# Degree Zero of Translation: Using Interlinear Baselines to Quantify Translator Intervention

**Maciej Rapacz**
AGH University of Kraków
mrapacz@agh.edu.pl

**Aleksander Smywiński-Pohl**
AGH University of Kraków
apohllo@agh.edu.pl

## Abstract

Literary translation is rarely a neutral act of linguistic transfer, but rather a continuous series of conscious interventions—restructuring, semantic shifts, and stylistic adaptations. While Translation Studies analyzes these shifts qualitatively, current computational methods focus primarily on quality evaluation (e.g., BLEU, COMET) or authorship attribution (e.g., stylometry), lacking a scalable metric to quantify the extent and character of the translator's intervention. We propose a novel method to measure the translator's signal by using Interlinear Translation – a strict word-for-word gloss – as a computational baseline representing translational "Degree Zero", i.e. a neutral form of source text devoid of any stylistic adaptation. We define the Intervention Vector as the semantic difference between a literary translation and its interlinear counterpart in a high-dimensional vector space. We validate this approach on a multilingual corpus of the Greek New Testament translations comprising 5 interlinear baselines and 74 literary translations across 5 languages – English (16), French (14), Italian (12), Polish (16), and Spanish (16). Our results demonstrate that the magnitude of the Intervention Vector effectively ranks texts along a spectrum from literal to paraphrase, aligning with established theoretical categories. We find that this magnitude consistently distinguishes between translation strategies, yielding significantly longer vectors for dynamic and paraphrase strategies compared to literal and formal ones. This framework provides a quantitative method for analyzing translator agency without the need for a comprehensive corpus of reference translations.

## 1 Introduction

Translation, especially in a literary context, is rarely a neutral act of linguistic transfer, but rather an act of rewriting (Lefevere, 2016) or mediation (Munday, 2013). The translator makes a continuous series of conscious decisions to adapt the source text to a new linguistic and cultural context, a specific audience, or a particular stylistic purpose. We propose a method to quantify these decisions, which manifest themselves in the collective set of syntactic restructuring, semantic shifts, and stylistic adaptations that distinguish a literary text from a mechanical, word-for-word gloss. Although this intervention is an inherent feature of human translation, quantifying it has lacked a reference-free, scalable method.

Current computational approaches do not capture this phenomenon. Natural Language Processing (NLP) has historically focused on Machine Translation (MT), where the objective is to optimize the output for correctness and fluency. Consequently, evaluation metrics – from n-gram matching such as BLEU (Papineni et al., 2002) to reference-free neural estimators such as COMET-QE (Rei et al., 2020) – are designed to correlate with human quality judgments (Fomicheva and Specia, 2016). These tools operate on an evaluative axis: they assign a scalar score to determine how "good" a translation is.

In contrast, Translation Studies (TS) originates from the analysis of human translation, where divergence from the source is viewed not as error but as a deliberate *strategy*. Theories such as those from the Manipulation School posit that "all translation is manipulation of the source text for a certain purpose" (Hermans, 1985). However, the field lacks the computational methodology to characterize this manipulation. Traditional approaches in Computational Stylometry, such as Burrows' Delta (Burrows, 2002), are primarily used for *attribution*—identifying lexical "fingerprints" to determine *who* translated a text. Yet attribution metrics are inherently relative; they indicate that *Translation A* is statistically similar to *Translation B*, but they do not measure the shift from the source text. They capture lexical tendencies but miss the broader semantic and syntactic restructuring per-

*Proceedings of SIGHUM 2026*, pages 227–240
March 28-29, 2026 ©2026 Association for Computational Linguistics

formed by the translator.

We propose using the *Interlinear Translation* as a computational baseline. Unlike standard literary translations, an interlinear version strictly preserves the order of words in the source language, placing the target-language words directly under the corresponding source words (Shuttleworth and Cowie, 2014). Drawing on Barthes (1953)'s concept of Writing Degree Zero, a neutral mode of writing, we hypothesize that this extremely literal form of translation represents a neutral representation of the source text – or as close to it as possible – completing the necessary switch from one language to another while minimizing adaptation for style, audience, or fluency. By calculating the difference between vector representations of a literary translation under study and this baseline, we isolate the translator's signal. We term this difference the *Intervention Vector*, and we posit that it captures both the character and the extent of the translator's intervention.

We validate this method on a multilingual corpus of New Testament translations in five languages (English, French, Italian, Polish, and Spanish). We selected this domain because interlinear translations are primarily available as study resources for ancient and religious texts and because the rich history of translation of the New Testament offers a diverse range of strategies, which we use to test the sensitivity of the metric.

Our contributions[1] are as follows:

1. **Dataset and Validation Framework:** We construct a multilingual parallel corpus aligning literary New Testament translations with interlinear baselines across five languages. We use this dataset to establish a testbed for evaluating quantitative measures of translator intervention against theoretical categories.

2. **Spectrum Recovery:** We demonstrate that the magnitude of the Intervention Vector recovers the theoretical translation spectrum without supervision, effectively sorting translation strategies from literal to paraphrase.

3. **Pairwise Sensitivity:** We validate the metric's granularity by showing that it consistently distinguishes intervention levels even in pair-wise comparisons of individual translations, proving it is sensitive enough to capture differences beyond broad strategy labels.

## 2 Related Work

**Machine Translation Evaluation**. Standard Machine Translation (MT) evaluation is based on reference-based metrics that penalize divergence. Neural metrics like COMET (Rei et al., 2020) improve on BLEU (Papineni et al., 2002) but still rely on some sort of reference, which might inadvertently introduce reference bias, penalizing valid creative variations Fomicheva and Specia (2016). Recent work has attempted to mitigate this by using multiple references (Wu et al., 2025), referenceless metrics or multi-agent LLM evaluators (Kim et al., 2025), but these approaches still stay within the evaluative paradigm, aiming to score the translation's quality, rather than quantify the translator's intervention.

**Computational Stylometry**. Stylometric research typically focuses on attribution and clustering. The most common method used here is Burrows' Delta (Burrows, 2002), which distinguishes authors through frequent-word distributions. There have been multiple attempts to identify translators in the past, with varying success. It was found that, especially when dealing with translations of multiple works, clustering algorithms are more likely to detect the author than the translator (Rybicki, 2012). Identifying multiple translators within a single work resulted in greater success (Rybicki and Heydel, 2013). N-gram-based methods together with gradient boosting have been found to have been successful in translator attribution, even in multi-work, multi-translator scenarios (Mohamed et al., 2023). However, none of these measures quantifies the translator's intervention. They rely either on relative distance (as with Burrows' Delta) – which merely indicates that Translation A is statistically similar to Translation B, rather than defining its intrinsic character – or on surface-level features like n-grams. While the latter can describe stylistic tendencies within the target language, none of these methods capture the degree of departure from the source text, nor do they quantify the specific semantic shifts introduced by the translator.

**Translationese and Corpus Linguistics**. Corpus-based Translation Studies identify universal features of translated text, such as simplification and explicitation (Baker, 1993). Research in this domain typically quantifies "translationese" – the

[1] The code and data are available at `https://github.com/mrapacz/sighum-interlinear-vector-baselines`

linguistic fingerprints distinguishing translated text from native writing – through surface-level features like type-token ratio, sentence length, or part-of-speech density (Volansky et al., 2015). While translationese is often viewed pejoratively as a mark of lower quality or lack of fluency (Wein and Schneider, 2024), we argue that in literary translation, the translator's signal stems from deliberate strategy rather than linguistic interference. Consequently, our definition of intervention is broader than translationese: it encompasses valid, creative restructuring in the target language that would not be captured by metrics designed solely to detect non-native artifacts.

**Vector Semantics in Digital Humanities**. The use of embeddings to model semantic change is well-established in diachronic linguistics. Researchers use distributional models to measure semantic drift over time (Kutuzov et al., 2018; Hamilton et al., 2016). In literary studies, recent work has employed BERT-based embeddings to analyze narrative structures (Wegmann et al., 2022) or compared BERT and ELMo embeddings in the study of diachronic shifts (Kutuzov et al., 2018). We rely on vector representations to study semantic shift; here the shift is not diachronic but the distance between a full translation and the interlinear baseline.

**Translator Intervention in a Quantitative Setting** To our knowledge, our work is the first to propose a computational operationalization of translator intervention.

## 3 Methodology

Interlinear translation, often cited as the "archetype or ideal of all translation" (Benjamin, 1923), maps target-language words directly onto the source syntax without reordering for fluency (Shuttleworth and Cowie, 2014). To establish a baseline of non-intervention, we draw from Roland Barthes' concept of Writing Degree Zero (Barthes, 1953), which describes a "colorless" and neutral style of writing that strips away all artistic flair to report reality as directly as possible. By analogy, we propose that interlinear translation is a "colorless" mode of translational practice that strips away all artistic flair to report the "reality" of the source text as directly as possible. Although no translation is entirely free of interpretation, even in extreme forms like the interlinear, we hypothesize that this state is as close to it as possible: a state where the compulsory language shift has occurred, but the specific adaptations for fluency, style, or target audience have not yet been applied.

### 3.1 Dataset Curation

The data consist of two text types: the translations under study and interlinear translations, which act as the point of reference.

**The Interlinear Baseline.** We scraped interlinear translations from five language-specific repositories: *BibleHub* (English), *Editeur BPC* (French), *Altervista* (Italian), *Oblubienica* (Polish) and *Bibliatodo* (Spanish). These texts preserve the syntactic order of the Ancient Greek source while mapping words to their literal target-language equivalents. We remove verse numbers, but retain original punctuation and capitalization.

**The Literary Corpus.** We utilized the *targum* corpus (Rapacz and Smywiński-Pohl, 2026), a multilingual collection containing 657 New Testament translations across our five target languages. To interpret the results effectively, we hand-picked a subset of 74 translations – English (16), French (14), Italian (12), Polish (16), and Spanish (16) – to have representatives in each of four orientational categories:

1. **Literal:** Strict word-for-word translation (e.g., Literal Standard Version).

2. **Formal:** Translations that adhere closely to source structure but permit necessary grammatical adjustments for fluency (e.g. New Revised Standard Version).

3. **Dynamic:** Translations that prioritize functional meaning, with substantial structural changes (e.g. New Living Translation).

4. **Paraphrase:** Texts with the highest allowance for adaptation, occasionally bordering on distinct literary works (e.g. The Message).

We emphasize that these labels are purely orientational and serve only to mark the varying degrees of allowance for adaptation; they do not imply any judgment on theological validity or fidelity.

Furthermore, we restricted ourselves to translations from 1900 onwards. We deliberately excluded earlier translations (e.g. from the 17th or 18th centuries) so as to compare translations synchronically and to avoid confounding effects from diachronic

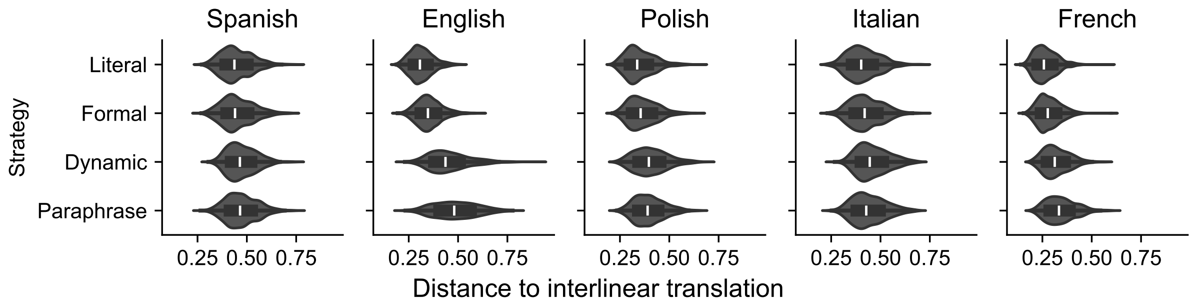

Figure 1: Violin plot showing distribution of intervention vector lengths for each chapter across languages and strategies.

language change. The complete list of translations is available in Appendix A.

### 3.2 Vectorization

We aggregate the text at the *Chapter Level* (260 units per translation). While verse-level segmentation offers finer granularity, it introduces alignment challenges (e.g., merged verses like "2-6a") and noise from short verses (e.g., "Jesus wept"). Pericope-level segmentation, while semantically coherent, lacks a standardized schema across translations. We generate embeddings using the `Qwen3-Embedding-8B` model[2]. We choose this model for its strong multilingual performance and the 32k token context window (Zhang et al., 2025), which fits full chapters of the New Testament without the need to segment or truncate the text.

### 3.3 The Intervention Vector

We define the *Intervention Vector* ($V_{int}$) as the difference between the vector representation of a literary translation ($V_{trans}$) and the Interlinear Baseline ($V_{base}$) for the same chapter:

$$V_{int} = V_{trans} - V_{base} \tag{1}$$

Using this vector, we can derive two straightforward metrics: magnitude (euclidean norm) and stability (standard deviation). Further, we also projected the vectors onto a plane using Principal Component Analysis (PCA) per language to visualize a simplified picture of the intervention characteristics.

## 4 Results

**The Spectrum of Intervention.** First, we evaluate whether the magnitude of the intervention vector ($||V_{int}||$) accurately reflects the theoretical degree of translator intervention.

Figure 1 presents the distribution of intervention vector lengths. The violins show that more literal strategies cluster closer to the origin while dynamic and paraphrase versions extend further to the right.

We test whether chapter-level distances differ between strategy groups using Mann–Whitney U tests with Bonferroni correction (see Appendix B for full group-level and pairwise comparisons). In English and French the four strategies form a strict order: Literal < Formal < Dynamic < Paraphrase (all $p < 0.001$). In Italian and Polish, Literal and Formal lie closer to the baseline than Dynamic and Paraphrase (cross-group comparisons $p < 0.001$, or Formal < Paraphrase at $p < 0.01$); we do not find significant differences between Literal and Formal, or between Dynamic and Paraphrase. In Spanish, Literal is closer than all other groups ($p < 0.001$) and Formal closer than Dynamic ($p < 0.05$). At the level of individual translations, one of the two is significantly closer to the interlinear baseline in 90% of English pairs, 69.2% in French, 55.8% in Polish, 48.3% in Spanish, and 51.5% in Italian (Mann–Whitney with Bonferroni, $p < 0.001$).

**The Range of Intervention.** Median chapter-level distance to the interlinear baseline varies by language and the ranges differ (e.g. English 0.31–0.48, French 0.258 – 0.334, Polish 0.34–0.39, Italian 0.40–0.45, Spanish 0.44–0.66). That these ranges differ may reflect typological differences between each target language and Ancient Greek: where the language forces more structural adaptation, translations may consistently lie further from the interlinear baseline. We further check the relation between the mean and the standard deviation of the intervention vector and find that in the case of English and French there is a clear rela-

[2] https://huggingface.co/Qwen/Qwen3-Embedding-8B

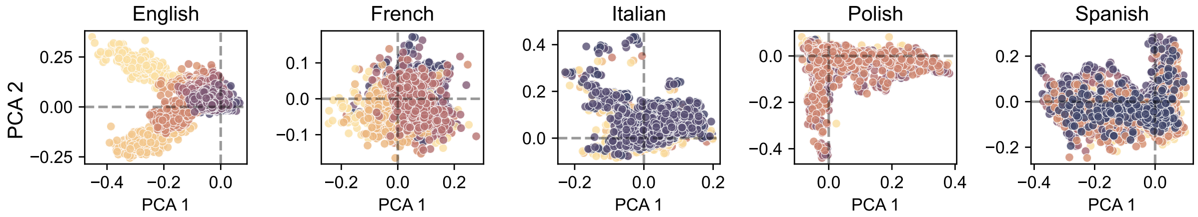

Figure 2: PCA projection of intervention vectors by language. PCA was calculated independently per language, so axes are not comparable. Dots represent chapters, colored by the translation's mean intervention magnitude (brighter colors indicate higher magnitude).

tionship, 0.83 and 0.88, respectively. For the other languages we find a weak relationship for Polish (0.428), weak negative for Spanish (-0.511) and no relationship for Italian (0.068), though we cannot distinguish whether these values stem from typological differences between the languages or from better representation in English and French.

**The Topology of Intervention.** Finally, we investigate the geometric shape of the intervention space. While magnitude measures *how much* a translator intervenes, it does not indicate *how*. To assess whether the intervention vectors contain some characteristics of the translator's strategy, we computed a PCA on the difference vectors ($V_{int}$) (independently for each language). We used PCA to further reduce dimensionality and to project it onto a 2D plane, as seen in Figure 2. We shift the coordinates so that the interlinear baseline is at the origin $(0, 0)$. In the figure, color intensity indicates mean distance from the interlinear baseline (brighter colors represent greater intervention); since PCA was performed separately for each language, the color scale is language-specific and not comparable across panels.

We observe distinct patterns in the intervention vectors for each language. For Polish, we find a clear, orthogonal v-shape. English[3] and French, while lacking such a clear pattern, place translations with overall higher intervention further away from the origin. For Italian and Spanish, the patterns are more complex, and PCA did not identify anything that would align with the magnitude of the intervention.

## 5 Conclusion

We propose a novel framework for quantifying the translator intervention. By computing the intervention vector, the vector difference between the translation under study and its interlinear counterpart, we form a representation that captures characteristics of the translator's intervention.

We demonstrate that the magnitude of the intervention vector can be used to rank translations according to their position on the translation spectrum, from literal to paraphrase. We also show that our method consistently yields longer vectors for higher-intervention strategies, both when comparing (literal, formal) vs. (dynamic, paraphrase) groups and when comparing individual translations; for the latter, we found a statistically significant difference in chapter-level distances in roughly half of the pairs (and in 90% of English pairs).

Finally, we show that the patterns in the intervention vectors differ from language to language. We observe different ranges of intervention vector lengths and different patterns when projecting them onto a plane using PCA. These patterns may arise from typological differences, representation quality, or sample composition.

Future work will extend this preliminary study in three directions. First, we would like to study how the metric behaves depending on parameters such as model choice (and model size) as well as the granularity of the text (chapter, verse, pericope). Second, in this work, we explicitly excluded the diachronic axis to isolate the method from drift; it would be useful to test how our method treats archaic texts. Third, we plan to validate this metric on distinct literary corpora beyond the New Testament to see how well these findings generalize to other domains.

Although we tested the method on a corpus of translations, the metric itself can be applied to a single translation. The concept of the interlinear baseline could also be applied in other domains, e.g. to isolate prosody from texts or musical interpretation from a very strict midi file.

[3] Erratum: English PCA results were not interpreted in the original submission due to an algorithmic error.

## Limitations

First, we focused our analysis on the New Testament, because for this corpus the interlinear translations were readily available on-line. However, biblical texts are heavily represented in LLM training corpora, often within specific theological contexts. We do not know how this method performs on modern, secular literary works, which might exhibit different embedding behaviors. Future work is needed to determine if the stability we observed generalizes to domains less saturated in the model's training data.

Second, we treat the interlinear translation as a neutral reference point. In reality, rendering Ancient Greek into the target language requires selecting specific words, which is already an act of interpretation. Our results might be skewed by this fact; a translation could appear distant from the baseline not because the translator's intervention was substantial, but simply because their lexical choices differed from those of the interlinear author.

Third, we did not account for differences in the underlying source text. The New Testament exists in multiple critical editions (e.g., *Textus Receptus* vs. *Nestle-Aland*). If a translation follows a different critical edition than our interlinear baseline, the resulting vector distance captures textual variance—such as added or deleted verses—rather than the translator's intervention. Future work would need to align the critical edition of the baseline with that of the target text to better isolate the translator's style.

Fourth, for the purpose of this study, we labeled translations with broad categories such as Literal, Dynamic, and Paraphrase. We acknowledge that this is a coarse taxonomy intended only to approximate the general extent of adaptation, not the nuance of its character. These labels serve purely as auxiliary benchmarks to verify whether our methodology can distinguish fundamental approaches—separating literal strategies from dynamic ones—and do not constitute a qualitative judgment. For instance, designating a text as a "paraphrase" describes the magnitude of its structural divergence from the source, not its legitimacy as a translation.

Fifth, we relied on data scraped from third-party websites. These sources contained formatting errors, which we fixed manually wherever our validation logic spotted them, and our own processing pipeline may have introduced others. Consequently, our metric might interpret these data artifacts as a translator intervention, even when such deviations are not present in the actual published text.

## Acknowledgments

The research was supported by the National Science Centre, Poland under the project number 2025/57/N/HS2/04961. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2026/019145. The research presented in this article was partially supported by the funds of the Polish Ministry of Science and Higher Education assigned to the AGH University of Kraków.

## References

Mona Baker. 1993. Corpus Linguistics and Translation Studies: Implications and Applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology: In Honour of John Sinclair*, pages 233–250. John Benjamins.

Roland Barthes. 1953. *Writing Degree Zero*. Hill and Wang, New York.

Walter Benjamin. 1923. The Task of the Translator. In Lawrence Venuti, editor, *The Translation Studies Reader*, pages 15–25. Routledge, London.

John Burrows. 2002. 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3):267–287.

Marina Fomicheva and Lucia Specia. 2016. Reference Bias in Monolingual Machine Translation Evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 77–82, Berlin, Germany. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Theo Hermans, editor. 1985. *The Manipulation of Literature: Studies in Literary Translation*. Croom Helm, London.

Junghwan Kim, Kieun Park, Sohee Park, Hyunggug Kim, and Bongwon Suh. 2025. MAS-LitEval : Multi-Agent System for Literary Translation Quality Assessment. *arXiv preprint*. ArXiv:2506.14199 [cs] version: 1.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

André Lefevere. 2016. *Translation, rewriting, and the manipulation of literary fame*. Routledge.

Emad Mohamed, Raheem Sarwar, and Sayed Mostafa. 2023. Translator attribution for Arabic using machine learning. *Digital Scholarship in the Humanities*, 38(2):658–666.

Jeremy Munday. 2013. *Style and Ideology in Translation: Latin American Writing in English*. Routledge, New York.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maciej Rapacz and Aleksander Smywiński-Pohl. 2026. Targum – A Multilingual New Testament Translation Corpus. *arXiv preprint*. ArXiv:2602.09724 [cs].

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Jan Rybicki. 2012. The great mystery of the (almost) invisible translator: Stylometry in translation. In Michael P. Oakes and Meng Ji, editors, *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research*, pages 231–248. John Benjamins Publishing Company.

Jan Rybicki and Magda Heydel. 2013. The stylistics and stylometry of collaborative translation: Woolf's Night and Day in Polish. *Literary and Linguistic Computing*, 28(4):708–717.

M. Shuttleworth and M. Cowie. 2014. *Dictionary of translation studies*. St. Jerome Publishing.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same Author or Just Same Topic? Towards Content-Independent Style Representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics. TODO.

Shira Wein and Nathan Schneider. 2024. Lost in Translationese? Reducing Translation Effect Using Abstract Meaning Representation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 753–765, St. Julian's, Malta. Association for Computational Linguistics.

Si Wu, John Wieting, and David A. Smith. 2025. Multiple References with Meaningful Variations Improve Literary Machine Translation. *arXiv preprint*. ArXiv:2412.18707 [cs].

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 [cs].

## A Complete list of translations used in the experiments

This appendix presents the complete list of translations used in the experiments across all five languages. For each translation, we provide its name, abbreviation, strategy classification (Literal, Formal, Dynamic, or Paraphrase), and year of publication. The abbreviations are used to label translations in the pairwise comparison figures in Appendix B.

Table 1: List of English translations included in the experiments.

| Translation | Abbr. | Strategy | Year |
|---|---|---|---|
| Disciples' Literal New Testament | DLNT | Literal | 2011 |
| Berean Literal Bible | BSB | Literal | 2016 |
| Literal Standard Version | LSV | Literal | 2020 |
| New American Standard Bible (2020) | NASB20 | Formal | 2020 |
| NRSV Updated Edition | NRSVue | Formal | 2021 |
| English Standard Version | ESV | Formal | 2025 |
| New World Translation | NWT | Formal | 2025 |
| Bible in Basic English | BBE | Dynamic | 1965 |
| Complete Jewish Bible | CJB | Dynamic | 1998 |
| New Living Translation | NLT | Dynamic | 2015 |
| Good News Translation | GNT | Dynamic | 2017 |
| First Nations Version | FNV | Dynamic | 2021 |
| EasyEnglish Bible | EASY | Dynamic | 2024 |
| The New Testament in Modern English | Phillips | Paraphrase | 1958 |
| The Message | MSG | Paraphrase | 2018 |
| Orthodox Jewish Bible | OJB | Paraphrase | 2023 |

Table 2: List of French translations included in the experiments.

| Translation | Abbr. | Strategy | Year |
|---|---|---|---|
| Bible Bovet-Bonnet | BBO | Literal | 1900 |
| Louis Segond (1910) | LSG | Formal | 1910 |
| Bible Catholique Crampon | BCC | Formal | 1923 |
| Nouvelle Édition de Genève | NEG79 | Formal | 1979 |
| Bible d'Ostervald | OST | Formal | 1996 |
| Bible King James Française | KJF | Formal | 2006 |
| Segond 21 | S21 | Formal | 2007 |
| La Bible des Peuples | BdP | Dynamic | 1994 |
| Bible en français courant | BFC | Dynamic | 1997 |
| La Bible du Semeur | BDS | Dynamic | 2015 |
| Parole Vivante | PVV | Paraphrase | 2013 |
| Parole de Vie | PDV | Paraphrase | 2017 |
| Traduction du monde nouveau | TMN | Paraphrase | 2025 |

Table 3: List of Italian translations included in the experiments.

| Translation | Abbr. | Strategy | Year |
|---|---|---|---|
| Riveduta (Luzzi) | RIV | Literal | 1925 |
| Nuova Riveduta | NR06 | Literal | 2006 |
| Nuova Riveduta 2020 | NR20 | Literal | 2020 |
| Riveduta (1927) | RIV27 | Literal | 2020 |
| Sacra Bibbia (Tintori) | TIN | Formal | 1931 |
| Sacra Bibbia (Ricciotti) | RIC | Formal | 1940 |
| Bibbia CEI (1974) | CEI74 | Formal | 1974 |
| La Nuova Diodati | LND | Formal | 1991 |
| La Parola è Vita | PEV | Dynamic | 2006 |
| Parola del Signore (TILC) | TILC | Dynamic | 2014 |
| La Bibbia della Gioia | BDG | Paraphrase | 2005 |
| Traduzione del Nuovo Mondo | TNM | Paraphrase | 2025 |

Table 4: List of Polish translations included in the experiments.

| Translation | Abbr. | Strategy | Year |
|---|---|---|---|
| Nowy Testament (Szczepański) | SZCZ | Literal | 1917 |
| Przekład Dąbrowskiego | DĄB | Literal | 1961 |
| Nowa Biblia Gdańska | NBG | Literal | 2012 |
| EIB Przekład Dosłowny | EIB-PD | Literal | 2019 |
| Biblia Jakuba Wujka | BJW | Formal | 1923 |
| Biblia Poznańska | BP | Formal | 1975 |
| Biblia Warszawska | BW | Formal | 1975 |
| Biblia Tysiąclecia (V wyd.) | BT5 | Formal | 2000 |
| EIB Przekład Literacki | EIB | Formal | 2016 |
| Biblia Króla Jakuba | BKJ | Formal | 2017 |
| Uwspółcześniona Biblia Gdańska | UBG | Formal | 2017 |
| Biblia Warszawsko-Praska | BWP | Dynamic | 1997 |
| Nowy Przekład Dynamiczny | NPD | Dynamic | 2021 |
| Przekład Mariawicki | MAR | Paraphrase | 1921 |
| Słowo Życia | SŻ | Paraphrase | 2016 |
| Przekład Nowego Świata | PNŚ | Paraphrase | 2025 |

Table 5: List of Spanish translations included in the experiments.

| Translation | Abbr. | Strategy | Year |
|---|---|---|---|
| Versión Moderna (Pratt) | VM | Literal | 1929 |
| Nueva Biblia de las Américas | NBLA | Literal | 1986 |
| La Biblia de las Américas | LBLA | Literal | 1997 |
| Reina-Valera 1960 | RVR60 | Formal | 1977 |
| Biblia del Jubileo | JBS | Formal | 2000 |
| La Biblia de América | BAME | Dynamic | 2010 |

Continued on next page

Table 5: List of Spanish translations included in the experiments.

| Translation | Abbr. | Strategy | Year |
|---|---|---|---|
| La Palabra (España) | BLP | Dynamic | 2010 |
| Nueva Traducción Viviente | NTV | Dynamic | 2010 |
| Nueva Versión Internacional | NVI | Dynamic | 2022 |
| Dios Habla Hoy | DHH | Dynamic | 2023 |
| Biblia Jünemann (Septuaginta) | JUN | Paraphrase | 1992 |
| Biblia Latinoamericana | BL95 | Paraphrase | 1995 |
| Traducción en Lenguaje Actual | TLA | Paraphrase | 2002 |
| Palabra de Dios para Todos | PDT | Paraphrase | 2015 |
| Biblia Peshitta en Español | PES | Paraphrase | 2017 |
| Traducción del Nuevo Mundo | TNM | Paraphrase | 2025 |

## B Pairwise Comparisons

This appendix presents pairwise Mann–Whitney U tests (one-sided, Bonferroni-corrected) on chapter-level distances to the interlinear baseline. Each heatmap cell indicates whether the row's chapter-level distances are significantly smaller than the column's: *** indicates $p < 0.001$, ** indicates $p < 0.01$, * indicates $p < 0.05$, and blank cells are not significant.

Figure 3 shows comparisons at the strategy-group level. Figures 4–8 show comparisons between individual translations, ordered by strategy category.

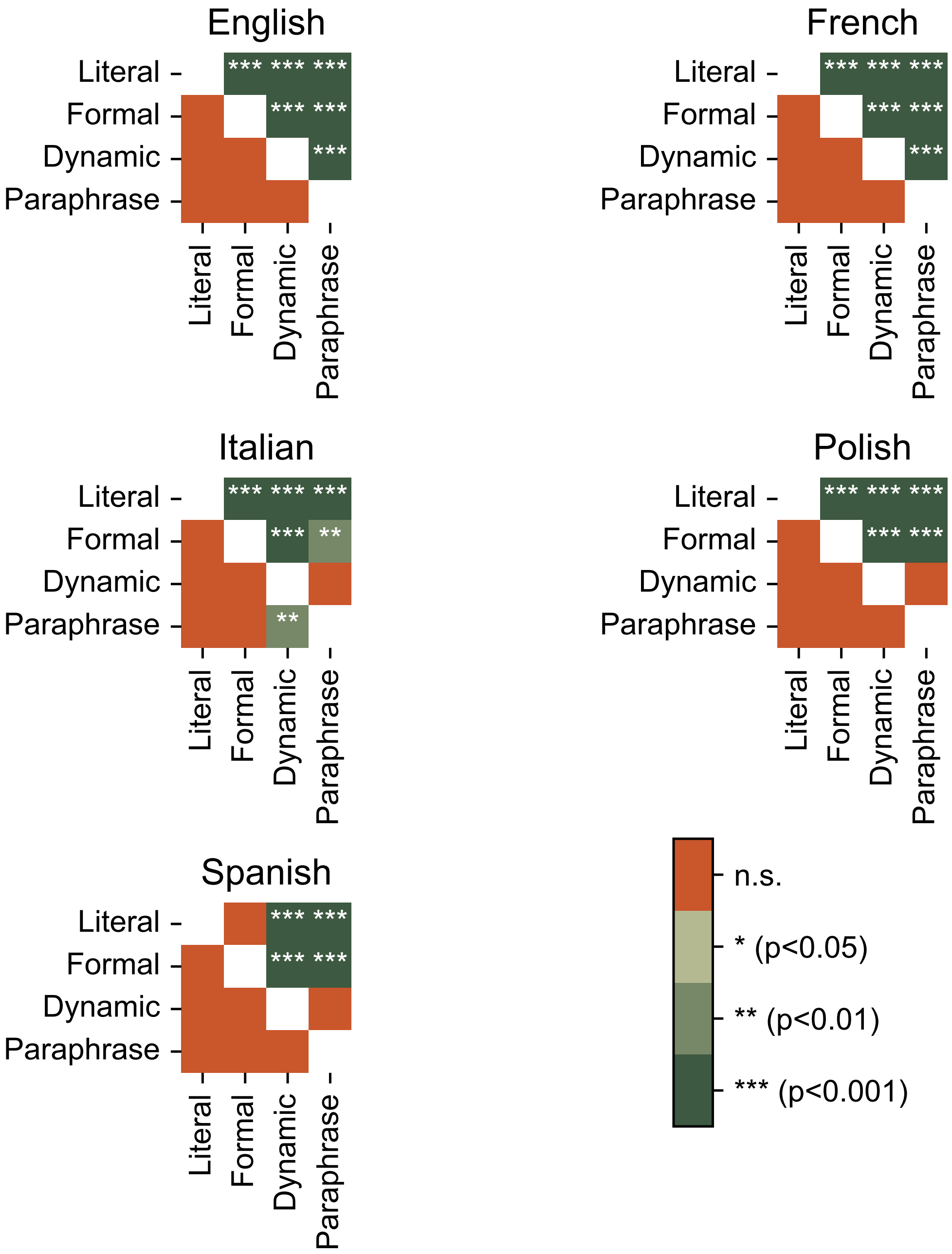

Figure 3: Strategy-group comparisons.

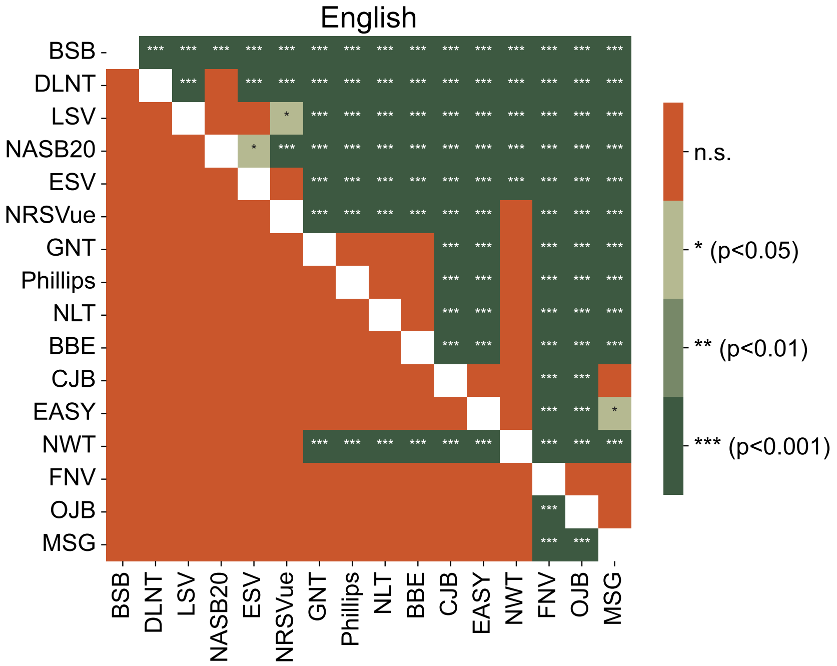

Figure 4: Pairwise comparisons of chapter intervention vector magnitudes for English translations.

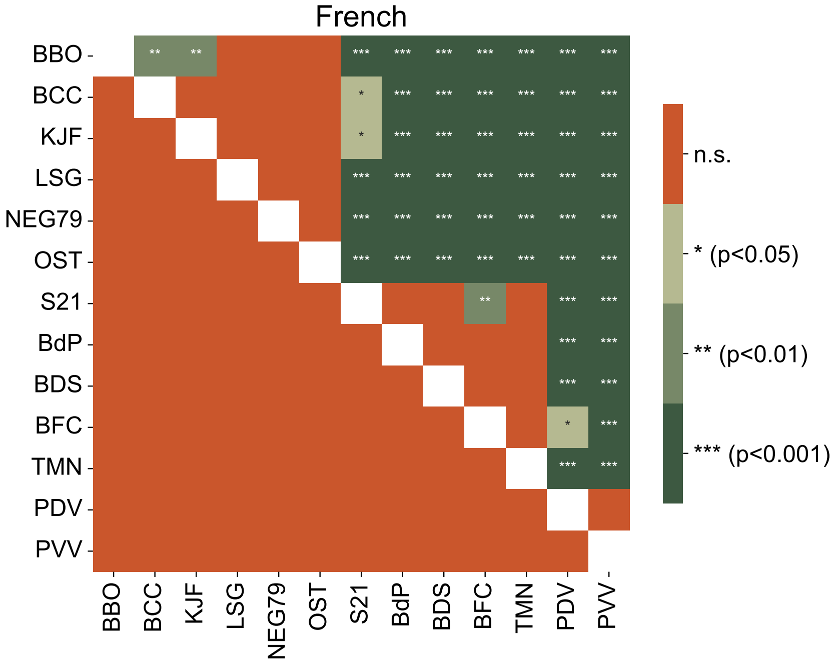

Figure 5: Pairwise comparisons of chapter intervention vector magnitudes for French translations.

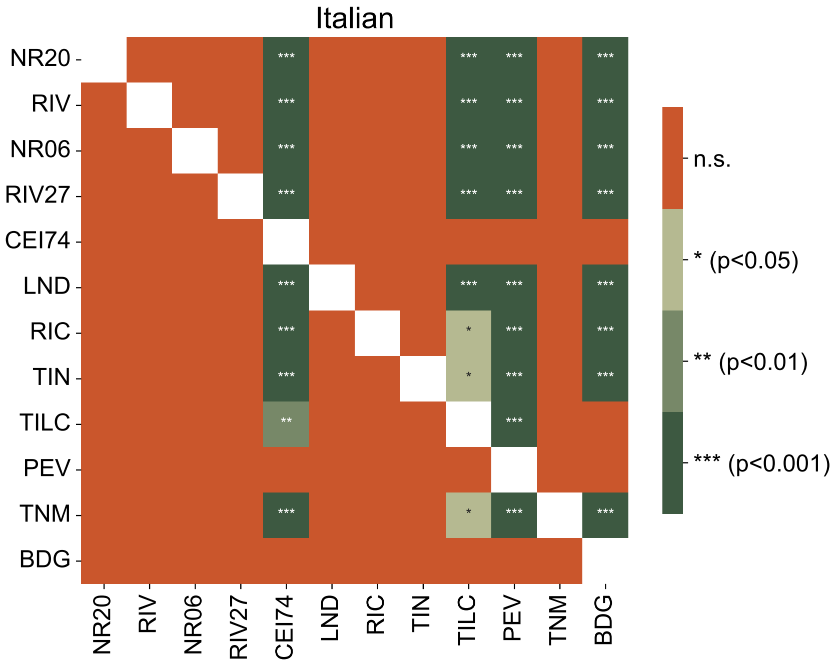

Figure 6: Pairwise comparisons of chapter intervention vector magnitudes for Italian translations.

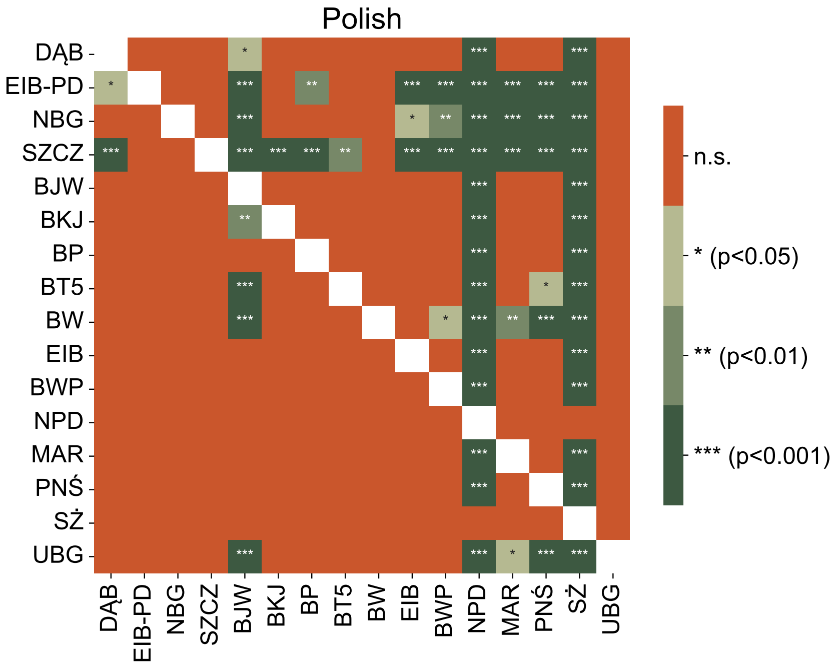

Figure 7: Pairwise comparisons of chapter intervention vector magnitudes for Polish translations.

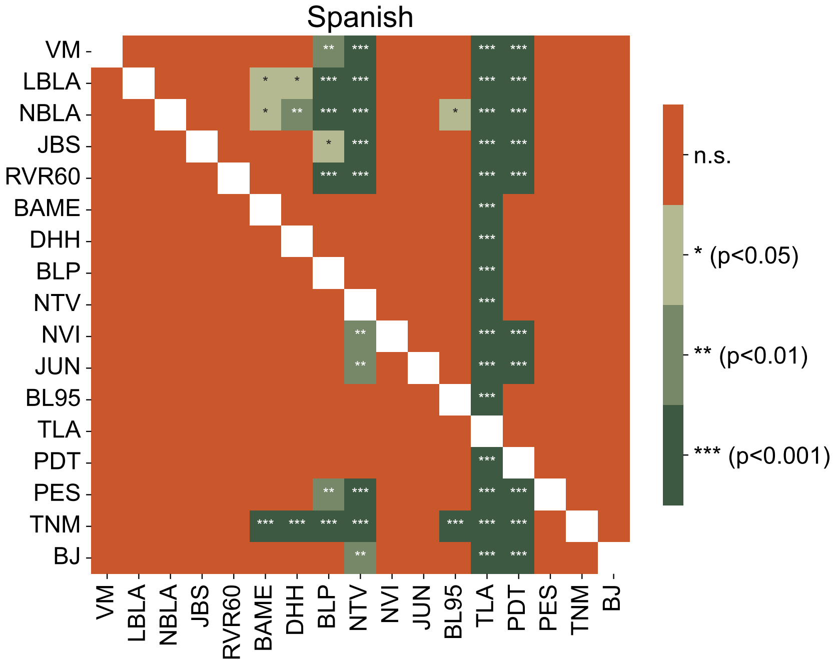

Figure 8: Pairwise comparisons of chapter intervention vector magnitudes for Spanish translations.