data

MDPI

Data Descriptor

Mixtec–Spanish Parallel Text Dataset for Language Technology Development

Hermilo Santiago-Benito [1], Diana-Margarita Córdova-Esparza [1,*], Juan Terven [2], Noé-Alejandro Castro-Sánchez [3], Teresa García-Ramirez [2], Julio-Alejandro Romero-González [2] and José M. Álvarez-Alvarado [4]

[1] Facultad de Informática, Universidad Autónoma de Querétaro, Av. de las Ciencias S/N, Campus Juriquilla, Querétaro 76230, Mexico; hsantiago13@alumnos.uaq.mx
[2] Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada—Unidad Querétaro, Instituto Politécnico Nacional, Cerro Blanco No. 141, Col. Colinas del Cimatario, Querétaro 76090, Mexico; jrtervens@ipn.mx (J.T.); teregar@uaq.mx (T.G.-R.); julio.romero@uaq.mx (J.-A.R.-G.)
[3] Centro Nacional de Investigación y Desarrollo Tecnológico, Tecnológico Nacional de México, Interior Internado Palmira S/N, Palmira, Cuernavaca 62493, Mexico; noe.cs@cenidet.tecnm.mx
[4] Facultad de Ingeniería, Universidad Autónoma de Querétaro, Querétaro 76010, Mexico; jmalvarez@uaq.edu.mx
* Correspondence: diana.cordova@uaq.mx

Abstract

This article introduces a freely available Spanish–Mixtec parallel corpus designed to foster natural language processing (NLP) development for an indigenous language that remains digitally low-resourced. The dataset, comprising 14,587 sentence pairs, covers Mixtec variants from Guerrero (Tlacoachistlahuaca, Northern Guerrero, and Xochapa) and Oaxaca (Western Coast, Southern Lowland, Santa María Yosoyúa, Central, Lower Cañada, Western Central, San Antonio Huitepec, Upper Western, and Southwestern Central). Texts are classified into four main domains as follows: education, law, health, and religion. To compile these data, we conducted a two-phase collection process as follows: first, an online search of government portals, religious organizations, and Mixtec language blogs; and second, an on-site retrieval of physical texts from the library of the Autonomous University of Querétaro. Scanning and optical character recognition were then performed to digitize physical materials, followed by manual correction to fix character misreadings and remove duplicates or irrelevant segments. We conducted a preliminary evaluation of the collected data to validate its usability in automatic translation systems. From Spanish to Mixtec, a fine-tuned GPT-4o-mini model yielded a BLEU score of 0.22 and a TER score of 122.86, while two fine-tuned open source models mBART-50 and M2M-100 yielded BLEU scores of 4.2 and 2.63 and TER scores of 98.99 and 104.87, respectively. All code demonstrating data usage, along with the final corpus itself, is publicly accessible via GitHub and Figshare. We anticipate that this resource will enable further research into machine translation, speech recognition, and other NLP applications while contributing to the broader goal of preserving and revitalizing the Mixtec language.

Dataset: https://doi.org/10.6084/m9.figshare.28557182.v1.

Dataset License: CC-BY-4.0

Keywords: Mixtec language; parallel corpus; low resource language; OCR

Academic Editor: Joaquín Torres-Sospedra

Received: 30 March 2025
Revised: 2 June 2025
Accepted: 20 June 2025
Published: 21 June 2025

Citation: Santiago-Benito, H.; Córdova-Esparza, D.-M.; Terven, J.; Castro-Sánchez, N.-A.; García-Ramirez, T.; Romero-González, J.-A.; Álvarez-Alvarado, J.M. Mixtec–Spanish Parallel Text Dataset for Language Technology Development. *Data* **2025**, *10*, 94. https://doi.org/10.3390/data10070094

1. Summary

This work presents the first publicly available preprocessed Spanish–Mixtec parallel corpus in Mexico, designed to facilitate the development of language technologies for this historically low-resource language [1,2]. The dataset, currently comprising 14,587 sentence pairs, was collected from digital sources such as official government websites, religious organization websites, Mixtec enthusiast blogs, and physical texts obtained by scanning printed books at the Autonomous University of Querétaro (UAQ). By including texts from diverse domains, such as education, law, health, and religion, and drawing on 13 Mixtec variants spread throughout Guerrero and Oaxaca, this corpus provides a rich and representative linguistic resource.

Mixtec is regarded as a low-resource language in a digital context, lacking large-scale documentation and linguistic tools [3–5]. According to the National Institute of Indigenous Languages (INALI), Mixtec belongs to the Oto-Manguean family, which includes Amuzgo, Triqui, Zapotec, and others [6]. It is one of the most spoken indigenous languages in Mexico, with 526,593 speakers [7]. This situation has motivated the development of technologies aimed at reducing language barriers across various domains. In education, projects such as Aprendamos Mixteco [8] have used gamified digital resources to teach Mixtec to migrant communities in the U.S., while mobile applications for teaching Nahuatl with voice recognition have proven effective in both urban and rural contexts [9]. In another context, automatic translators between Spanish and indigenous languages (such as Nahuatl and Mixtec) offer new possibilities to improve initial interpretation in legal proceedings where speakers do not understand Spanish [10]. Recent projects have also explored the use of multilingual models and synthetic data to translate medical instructions into Mixtec [11]. However, it faces significant challenges in computational processing due to fragmented documentation and regional variants.

In response to these challenges, this dataset underwent an extensive methodology that included the systematic search and collection of bilingual texts, OCR-based digitization, and manual and semi-automatic correction to normalize special Mixtec glyphs [12–15], along with data cleaning to remove duplicates, empty lines, and nonparallel fragments. The final corpus adheres to UTF-8 encoding, ensuring broad usability across multiple NLP pipelines.

This project provides the basis for future linguistic research and technology development in Mexico. The resource has immediate applications in machine translation (including multimodal translation [16]), optical character recognition [17,18], information retrieval [19], and voice recognition [20,21]. In addition, it can serve as an essential step towards the preservation of the Mixtec language and culture [22,23] by supporting educational materials, dictionaries, and other language revitalization tools. Preliminary experiments with GPT-based models, as well as open source frameworks such as mBART-50 and M2M-100, have shown promising but still limited translation quality, highlighting the need for further research and larger corpora.

By publicly releasing and describing the dataset, our aim is to empower researchers, developers, and community members to collaborate and advance language technologies for Mixtec, ultimately contributing to both linguistic innovation and cultural preservation. The usage instructions are available on the following GitHub repository: https://github.com/hermilocap/ParalellCorpusMixtec-Spanish (accessed on 18 June 2025).

2. Background

A language is considered digitally low-resourced when it lacks sufficient documentation, either in digital format or otherwise, to support the development of language technologies [3–5]. In many cases, these languages are also classified as endangered due

to a declining number of speakers and limited resources (e.g., dictionaries, textbooks) available either online or in print. Consequently, both the scarcity of digital data and the small population of native speakers hamper the creation of natural language processing (NLP) tools such as tokenizers, morphological analyzers, and spell checkers.

Based on our preliminary investigations, the Mixtec language (which belongs to the Oto-Manguean family) can be classified as a low-resource digital language. Despite the existence of written materials in various dialects, these materials are scattered across physical libraries, such as those of local universities and social organizations, and they are not readily accessible in digital form. Unlike Nahuatl, which has 1,651,958 speakers [24] and historical corpora (38 Nahuatl–Spanish texts, multiple dictionaries, and extensive literature [25]), Mixtec lacks digital resources, and its 81 varieties [26] hinder the development of language technologies. While languages such as Quechua and Guarani have made significant advances due to larger corpora and a degree of orthographic standardization [27], Mixtec has required alternative strategies, including the generation of synthetic data, the manual construction of parallel corpora, and the use of neural network-based translation, which can achieve high accuracy even with limited corpora in controlled contexts [28]. This lack of digital documentation severely restricts the development of core NLP tasks and language applications.

At present, there is no automatic tokenizer or morphosyntactic parser available for Mixtec. The absence of essential NLP tools, compounded by the shortage of domain-diverse corpora, makes it challenging to build robust language technologies. Moreover, the fragmentation of data across multiple dialects, each with its orthographic conventions, adds further complexity to any standardization efforts.

In light of these challenges, our primary motivation is to contribute to the preservation and promotion of the Mixtec language by compiling a parallel corpus that encompasses various domains (e.g., health, legal, and educational) and dialectal variants. This resource can provide the foundation for critical NLP technologies, including text classifiers for language identification, speech recognizers, and multimodal translation systems. These tools not only serve researchers and developers, but they can also address urgent community needs in domains such as public health, legal services, and education.

This work represents a substantial endeavor to collect, digitize, and manually correct parallel texts in Mixtec. Notably, the process required scanning and converting physical materials into editable digital formats, followed by meticulous verification and the alignment of Mixtec–Spanish sentences. The result is a novel resource that can be used for academic research, educational initiatives, and the broader development of language technologies intended to improve the vitality of Mixtec.

3. Data Description

According to INALI, Mexico is home to 68 linguistic groups, 364 variants, and 11 linguistic families [6]. Among these families are Oto-Manguean, Yuto-Nahua, Seri, and Mixe-Zoque. Mixtec belongs to the Oto-Manguean family, sharing various linguistic similarities with languages such as Amuzgo, Triqui, Zapotec, and Tlapanec. The states of Guerrero, Oaxaca, and Puebla have the largest number of Mixtec speakers. Official estimates suggest that there are more than 81 distinct Mixtec variants in Mexico, although this number may be higher if small communities and municipalities with their own dialectal differences are considered.

Based on data from the National Institute of Statistics and Geography (INEGI), the Mixtec language is spoken by 526,593 individuals in Mexico, making it the fourth most spoken indigenous language in the country [7]. Despite the significant number of speakers,

the resources necessary for computational analysis remain fragmented or limited to physical formats, thereby hindering the creation of robust language technologies.

In this section, we describe the data generation process, dataset statistics, and final data organization.

3.1. Dataset Generation Process

The following steps were followed to generate the parallel corpus.

1. We visited university libraries and official government websites and had direct contact with Mixtec advocates who provided support on data sources.
2. We selected texts that had translations in Mixtec and Spanish and were in the public domain, discarding those that did not meet these criteria.
3. We collected texts from various categories whenever possible, including educational, health, legal, and religious texts.
4. Each Mixtec–Spanish pair was rigorously reviewed to ensure the correct alignment.
5. We corrected spelling errors and removed empty and duplicate sentences.

In generating the dataset, we identified that the parallel corpus was translated by native Mixtec speakers from Mixtec communities who were invited by educational institutions, government agencies, and private companies to collaborate in the translation of the texts. The texts generated by these institutions are intended to communicate and inform the Mixtec community, provide healthcare recommendations, and offer educational resources in the Mixtec language.

3.2. Mixtec Variants and Domains

By analyzing the collected texts, we identified the following 13 Mixtec variants: Western Coast, Southern Lowland, Santa María Yosoyúa, Central, Lower Cañada, Western Central, San Antonio Huitepec, Upper Western, Southwestern Central, Tlacoachistlahuaca, Northern Guerrero, and Xochapa. Figure 1 shows the total number of documents per variant. In particular, the Santa María Yosoyúa variant from the state of Oaxaca is represented by 12 documents, while two documents represent Northern Guerrero.

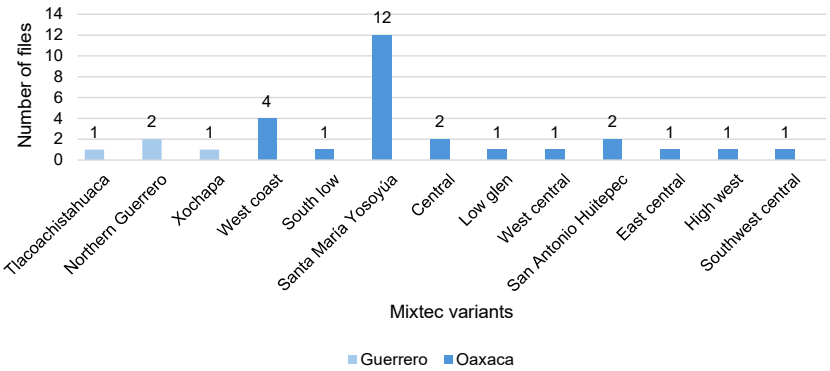

Figure 1. Corpus classification according to Mixtec variants. Santa María Yosoyúa from Oaxaca shows the largest number of collected documents (12), whereas Northern Guerrero has the largest share in the state of Guerrero (2).

Depending on the type of document, we organize the corpus into the following four main domains: education, law, health, and religion. Figure 2 illustrates the categories and their proportion from the total documents. The largest category is educational texts (80%), followed by legal texts (14%), health-related texts (3%), and religious texts (3%). Health materials primarily include patient care guidelines and patient rights documentation.

Although this distribution is unavoidably skewed, reflecting the real-world availability of written Mixtec, machine translation models can readily compensate for such bias through lightweight strategies such as temperature-based sampling or dynamic data weighting during fine-tuning [29].

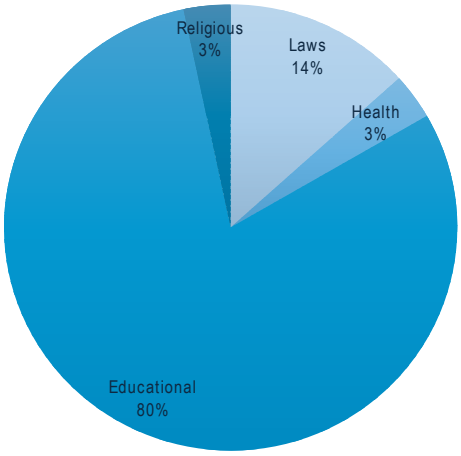

Figure 2. Proportion of documents by category. Educational texts form the majority of the corpus, while health-related texts represent the smallest portion.

3.3. Corpus Statistics

Table 1 provides an overview of the corpus at the sentence and word level by category. The religious category (primarily consisting of Bible passages) contains the highest number of sentences, 16,434, while health-related texts comprise only 22 sentences, limiting the representativeness of the parallel corpus in this domain.

Table 1. Number of sentences and words in each corpus category. Religious texts largely derive from the Bible, while legal texts include the Mexican Constitution and official documents pertaining to indigenous language rights.

Category	Number of Mixtec–Spanish Sentences	Total Words (Mixtec–Spanish)	Unique Words (Mixtec–Spanish)
Educational	9220	42,654	6971
Laws	3498	180,236	4634
Health	22	278	70
Religious	16,434	461,234	5035

3.4. Organization of the Repository

All parallel texts are available through the project repository. The main folder, data, contains two subfolders as follows: train and test. The train folder includes mixtec-train.txt and spanish-train.txt, while the test folder contains mixtec-test.txt and spanish-test.txt. Each file corresponds line by line to its parallel counterpart as follows:

- mixtec-train.txt: Used for training; each line in Mixtec aligns with the corresponding line in spanish-train.txt.
- spanish-train.txt: Contains the Spanish version of each sentence, matching each line from mixtec-train.txt.
- mixtec-test.txt: Used for testing; each Mixtec sentence aligns with the corresponding line in Spanish.
- spanish-test.txt: Contains the Spanish version of each sentence in mixtec-test.txt.

Formally, if a Mixtec file x contains sentences x_i for $i = 1 \ldots n$, the corresponding Spanish file y contains sentences y_i aligned line by line with x_i.

4. Methods

This section details the methodology used to create the dataset. We followed a three-phase approach for building the Spanish–Mixtec parallel corpus. This same approach can be extended to other digitally low-resourced languages to create robust and relevant parallel corpora that can support various language technologies. Figure 3 illustrates the methodology, which includes the following phases: (1) systematic search for parallel texts, (2) scanning of physical materials and semi-automatic correction, and (3) preprocessing including deduplication, removing empty fragments, and manual classification of texts [12–15].

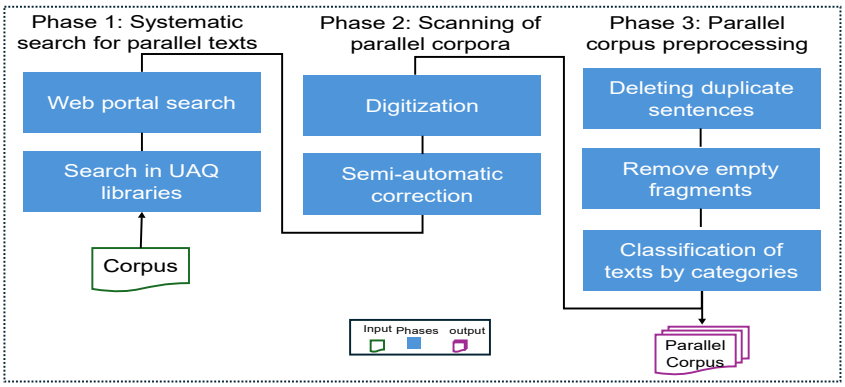

Figure 3. Methodology for constructing a Spanish–Mixtec parallel corpus. The input consists of books, flyers, posters, and webpages. After selecting parallel texts, the documents are scanned, preprocessed, and eventually classified into relevant categories. The final outcome is a clean, parallel corpus.

In the remainder of this section, we detail each phase.

4.1. Phase 1: Systematic Search for Parallel Texts

The first phase involves searching for parallel materials (Spanish–Mixtec), carried out in two main steps as follows: an online exploration and in-person visits to the libraries of the Autonomous University of Querétaro (UAQ). The goal is to discover and collect bilingual texts suitable for constructing the corpus. We apply inclusion and exclusion criteria to ensure that only relevant and high-quality documents are retained.

In Step 1, we began by identifying potential online portals that could host parallel texts, focusing on government resources such as the government website of Mexico, the INALI website, and the National Institute of Indigenous Peoples (INPI), as well as Mixtec language blogs and websites maintained by the Summer Institute of Linguistics (ILV). After visiting each platform, we downloaded texts that explicitly contained both Mixtec and Spanish content. This step yielded 22 digital books that contained parallel segments.

In Step 2, we then conducted a targeted search of the UAQ's virtual library catalog to identify additional bilingual materials. Following this digital exploration, we visited the libraries of the Faculty of Philosophy and the Faculty of Social Sciences in person. Through these visits, we obtained six additional physical books that contained parallel texts.

4.1.1. Inclusion Criteria

During the selection of relevant resources, the following inclusion criteria were used:

- The title of the book or document explicitly indicates that it includes both Mixtec and Spanish.

- There is a clear correspondence between a segment in Mixtec and its Spanish translation, at least at the paragraph or sentence level.
- For texts with long paragraphs, we confirmed that extensive sections in Mixtec are accurately matched with their Spanish equivalents.

4.1.2. Exclusion Criteria

The following factors led to the exclusion of certain materials as follows:

- Texts in which the Mixtec segment was disproportionately short compared to a very long Spanish passage (or vice versa). Long texts were on the order of 7000 words.
- Multilingual materials containing Mixtec, Spanish, and English, where the English component was not essential or only partially parallel.
- Poetry, songs, or lyrical verses, even if they contained some parallel segments, were omitted to maintain consistency in prose-based alignments.

4.2. Phase 2: Scanning of Physical Parallel Corpora

In this phase, we focus on digitizing the physical materials and ensuring their accuracy through a semi-automatic correction process.

- Digitization. Physical books from the UAQ libraries were manually scanned. Each page was scanned using a flatbed scanner, resulting in a digital file in either PDF or JPEG format. A total of eight books were digitized. This process is especially significant given the scarcity of digital resources for Mixtec, as it opens these texts to broader usage by researchers and computational linguists.
- Semi-automatic correction. To convert the scanned pages into machine-readable text, we used standard optical character recognition (OCR) software (e.g., ABBYY FineReader Engine 11). Since Mixtec is not natively supported by most commercial OCR tools, we used a Spanish-based recognition model and then customized it with a new Mixtec character set. Although ABBYY FineReader supports various high-resource languages, our review did not find direct support for Mixtec. Therefore, we constructed a special dictionary that contains relevant Mixtec glyphs. We ran the recognition process, reviewed the output, and corrected misread characters caused by faded printing and typed text. This iterative procedure was crucial in generating reliable digital text. Mixtec glyphs are those characters that OCR has difficulty recognizing [28]: A̱, a̱, E̱, e̱, I̱, i̱, O̱, o̱, U̱, u̱, Ā, ā, Ē, ē, Ī, ī, Ō, ō, Ū, ū. The incorrect words appear like x§én?, TahvT=tl, and caj o, while the correct forms are xēēn, Táhvi-ti, and caj-o. The blurred printouts that the OCR was unable to correct automatically were resolved by manually completing the characters.
  To validate the linguistic quality of the corpus, the texts were manually reviewed by a native Mixtec expert. After correction with ABBYY FineReader, the texts were reviewed to identify and correct spelling errors. A manual review of Mixtec grammar was also performed throughout the corpus.

4.3. Phase 3: Preprocessing of the Parallel Corpus

Preprocessing is indispensable in any NLP pipeline, with the goal of producing a clean, well-structured corpus for subsequent tasks, such as machine translation or language modeling. We followed the six steps described below.

1. Removing duplicate sentences. We identified and deleted duplicate segments that appeared multiple times within our corpus. These duplicates often arise from repeated passages in source documents or OCR artifacts. As duplicates can bias statistical models or neural networks, we removed them using a regular expression-based script that searched for repeated lines and deleted them in both the Mixtec and Spanish

versions simultaneously. Sentences such as “Te tu cuni ñuhün-un cahmu yunu xíi”, “Furthermore, with your spirit”, “—Vaha —cáchí da”, “T. Amen” were removed. Figure 4 shows our duplicate removal workflow.

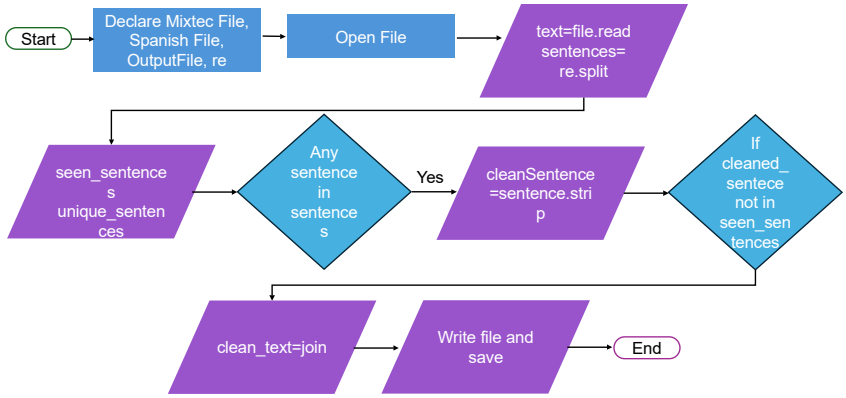

Figure 4. Flowchart for removing duplicate fragments. The algorithm opens the file, splits the text to identify unique sentences, trims leading and trailing spaces, removes duplicates, and saves the cleaned data.

2. Removing empty fragments. We found fragments containing only numbers or whitespace, plus non-textual metadata such as editorial information and printing addresses. We removed these lines from both language files, since they do not contribute meaningful linguistic content. Words such as “nan”, “- 21 -”, “- 3k -”; section titles from documents such as “Section 1: who you are”, “Section 2: I LIVE IN THIS TOWN”; and section descriptions such as “CURRENT PRESENT. IDENTIFYING PHRASE. RELATIONSHIP” and “WAY OF EXPRESSING THE SEARCH” were removed. Algorithm 1 outlines the process of detecting and deleting empty fragments.

Algorithm 1 Removing empty fragments or non-linguistic lines from the Mixtec–Spanish parallel corpus.

Require: MixtecFile: plain-text file in Mixtec
SpanishFile: plain-text file in Spanish
Ensure: Cleaned text files (*Mixtec* and *Spanish*)
1: Identify every empty fragment in both files {blank lines or lines with only whitespace}
2: Remove lines that contain no valid characters **or** only digits
3: Write the cleaned data into new text files

3. Removing non-parallel fragments. We also discarded segments where the Mixtec or Spanish part was disproportionately longer, indicating a lack of true parallel alignment. Disproportionately long fragments are those where the Mixtec text and its Spanish translation were present, but there was no division into alignment pairs. Sometimes the Mixtec text is shorter than the Spanish text, and there are no correspondences between the Mixtec and Spanish sentences.
4. Removing incorrect parallel segments. Some lines included Spanish text in the Mixtec file and vice versa. We manually aligned and corrected them to ensure each line in Mixtec corresponds to the correct line in Spanish.
5. Adding missing segments. We conducted a final review to identify incomplete lines, often caused by outdated translations or editorial omissions. In such cases, we referred back to the original source. This step was especially important for the Mexican Constitution’s translations, which had changed over time in Spanish but not always in Mixtec.

6. Categorizing texts. We completed the process by classifying the cleaned corpus into the following four major domains: education, law, and religious health. Figure 2 shows the number of documents per category, and Table 1 reports the total number of sentences and words in each category.

4.4. Exploratory Domain Analysis

To verify that the four domains are reflected in the lexical-semantic space of the corpus, we performed a lightweight, unsupervised analysis as follows:

1. All Spanish sentences and their Mixtec counterparts were embedded with LaBSE [30], yielding 768-dimensional vectors that are language-agnostic yet semantically aligned.
2. We projected the vectors to two dimensions using UMAP (cosine metric, $n_{\text{neighbors}} = 15$, min_dist $= 0.1$) to visualize the global structure while preserving local neighborhoods.
3. We plotted every sentence as a point, colored by its label, as shown in Figure 5 (Spanish) and Figure 6 (Mixtec). Dense, well-separated regions emerge for the Religious and Law domains, whereas Education and Health partially overlap, consistent with the class imbalance in Table 1. A Davies–Bouldin Index calculated in the original 768-D space yields 5.5 for Spanish and 5.3 for Mixtec. These high values (lower is better, with values less than one considered well separated) suggest appreciable overlap among the four topical clusters, consistent with the broad, partly interrelated nature of the domains.

This exploratory analysis reveals that lexical cues imprint an identifiable but diffuse domain structure, providing the corpus with a realistic and moderately challenging profile for downstream text classification tasks.

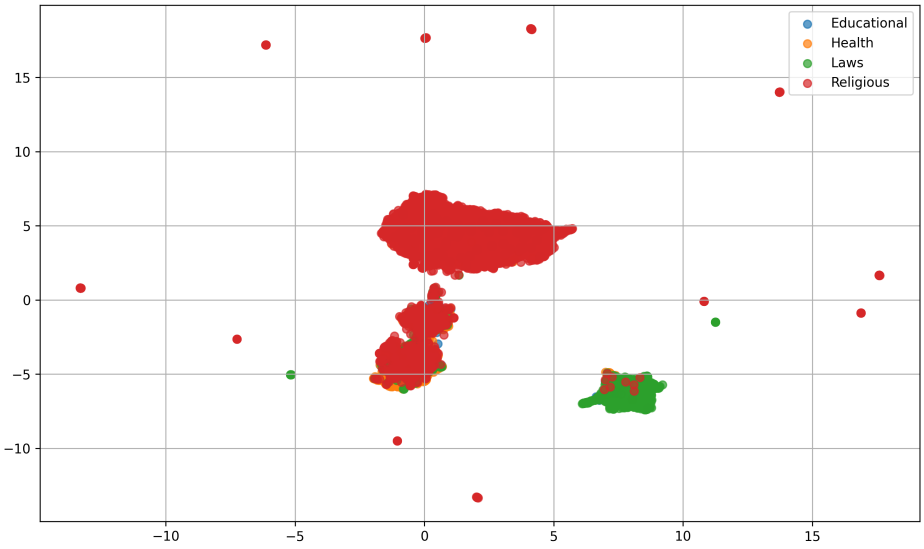

Figure 5. UMAP projection of Spanish sentences colored by domain. Clear clusters form for the "Religious" and "Laws" categories.

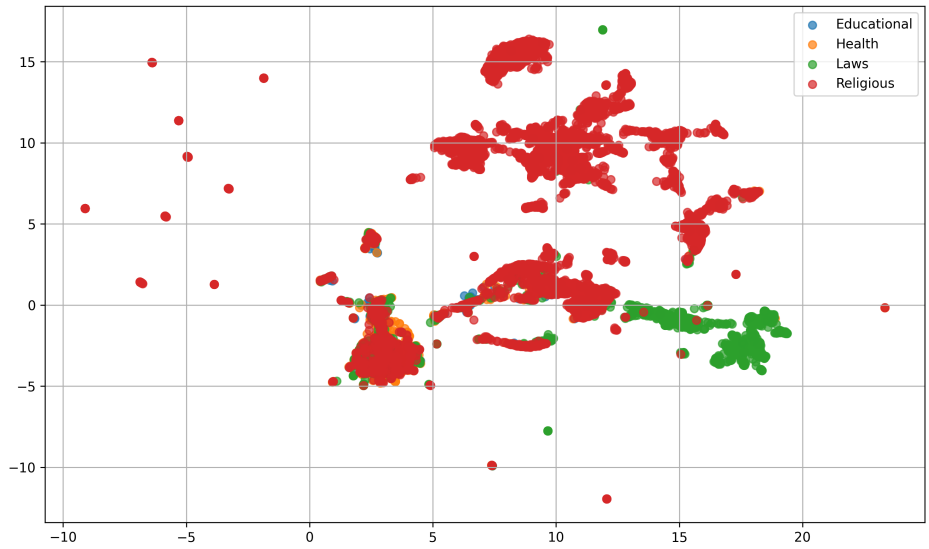

Figure 6. UMAP projection of Mixtec sentences. The global pattern mirrors the Spanish side, confirming cross-lingual consistency.

5. Translation Experiments

We used BLEU [31] and TER [32] as metrics in our experiments. BLEU evaluates translation quality from 0 to 1, using modified n-gram accuracy and a brevity penalty to align candidate translations with reference lengths in terms of word choice and order. This penalty applies corpus-wide for sentence flexibility. The reference length r is matched to the candidate sentence lengths, decreasing exponentially r/c. c indicates the total translation length of the candidate.

The BLEU metric is calculated by multiplying the geometric mean of the precision of modified n-grams by the brevity penalty, using formulas in Equations (1) and (2). The precision is derived for n-grams of length N with weights w_n. Here, c indicates the candidate translation length and r the reference corpus length.

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n \log P_n) \tag{1}$$

$$BP = \begin{cases} 1 & \text{if c>r} \\ e^{(1-r/c)} & \text{if c} \leq \text{r} \end{cases} \tag{2}$$

TER, on the other hand, measures the error rate in the prediction and the percentage of necessary corrections in the translation relative to the reference translation. The formula for obtaining the TER metric is shown in Equation (3).

$$TER = \frac{number\ of\ edits}{average\ number\ of\ reference\ word} \tag{3}$$

We used 14,587 Mixtec–Spanish sentence pairs divided into 80% training and 20% test sets. The experiments validated the usability of the data for the Spanish–Mixtec and Mixtec–Spanish translation using fine-tuned GPT, mBART-50, and M2M-100 models. The test set served as a reference for evaluation, with predictions reviewed by a native Mixtec expert and assessed using BLEU and TER metrics.

Tables 2 and 3 summarize the evaluation in terms of the BLEU and TER metrics.

Table 2. Evaluation of the Spanish–Mixtec dataset. Higher BLEU scores in the range of 0–100 indicate closer matches with the reference translation, whereas lower TER scores indicate fewer edits needed to match the reference.

Model	BLEU	TER
GPT-4o (zero-shot)	0.10	143.35
GPT-4o-mini (zero-shot)	0.07	131.32
Fine-tuned GPT-4o-mini	0.22	122.86
Fine-tuned mBART-50	4.2	98.99
Fine-tuned M2M-100	2.63	104.87

The results show that mBART-50 achieved the best results with a BLEU score of 4.2 and requires fewer edits (TER = 98.99).

Table 3. Evaluation of the Mixtec–Spanish dataset. The fine-tuned gpt-4o-mini-2024-07-18 model yields a higher BLEU score than M2M-100, although mBART-50 achieves better overall performance in terms of BLEU.

Model	BLEU	TER
GPT-4o (zero-shot)	0.46	182.28
GPT-4o-mini (zero-shot)	0.14	453.48
Fine-tuned GPT-4o-mini	0.23	121.01
Fine-tuned mBART-50	2.87	111.26
Fine-tuned M2M-100	0.05	112.57

For Mixtec–Spanish translation, mBART-50 also performs best, achieving a BLEU score of 2.87 and TER of 111.26. These experiments were conducted using five epochs and a learning rate of 3×10^{-5} on Google Colab with an A100 GPU, monitored using the Weights & Biases (Wandb) platform.

6. Discussion

In this section, we compare the advantages of our dataset with those of related works focused on parallel data collection for digitally low-resource languages.

The primary advantage of our dataset compared to the work of Tonja et al. [1] is its completeness. Every Spanish sentence in our corpus has a corresponding Mixtec sentence, eliminating the need for researchers to review and fill in missing data manually.

Another advantage compared to previous efforts [2,33] is that our corpus is thoroughly preprocessed, publicly available, and clean. We conducted rigorous preprocessing, including the removal of duplicate sentences, and we made the dataset openly accessible through FigShare and the code on GitHub.

Table 4 compares the size and processing properties of our corpus with related datasets.

Table 4. Comparison of corpus size with related works. A check-mark ✓indicates that the authors released a dataset that is already pre-processed or manually corrected, whereas a cross × indicates that such resource is not available.

Parallel Corpus	Number of Sentences	Languages	Preprocessed Datasets	Manual Corrected Datasets
Gaustad et al. [34]	110,367	English, Siswati	✓	×
Chiruzzo et al. [35]	14,500	Guarani, Spanish	✓	×
Oliver et al. [36]	173,530	South Tyrolean German, Italian	✓	×
Zhu et al. [37]	20,106	Chinese, Lao	×	✓
Ours	14,587	Mixtec, Spanish	✓	✓

A significant advantage of our dataset compared to other studies is its compilation of a diverse corpus sourced from multiple domains. This includes educational texts that can serve in the development of language technologies applicable to the educational sector, such as educational data mining [38] for identifying patterns that facilitate teaching and learning, Mixtec dialect classification [39], and correction of spelling errors [40].

Furthermore, several studies have shown that techniques such as translation [41], transfer learning [42], and the use of multilingual models [43] enable the development of functional tools even with limited available data. In the field of education, for example, the Lesan system has enabled translation between low-resource languages, facilitating access to educational content in local languages [44]. Mobile applications have also been developed to teach indigenous languages, which have proven useful in communities with limited connectivity [24]. In the legal domain, tools have been designed to help people access legal information through the training of models that summarize court rulings in various languages [43]. Therefore, adapting NLP models not only improves access to educational and legal domains but also extends to the healthcare sector [45].

7. Conclusions

This study outlines a method and dataset for creating a Spanish–Mixtec parallel corpus, addressing the challenges of digitizing bilingual materials for the low-resource Mixtec language. Using systematically online repositories and libraries, we collected texts from educational, legal, health, and religious fields. We digitized physical documents using manual scanning and semi-automatic OCR corrections, customized for Mixtec glyphs not supported by regular software.

The processed corpus, free of duplicates and non-parallel segments, serves as a reliable resource. Automatic translation experiments with the GPT, mBART-50, and M2M-100 models showed their potential to advance NLP technology.

Our work has some limitations. We did not address automatic text alignment between Spanish and Mixtec, and there is a disparity in sentence numbers across categories, with limited sentences in the health category. This affects the corpus's representativeness for technology development in this domain. We suggest expanding the corpus and working with Mixtec communities for translations. Finally, grammatical analysis of Mixtec was outside the scope of our project, and our corpus focuses only on Mixtec variants from Guerrero and Oaxaca.

This Spanish–Mixtec parallel dataset constitutes an essential resource for future research, facilitating the creation of tokenizers, morphological analyzers, machine translation systems, speech recognition, and other essential language technology tools that can significantly benefit the Mixtec community. We anticipate that publicly releasing this dataset will promote collaborative research efforts and practical applications, contributing meaningfully to the digital preservation and revitalization of the Mixtec language. In the future, we plan to expand the corpus by incorporating additional linguistic variants and various application domains while refining the automatic processing methods specifically adapted to the distinct typological characteristics of Mixtec.

Author Contributions: Conceptualization, H.S.-B., D.-M.C.-E. and N.-A.C.-S.; methodology H.S.-B. and D.-M.C.-E.; software, H.S.-B.; validation, H.S.-B., D.-M.C.-E. and N.-A.C.-S.; formal analysis, H.S.-B., D.-M.C.-E. and N.-A.C.-S.; investigation, H.S.-B., D.-M.C.-E. and N.-A.C.-S.; resources H.S.-B., D.-M.C.-E., N.-A.C.-S., J.-A.R.-G., T.G.-R. and J.T; writing—original draft preparation, H.S.-B., D.-M.C.-E., and N.-A.C.-S.; writing—review and editing, H.S.-B., D.-M.C.-E., N.-A.C.-S., J.-A.R.-G., T.G.-R., J.T. and J.-M.A.-A; visualization, H.S.-B., D.-M.C.-E., N.-A.C.-S., J.-A.R.-G., T.G.-R., J.T. and J.M.Á.-A.; supervision, D.-M.C.-E. and N.-A.C.-S.; project administration, H.S.-B., D.-M.C.-E.,

N.-A.C.-S., J.-A.R.-G., T.G.-R., J.T. and J.-M.A.-A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in FigShare at https://doi.org/10.6084/m9.figshare.28557182.v1.

Acknowledgments: We thank the students and academics from the Autonomous University of Querétaro (UAQ), the National Research Center (CENIDET), and the Center for Research in Applied Science and Advanced Technology (CICATA) who participated in generating the dataset for this article. Additionally, we acknowledge the use of two AI tools: Grammarly Assistant (Version 1.0, Grammarly Inc., San Francisco, CA, USA) to improve the grammar, clarity, and overall readability of the manuscript and GPT-4o (OpenAI, San Francisco, CA, USA) to assist with the wording and proofreading of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tonja, A.L.; Maldonado-sifuentes, C.; Mendoza Castillo, D.A.; Kolesnikova, O.; Castro-Sánchez, N.; Sidorov, G.; Gelbukh, A. Parallel Corpus for Indigenous Language Translation: Spanish-Mazatec and Spanish-Mixtec. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*; Mager, M., Ebrahimi, A., Oncevay, A., Rice, E., Rijhwani, S., Palmer, A., Kann, K., Eds.; Association for Computational Linguistics: Toronto, ON, Canada, 2023; pp. 94–102. [CrossRef]
2. Montaño, C.; Sierra Martínez, G.; Bel-Enguix, G.; Gomez, H. A Parallel Corpus Mixtec-Spanish. In *Proceedings of the 2019 Workshop on Widening NLP*; Axelrod, A., Yang, D., Cunha, R., Shaikh, S., Waseem, Z., Eds.; Association for Computational Linguistics: Florence, Italy, 2019; pp. 157–159.
3. Khan, M.; Ullah, K.; Alharbi, Y.; Alferaidi, A.; Alharbi, T.S.; Yadav, K.; Alsharabi, N.; Ahmad, A. Understanding the Research Challenges in Low-Resource Language and Linking Bilingual News Articles in Multilingual News Archive. *Appl. Sci.* **2023**, *13*, 8566. [CrossRef]
4. Magueresse, A.; Carles, V.; Heetderks, E. Low-resource Languages: A Review of Past Work and Future Challenges. *arXiv* **2020**, arXiv:2006.07264.
5. Cieri, C.; Maxwell, M.; Strassel, S.; Tracey, J. Selection Criteria for Low Resource Language Programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*; Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., et al., Eds.; European Language Resources Association (ELRA): Portorož, Slovenia, 2016; pp. 4543–4549.
6. INALI. Catálogo de las Lenguas Indígenas Nacionales. 2008. Available online: https://site.inali.gob.mx/pdf/catalogo_lenguas_indigenas.pdf (accessed on 4 September 2024).
7. INEGI. Hablantes de Lengua indíGena. 2020. Available online: https://beta.cuentame.inegi.org.mx/descubre/poblacion/hablantes_de_lengua_indigena/ (accessed on 4 September 2024).
8. Ventayol-Boada, A.; Cano, J.; Martínez, C.H.; Campbell, E.W. Digital free-to-use technologies for language maintenance in California's Central Coast Ñuu Savi (Mixtec) diaspora. *Living Lang.* **2024**, *3*, 18–52.
9. Gutierrez-Vasques, X.; Pugh, R.; Mijangos, V.; Barriga Martínez, D.; Aguilar, P.; Segura, M.; Innes, P.; Santillan, J.; Montaño, C.; Tyers, F. Py-Elotl: A Python NLP package for the languages of Mexico. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*; Mager, M., Ebrahimi, A., Pugh, R., Rijhwani, S., Von Der Wense, K., Chiruzzo, L., Coto-Solano, R., Oncevay, A., Eds.; Association for Computational Linguistics: Albuquerque, NM, USA, 2025; pp. 38–47.
10. Meque, A.G.M.; Angel, J.; Sidorov, G.; Gelbukh, A. Traducción automática entre lenguas indıgenas de México y el espanol. *Res. Comput. Sci.* **2023**, *152* , 329–337.
11. Shi, J.; Amith, J.D.; Chang, X.; Dalmia, S.; Yan, B.; Watanabe, S. Highland Puebla Nahuatl Speech Translation Corpus for Endangered Language Documentation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*; Mager, M., Oncevay, A., Rios, A., Ruiz, I.V.M., Palmer, A., Neubig, G., Kann, K., Eds.; Association for Computational Linguistics: Mexico City, Mexico, **2021**; pp. 53–63. [CrossRef]
12. Masua, B.; Masasi, N. In the heart of Swahili: An exploration of data collection methods and corpus curation for natural language processing. *Data Brief* **2024**, *55*, 110751. [CrossRef]

13. Carolina, E.; Cerbón, V.; Gutierrez-Vasques, X. Compilation of a parallel electronic corpus for a minority language: The case of Spanish–Nahuatl. In *Proceedings of the Primer Congreso Internacional El Patrimonio Cultural y las Nuevas Tecnologías*; Ciudad de México, Mexico, 5–30 August 2014; pp. 157–159.
14. Chen, X.; Ge, S. The Construction of English-Chinese Parallel Corpus of Medical Works Based on Self-Coded Python Programs. *Procedia Eng.* **2011**, *24*, 598–603. [CrossRef]
15. Allaberdiev, B.; Matlatipov, G.; Kuriyozov, E.; Rakhmonov, Z. Parallel texts dataset for Uzbek-Kazakh machine translation. *Data Brief* **2024**, *53*, 110194. [CrossRef]
16. Li, L.; Tayir, T.; Han, Y.; Tao, X.; Velásquez, J.D. Multimodality information fusion for automated machine translation. *Inf. Fusion* **2023**, *91*, 352–363. [CrossRef]
17. Ignat, O.; Maillard, J.; Chaudhary, V.; Guzmán, F. OCR Improves Machine Translation for Low-Resource Languages. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022*; Muresan, S., Nakov, P., Villavicencio, A., Eds.; Association for Computational Linguistics: Dublin, Ireland, 2022; pp. 1164–1174 [CrossRef]
18. Li, K.; Batjargal, B.; Maeda, A. A Prototypical Network-Based Approach for Low-Resource Font Typeface Feature Extraction and Utilization. *Data* **2021**, *6*, 134. [CrossRef]
19. Huang, Z.; Yu, P.; Allan, J. Improving Cross-lingual Information Retrieval on Low-Resource Languages via Optimal Transport Distillation. In Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, WSDM'23, Singapore, 27 February–3 March 2023; pp. 1048–1056. [CrossRef]
20. Chen, Y.; Yang, X.; Zhang, H.; Zhang, W.; Qu, D.; Chen, C. Meta adversarial learning improves low-resource speech recognition. *Comput. Speech Lang.* **2024**, *84*, 101576. [CrossRef]
21. Carnaz, G.; Antunes, M.; Nogueira, V.B. An Annotated Corpus of Crime-Related Portuguese Documents for NLP and Machine Learning Processing. *Data* **2021**, *6*, 71. [CrossRef]
22. Nilaphruek, P.; Charoenporn, P. Knowledge Discovery and Dataset for the Improvement of Digital Literacy Skills in Undergraduate Students. *Data* **2023**, *8*, 121. [CrossRef]
23. Wolfer, S.; Koplenig, A.; Kupietz, M.; Müller-Spitzer, C. Introducing DeReKoGram: A Novel Frequency Dataset with Lemma and Part-of-Speech Information for German. *Data* **2023**, *8*, 170. [CrossRef]
24. Santiago, C.A.A.; Zúñiga, H.A.G. Tecnologías del lenguaje aplicadas al procesamiento de lenguas indígenas en México: Una visión general. *Lingüíst. Lit.* **2023**, *44*, 79–102. . [CrossRef]
25. Mager, M.; Gutierrez-Vasques, X.; Sierra, G.; Meza, I. Challenges of language technologies for the indigenous languages of the Americas. *arXiv* **2018**, arXiv:1806.04291.
26. INALI. *Catálogo de las Lenguas indíGenas Nacionales: Variantes lingüíSticas de México con sus Autodenominaciones y Referencias geoestadíSticas*; Instituto Nacional de Lenguas Indígenas (INALI): Ciudad de Mexico, Mexico, 2009.
27. Romero, M.; Gómez-Canaval, S.; Torre, I.G. Automatic Speech Recognition Advancements for Indigenous Languages of the Americas. *Appl. Sci.* **2024**, *14*, 6497. [CrossRef]
28. Santiago-Benito, H.; Córdova-Esparza, D.M.; Castro-Sánchez, N.A.; García-Ramirez, T.; Romero-González, J.A.; Terven, J. Automatic Translation between Mixtec to Spanish Languages Using Neural Networks. *Appl. Sci.* **2024**, *14*, 2958. [CrossRef]
29. Arivazhagan, N.; Bapna, A.; Firat, O.; Lepikhin, D.; Johnson, M.; Krikun, M.; Chen, M.X.; Cao, Y.; Foster, G.; Cherry, C.; et al. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv* **2019**, arXiv:1907.05019.
30. Feng, F.; Yang, Y.; Cer, D.; Arivazhagan, N.; Wang, W. Language-agnostic BERT sentence embedding. *arXiv* **2020**, arXiv:2007.01852.
31. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*; Isabelle, P., Charniak, E., Lin, D., Eds.; Association for Computational Linguistics: Philadelphia, PA, USA, 2002; pp. 311–318. [CrossRef]
32. Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, Cambridge, MA, USA, 8–12 August 2006; pp. 223–231.
33. Sierra Martínez, G.; Montaño, C.; Bel-Enguix, G.; Córdova, D.; Mota Montoya, M. CPLM, a Parallel Corpus for Mexican Languages: Development and Interface. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*; Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., et al., Eds.; European Language Resources Association: Marseille, France, 2020; pp. 2947–2952.
34. Gaustad, T.; McKellar, C.A.; Puttkammer, M.J. Dataset for Siswati: Parallel textual data for English and Siswati and monolingual textual data for Siswati. *Data Brief* **2024**, *54*, 110325. [CrossRef]
35. Chiruzzo, L.; Amarilla, P.; Ríos, A.; Giménez Lugo, G. Development of a Guarani-Spanish Parallel Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*; Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., et al., Eds.; European Language Resources Association: Marseille, France, 2020; pp. 2629–2633.

36. Oliver, A.; Alvarez-Vidal, S.; Stemle, E.; Chiocchetti, E. Training an NMT system for legal texts of a low-resource language variety South Tyrolean German-Italian. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*; Scarton, C., Prescott, C., Bayliss, C., Oakley, C., Wright, J., Wrigley, S., Song, X., Gow-Smith, E., Bawden, R., Sánchez-Cartagena, V.M., et al., Eds.; European Association for Machine Translation (EAMT): Sheffield, UK, 2024; pp. 573–579.
37. Zhu, E.; Huang, Y.; Xian, Y.; Zhu, J.; Gao, M.; Yu, Z. Enhancing distant low-resource neural machine translation with semantic pivot. *Alex. Eng. J.* **2025**, *116*, 633–643. [CrossRef]
38. Zhang, Y.; Qu, X.; Liu, S.; Pang, Y.; Shang, X. Multiscale Weisfeiler-Leman Directed Graph Neural Networks for Prerequisite-Link Prediction. *IEEE Trans. Knowl. Data Eng.* **2025**, *37*, 3556–3569. [CrossRef]
39. Joshi, A.; Dabre, R.; Kanojia, D.; Li, Z.; Zhan, H.; Haffari, G.; Dippold, D. Natural Language Processing for Dialects of a Language: A Survey. *ACM Comput. Surv.* **2025**, *57*, 1–37. [CrossRef]
40. Sharma, U.; Bhattacharyya, P. Hi-GEC: Hindi Grammar Error Correction in Low Resource Scenario. In *Proceedings of the 31st International Conference on Computational Linguistics*; Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S., Eds.; Association for Computational Linguistics: Abu Dhabi, United Arab Emirates, 2025; pp. 6063–6075.
41. Ranathunga, S.; Lee, E.S.A.; Prifti Skenduli, M.; Shekhar, R.; Alam, M.; Kaur, R. Neural Machine Translation for Low-resource Languages: A Survey. *ACM Comput. Surv.* **2023**, *55*, 1–37. [CrossRef]
42. Moro, G.; Piscaglia, N.; Ragazzi, L.; Italiani, P. Multi-language transfer learning for low-resource legal case summarization. *Artif. Intell. Law* **2024**, *32*, 1111–1139. [CrossRef]
43. Ghosh, S.; Evuru, C.K.; Kumar, S.; Ramaneswaran, S.; Sakshi, S.; Tyagi, U.; Manocha, D. Dale: Generative data augmentation for low-resource legal nlp. *arXiv* **2023**, arXiv:2310.15799.
44. Hadgu, A.T.; Aregawi, A.; Beaudoin, A. Lesan–machine translation for low resource languages. In Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track, Online, 6–14 December 2021; pp. 297–301.
45. Bui, N.; Nguyen, G.; Nguyen, N.; Vo, B.; Vo, L.; Huynh, T.; Tang, A.; Tran, V.N.; Huynh, T.; Nguyen, H.Q.; et al. Fine-tuning large language models for improved health communication in low-resource languages. *Comput. Methods Programs Biomed.* **2025**, *263*, 108655. [CrossRef]