QomL'aqtaqa: A Qom–Spanish Parallel Corpus for Natural Language Processing with Machine Translation Evaluation

Viviana Cotik[1,2], **Aleksei Korablev**[3], **Paola Cúneo**[4,5] , **Pablo Laciana**[2]

[1] Universidad de Buenos Aires. FCEN. Departamento de Computación. Argentina.
[2] CONICET-UBA. Instituto de Ciencias de la Computación (ICC). Buenos Aires, Argentina.
[3] Minerva University
[4] CONICET, Argentina
[5]Universidad de Buenos Aires. FFyL. Instituto de Lingüística, Argentina

vcotik@dc.uba.ar

Abstract

Qom, a language of the Guaycuruan family, is a low-resource language for NLP and speech processing. We present the first parallel Qom–Spanish corpus in a computationally usable format, comprising 33,392 parallel segments, totaling 1,469,905 Qom tokens and 891,344 Spanish tokens. A subset of 2,943 segments excludes Bible-derived content. It includes alignments at different levels: sentences, sentence fragments, and paragraphs, and is compiled from multiple sources, both previously available and newly collected. We also present bidirectional neural machine translation baselines based on NLLB-200, achieving competitive performance in both translation directions on the full dataset, and lower performance on the non-Bible subset. An ablation study shows that training exclusively on biblical data reduces performance on non-biblical text, highlighting the importance of domain diversity in low-resource machine translation.

1 Introduction

Qom (also known as Toba) is an under-resourced Indigenous language of the Guaycuruan family spoken by one of the Indigenous peoples of Argentina, who number over 80,000 and mainly inhabit the Gran Chaco region, although migration to urban centers has diversified their sociolinguistic situation. While Qom remains vital in rural areas, language shift toward Spanish is increasing in urban contexts, particularly in the Buenos Aires metropolitan area.

In this context, the development of parallel corpora for machine translation is relevant not only for advancing low-resource NLP research but also for supporting language revitalization efforts, as computational resources can contribute to documentation, visibility and intergenerational transmission of ancestral languages.

However, to the best of our knowledge, parallel Qom–Spanish corpora suitable for machine translation are not yet available.

As a traditionally oral language with relatively recent written forms, Qom lacks a universally adopted orthographic standard, resulting in variation across communities that introduces additional challenges for computational processing. Furthermore, its morphological complexity and syntactic divergence from Spanish make it a particularly challenging and valuable case study for machine translation research.

Our key contributions are the following:

- the creation of Qom l'aqtaqa,[1] a parallel Qom–Spanish corpus by converting existing resources, originally available only as PDFs, into a machine-readable format and aligning additional Qom and Spanish texts that were previously unpaired,
- bidirectional machine translation baselines for Qom–Spanish and Spanish–Qom.

The corpus is accompanied by a comprehensive Data Statement that provides detailed documentation of its composition, creation process, and limitations (Bender and Friedman, 2018; Gebru et al., 2021). The resulting dataset is ready for computational use. Its public release is planned for a future version, which will incorporate additional data collected through ongoing recording, transcription, and translation efforts. The code will be available on GitHub upon publication.

The rest of the paper is organized as follows: Section 2 provides linguistic background on Qom. Section 3 reviews prior work on data and low-resource machine translation for Latin American Indigenous

[1] *Qom l'aqtaqa* (lit. "the people's words") is often translated as "the Qom language" but its semantic scope extends beyond a strictly linguistic or grammatical notion. More broadly, speakers understand *qom l'aqtaqa* as encompassing Qom ways of speaking, including not only referential meaning but also the intentional and agentive force attributed to words, and to play a central role in valued practices such as oratory and healing (Messineo, 2014).

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 186–202

languages, with a focus on Qom. Section 4 presents the construction and characteristics of the Qom–Spanish parallel corpus. Section 5 describes the machine translation baseline and reports experimental results. Section 6 concludes the paper and outlines future work. Additional details are provided in Appendices A.1 (Qom phonological system) , A.2 (Bible corpus coverage), A.3 (source-specific processing decisions), A.4 (alignment quality), A.5 (lexical analysis), A.6 (translation examples), and A.7 (Data Statement).

2 The Qom Language

The Qom language[2] presents several typological features that are relevant for machine translation. It is a morphologically rich language with polysynthetic and agglutinative tendencies, such that single words may encode information corresponding to an entire clause in English or Spanish. Morphemes are relatively segmentable and tend to be formally invariant, as illustrated in example (1):[3]

Example (1)

N'axaỹaxangui
n-ʔaGay-aGan-gi
3M-listen-CAUS-DIR

añi	*lquiyaqte*	*so*	*ỹaxaiquiolec*
a-ɲi	l-kizaqte	so	yaGaiki-ole-k
F-D:trid	3POSS-heart	D:DIST	old.man-DIM-MASC

'(He/she) was listening to the little old man's heartbeat.'

Both nominal and verbal morphology are highly complex: nouns inflect for number and gender and combine with deictic classifiers encoding positional and spatial distinctions, while possessed nouns distinguish between alienable and inalienable possession through different morphological marking strategies; for instance, in (1), the prefix *l-* marks a third-person possessor in *lquiyaqte* 'heart', yielding a possessive construction equivalent to that 'little old man's heartbeat'. In the same example, *ñi* (tridimensional) occurs with 'heart', while *so* (distal) modifies 'little old man' in a past narrative context. The verbal morphology is particularly elaborate, including three distinct sets of person prefixes conditioned by semantic roles and participant affectedness (active, middle, and inactive paradigms). For instance, in (1) the prefix *n-* indexes a middle person paradigm, marking an affected agent (i.e., the participant who listens). Verbs also encode aspectual distinctions and may take suffixes expressing direction, position, reflexivity, and reciprocity, and causation, as seen in (1) (*-aGan* 'CAUS', *-gi* 'DIR'). In addition, Qom lacks copular verbs. Basic constituent order differs from Spanish and English: transitive clauses generally follow Subject–Verb–Object order, whereas intransitive clauses tend toward Verb–Subject, and certain pronominal objects precede the verb.

These typological characteristics pose well-known challenges for machine translation. Differences in constituent order require substantial structural reordering during translation, while rich morphology necessitates modeling below the word level, typically through subword segmentation approaches (Jurafsky and Martin, 2025). Furthermore, the high degree of morphological synthesis complicates alignment between Qom lexical units and those of less synthetic target languages.

As a traditionally oral language, Qom has only relatively recently developed written forms, and no single orthographic standard has been universally adopted across communities. As a result, variation exists not only across dialects but also across domains of literacy use, including schools, churches, health services, and community organizations, each of which may promote distinct writing conventions. Consequently, multiple orthographic variants coexist in contemporary written Qom, reflecting different community-level agreements regarding the representation of particular phonemes. This variation affects several phonological segments. For instance, the palatal glide may be represented by multiple equivalent graphemes, including ỹ, ŷ, ý, ȳ, and ÿ. Other frequent alternations include d/r (e.g., *doqshe/roqshe* 'non-Indigenous person'), h/j (e.g., *hec/jec* 's/he goes'), and e/i (e.g., *nache/nachi* 'then'). Special mention should be made of the apostrophe ('), which represents the glottal stop, a phoneme absent in Spanish and inconsistently represented in writing (e.g., *do'onataxan/d'onataxan*).

A complete overview of the Qom phonological system and its orthographic variants is provided in Table 5 in Appendix A.1.

[2]ISO-639-3: tob. Glottocode: toba1269 https://glottolog.org/resource/languoid/id/toba1269)

[3] The first line gives the Qom text, the second its phonological representation, the third the morphological glosses, and the final line the English translation. Abbreviations in the glosses conform to the Leipzig Glossing Rules: 3 (third person), CAUS (causative), D (deictic classifier), DIM (diminutive), DIR (directional), DIST (distal), F (feminine), M (middle person marker), MASC (masculine), POSS (possessive), TRID (tridimensional).

3 Related Work

While efforts to develop computational resources for Latin American Indigenous languages have increased in recent years (Huamán-Águila et al., 2024; Ortiz Coronel et al., 2024; Tonja et al., 2024), these languages are still widely considered underrepresented. These efforts include initiatives such as the AmericasNLP shared tasks on machine translation (MT)[4] between Indigenous languages and Spanish, organized from 2021 to 2025, that have progressively expanded to 14 languages in recent editions and cover both translation directions. However, most MT benchmarks and corpora focus on a limited set of languages, many of which are primarily represented by varieties from countries such as Mexico, Paraguay, Bolivia, and Peru. In contrast, Indigenous languages spoken in Argentina remain largely underrepresented (Ticona et al., 2025)[5].

Several parallel resources for Qom–Spanish exist, covering a range of linguistic domains and text types (e.g. Messineo (2014); Martínez et al. (2013)). However, these materials are not available in formats readily usable for computational processing and require substantial effort for cleaning, alignment, or normalization.

To our knowledge, there are no parallel or text-based corpora available for machine translation in Qom that are directly suitable for NLP applications. Computational work on Qom is very limited, with prior studies restricted to spoken language identification (Garber and Riera, 2022) and a morphological description based on a linear context-free grammar (Porta, 2010). A small spontaneous speech dataset has been released through Mozilla Common Voice[6], but it is designed for Automatic Speech Recognition and is not suitable for MT. More broadly, computational resources for other Guaycuruan languages such as Pilagá and Mocoví are, to our knowledge, also absent.

Low-resource machine translation studies rely on relatively small parallel datasets, although sizes vary considerably across languages and tasks, ranging from a few hundred to over 100k sentence pairs, with a median of approximately 14.6k in AmericasNLP (de Gibert et al., 2025).

MT systems commonly evaluate performance using BLEU (Papineni et al., 2002) and ChrF++ (Popović, 2017). These metrics are widely adopted in low-resource settings, including shared tasks such as AmericasNLP (de Gibert et al., 2025), which also explore additional evaluation metrics. In extremely low-resource scenarios, Biblical or religious corpora are sometimes used as training data, although their impact and prevalence vary across languages and regions.

Recent work on low-resource machine translation increasingly relies on pretrained multilingual models such as NLLB-200, where joint multilingual fine-tuning across language pairs has been shown to outperform per-language approaches (Gow-Smith and Sánchez Villegas, 2023; de Gibert et al., 2025; NLLB Team et al., 2022).

Prior work suggests that transfer learning benefits from typological similarity between languages, supporting the use of related languages as proxies for low-resource settings (Moreno et al., 2024).

4 Corpus Collection and Curation

The corpus used in this work consists of a collection of bilingual Qom–Spanish texts spanning multiple discourse genres, including oral narratives, educational materials, and literary and religious translations. The dataset combines pre-existing parallel resources with texts that were aligned specifically for this study. For the former, we converted already aligned bilingual materials into a machine-readable format. For the latter, we performed the alignment of previously unpaired Qom–Spanish texts.

In this section, we present an overview of the corpus (4.1), describe the data collection and processing steps (4.2), the quality assessment of the alignments (4.3), and finally detail the data filtering and deduplication steps (4.4) and corpus statistics (4.5). A full Data Statement documenting the Qom–Spanish parallel corpus is provided in Appendix A.7.

4.1 Overview of the Corpus

We first present the pre-existing parallel resources, and then those aligned in the present study.

Pre-existing parallel resources

Arte verbal Qom: consejos, rogativas y relatos de El Espinillo (Chaco)[7] (Messineo, 2014) (here-

[4] AmericasNLP: https://americasnlp.org/, last accessed Apr. 2026.

[5] Although some languages spoken in Argentina—such as Quechua, Aymara, and Guaraní—are included in initiatives such as AmericasNLP, available datasets for these and other languages spoken in the country, such as Mapudungun, generally correspond to varieties from other countries and do not reflect those used within Argentina.

[6] https://mozilladatacollective.com/datasets/cmn1ppks200u7o1078xxw5izl

[7] [Qom Verbal Art: Advice, Prayers, and Narratives from

after, ***Arte Verbal***) is a compilation of 85 bilingual Qom–Spanish texts comprising advice, ritual speech, and traditional narratives. It represents a valuable source of formalized oral discourse, exhibiting significant pragmatic and stylistic diversity. The texts were originally produced orally in Qom, recorded, and later transcribed and translated into Spanish through collaborative fieldwork. The Spanish translations are subordinate to the Qom source and aim to remain closely aligned in structure and meaning.

Educación Sanitaria Intercultural: Manual de promoción de la salud entre los tobas (qom) del Chaco Central – Comunidades Tobas del Río Bermejito, Chaco (Argentina)[8] (Martínez et al., 2013) (hereafter, ***Manual de Salud***) is a bilingual manual for health promotion among the Qom communities of the Central Chaco. This resource provides expository health content and domain-specific terminology in both Qom and Spanish.

Las aventuras de Copaic, el gato montés.[9] (Haddad, 2022) (hereafter, ***Copaic***) is a short illustrated bilingual story intended for children. It provides examples of simple narrative structures and relatively controlled syntax.

Materiales del Taller de Lengua y Cultura Toba[10] (Messineo and Dell'Arciprete, 2005) (hereafter, ***Taller Derqui***) is a collection of materials for language and cultural outreach, including 10 texts, 6 songs, and Article 75 of the Argentine National Constitution. The materials were developed over four years of community-based workshops held in an urban Indigenous neighborhood, with the aim of strengthening the Qom language and fostering its transmission and visibility within the community.

Resources aligned in this work

La ***Declaración Universal de los Derechos Humanos***[11] (hereafter, ***UDHR***) (OHCHR, 1948a,b) is an internationally recognized legal instrument available in bilingual Qom–Spanish format, containing 4,037 words in Qom and 2,085 words in Spanish.

El Principito[12] is the short novel (96 pages) by Saint-Exupéry, which was manually aligned primarily at the paragraph level, with some alignments at the sentence level between the Spanish version (Saint-Exupéry, 1943) and its Qom translation (Saint-Exupéry, 2005).

La Biblia[13] (hereafter, ***The Bible***) is a collection of texts from both the Old and New Testament, available in bilingual Qom–Spanish format. The Qom version (Sociedad Bíblica Argentina, 2013) and Spanish version (Sociedades Bíblicas Unidas, 1992) were aligned through a custom-developed program at the verse level, where each verse typically corresponds to at least one full sentence, but may sometimes extend to a paragraph.

The corpus preserves the orthographic variation present in the original sources, as it reflects naturally occurring written materials produced by speakers from different communities and sociolinguistic backgrounds rather than a normalized standard (see Section 2). It comprises heterogeneous textual units, including sentence fragments, full sentences, and paragraphs, which may co-occur within the same text. This variation in granularity reflects the original parallelization of the materials. For further details, see Table 1.

4.2 Data Creation and Processing

Both the PDF-aligned text and the version aligned for this work were produced by two authors with a background in computer science, who are Spanish speakers (one native and one non-native) and do not speak Qom. The final alignments were subsequently reviewed by a linguist, a native Spanish speaker and Qom specialist, with over twenty years of experience in the documentation and linguistic description of this Indigenous language (see Section 4.3). The entire process was overseen by a PhD in computer science with experience in NLP, corpus creation, and annotation, who is a native Spanish speaker and does not speak Qom. For further details, see Appendix A.7.

The four originally bilingual sources were available only as PDFs, so the first step was to convert them into a machine-readable tabular format. Extraction strategies varied depending on PDF quality and document structure, combining automated methods (including large language model (LLM)-assisted text reconstruction (e.g., GPT-4) in cases

El Espinillo (Chaco)]

[8] *Paxaguenaxac da qantela'a da chalataxac yalexa't da nataxac. Lma' na qom tala Bermejito* [*Intercultural Health Education: A Health Promotion Manual among the Toba (Qom) of the Central Chaco – Toba Communities of the Bermejito River, Chaco (Argentina)*]

[9] *Lmitaxamaxac so copaic* [*The Adventures of the Wildcat*]

[10] *Lo'onatacpi na qom Derqui l'ecpi* [*Qom Language and Culture Workshop Materials (Derqui)*]

[11] *Na nqataxacpi na ỹotta'a't shiỹaxauapi mayi netalec ana 'alhua* [*Universal Declaration of Human Rights*] (United Nations)

[12] *So shiỹaxauolec nta'a* [*The Little Prince*]

[13] *La'aqtaqa Ñim Lo'onatac 'Enauacna* [*The Bible*]

of corrupted encodings) with manual correction where necessary.

Across sources, we filtered out non-parallel or document-structural elements, such as tables without complete sentences, captions, speaker labels, cross-references, and other metadata not corresponding to translational content. Bracketed content was also excluded, as it typically marks neologisms without direct Spanish counterparts.

Texts were then aligned at the paragraph, sentence, or verse level, depending on the structure of each source. Alignment relied on punctuation cues and document structure, and was complemented by manual inspection.

While some sources required additional preprocessing—such as reconstruction from corrupted encodings, exclusion of non-bilingual sections, or normalization of orthographic inconsistencies—others presented relatively clean parallel structures. Detailed, source-specific processing decisions are provided in Appendix A.3.

For texts obtained from web sources (e.g., *UDHR*), Qom and Spanish versions were aligned manually from the outset. In the case of *El principito*, minor discrepancies in layout required paragraph-level restructuring prior to sentence alignment.

We used the Qom version of *the Bible* Sociedad Bíblica Argentina (2013), available at Bible.com (YouVersion)[14], under the code LÑLE13 (*La'aqtaqa Ñim Lo'onatac 'Enauacna*, 2013). The platform allows parallel visualization at the verse level, although the provenance of the Spanish source text is not specified. Among the more than 25 Spanish versions available, we selected *Dios Habla Hoy, Standard Edition (DHHS94)* (Sociedades Bíblicas Unidas, 1992), as it most closely matches the Qom text in style and content. There is a clear correspondence between the two languages at the book, chapter, and verse levels,[15] which allowed us to systematically extract each verse from every chapter and book and store them in parallel format in a CSV file for subsequent analysis. The corpus presents a high level of alignment, but we identified differences in completeness between the two versions, that are discussed in Appendix A.2.

4.3 Alignment Quality Assessment

A full revision of the corpus was conducted by a native Spanish-speaking linguist with expertise in Qom, comparing the original Qom texts with the aligned CSV version. This process included correcting orthographic inconsistencies, errors introduced during PDF-to-CSV conversion (e.g., extra spaces, character confusions), and occasional errors in the source texts. Potentially problematic segments identified during the parallelization stage were also reviewed.

Spelling was standardized across the dataset to better reflect the original forms. Ambiguous cases identified during preprocessing, including misalignments, were carefully checked against both the Qom and Spanish originals. For more details, refer to Appendix A.4.

4.4 Data Preprocessing

After extraction and alignment, all sources underwent a common automated filtering pipeline. Filtering proceeded in three steps: (i) exact duplicate pairs, matched on both the Qom and Spanish sides within the same source document, were removed; (ii) pairs with an empty side were discarded; and (iii) pairs were retained only if the token-length ratio $(|\text{ES}|+1)/(|\text{QOM}|+1)$ fell within $[0.15,\ 6.0]$, ensuring that neither side is more than roughly six times longer than the other, and if both sides contained at least two tokens, to remove extreme length mismatches likely to reflect alignment errors. Apostrophe-like characters were also normalized to U+0027 across all sources, since the glottal stop phoneme is represented by multiple visually similar Unicode codepoints. The segment counts in Table 1 reflect the corpus after these steps.

4.5 Corpus Statistics

We define three corpus variants: *QomL'aqtaqa-Base* (all sources except UDHR[16] and the Bible), *QomL'aqtaqa-Bible*, and *QomL'aqtaqa-Base+Bible*. We will sometimes use *QomL* as an abbreviation of *QomL'aqtaqa*. Table 1 shows the statistics of our corpus.

To account for variability in textual granularity, we report dataset size using segment-based statistics. These segments include full sentences, sentence fragments, and paragraphs. The reported

[14] Bible.com (YouVersion) is a free digital platform to read, listen to, and study the Bible. It offers multiple translations and audio versions, among others: https://www.bible.com/bible.

[15] See https://www.bible.com/bible/574/GEN.1.DHHS94?parallel=128

[16] This material was not included, as it became available after processing had begun.

statistics include the total number of segments, total token counts for each language, average segment length, and the standard deviation of segment lengths. The high standard deviation in some sources reflects heterogeneous segmentation granularity within those documents.

Token counts are based on orthographic segmentation (i.e., units separated by whitespace and punctuation).[17]

5 Machine Translation Baseline

We fine-tuned a pretrained neural machine translation model based on NLLB-200 on the Qom–Spanish parallel corpus. This system serves as a baseline for future work.

Since NLLB-200 does not include Qom (ISO 639-3: `tob`), we used the Guaraní tag (`grn_Latn`) as a proxy, following expert validation in Chacoan typology, which identified it as the most suitable available option. This choice enables a warm-start initialization from a typologically related language, allowing the model to adapt pretrained representations during fine-tuning (Moreno et al., 2024).

5.1 Model and Training Setup

We fine-tuned `facebook/nllb-200-distilled-600M` (NLLB Team et al., 2022), a 600M-parameter distilled variant of the NLLB-200 model covering 200 languages. For fine-tuning, we used the Hugging Face `Seq2SeqTrainer` with the Adafactor optimizer (learning rate 5×10^{-4}), an effective batch size of 16 (2 per device with 8 gradient accumulation steps), fp16 precision, and 10 epochs. Experiments were run on a single NVIDIA T4 GPU (16 GB) via Kaggle (12-hour session limit).

5.2 Corpus Configurations and Data Splits

We experimented with two corpus configurations: **Base** (*QomL'aqtaqa-Base*) and **Base+Bible** (*QomL'aqtaqa-Base+Bible*). The latter incorporates Biblical data, making the corpus roughly 11 times larger (from 2,943[18] to 33,392 pairs). This mirrors a common practice in low-resource machine translation. For instance, Chiruzzo et al. (2022) augmented their Guaraní–Spanish system with Bible data and reported consistent improvements, even though the text differs substantially from contemporary usage.

For each configuration, we applied two split strategies. The **random** split assigns pairs to train/development/test sets uniformly at random (approximately 66/17/17% for *QomL-Base*, and 80/10/10% for *QomL-Base+Bible*). The **stratified** split assigns pairs by source document, ensuring that each document contributes proportionally to every partition and thus maintaining a balanced representation of sources across splits; this results in an approximate 86/7/7% split for *QomL-Base* and 80/10/10% for *QomL-Base+Bible*. The smaller test set in the stratified *QomL-Base* configuration reflects the limited corpus size while retaining a representative evaluation sample. In both cases, the splits are disjoint: no sentence pair appears in more than one partition. Table 2 shows the resulting sizes of the dataset splits.

5.3 Results and Discussion

We report ChrF++ as the primary metric and BLEU as a secondary reference for comparability with prior work, following the evaluation protocol of the AmericasNLP shared tasks (de Gibert et al., 2025), with both metrics computed using SacreBLEU (Post, 2018). Results are shown in Table 3.

QomL-Base+Bible yields substantially higher scores. This is likely due to the fact that the Bible constitutes approximately 91% of its training data and is highly represented in the test sets under both split strategies. Therefore, these scores primarily reflect Bible-register performance rather than general Qom–Spanish translation quality. Given that the test sets of both configurations differ in size and composition—particularly since *QomL-Base+Bible* is dominated by biblical text—absolute scores are not directly comparable across configurations and should therefore be interpreted cautiously.

Under *QomL-Base*, stratified split scores are uniformly higher than random split scores across both directions (e.g., ChrF++ 30.89 vs. 28.81 for ES→QOM). This is the opposite of what is sometimes expected and requires explanation. Under the random split, the test set is dominated by *Arte Verbal* sentences (the largest *QomL-Base* source), which also dominate training; the model is thus largely evaluated on in-distribution examples. The stratified split ensures that each source document contributes proportionally to every partition, including *Taller Derqui* and *Manual de Salud*, which

[17] While Spanish shows moderate morphological complexity, Qom displays a higher degree of synthesis, with single tokens often encoding information that corresponds to multiple words in Spanish. Accordingly, "tokens" are treated here as practical units for corpus comparison rather than as strict morpholexical units.

[18] 2,882 without UDHR.

Title	Reference	Segs	Seg. unit	Tokens		Avg seg. len. (tokens)	
				Qom	ES	Qom (avg ± std)	ES (avg ± std)
Arte verbal Qom	Messineo (2014)	1,831	fragment	16,242	15,009	8.87 ± 5.44	8.20 ± 4.67
Educación Sanitaria Intercultural	Martínez et al. (2013)	212	paragraph; sentence; fragment	7,438	6,343	35.08 ± 35.18	29.92 ± 30.58
Materiales del Taller de Lengua y Cultura Toba	Messineo and Dell'Arciprete (2005)	273	sentence; fragment	2,299	1,946	8.42 ± 4.97	7.13 ± 3.42
Las Aventuras de Copaic	Haddad (2022)	16	paragraph; sentence; fragment	560	484	35.00 ± 26.95	30.25 ± 20.95
El Principito	Saint-Exupéry (1943, 2005)	550	paragraph; sentence; fragment	16,817	15,793	30.58 ± 25.98	28.71 ± 26.44
La Declaración Universal de los Derechos Humanos	OHCHR (1948b)	61	paragraph; sentence	4,037	2,085	66.18 ± 43.44	34.18 ± 22.47
Biblia†	Sociedad Bíblica Argentina (2013); Sociedades Bíblicas Unidas (1992)	30,449	paragraph; sentence	1,422,512	849,684	46.72 ± 21.14	27.91 ± 12.10
Subtotal QomL-Base		*2,943*		*47,393*	*41,660*		
Total QomL-Base+Bible		*33,392*		*1,469,905*	*891,344*		

Table 1: *QomL'aqtaqa* parallel corpus sources and statistics. † Source added in the extended (*QomL-Base+Bible*) configuration. *Segs* counts parallel segment pairs; each segment may correspond to a fragment, a full sentence, or a paragraph, depending on the granularity of the source text. *Fragments* are sub-sentential units reflecting the prosodic segmentation of the original texts, where a single utterance may span multiple lines. Token counts are based on orthographic segmentation. Average segment length and standard deviation are reported in tokens.

Config	Split	Train	Dev	Test
QomL-Base	Random	1,912	485	485
	Stratified	2,488	197	197
QomL-Base+Bible	Random	26,802	3,256	3,273
	Stratified	26,582	3,335	3,414

Table 2: Dataset split sizes for each corpus configuration.

Config	Direction	Split	BLEU	ChrF++
QomL-Base	ES→QOM	Random	4.42	28.81
		Stratified	4.61	30.89
	QOM→ES	Random	4.02	23.05
		Stratified	5.50	23.88
QomL-Base+Bible	ES→QOM	Random	23.55	53.97
		Stratified	23.37	53.45
	QOM→ES	Random	24.73	46.42
		Stratified	22.80	45.02

Table 3: MT results for both corpus configurations and split strategies.

exhibit distinct vocabulary, register, and sentence structure. The slightly higher stratified scores likely reflect the contribution of *El Principito* to the test set under stratification, whose narrative register aligns well with *Arte Verbal* training examples. The QOM→ES direction shows a smaller gap (23.05 vs. 23.88 ChrF++) because generating Spanish benefits from the model's strong Spanish decoder regardless of the Qom source domain.

Although our scores are not directly comparable to those reported in the AmericasNLP 2025 shared task for low-resource Indigenous language MT (de Gibert et al., 2025) due to differences in languages, test sets, and, in our case, a Bible-dominated evaluation set, we provide the shared-task results for contextualization. The best-performing reference systems for each translation direction achieved 47.81 ChrF++ for Indigenous-to-Spanish translation and 36.76 for Spanish-to-Indigenous translation, while our models obtain 46.42 and 53.97 ChrF++, respectively.

5.4 Ablation: Data Composition vs Domain Specificity

The improvements observed in *QomL-Base+Bible* are likely driven by the substantially larger amount of training data. However, its strong performance may be influenced by the high proportion of Bible data (91% of the training data), raising questions about generalization beyond this specific domain.

To investigate the effect of training data composition, we train a **Bible-only** (with *QomL-Bible*) model and compare it against the *QomL-Base* model. Both models are evaluated on a shared held-out test set: the 158-pair *QomL-Base* strati-

fied test set. This provides a controlled setting to compare a model trained exclusively on a single domain (Bible) against a model trained on more diverse non-Bible sources under the same evaluation conditions.

Table 4 reports ChrF++ scores for both models on this test set.

Model	Direction	ChrF++
QomL-Base	ES→QOM	38.25
QomL-Bible	ES→QOM	33.78
QomL-Base	QOM→ES	30.88
QomL-Bible	QOM→ES	20.31

Table 4: Ablation comparing a Bible-only model (*QomL-Bible*) and a model trained on non-Bible data (*QomL-Base*), evaluated on the same 158-pair *QomL-Base* stratified test set.

The *QomL-Bible* model performs consistently worse than the *QomL-Base* model when evaluated on non-Bible content. The larger degradation in the QOM→ES direction (−10.57 ChrF++ points) suggests that the Spanish decoder overfits to a narrow biblical register, which does not transfer well to general Qom–Spanish translation. The smaller gap in the ES→QOM direction (−4.47 ChrF++ points) indicates that lexical knowledge from the Bible may still be partially useful for Qom generation, likely due to shared morphological patterns across registers.

A stratified set of 28 translations per direction from the test set of *QomL-Base*, produced by the *QomL-Base+Bible* model, was evaluated by a linguist with expertise in Qom. Results show 82% correct for Qom→Spanish, and for Spanish→Qom, 89% correct, 7% incorrect, and the remainder uncertain. Evaluation focused on preserving semantic and syntactic integrity with respect to the source, assessing model outputs independently of the references, which were sometimes less accurate. A brief descriptive analysis of translation errors is presented in Appendix A.6.

5.4.1 Lexical Analysis of Generated Output

To characterize the register difference qualitatively, we analyzed the most frequent content words (excluding Spanish stopwords) in the Spanish outputs generated by each model on the *QomL-Base* test set. See Appendix A.5 for word clouds and top-10 content words for each model.

A clear contrast is observed. The *QomL-Base* model produces domain-appropriate vocabulary: *paredes* (walls)[19], *indicador* (indicator), *parásito* (parasite), *árboles* (trees), *agua* (water), and *tobas* reflect the health and educational sources in the corpus. The Bible-only model's output is markedly sparser—its top content word appears only 10 times, compared to 93 for the *QomL-Base* model—and its vocabulary (*comen* (eat), *puede* (can), *lugar* (place), and *lengua* (language)). This confirms that training exclusively on the Bible leads to a model that generates narrow, domain-inappropriate Spanish for general Qom–Spanish translation.

6 Conclusion

We have presented *QomL'aqtaqa*, the first parallel Qom–Spanish corpus in a computationally usable format, along with bidirectional MT baselines fine-tuned from pretrained NLLB-200 models. The corpus combines seven sources spanning oral narratives, educational materials, and literary and religious translations, totaling 33,392 aligned segment pairs in its full configuration. Our ablation confirms that the Bible, while a valuable source of parallel data, introduces a strong register bias: a model trained exclusively on biblical text shows a drop of 10.57 ChrF++ points for QOM→ES and 4.47 for ES→QOM when evaluated on non-biblical content, and generates lexically narrow Spanish dominated by nature and cosmological vocabulary. Overall, the results indicate that the proposed approach is effective for bidirectional Qom–Spanish machine translation.

Several directions remain for future work. We plan to compare the use of the Guaraní proxy code (`grn_Latn`) with adding a new `tob_Latn` embedding, evaluate *QomL-Base+Bible* models on *QomL-Base* test sets to better assess out-of-domain generalization, and explore a two-stage training strategy with multilingual fine-tuning followed by Qom–Spanish adaptation. Future work should also revisit evaluation metrics for Qom given its polysynthetic morphology, and incorporate additional resources such as the *Vocabulario Toba* (Buckwalter and Litwiller de Buckwalter, 2013). Finally, we will expand the corpus with ongoing data collection efforts and develop a lightweight web interface for interactive translation.

[19]The high frequency of *paredes* (walls) reflects its domain-specific use in Chagas prevention materials, where walls are a key site for detecting and controlling vector infestation (e.g., cracks, stains, and hiding places of triatomine bugs).

Limitations

Several limitations should be considered. First, segment heterogeneity (including full sentences, fragments, and paragraphs) may affect both training and evaluation of machine translation systems; however, prior work suggests that sub-sentential fragments can still provide useful translational signal in low-resource settings (Steingrímsson et al., 2023). Second, due to resource constraints, alignment quality was not comprehensively assessed across the entire corpus. Third, the evaluation relies on automatic metrics such as BLEU and chrF++, which may not fully capture translation quality in low-resource, morphologically rich languages. Fourth, we did not evaluate the Bible-only model on a held-out Bible test set, limiting direct comparison between in-domain and out-of-domain performance; we leave this controlled comparison to future work.

Regarding data construction, the splits are group-disjoint at the discourse-unit level, meaning that train, development, and test partitions are defined over complete discourse units (i.e., paragraphs, sentence fragments, or sentences), rather than individual sentence pairs. All sentences belonging to the same discourse unit are assigned to the same partition, ensuring that no unit is split across different subsets and that no exact duplicate pairs appear across partitions. However, substring overlap across units was not explicitly checked, although it is structurally unlikely. In addition, results for *QomL-Base+Bible* are heavily influenced by Bible data (approximately 91% of the training set), and thus primarily reflect performance on this specific register rather than general Qom–Spanish translation quality, and should be interpreted accordingly. Finally, the use of Guaraní (`grn_Latn`) as a proxy language tag for Qom constitutes a practical workaround in the absence of a dedicated code, but introduces a confound in the learned representations. Future work should address this limitation by introducing a dedicated `tob_Latn` embedding.

Ethical considerations

This work builds on prior efforts to collect and parallelize Qom–Spanish data, in which native speakers and community members have been actively involved. In line with established guidelines for working with Indigenous language communities (Bird, 2020), we maintain communication with collaborators and speakers regarding the present work and aim to respect community perspectives and cultural context. Although the primary source materials are mostly publicly available, the processed resources reported in this study are not yet released. We plan to make them available in formats that are accessible and useful to the community in future work.

Acknowledgments

This work was partially supported by the Lacuna Fund Natural Language Processing 2024 grant "Corpus Lengua y Cultura Qom" (Grantee 109).

References

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6:587–604.

Steven Bird. 2020. Decolonising speech and language technology. In Proceedings of the 28th International Conference on Computational Linguistics, pages 3504–3519.

Alberto S. Buckwalter and Lois Litwiller de Buckwalter. 2013. Vocabulario Toba–Castellano y Castellano–Toba. Equipo Menonita. `https://chacoindigena.net/materiales/`. Accessed: 2026-04-07.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2022. Jojajovai: A parallel guaraní–Spanish corpus for MT evaluated with humans and automatic metrics. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 2098–2107. European Language Resources Association.

Ona de Gibert, Robert Pugh, Ali Marashian, Raúl Vázquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan C. Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno Veliz, Ángel Lino Campos, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 134–152. Association for Computational Linguistics.

Leandro Martín Garber and Pablo Ernesto Riera. 2022. Sistema de identificación de idioma (LID) para grabaciones de entornos naturales bilingües en comunidades qom. Ph.D. thesis, Master thesis, Universidad de Buenos Aires.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. Communications of the ACM, 64(12):86–92.

Edward Gow-Smith and Danae Sánchez Villegas. 2023. Sheffield's submission to the AmericasNLP shared task on machine translation into indigenous languages. In Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 192–199. Association for Computational Linguistics.

María del Rosario Haddad. 2022. Las aventuras de Copaic, el gato montés: ¿Dónde está mi música? Number 1 in Cuentos originarios ilustrados. María del Rosario Haddad and Instituto de Investigación en Etnomusicología, Ciudad Autónoma de Buenos Aires. Illustrated by Carlus Rodríguez. Translated to Qom by Amada Farías et al. Bilingual edition: Spanish–Qom.

Óscar Huamán-Águila, C. E. Fernández-García, and C. R. Gonzales García. 2024. Uso de la transcripción fonética para lograr que avatares de inteligencia artificial pronuncien discurso en lengua quechua: caso illariy. Lengua y Sociedad, 23(2):833–851.

Daniel Jurafsky and James H. Martin. 2025. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript draft released August 24, 2025.

Gustavo J. Martínez, Paola Cúneo, and Mauricio Maidana. 2013. Educación Sanitaria Intercultural. Manual de promoción de la salud entre los tobas (qom) del Chaco Central - Comunidades Tobas del Río Bermejito, Chaco (Argentina). Paxaguenaxac da qantela'a da chalataxac yalexa't da nataxac. Lma' na qom tala Bermejito, Chaco (Argentina). Museo de Antropología, Universidad Nacional de Córdoba, Córdoba, Argentina. Edición bilingüe Qom–Español.

Cristina Messineo. 2003. Lengua toba (guaycurú): aspectos gramaticales y discursivos. Number 48 in LINCOM Studies in Native American Linguistics. LINCOM Europa, Munich, Germany.

Cristina Messineo. 2014. Arte verbal Qom: consejos, rogativas y relatos de El Espinillo (Chaco). Textos y comentarios de Mauricio Maidana. Asociación Civil Rumbo Sur / Ethnographica, Buenos Aires, Argentina. Contains 85 bilingual Qom–Spanish texts compiled with contributions from Mauricio Maidana.

Cristina Messineo and Ana Dell'Arciprete. 2005. Lo'onatacpi na qom Derquil'ecpi: materiales del taller de lengua y cultura toba, 1 edition. Comunidad Toba Daviaxaiqui, Buenos Aires.

Oscar Moreno, Yanua Atamain, and Arturo Oncevay. 2024. Awajun-OP: Multi-domain dataset for Spanish–Awajun machine translation. In Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024), pages 112–120, Mexico City, Mexico. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

OHCHR. 1948a. La Declaración Universal de los Derechos Humanos. Proclamada por la Asamblea General de las Naciones Unidas, Resolución 217 A (III), 10 de diciembre de 1948. Accessed: 2026-04-07.

OHCHR. 1948b. Na nqataxacpi na Ỹotta'a't shiỹaxauapi mayi netalec ana 'alhua [Universal Declaration of Human Rights in Toba/Qom]. Toba (Qom) translation of the 1948 declaration; translation date not recorded by OHCHR. Accessed: 2026-04-07.

D. Ortiz Coronel, R. Lacerda de Sá, L. Trigo, and J. R. Pichel Campos. 2024. Proyecto guaná ia: innovación en la enseñanza de la lengua en riesgo. Lengua y Sociedad, 23(2):945–961.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In Proceedings of the Second Conference on Machine Translation, pages 612–618. Association for Computational Linguistics.

Andrés Osvaldo Porta. 2010. The use of formal language models in the typology of the morphology of amerindian languages. In Proceedings of the ACL 2010 Student Research Workshop, pages 109–114.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191. Association for Computational Linguistics.

Antoine de Saint-Exupéry. 1943. El Principito. Reynal & Hitchcock; Gallimard, Estados Unidos / Francia. Publicada originalmente en francés; existen muchas traducciones posteriores, incluyendo al español y al qom.

Antoine de Saint-Exupéry. 2005. So Shiyaxauolec Nta'a. AEAC Editores, Argentina. Traducción de El Principito al idioma Qom (Toba).

Sociedad Bíblica Argentina. 2013. La'aqtaqa Ñim Lo'onatac 'Enauacna: Qom (Toba) Bible. Accessed: 2026-04-07.

Sociedades Bíblicas Unidas. 1992. Dios Habla Hoy: Versión Española. Spanish Bible (DHH, Versión Española). Accessed: 2026-04-07.

Steinþór Steingrímsson, Pintu Lohar, Hrafn Loftsson, and Andy Way. 2023. Do not discard – extracting useful fragments from low-quality parallel data to improve machine translation. In Proceedings of the Second Workshop on Corpus Generation and Corpus Augmentation for Machine Translation, pages 1–13, Macau SAR, China.

Belu Ticona, Fernando Martín Carranza, and Viviana Cotik. 2025. Indigenous languages spoken in argentina: a survey of nlp and speech resources. In Proceedings of the 31st International Conference on Computational Linguistics, pages 8449–8461.

Atnafu Lambebo Tonja, Fazlourrahman Balouchzahi, Sabur Butt, Olga Kolesnikova, Hector Ceballos, Alexander Gelbukh, and Thamar Solorio. 2024. NLP Progress in Indigenous Latin American Languages. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 6972–6987.

A Appendix

A.1 Qom phonological system

Below, we present the Qom phonological inventory, based on Messineo (2003). Table 5 lists the consonant phonemes along with their corresponding graphemic representations, while Table 6 shows the vowel phonemes.

A.2 Details on Coverage and Missing Data in the Parallel Bible Corpus

The corpus comprises 35,173 aligned verses from the LNLE13 (Qom) and DHHS94 (Spanish) versions, with high correspondence at the book, chapter, and verse levels. However, completeness differs: LNLE13 reaches 87.9% coverage (4,247 missing verses), mainly due to the absence of the seven deuterocanonical books (Sirach, 1–2 Maccabees, Wisdom, Judith, Tobit, and Baruch), as well as additional partial gaps in several books (e.g., Matthew, Mark, Psalms, Romans, and Luke) and 139 entirely empty chapters. In contrast, DHHS94 is nearly complete (99.0% coverage, 346 missing verses), with only minor, distributed gaps and two empty chapters (Psalms 42–43).

A.3 Source-Specific Processing Details

This appendix provides detailed, source-level descriptions of the extraction, cleaning, and alignment procedures summarized in Section 4.2.

Arte Verbal. This source presented significant challenges due to the use of a legacy font encoding, which resulted in systematically corrupted text under standard extraction methods. To address this, the document was processed in segments of up to six pages. Non-parallel introductory and commentary sections were removed prior to processing. Each segment was then reconstructed using a large language model (GPT-4), producing a machine-readable version of the text. Two page ranges could not be recovered in usable form and were excluded (pp. 84–86 and pp. 95–96 following section Ro.15). Additional post-processing included deduplication and removal of bracketed content marking neologisms without Spanish equivalents.

Taller Derqui. The PDF quality of this source was very low, making automated extraction unreliable. As a result, the full text was manually reviewed and corrected. The document contains word lists presented as grammatical examples; these were excluded, and only the parallel narrative texts were retained.

Manual de Salud. This source required selective extraction due to its heterogeneous structure. Chapters 1 and 5 contained no bilingual content and were excluded. Within the remaining chapters, the following elements were removed: tables lacking complete sentences, figure captions, cross-references to figures, and speaker-role indicators (e.g., *Lta'a (Padre)*, *Qa'ñole (Jovencita)*), as these correspond to document structure rather than translational content. Bracketed content was also excluded, as it marks Qom neologisms without direct Spanish counterparts.

Copaic. This source required minimal processing. The text is short, the layout is clean, and the parallel structure is explicit. No major structural filtering was necessary beyond normalization of apostrophes.

UDHR. The Qom and Spanish versions were obtained from web sources and aligned manually at the article level. Due to the clean and consistent structure of the texts, no additional preprocessing was required.

El Principito. Both the Qom and Spanish versions were processed. Although the placement of illustrations is largely consistent across versions,

Consonants	Labial	Alveolar	Palatal	Velar	Uvular	Laryngeal / Glottal
Plosive	/p/ *p*	/t/ /d/ *t* *d, r*	/tʃ/ *ch*	/k/ /g/ *c, qu* *g, gu*	/q/ /G/ *q* *x*	/ʔ/ *'*
Fricative		/s/ *s*	/ʃ/ /ʒ/ *sh* *y*			/h/ *j*
Nasal	/m/ *m*	/n/ *n*	/ɲ/ *ñ*			
Lateral		/l/ *l*	/ʎ/ *ll*			
Glide	/w/ *hu, u, v*		/y/ *ỹ*			

Table 5: Qom consonant inventory with corresponding grapheme representations, adapted from Messineo (2003).

Vowels	Front	Central	Back
Close	*i*		*o*
Mid	*e*		
Open		*a*	

Table 6: Qom vowel inventory, adapted from Messineo (2003).

minor discrepancies were observed. The text was first reorganized to achieve paragraph-level alignment, guided by illustration placement and manual inspection. Subsequently, sentence- or paragraph-level alignment was performed using punctuation cues, such as full stops and colons.

A.4 Details on Alignment Quality

As mentioned in the body of the paper, the alignment was reviewed by a native Spanish-speaking linguist with extensive expertise working on Qom. While a subset of potentially problematic segments had been identified during the parallelization stage, the revision was performed over the entire corpus.

The review process involved a systematic comparison between the original Qom texts and the aligned CSV version. Several types of issues were identified and corrected:

- **Errors introduced during format conversion (PDF to CSV)**, including:
  - extra or missing whitespace,
  - character confusions (e.g., i instead of l),
  - incorrect handling of special characters representing the glottal stop (phoneme /ʔ/), written with an apostrophe-like symbol (') (see Table 6), and
  - segmentation inconsistencies.
- **Orthographic inconsistencies in the original texts**, which were revised to improve internal consistency while preserving the original forms as much as possible.
- **Occasional errors in the source material**, which were corrected when clearly identifiable.
- **Alignment issues**, such as shifted or mismatched segments, which required cross-checking both the original and processed Qom and Spanish versions.

In addition, previously flagged ambiguous cases were carefully re-examined in context. For *Taller Derqui*, the low PDF resolution meant that the two orthographic variants $\tilde{y}$ and $\bar{y}$, which are equivalent in Qom (see Table 6), could not always be distinguished visually; all instances were unified to $\tilde{y}$ to ensure internal consistency.

The Bible corpus was not reviewed in its entirety; instead, attention was restricted to previously identified doubtful segments.

A.5 Lexical Analysis

Figure 1 and Table 7 present word clouds and the top-10 content words from Spanish outputs generated by the *QomL-Base* and *QomL-Bible* models, excluding stopwords. Analysis is provided in Section 5.4.1.

A.6 Translation Examples

A brief overview of selected translation examples (see Table 8) from both directions of the *QomL-Base+Bible* corpus follows, illustrating the types

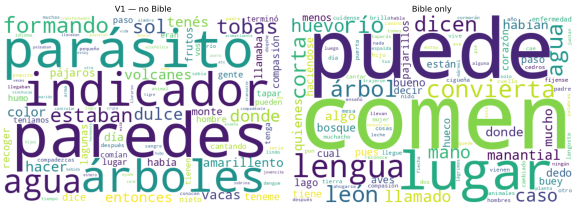

Figure 1: Word clouds of content words in Spanish outputs on the *QomL-Base* test set. Left: *QomL-Base* model (no Bible). Right: *QomL-Bible* model. Word size reflects frequency. For *QomL-Base* the most frequent words are: *paredes* (walls), *indicador* (indicator), *parásito* (parasite), *árboles* (trees), *agua* (water), and textittobas (Tobas). For *QomL-Bible* the most frequent words are: *comen* (eat), *puede* (can), *lugar* (place), *lengua* (language).

QomL-Base		***QomL-Bible***	
Word (EN)	**Count**	**Word (EN)**	**Count**
paredes (walls)	93	comen (eat)	10
parásito (parasite)	66	puede (can)	10
indicador (indicator)	52	lugar (place)	8
árboles (trees)	48	lengua (tongue)	8
agua (water)	46	convierta (turn)	8
formando (forming)	35	león (lion)	7
tobas (Tobas)	30	árbol (tree)	7
estaban (were)	27	río (river)	7
sol (sun)	26	agua (water)	7
donde (where)	20	dicen (say)	6

Table 7: Top-10 content words (stopwords removed) in Spanish outputs generated by each model on the *QomL-Base* stratified test set. English translations are provided for readability only. The *QomL-Base* model produces domain-specific vocabulary from health and educational texts; the *QomL-Bible* model generates sparse output dominated by lower-frequency biblical vocabulary.

of errors identified in the evaluation. The model produces recognizable Qom morphological structure in some cases (example 4, ES→QOM, where the output is verbatim correct), but also exhibits characteristic low-resource errors grounded in specific generated translations.

The Qom→Spanish outputs reveal two distinct types of errors. On the one hand, some cases can be attributed to relatively shallow **lexical substitutions or hallucination-like behavior** (see example 2), where the system selects semantically inappropriate equivalents despite otherwise well-formed structure—for instance, producing *nido de guazuncho y cuero* ('nest of deer and leather') in place of *gallineros y corrales* ('chicken coops and corrals').

On the other hand, more problematic cases involve a deeper **misanalysis of syntax and information structure** (see example 1). In particular, the model fails to correctly interpret discourse-fronted constituents such as *para los qom* ('for the Qom people'), reanalyzing them as canonical subjects (*esa gente toba*, 'those Toba people'). This results in translations that are not only pragmatically infelicitous but also syntactically distorted, suggesting that the system does not adequately capture the interaction between word order and discourse functions in Qom.

In the Spanish→Qom direction, errors tend to cluster around **morphosyntactic constraints** rather than lexical choice. For instance, in example 4, the system produces *Ndoteec da ayem 'auchoxoden*, where *ndoteec* ('only/just') appears in a position that is not grammatically licensed, even though its meaning is appropriate. This indicates partial knowledge of lexical semantics without full control over distributional restrictions. At the same time, there are encouraging signs at the morphological level: the model appears capable of identifying and generating derivational morphology such as diminutives, as in *-olec* (DIM) in *nsoqolec* ('young man', see example 3).

A.7 Data Statement

We present a data statement for *QomL'aqtaqa* corpus following Bender and Friedman (2018)'s data statements and incorporating elements from Gebru et al. (2021)'s Datasheets for datasets.

A. **Curation Rationale**

The primary criterion for constructing the corpus was the inclusion of texts that are publicly available, and accessible in both Qom and Spanish. The corpus includes the following materials:

(a) Arte verbal Qom: consejos, rogativas y relatos de El Espinillo (Chaco) [Qom Verbal Art: Advice, Prayers, and Narratives from El Espinillo (Chaco)] Messineo (2014).

(b) Educación Sanitaria Intercultural: Manual de promoción de la salud entre los tobas (qom) del Chaco Central. Comunidades Tobas del Río Bermejito, Chaco (Argentina) / *Paxaguenaxac da qantela'a da chalataxac yalexa't da nataxac. Lma' na qom tala Bermejito* [Intercultural Health Education: A Health Promotion Manual among the Toba (Qom) of the Central Chaco. Toba Communities of the Bermejito River, Chaco (Argentina)] Martínez et al. (2013).

(c) Las aventuras de Copaic, el gato montés /

	Corpus	Source	Reference	Hypothesis
QOM→ES				
1	*Arte Verbal*	qaq ne'ena qom nache ñi'ipiaqa' taqa'en yi'iyi	y para los qom ese era un lugar de marisca.	y esa gente toba es un lugar de marisca
2	*Manual de Salud*	Da qaỹo'ot ca lauo' ana 'olegaxa qataq nlo laỹi' qoỹoqta que'eca noỹic. Ỹaqto saishet da taigui ca noỹic aca chiincha qataq ca llalec huetaigui ca 'olegaxa lauo' qataq nlo laỹi'.	Poner los gallineros y corrales lo más lejos posible de la vivienda, para que no entren en la casa las vinchucas y crías que estén instaladas en estos lugares.	Haz un nido de guazuncho y cuero lejos de la vivienda, para que no entren en la vivienda las vinchucas y sus crías.
ES→QOM				
3	*Arte Verbal*	"Vos, jovencito, cuando vas al monte	am nsoq da 'anqo'ona ada aviaq	Am nsoqolec da huo'o ca taigui ada'ada aviaq
4	*Taller Derqui*	Solamente te pido que me tengas compasión	Chaqdata da aỹem auchoxonnec qome	Ndoteec da aỹem 'auchoxoden

Table 8: Selected translation examples from the *QomL-Base+Bible* non-stratified model. Examples 1–2 are QOM→ES; examples 3–4 are ES→QOM.

Lmitaxamaxac so copaic [The Adventures of the Wildcat] (Haddad, 2022).

(d) Materiales del Taller de Lengua y Cultura Toba / *Lo'onatacpi na qom Derqui l'ecpi* [Qom Language and Culture Workshop Materials (Derqui)] (Messineo and Dell'Arciprete, 2005).

(e) Declaración Universal de los Derechos Humanos / *Na nqataxacpi na ỹotta'a't shiỹaxauapi mayi netalec ana 'alhua* [Universal Declaration of Human Rights] (OHCHR, 1948a,b).

(f) El Principito / *So shiỹaxauolec nta'a* [The Little Prince] (Saint-Exupéry, 1943, 2005).

(g) La Biblia / *La'aqtaqa Ñim Lo'onatac 'Enauacna* [The Bible] (Sociedades Bíblicas Unidas, 1992; Sociedad Bíblica Argentina, 2013).

Some of these texts *(a-d)* were already aligned, but in PDF format and were converted into a machine-readable format as part of this work, other texts *(e-g)* were previously unpaired. In these cases the alignment constitutes a contribution of our team. The selection strategy combined two complementary types of texts: First, a subset of the corpus (notably *a, c* and *d*) is of particular relevance because these texts were originally produced in the Qom language. They represent culturally grounded genres and styles, reflecting Qom verbal art, worldview, and communicative practices. These materials provide valuable insight into endogenous linguistic structures and discourse patterns. Within this group, *(c) Copaic* is especially notable as it targets a child audience, thus contributing variation in register and genre. Second, the corpus also includes translations from Spanish into Qom, such as widely disseminated texts like *(e) UDHR*, *(f) El Principito*, and *(g) The Bible*. These texts are commonly available across multiple languages and contribute to cross-linguistic comparability. Importantly, they expand the lexical and thematic coverage of the corpus, incorporating domains beyond poetic expression, including educational, health-related, legal, and institutional discourse.

Overall, the curation rationale balances cultural representativeness (through original Qom texts) and domain diversity and comparability (through translated materials). This combination supports broader generalization potential for systems trained on the dataset, while maintaining a strong grounding in authentic Qom language use.

B. **Language Varieties**

Qom, also known as Toba. Guaycuruan family. Area: Gran Chaco, South America. Typological features: polysynthetic and agglutinative tendencies; rich morphology; nouns inflect for number and gender; possessed nouns distinguish between alienable and inalienable; deictic classifiers; verbal morphology includes three distinct sets of person prefixes and suffixes enconding aspectual distinctions, direction, position, reflexivity, and reciprocity; no

copular verb; no adpositions; basic word order is SVO and VS. ISO-639-3: tob. Glottocode: toba1269 (https://glottolog.org/resource/languoid/id/toba1269). Language tag: tob_Latn.

Spanish. Rioplatense Spanish as used in Argentina. Standard Indo-European Romance language with relatively analytic morphology and SVO word order. Used as the target/source language in all translation pairs. Language tag: es_Latn.

C. **Speaker Demographic**

We do not have complete sociolinguistic metadata for all contributors involved in the production of the Qom texts, as such information is not consistently documented in the original materials. Below we report the available information:

(a) Arte Verbal and *(b) Manual de Salud*: Produced by a native Qom speaker (male, 60+ years old) from the western region of Chaco (*dapigueml'ec* variety). The speaker has extensive experience in linguistic–anthropological work, as well as in teaching and translation related to Qom language and culture.

(c) Copaic: Produced collaboratively by a group of Qom speakers, including both men and women, aged 25-45. The contributors are native speakers with strong knowledge of their cultural traditions and oral narratives.

(d) Taller Derqui: Developed collaboratively by adult Qom speakers of diverse ages, genders, and regional origins, all residing in an Indigenous neighborhood in Buenos Aires. All Qom participants are native speakers. The process also involved researchers and students (primarily from linguistics and anthropology) affiliated with the University of Buenos Aires.

(f) El Principito: Translated collectively by multiple adult Qom speakers from the provinces of Chaco and Formosa. The work was carried out across different stages, workshops, and versions. All contributors are native speakers.

(e) UDHR and *(g) The Bible*: No speaker-level metadata available.

Integrated Sociolinguistic Overview:

Age: When available, contributors range from approximately 25 to 60+ years old.

Gender: Both male and female contributors are represented in the subset of texts with available metadata.

Race/ethnicity: Contributors include Qom (Indigenous) speakers and non-Indigenous collaborators.

Native language: All identified Qom contributors are native speakers of Qom; materials also involve non-native collaborators.

Socioeconomic status: All Qom contributors are characterized as low socioeconomic status (low SES).

Number of speakers: The corpus includes both single-author texts and collaboratively produced materials involving multiple contributors; however, the total number of distinct speakers cannot be fully determined due to incomplete records.

Disordered speech: There is no evidence or documentation of disordered speech (e.g., dysarthria) in any part of the corpus.

D. **Annotator Demographic**

Annotation and data processing (including PDF-to-CSV conversion) were carried out by two authors with a background in computer science. Both are Spanish speakers (one native, one non-native) and do not speak Qom. The final alignments and evaluation were reviewed by a trained linguist, a native Spanish speaker with in-depth knowledge of Qom, with over 25 years of experience working with Qom communities. The entire process was overseen by a PhD in computer science with experience in NLP, corpus creation, and annotation, who is a native Spanish speaker and does not speak Qom.

All contributors are adult, non-Indigenous researchers (two men and two women) with higher education.

We acknowledge that the exclusively non-Indigenous composition of the team may influence both the annotation process and the interpretation of the data, representing a potential limitation. At the same time, the long-term collaborative experience of the linguist with Qom communities provides sustained sociolinguistic expertise.

E. **Speech Situation**

We have partial information regarding the speech situation and production context of the materials included in the corpus. The available details are summarized below.

Time and place:

A subset of the materials —specifically *(a) Arte Verbal*, *(c) Copaic*, and *(d) Taller Derqui*—

were collected from the early 2000s onward. These texts originate from Qom-speaking communities primarily in the Chaco region, as well as from collaborative workshop settings in Buenos Aires.

For other materials, such as *(e) UDHR* and *(g) The Bible*, precise information about the time and place of translation into Qom is not available. The health material *(b) Manual de Salud* and *(f) El Principito* were produced/translated during the early 2000s in Buenos Aires, Chaco and Formosa.

Modality (spoken vs. written):

Materials *(a)*, *(c)*, and *(d)* originate in oral production in Qom, recorded in audio, and later transcribed and translated into Spanish through collaborative fieldwork, thus preserving features of naturally occurring speech while providing aligned bilingual material. The transcription is segmented into discourse lines based on prosodic units rather than strictly syntactic or semantic sentence boundaries, reflecting the organization of the original oral performance. Moreover, the Spanish version is fully subsidiary to the Qom source text, offering a free translation intended to remain closely aligned with the structure and meaning of the original rather than functioning as an independent adaptation.

In contrast, *(b)*, *(e)*, *(f)*, and *(g)* are written texts, either originally composed or translated in written form.

Scripted/edited vs. spontaneous:

The orally based materials *(a, c, d)* are relatively spontaneous, although later subject to transcription and little degree of editing.

The translated and institutional materials *(b, e, f, g)* are scripted and edited, as they derive from established written sources and underwent controlled translation processes.

Synchronous vs. asynchronous interaction:

The original oral productions *(a, c, d)* were generated in synchronous communicative contexts (e.g., storytelling, workshops), though the corpus itself represents their later, asynchronous textual form. The written and translated materials *(b, e, f, g)* are asynchronous, as they were produced without real-time interaction between participants.

Intended audience:

(a) Arte verbal: primarily community-internal audiences, with cultural and narrative functions, as well as the explicit aim of contributing to the transmission and dissemination of Qom language and culture.

(c) Copaic: oriented toward children, reflecting an educational and narrative purpose.

(d) Taller Derqui: mixed audience, including Qom community members and learners in educational contexts, and a role in the maintenance, transmission, and dissemination of Qom language and cultural practices.

(b) Manual de Salud: aimed at community health communication, targeting Qom-speaking populations.

(e) UDHR and *(g) The Bible:* broad and general audiences, as part of widely disseminated multilingual texts; in the case of the Bible, it has also historically been used as a tool for evangelization in Indigenous communities.

(f) El Principito: literary audience, including both educational and general readership.

Overall, the corpus reflects a combination of oral, culturally grounded productions and written, translated materials.

F. **Text Characteristics**

The corpus encompasses a broad and diverse range of genres and discourse types, which directly influence both lexical selection and structural patterns.

First, it includes genres that are intrinsic to Qom verbal tradition, particularly represented in *(a) Arte Verbal*, as well as in *(c) Copaic* and *(d) Taller Derqui*. These comprise narrative genres (e.g., traditional stories), as well as persuasive and ritual forms such as advice, exhortations, prayers, and chants. These texts reflect culturally specific communicative practices and exhibit distinctive stylistic and discourse features.

Second, the corpus incorporates health-related genres in *(b) Manual de Salud*, including descriptions of diseases and treatments, as well as simulated or reported medical interactions. These texts introduce domain-specific terminology and explanatory discourse structures.

Third, legal and institutional genres are represented in *(e) UDHR*, as well as in parts of *(d)*, which include the translation of an article of the Argentine Constitution. These materials are characterized by highly specialized vocabulary and formal, technical registers.

In addition, the corpus includes Western literary narrative genre, specifically the short fic-

tional work *(f) El Principito*, which reflects stylistic conventions of European literary tradition.

Finally, *(g) The Bible* represents biblical/religious genre, with its own highly conventionalized structures and specialized vocabulary.

Overall, this diversity of genres and topics results in a corpus that captures a wide spectrum of linguistic variation, from culturally embedded oral discourse to formal, technical, and literary registers. This heterogeneity should be taken into account when interpreting linguistic patterns and evaluating generalization capacity.

G. **Recording Quality**

N/A

H. **Other**

We have obtained permission to use the following resources for building a computationally usable corpus, to be released in the future, and for supporting the development of the translation system: *(a) Arte verbal*, *(b) Manual de Salud*, *(c) Las aventuras de Copaic*, *(d) Taller Derqui*, and *(f) El Principito*.