

Targum — A Multilingual New Testament Translation Corpus

Maciej Rapacz, Aleksander Smywiński-Pohl

AGH University of Kraków
{mrapacz, apohllo}@agh.edu.pl

Abstract

Many European languages possess rich biblical translation histories, yet existing corpora — in prioritizing linguistic breadth — often fail to capture this depth. To address this gap, we introduce a multilingual corpus of 651 New Testament translations, of which 334 are unique, spanning five languages with 2.4–5.0× more translations per language than any prior corpus: English (194 unique versions from 390 total), French (41 from 78), Italian (17 from 33), Polish (29 from 48), and Spanish (53 from 102). Aggregated from 12 online biblical libraries and one preexisting corpus, each translation is annotated with metadata that maps the text to a standardized identifier for the work, its specific edition, and its year of revision. This canonicalization allows researchers to define "uniqueness" for their own needs: they can perform micro-level analyses on translation families, such as the KJV lineage, or conduct macro-level studies by deduplicating closely related texts. By providing the first multilingual resource with sufficient depth per language for flexible, multilevel analysis, the corpus fills a gap in the quantitative study of translation history.

Keywords: Multilingual Parallel Corpus, Bible Translation, Translation Studies, Diachronic Analysis, Digital Humanities

1. Introduction

In the data-scarce era of early Natural Language Processing, the Bible was a widely used parallel corpus. Its utility for methods such as statistical machine translation (SMT) was based on the assumption that the verses between translations were semantically equivalent (Mueller et al., 2020). The current landscape of NLP, characterized by terabyte-scale datasets and large language models, redefines this utility. The Bible's volume is negligible for training modern foundation models, and its specific thematic and stylistic registers are not representative of general language use. This shift reframes the main application of the corpus: from being data *for* AI systems to being an object of study *by* AI systems, enabling large-scale quantitative analysis of textual choices that reflect its cultural and theological history. Far from a relic of the past, Bible translation remains an active enterprise: a substantial share of the translations in our corpus date from the 21st century, underscoring the continued demand for new renderings across confessional and stylistic lines.

We introduce *Targum*,¹ a corpus designed to prioritize depth over linguistic breadth. *Targum* provides coverage for five languages with extensive translation histories — English, French, Italian, Polish, and Spanish. This depth supports computational analysis of a wide spectrum of historical periods and confessional traditions, treating translation variants that heuristic methods previously discarded as a primary object of study.

¹<https://github.com/mrapacz/targum-corpus>

The corpus is named after the Targum (תרגום, "translation"), the tradition of ancient Aramaic Bible translations that combined literal rendering with dynamic expansion to serve their audience (Alexander, 1992; Schühlein, 1912). Our corpus is designed to span this same spectrum — from the most formal, literal renderings to the most dynamic paraphrases.

The main contributions of this paper are two-fold. First, we introduce *Targum*, a corpus of 651 New Testament translations (334 unique) across five European languages, providing 2.4–5.0× the depth of any prior resource per language. Second, we verify and normalize the metadata for each translation, including its publication history and edition identity. We first position our work with respect to existing corpora, then detail the methodology for corpus construction and metadata normalization, describe the resulting corpus and its distribution format, present a statistical analysis of the resource, and conclude by discussing new research directions.

2. Related Work

The Bible is a frequently used text for creating multilingual corpora due to its verse structure, which provides de facto parallel sentence-level alignment in a large number of languages (Mueller et al., 2020). Research using these corpora has followed two main paradigms. The most common has been *instrumental*, using the text as a convenient dataset to bootstrap NLP tools for low-resource languages, such as training machine translation systems (Guzmán et al., 2019) or creating part-of-speech taggers via annotation projection (Nicolai

Language	English	French	Italian	Polish	Spanish	Total
Corpus						
Christodouloupoulos and Steedman (2015)	2	1	1	1	1	6
Akerman et al. (2023)	24	4	2	2	5	37
Mayer and Cysouw (2014)	39	17	7	6	22	91
Targum	390	78	33	48	102	651
Targum (no duplicates)	194	41	17	29	53	334

Table 1: The number of translations for each language in the Bible translation corpora.

and Yarowsky, 2019; Nicolai et al., 2020). In contrast, a more recent paradigm, often aligned with the digital humanities, treats the biblical text as the *primary object of study*. This line of research applies computational methods to perform a diachronic analysis of language change or to study stylistic variation between translations. Our work is situated within this second paradigm.

A significant body of work has focused on creating biblical corpora that maximize linguistic breadth. Over time, a series of progressively larger collections were introduced, starting with foundational resources such as those from Christodouloupoulos and Steedman (2015) and Mayer and Cysouw (2014). This trend culminated in the Johns Hopkins University Bible Corpus, which expanded coverage to over 1,600 languages to support tasks such as typological analysis and low-resource NLP (McCarthy et al., 2020). The Parallel Bible Corpus (Mayer and Cysouw, 2014) has also served as a foundation for typological studies, such as the computational investigation of tense marking across over 1,000 languages (Asgari and Schütze, 2017). The state of these digital resources, however, is often dynamic; for instance, while some corpora like that of Mayer and Cysouw (2014) are actively maintained and have grown since their initial publication, the Johns Hopkins corpus is no longer publicly available. A more recent effort, the eBible Corpus, provides another standardized and accessible resource (Akerman et al., 2023). While these collections offer wide breadth, their depth for any single language remains limited. Table 1 quantifies this gap: against the largest prior resource for each language (Mayer and Cysouw, 2014), Targum contains 194 unique English versions (5.0× increase), 41 French (2.4×), 17 Italian (2.4×), 29 Polish (4.8×), and 53 Spanish (2.4×).

Although most work has focused on breadth, some projects have curated collections with depth for specific scholarly questions. A prime example is the EDGeS Diachronic Bible Corpus, a resource containing 36 translations across four Germanic languages designed to enable the “longitudinal and contrastive study of complex verb constructions” (Bouma et al., 2020). Similarly, Carlson et al. (2018) compiles 34 English versions for the spe-

cific task of evaluating the transfer of prose style. Our work extends this depth-focused approach to a multilingual setting. Although these prior corpora were built for predefined research questions, Targum is designed as a *general-purpose resource for comparative analysis*, letting researchers construct subcorpora tailored to specific inquiries across multiple languages and traditions.

3. Methodology

3.1. Data Acquisition and Processing

The construction of the corpus began with the aggregation of translations from a diverse set of sources, including 12 online biblical libraries and one preexisting digital corpus. The latter comprises translations published by eBible.org (Akerman et al., 2023), which we sourced from its public GitHub repository; in our source table (Table 2) these appear under the *ebible.org* site. We identified well-known, large-scale aggregators (e.g., Bible Gateway, Bible Hub) to ensure broad coverage, and supplemented these with language-specific repositories (e.g., *bibliepolskie.pl* for Polish, *laparola.net* for Italian) to achieve deep coverage in our target languages. The acquisition pipeline consists of three stages.

First, an *indexer* inventories all available translations on each source website, capturing metadata such as source-specific identifiers and translation names, producing a stable, timestamped snapshot of each site’s collection.

Second, a *scraper* uses this index to download the full text of each translation, caching the raw source files (e.g., HTML or JSON) for each of the New Testament’s 260 chapters without modification. Separating download from parsing allows iterative refinement of extraction logic offline, prevents repeated requests to source websites, and guarantees reproducibility.

Third, a *parser* processes the cached files to extract verse-level text. A key challenge at this stage is the structural variation in verse notation. While many translations align to a simple one-to-one mapping, more dynamic versions often introduce non-standard segments (e.g., 6a, 6b) or con-

Language Source	English	French	Italian	Polish	Spanish	Total
bible.audio	3	15	0	0	1	19
bible.com	75	19	7	8	28	137
bible.is	18	8	1	2	15	44
biblegateway.com	64	4	5	3	19	95
biblehub.com	46	0	0	0	0	46
bibles.org	72	6	3	3	14	98
biblestudytools.com	37	0	0	0	0	37
bibliepolskie.pl	0	0	0	20	0	20
crossbible.com	12	1	0	0	1	14
ebible.org	24	4	2	2	5	37
jw.org	1	1	1	1	1	5
laparola.net	9	3	11	0	2	25
obohu.cz	29	17	3	9	16	74
Total	390	78	33	48	102	651

Table 2: Number of translations sourced from each site for each language in the `Targum` corpus.

solidated ranges (e.g., 6–8). Our parser handles these cases. The final output is stored as a newline-delimited JSON file (JSONL) that preserves the detailed verse segmentation; the structure of these files is documented in Appendix B.

3.2. Canonicalization and Metadata Normalization

The scraped data is inconsistent across sources; the same translation often appears with different names and inconsistent metadata (e.g., publication year). We resolved this by mapping each scraped instance to a grouping identifier that unifies all instances of the same translation work. This metadata is stored in YAML files, where each translation is assigned a `canonical_id` and a `canonical_revision_year`. The `canonical_id` groups together all instances that represent the same translation work regardless of their source, but it does not model the full complexity of translation lineage. For example, all scraped copies of the King James Version share one `canonical_id`, while the KJV’s derivative translations (e.g., NKJV, RV, ASV) receive separate identifiers. We also record a `canonical_version` string that identifies the specific edition: while it often reduces to the year (e.g., "1769"), it sometimes encodes richer distinctions such as "2nd-edition", "1989-anglicised-catholic", or "1989-americanized". Two instances sharing the same `canonical_id` and `canonical_version` from different sources should be near-identical texts (the same edition scraped from multiple sites). Within each group, the `canonical_revision_year` is an integer that places the edition on the time axis. For both the copyright status and version fields we store

accompanying reasoning fields that document the evidence behind each decision. We acknowledge that some translation histories are highly complex, involving works that draw from multiple predecessors; we plan to address these with more granular metadata in future work.

Determining these values required extensive external research into publication histories, prefaces, and publisher catalogs. While we initially explored AI assistants for automated metadata extraction, they proved unreliable under manual review, often failing to distinguish between closely-related revisions or misidentifying publication details. We therefore performed the normalization manually. We note that this task is factual cataloguing — mapping a scraped text to its real-world publication identity — rather than a subjective linguistic judgment. Each case is resolved by tracing documentary evidence (title pages, publisher records, preface dates), and the reasoning is recorded alongside each annotation to support future reassessment. The cross-source validation described in the following section provides an additional safeguard: translations grouped under the same identifier are checked for textual consistency using similarity metrics.

3.3. Quality Control via Cross-Source Validation

The aggregation process naturally yields multiple copies of the same canonical text, particularly for widely-distributed versions. Instead of viewing these as duplicates to be discarded, we treat their presence as a central feature of our quality control methodology.

Once scraped translations are assigned the same `canonical_id` and `canoni-`

Language Copyright Status	English	French	Italian	Polish	Spanish	Total
Public Domain	151	42	13	13	18	237
Open License	40	2	0	12	11	65
Copyrighted	199	34	20	23	73	349
Total	390	78	33	48	102	651

Table 3: Copyright status distribution of translations in the Targum corpus by language.

cal_version, we conduct a systematic, side-by-side comparison of their texts using lexical similarity metrics. Such metrics are sensitive to the single-character variations that distinguish editions or reflect transcription errors. This comparison reliably spots discrepancies, which expose one of two issues: a subtle error in our parsing logic, or a pre-existing data error on the source website, such as a typo or a missing verse. To resolve these differences and identify the correct edition, we employ several verification techniques, including comparing the text against scanned historical documents and using search engines to find quotations of a specific divergent phrase, which can often be traced back to a particular publication. This iterative process allows us to correct our parsers and improve the textual accuracy of the entire collection.

4. The Targum Corpus

4.1. Sources and Scale

Table 2 details the provenance of the 651 translations collected by the pipeline described above. Coverage is distributed across both large-scale international aggregators and specialized, language-specific repositories. English is well-represented across numerous sites, with *bible.com*, *biblegateway.com*, and *bibles.org* as the largest sources. *bible.com* also serves as the single largest source for Spanish translations. For other languages, a combination of sources proved most effective; French, for instance, drew heavily from *bible.com*, *bible.audio*, and *obohu.cz*. In contrast, the collections for Polish and Italian are led by their respective national sites, *bibliepolskie.pl* and *laparola.net*. Achieving broad coverage required combining international aggregators with language-specific repositories.

4.2. Corpus Structure and Distribution Format

The corpus is not a single, pre-filtered dataset where duplicates have been removed. Instead, we deliver the complete archive of all collected translations, accompanied by a metadata index recording

the provenance and specific edition of each text.

This structure gives researchers full control rather than imposing a pre-defined notion of what constitutes a “duplicate.” The rich metadata allows users to define uniqueness on their own terms, tailored to their specific research question. For example, a scholar studying Polish translations might prioritize copies from the language-specific *bibliepolskie.pl* source. A researcher analyzing the evolution of the KJV family would treat each revision as a distinct entity. Conversely, a user performing a broad, cross-lingual study might deduplicate aggressively, keeping only one instance per `canonical_id`.

Each translation instance is described by: source provenance (`site`), language (`iso`), source-specific identifiers (`translation_id`, `translation_name`, `translation_abbr`), canonical identity (`canonical_id`, `canonical_version`, `canonical_year`), completeness statistics (`num_chapters`, `num-verses`, `num_words`), and `copyright_status` with documented reasoning. A complete field reference with descriptions and examples is provided in Appendix A; the full file hierarchy is documented in Appendix B.

In addition to the text files and metadata index, the corpus distributes several pre-computed resources: a *book coverage* file recording which of the 27 New Testament books are present in each translation; chapter- and verse-level vector embeddings using the Qwen3-Embedding model (Zhang et al., 2025) in two sizes (0.6B and 8B), selected for its multilingual performance and long-context capabilities; and pre-computed pairwise intra-lingual similarity scores for all translation pairs within each language, covering both semantic similarity (cosine distance on the 8B embeddings) and lexical similarity (Levenshtein), computed at the chapter level. These pre-computed resources support immediate downstream analysis without expensive recomputation, and allow research on copyrighted translations without distributing the source texts. Table 3 presents the copyright status distribution of translations across languages.

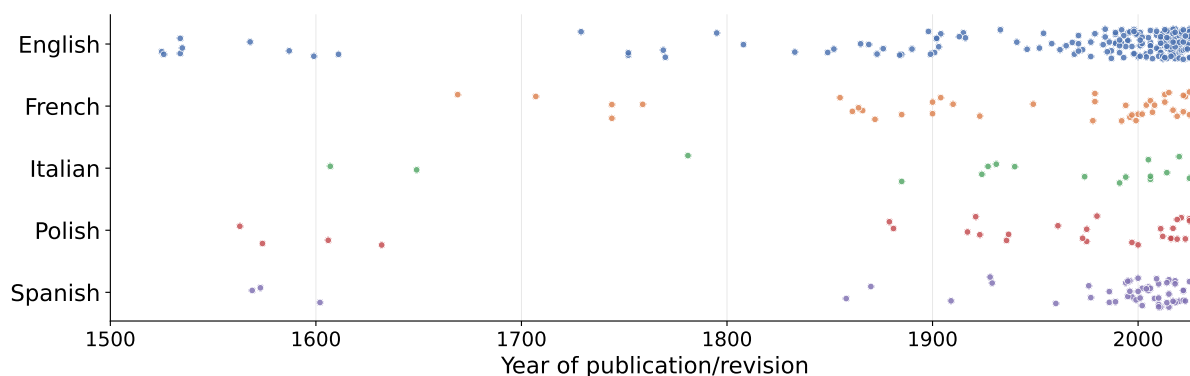


Figure 1: Diachronic distribution of translations per language. Each dot represents a distinct $(\text{canonical_id}, \text{canonical_version})$ pair.

5. Corpus Analysis

5.1. Diachronic Distribution

The depth of the corpus is not only numerical but also historical. Figure 1 illustrates the temporal distribution of the unique translations based on their revision year. Across all five languages, the distribution is skewed to the right, with a significant concentration of translations published in the last 50 years. The historical scope of the corpus begins in the 16th-century Reformation and Counter-Reformation period, with English showing the most continuous and extensive history from that point forward. The plot also reveals notable historical gaps. For Polish and Spanish, there is a long period of silence between a small cluster of early 17th-century translations and a renewed burst of activity in the second half of the 19th century. Similarly, while our Italian collection includes a few early 17th-century texts, this period remains scarce, and the earliest French translation in the corpus is from the latter half of the 17th century.

5.2. Book Length Distribution

To characterize the structural properties of the corpus, we examined the distribution of book lengths (in characters) across the 334 unique translations (deduplicated by edition). For each of the 27 New Testament books, we aggregated the chapter-level text for every unique translation and computed the total character count. Figure 2 presents these distributions as a ridgeline plot, with one density curve per language overlaid for each book. Matthew, Luke, John, and Acts are the longest books across all languages, while the shorter Epistles (e.g., Philemon, 2 John, 3 John, Jude) show tightly concentrated distributions. Notably, for every book the distributions of all five target languages are shifted to the right of the Greek source text (SBLGNT), indicated by the dashed vertical line, reflecting the

expansion that naturally occurs in translation from a highly synthetic source language. While the distributions appear broadly similar and overlap substantially across languages, visual inspection suggests that French translations tend to be somewhat longer and Polish translations somewhat shorter than the other groups. To test this, we performed pairwise Mann-Whitney U tests over all $\text{book} \times \text{translation}$ character counts with Bonferroni correction ($n = 10$ comparisons). The analysis confirms that French is significantly longer than English ($p_{\text{adj}} = 0.015$), Polish ($p_{\text{adj}} = 0.012$), and Spanish ($p_{\text{adj}} = 0.038$). The French–Italian difference does not survive correction ($p = 0.030$, $p_{\text{adj}} = 0.30$). All other pairwise comparisons are non-significant after correction. Because the pooled test treats every $\text{book} \times \text{translation}$ observation as independent, we also performed a secondary analysis on per-translation mean character counts (one observation per translation, N ranging from 17 to 194). This test, which removes within-translation book-to-book variance, yields 7 significant pairs: French is significantly longer than all four other languages (including Italian, $p_{\text{adj}} = 0.000$), and Polish is significantly shorter than English ($p_{\text{adj}} = 0.000$), French, Italian ($p_{\text{adj}} = 0.012$), and Spanish ($p_{\text{adj}} = 0.001$). English, Italian, and Spanish do not differ significantly from each other under either test. These differences likely reflect morphological and orthographic properties of each language (e.g., French tends to use more articles, prepositions, and accented characters; Polish uses fewer function words due to its rich case system) rather than translation philosophy.

5.3. Pairwise Similarity Structure

Given that the corpus contains numerous translations for each language, including multiple copies of widely-distributed versions, we quantify their similarity across all New Testament books using two comparison methods: semantic and lexical.

For the semantic comparison, we use the

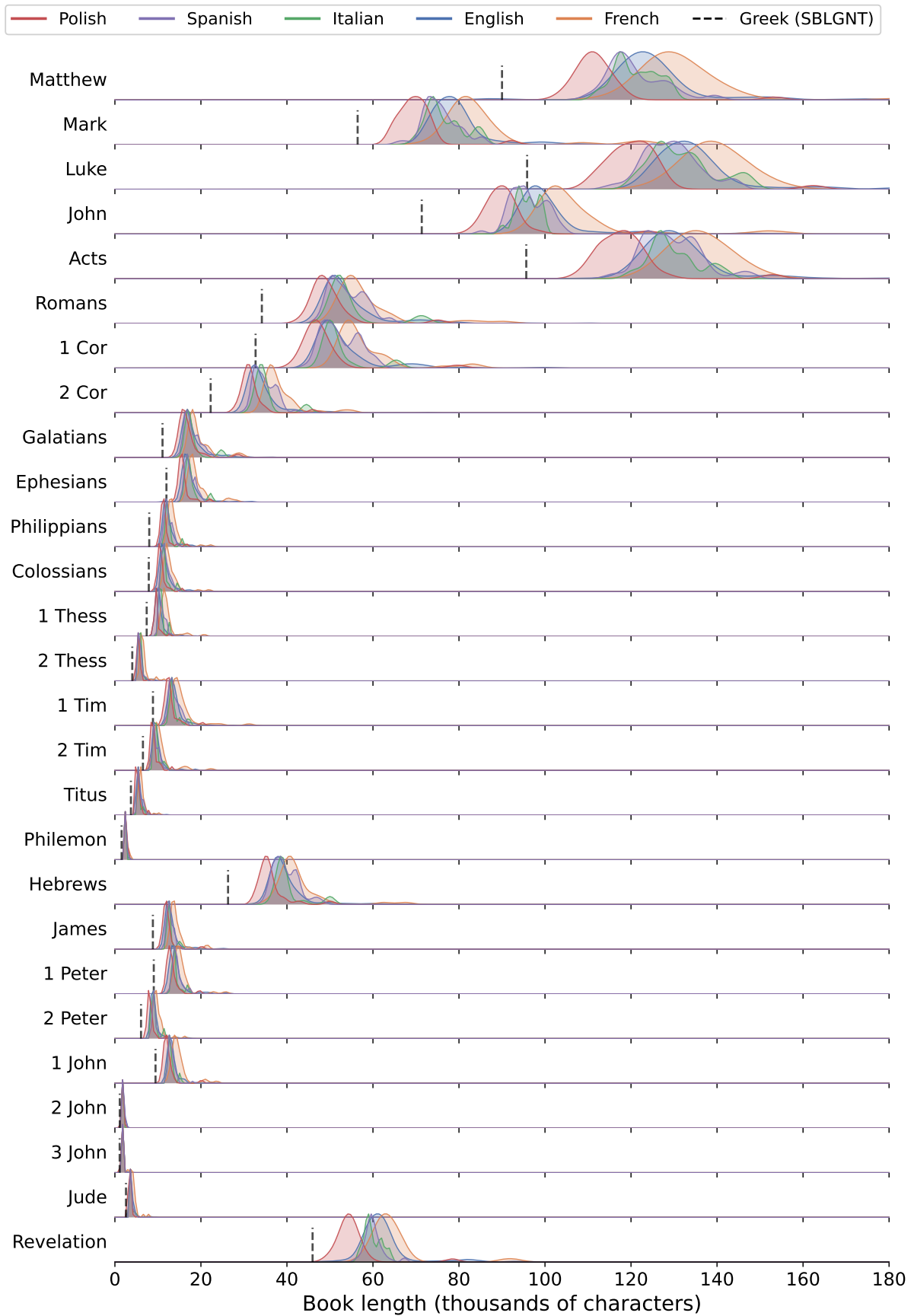


Figure 2: Distribution of book lengths (in characters) across the 334 unique translations for each of the 27 New Testament books, broken down by language. Each curve shows the kernel density estimate for one language. The dashed vertical line marks the length of the Greek source text (SBLGNT).

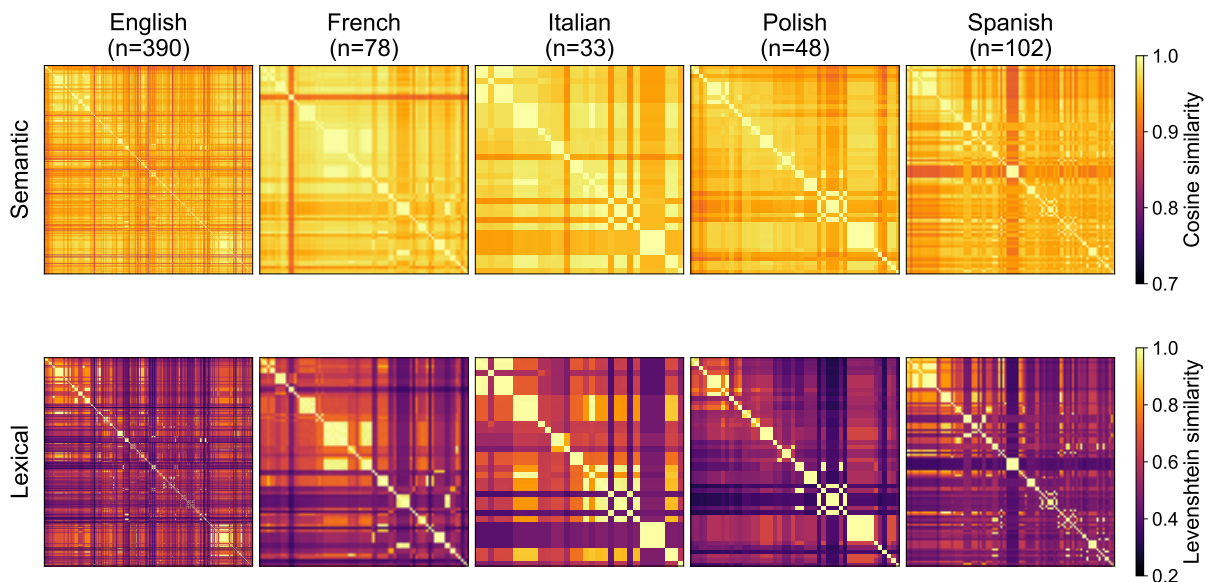


Figure 3: Pairwise similarity matrices between translations for each language, ordered by year of publication. Top row: semantic similarity (cosine on Qwen3-Embedding-8B chapter embeddings). Bottom row: lexical similarity (Levenshtein). Computed at the chapter level across all New Testament books.

chapter-level Qwen3-Embedding-8B vectors distributed with the corpus (§4.2). For each pair of translations sharing a chapter, we compute the cosine similarity of their embeddings and average across all shared chapters to produce a final score.

While effective for capturing broad stylistic differences, semantic similarity can fail to register minor textual variations. We therefore also use the pre-computed lexical similarity scores, based on the Levenshtein similarity defined as $1 - \frac{LD(x_1, x_2)}{\max(\text{len}(x_1), \text{len}(x_2))}$, where $LD(x_1, x_2)$ is the Levenshtein distance.

The pairwise similarity matrices in Figure 3 reveal the corpus’s internal structure. From the semantic matrices, most translation pairs score above 0.8 in cosine similarity, reflecting their shared source text and narrative content.

Two types of clusters are visible. The most prominent are the tight, high-similarity blocks along the diagonal, representing near-identical editions or minor revisions of the same translation work. We also observe distinct clusters of translations that are internally cohesive but markedly dissimilar from the main group, often corresponding to dynamic or paraphrastic translations that employ a different stylistic register.

These clusters appear in both matrices, but their distinctiveness is far more pronounced in the lexical view. The lexical metric, sensitive to specific word choices and sentence structures, better captures the stylistic strategies of dynamic versions — differences that the semantic embeddings partially smooth over.

Figure 4 extends this analysis across all five lan-

guages. Within-language blocks along the diagonal exhibit the highest mean cosine similarity (0.955), ranging from 0.943 (English, whose large and stylistically diverse collection lowers the average) to 0.965 (French and Italian). Cross-language similarity is lower overall (mean 0.823), yet the block structure reveals a clear typological pattern. Among the Romance languages, Italian–Spanish reaches 0.863, the highest cross-language pair, though English–French is comparable at 0.862. Polish, the sole Slavic language, is the most distant from every other language in the corpus: its cross-language means range from 0.766 (French–Polish, the lowest pair) to 0.803 (Polish–Spanish). English occupies an intermediate position, reflecting both its Germanic substrate and its Romance-derived vocabulary.

5.4. Pairwise Variation

We use the chapter-level Levenshtein scores to validate our canonicalization and quantify inter-edition variation, distinguishing three tiers of relatedness: (a) all within-language pairs, (b) pairs sharing a `canonical_id` (same translation family), and (c) pairs sharing both `canonical_id` and `canonical_version` (same edition from different sources).

Figure 5 presents kernel density estimates for these three groups, pooled across all five languages. All within-language pairs center at a median of 0.531. Same-`canonical_id` pairs shift rightward (median 0.990), and same-version pairs cluster near 1.0 (median 0.996). Per-language

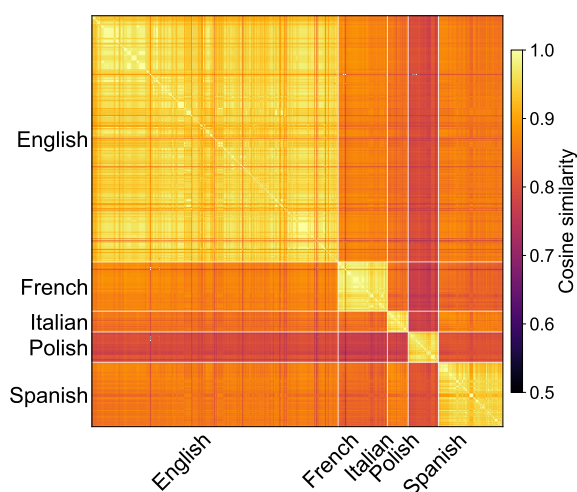


Figure 4: Cross-language semantic similarity matrix for all 651 translations, sorted by language and year. Within-language blocks (along the diagonal) exhibit high similarity; off-diagonal blocks show cross-language semantic correspondence.

percentiles (P10, P50, P90) for the two matched groups are reported in Appendix C. The per-language breakdown confirms that our canonicalization scheme reliably groups near-identical editions: same-version pairs achieve a P10 above 0.83 in every language, with median similarity at least 0.988. For same-`canonical_id` pairs, the spread is wider (P10 ranging from 0.755 for Spanish to 0.986 for Polish), reflecting the varying degree of textual revision within translation families across languages.

Beyond validating the annotation, the pairwise scores allow us to test whether translation diversity has increased over time. For this analysis, we use the 334 unique translations (deduplicated by edition) to avoid pseudoreplication from cross-site copies. We regressed pairwise similarity on the midpoint year of each pair, controlling for temporal distance (both predictors standardized). The analysis covers all non-same-`canonical_id` pairs within each language. A negative mid-year coefficient indicates that translations published in the same era are more diverse in recent centuries. The mid-year coefficient is significantly negative in every language for both lexical and semantic similarity (all $p \leq .011$): lexical β ranges from -0.03402 (English) to -0.10069 (Spanish); semantic β ranges from -0.00340 (English) to -0.00987 (Italian). Full per-language regression results are reported in Appendix D. The effect sizes are modest ($R^2 = 0.02$ – 0.28 ; see Appendix D), indicating that the temporal trend, while reliable, accounts for only a fraction of pairwise similarity variance — most of which is driven by translation philosophy and other factors. Nevertheless, the consistent direction across all

languages and both metrics indicates that modern translations are both lexically and semantically more diverse than historical ones, consistent with a broadening of translation strategies over time — from narrower stylistic conventions in earlier centuries to the wide spectrum of approaches (literal, dynamic, paraphrase) seen today.

6. Discussion

6.1. New Research Directions

The `Targum` corpus makes tractable research questions that were previously difficult to address at scale. These avenues can be broadly grouped into *micro-* and *macrolevel* analyses.

6.1.1. Micro-Level Analysis

First, `Targum` facilitates *micro-level analyses* by allowing researchers to select and compare focused subsets of the corpus based on specific metadata criteria. For example, scholars can:

- **Conduct diachronic studies** by tracing linguistic changes across decades of revisions within a single translation family (e.g., the King James Version).
- **Perform synchronic comparisons** of contemporary versions to isolate the effects of confessional tradition or individual translator style.
- **Track translation lineage** by following the propagation of theological and stylistic choices through indirect translation chains, such as those mediated by a Latin Vulgate source or, in modern institutional cases (e.g., Jehovah’s Witnesses, Biblica), by an English pivot translation.

Furthermore, the corpus can serve as a *quantitative complement to traditional qualitative evaluation*. While an expert might label a translation as “controversial” or “overly literal” based on key passages, `Targum` allows researchers to measure the pervasiveness of these features across the entire New Testament, framing such judgments within the full scope of the text.

6.1.2. Macro-Level Analysis

Second, on a *macro-level*, `Targum` can function as a *benchmark* for any new translation. By generating a vector embedding for a new version, whether produced by humans or machines, it can be situated within the semantic space of the entire corpus. Its stylistic and theological profile can then be quantitatively compared against the full space of existing

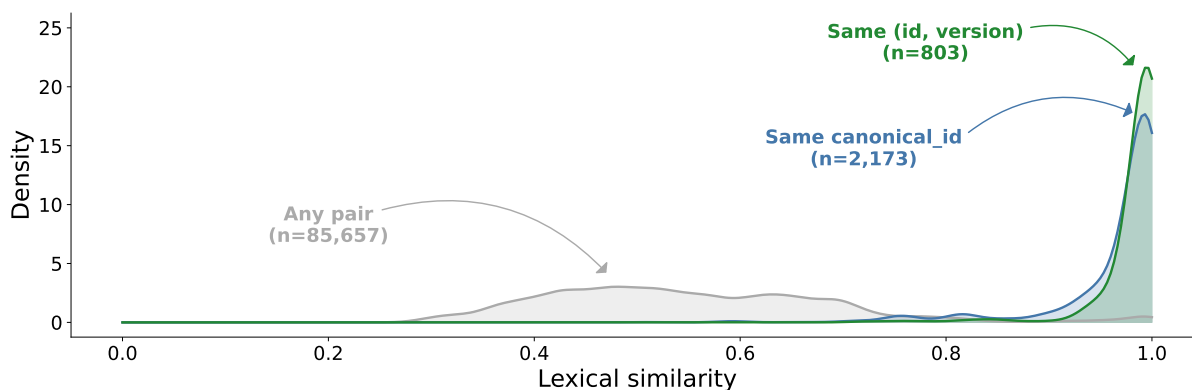


Figure 5: Lexical similarity distributions for three tiers of translation relatedness, pooled across all five languages: all within-language pairs (“Any pair”), pairs sharing a `canonical_id`, and pairs sharing both `canonical_id` and `canonical_version`.

translations — answering questions such as: “Is this new translation stylistically closer to the KJV or the NIV family?” or “Does its interpretation of key theological terms align more with historical Catholic or Protestant traditions?”

6.2. Future Work

We plan to extend `Targum` along three axes: more translations, richer metadata, and broader language coverage.

More translations. We intend to add translations not yet captured by the current scraping scope, pursuing this along two lines. First, additional translations increase the historical representativeness of the corpus within each language, filling gaps in the NT translation record that the current source set leaves open. Second, adding further copies of translations already in the corpus — sourced from sites not yet scraped — functions as a cross-validation mechanism. Our site-specific parsing scripts are tested by comparing copies of the same translation that, by manual annotation, should be near-identical. When pairwise lexical similarity between such copies falls below the expected range, this signals a parsing error in one of the sources. Resolving these anomalies improves the accuracy of the affected site’s parser, which in turn raises quality not only for translations with multiple copies but also for single-copy translations drawn from the same site, where analogous errors would otherwise go undetected.

Richer metadata. The current annotation covers canonical identity, revision year, and copyright status. We plan to add four fields. First, *source text provenance* will record whether a translation was produced directly from the Greek, or mediated through the Latin Vulgate or an English inter-

mediary — a distinction that affects how closely a translation can be expected to track the Greek original. Second, *translation lineage* will map dependency relationships between versions (e.g., which prior translations a given rendering explicitly acknowledges), enabling the modeling of complex genealogical flows that a flat canonical ID cannot represent. Third, *confessional tradition* will classify each translation’s institutional or theological affiliation. Fourth, *critical edition* will identify the Greek source text used (e.g., Nestle-Aland 28th edition, Textus Receptus, Majority Text) where this can be established; this is feasible for most modern translations and less so for historical ones.

More languages. Extending coverage to other languages with rich NT translation traditions (e.g., German, Portuguese) is technically straightforward, but requires annotators with sufficient domain knowledge to verify canonical identities and assess translation history. We are open to expanding the language set, contingent on interest from the research community and the availability of collaborators who can contribute that expertise.

7. Conclusion

`Targum` addresses a key gap in existing biblical resources by providing a multilingual New Testament corpus focused on the depth of translation within each language. With extensive coverage for five European languages and 334 unique translations enriched with verified metadata, `Targum` provides the infrastructure for a new class of computational research — addressing long-standing questions about the interplay of theology, style, and culture in translation using scalable quantitative methods.

8. Limitations

We acknowledge several limitations in the current version of the `Targum` corpus.

First, the corpus is necessarily constrained by the availability of digitized translations on public websites. This introduces a potential “digitization bias”: the corpus reflects the history of translations that have been made digitally accessible, which may not perfectly align with the complete publication history. For instance, German was initially considered for inclusion due to its rich translation tradition but was ultimately excluded because of the relative scarcity of online versions compared to the five languages that were selected. Therefore, the distribution of translations in `Targum` should be understood as representative of the digital landscape, not the complete landscape.

Second, the manual normalization process has inherent limitations. Although our methodology is systematic, the final classification is still subject to human error. A key challenge arose when trying to pinpoint the exact revision year for certain translations. This difficulty was most pronounced not with historical texts, which often have well-documented print editions, but with modern, digital-native translations distributed primarily online. Such versions can be revised multiple times per year without formal versioning, and their release history is often undocumented. In these cases, we made a best-effort judgment based on the available textual and contextual evidence (e.g., copyright dates on the webpage) to determine the edition. A complementary difficulty arises at the opposite end of the time axis: historical translations that have undergone modernized re-editions. For example, Wycliffe’s Bible received a modernized-spelling edition for its 500th anniversary, and many other historical translations have been periodically revised over the centuries. In such cases the translation philosophy is historical while the orthography is contemporary, making it difficult to assign a single revision year that captures both dimensions. When the time span between the original and the revision is large the distinction is clear, but for translations with recurrent, incremental updates the boundary is less obvious.

9. Ethics Statement

We aggregated data for the `Targum` corpus from publicly accessible websites. The authors do not claim copyright on any of the translated texts. As shown in Table 3, the copyright status varies across languages. Our distribution plan respects the intellectual property rights of the copyright holders while maximising the utility of the resource for the research community. Translations in the public

domain or behind open licenses, along with the pre-computed embeddings and pairwise similarity scores for all translations, will be made fully publicly available (at <https://github.com/mrapacz/targum-corpus>). Copyrighted texts will not be publicly distributed in full; instead, they will be made available upon reasonable request for noncommercial research purposes. We will remove any translation from the corpus upon a direct request from a verified copyright holder.

In the preparation of this work, AI assistants were used as productivity tools. During initial development, Gemini 2.5 Pro (via Cursor and the Gemini web interface) was used for tasks such as generating boilerplate code for web scraping and helping with the editing and formatting of the submitted manuscript. For the camera-ready version, Claude Opus 4.6 and Sonnet 4.6 (via GitHub Copilot) were used for agentic engineering tasks such as generating code for data analysis and validation of data prior to publishing, and helping with the editing and formatting of this article.

10. Acknowledgments

This research was supported by the National Science Centre, Poland, under project number 2025/57/N/HS2/04961. We gratefully acknowledge the Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computational facilities and support under grant no. PLG/2026/019145.

11. Bibliographical References

- Vesa Akerman, David Baines, Damien Daspit, Ulf Hermjakob, Taeho Jang, Colin Leong, Michael Martin, Joel Mathew, Jonathan Robie, and Marcus Schwarting. 2023. The ebible corpus: Data and model benchmarks for bible translation for low-resource languages. *arXiv preprint arXiv:2304.09919*.
- Philip S. Alexander. 1992. Targum, targumim. In David Noel Freedman, editor, *The Anchor Bible Dictionary*, volume 6, pages 320–331. Doubleday, New York.
- Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124, Copenhagen, Denmark. Association for Computational Linguistics.
- Gerlof Bouma, Evie Coussé, Trude Dijkstra, and Nicoline van der Sijs. 2020. The edges diachronic bible corpus.

- Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating prose style transfer with the bible. *Royal Society open science*, 5(10):171920.
- Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.
- Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc’Aurelio Ranzato. 2019. [The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.
- Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. *Oceania*, 135(273):40.
- Arya D McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The johns hopkins university bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892.
- Aaron Mueller, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu, and David Yarowsky. 2020. [An analysis of massively multilingual neural machine translation for low-resource languages](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3710–3718, Marseille, France. European Language Resources Association.
- Garrett Nicolai, Dylan Lewis, Arya D. McCarthy, Aaron Mueller, Winston Wu, and David Yarowsky. 2020. [Fine-grained morphosyntactic analysis and generation tools for more than one thousand languages](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3963–3972, Marseille, France. European Language Resources Association.
- Garrett Nicolai and David Yarowsky. 2019. [Learning morphosyntactic analyzers from the Bible via iterative annotation projection across 26 languages](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1765–1774, Florence, Italy. Association for Computational Linguistics.
- Franz Schühlein. 1912. [Targum](#). In Charles G. Herbermann, Edward A. Pace, Condé B. Pallen, Thomas J. Shahan, and John J. Wynne, editors, *The Catholic Encyclopedia*, volume 14. Robert Appleton Company, New York.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A. Corpus Metadata Fields

Each translation in the Targum corpus is described by the metadata fields listed in Table 4. These fields are derived from the source websites and our manual annotation process.

Field	Description	Example
site	Source website from which the translation was acquired.	laparola.net
iso	ISO 639-3 language code.	eng
translation_id	Site-specific identifier assigned by the source.	asv
translation_name	Full human-readable name.	American Standard Version
translation_abbr	Common abbreviation.	ASV
canonical_id	Groups all instances of the same translation work across sites and editions.	king-james-version
canonical_version	Edition label; often a year but may encode richer distinctions (e.g. 1989-anglicised-catholic).	1769
canonical_year	Integer revision year; places the edition on the time axis.	1769
num_chapters	Number of New Testament chapters present.	260
num_verses	Number of New Testament verses present.	7957
num_words	Total word count across all chapters.	179836
copyright_status	Copyright classification; accompanied by a reasoning field (not shown).	public_domain

Table 4: Metadata fields available for each translation in the Targum corpus.

B. Corpus File Hierarchy

The corpus repository has the following top-level structure:

```
targum-corpus/  
  index.tsv  
  copyrights.tsv  
  book_coverage.tsv  
  manifest.json  
  corpus/  
    {site}/  
      {iso}/  
        {id}.jsonl  
  similarity/  
    lexical/  
      levenshtein/chapter/{iso}.parquet  
    semantic/  
      cosine/  
        {model}/chapter/{iso}.parquet  
        {model}/chapter/cross_language.parquet  
  embeddings/  
    {model}/  
      language={iso}/  
        site={site}/  
          translation={id}/  
            granularity={chapter|verse}/  
              data.parquet
```

The placeholders expand as follows: {site} is the source library (e.g. ebible.org, biblegateway.com); {iso} is the ISO 639-3 language code (eng, fra, ita, pol, spa); {id} is the source-specific translation identifier listed in index.tsv; and {model} encodes the embedding model with the separator XxX (e.g. QwenXxXQwen3-Embedding-8B).

Root-level metadata. Four files at the repository root describe the corpus as a whole. `index.tsv` lists every translation instance with the metadata fields described in Appendix A. `copyrights.tsv` records the copyright text and status for each translation. `book_coverage.tsv` indicates which of the 27 New Testament books each translation contains. `manifest.json` provides summary statistics (total counts, per-language breakdowns).

Translation texts (corpus/). Each translation is stored as a JSONL file (one JSON object per verse) under `corpus/{site}/{iso}/`. Each line contains four fields, as illustrated below with the first three verses of the King James Version (Matthew 1:1–3):

```
{ "book": "MAT", "chapter": 1, "verse": "1",
  "text": "The book of the generation of
  Jesus Christ, the son of David, the
  son of Abraham." }
{ "book": "MAT", "chapter": 1, "verse": "2",
  "text": "Abraham begat Isaac; and Isaac
  begat Jacob; and Jacob begat Judas and
  his brethren;" }
{ "book": "MAT", "chapter": 1, "verse": "3",
  "text": "and Judas begat Phares and Zara
  of Tamar; and Phares begat Esrom; and
  Esrom begat Aram;" }
```

The `book` field uses USFM three-letter codes (e.g. MAT, MRK, REV); `chapter` is an integer; `verse` is a string to accommodate non-standard segmentation (e.g. "6a", "6-8").

Pairwise similarity scores (similarity/). Pre-computed chapter-level pairwise similarity scores are stored as Parquet files, one per language. Two metrics are provided: Levenshtein (lexical) and cosine distance over Qwen3-Embedding-8B vectors (semantic). The semantic directory also contains a `cross_language.parquet` file with all inter-lingual pairs.

Text embeddings (embeddings/). Embeddings are stored as Hive-partitioned Parquet files. Two models are included (Qwen3-Embedding-0.6B and Qwen3-Embedding-8B), each at chapter and verse granularity. Each `data.parquet` file contains two columns: `key` (e.g. MAT 1 for chapter, MAT 1:1 for verse) and `value` (the embedding vector).

C. Similarity Thresholds

Table 5 reports per-language lexical similarity percentiles for translation pairs at three levels of relatedness: any within-language pair, pairs sharing a canonical identity, and pairs sharing both canonical identity and version.

Language	Any pair			Same <code>canonical_id</code>			Same (id, version)		
	P10	P50	P90	P10	P50	P90	P10	P50	P90
English	0.386	0.535	0.710	0.944	0.991	1.000	0.963	0.995	1.000
French	0.419	0.552	0.732	0.903	0.986	1.000	0.963	0.988	1.000
Italian	0.423	0.601	0.893	0.835	0.991	1.000	0.835	0.998	1.000
Polish	0.355	0.492	0.648	0.986	0.998	1.000	0.993	0.999	1.000
Spanish	0.411	0.491	0.750	0.755	0.956	1.000	0.990	0.999	1.000

Table 5: Lexical similarity percentiles (P10, P50, P90) for within-language translation pairs: any pair, pairs sharing a `canonical_id`, and pairs sharing both `canonical_id` and `canonical_version`.

D. Diachronic Diversity Regressions

Table 6 reports the OLS regression results for the diachronic diversity analysis. For each language, we regressed pairwise similarity (lexical and semantic) on the midpoint year of the pair, controlling for temporal distance between the two translations (both predictors standardized). Only non-same-`canonical_id`

pairs among the 334 unique translations (deduplicated by edition) are included. A negative mid-year coefficient (β) indicates that translations produced in the same era are more diverse in more recent centuries.

Language	Lexical				Semantic			
	<i>N</i>	β	<i>p</i>	R^2	<i>N</i>	β	<i>p</i>	R^2
English	18,528	-0.03402	< .001	0.0197	18,528	-0.00340	< .001	0.0537
French	814	-0.05848	< .001	0.1341	814	-0.00431	< .001	0.0297
Italian	129	-0.05390	= 0.011	0.0585	129	-0.00987	< .001	0.1198
Polish	402	-0.07290	< .001	0.2367	402	-0.00390	= 0.001	0.2790
Spanish	1,348	-0.10069	< .001	0.1169	1,348	-0.00805	< .001	0.0915

β = standardized mid-year coefficient; negative values indicate increasing diversity over time.

Table 6: OLS regression of pairwise similarity on mid-year for non-same-canonical_id pairs, per language. Negative β indicates increasing diversity over time.