 2,417 | 14 |
| PHR | 2,208 | 15 |
| POL | 1,750 | 16 |

Table 4: Tokens' number (#) per topic.

| Dialectal varieties | | |
|---|---|---|
| **Acronym** | **#** | **Num.** |
| HV | 1 307,351 | 1 |
| MIX | 938,825 | 2 |
| HP | 833,297 | 3 |
| CEA | 665,008 | 4 |
| CV | 457,333 | 5 |
| GUE | 453,994 | 6 |
| CEO | 242,554 | 7 |
| IST | 189,767 | 8 |
| SNNP | 185,564 | 9 |
| OAX | 170,483 | 10 |
| SOP | 156,861 | 11 |
| SNP | 155,944 | 12 |
| ANP | 142,385 | 13 |
| SNEP | 89,127 | 14 |
| NAW | 64,271 | 15 |
| NOC | 62,406 | 16 |
| HH | 32,581 | 17 |
| H | 25,204 | 18 |
| CEP | 12,746 | 19 |
| TEM | 7,220 | 20 |
| ORI | 4,263 | 21 |
| CEN | 2,273 | 22 |
| SNSP | 1,618 | 23 |
| TLA | 972 | 24 |
| CEB | 735 | 25 |
| ORP | 728 | 26 |

Table 5: Tokens' number (#) per dialectal variety.