Spanish Corpus and Provenance with Computer-Aided Translation for the WMT24 OLDI Shared Task

Jose Cols
Department of Linguistics, University of Washington
jcols@uw.edu

Abstract

This paper presents the SEED-CAT submission to the WMT24 Open Language Data Initiative shared task. We detail our data collection method, which involves a computer-aided translation tool developed explicitly for translating SEED corpora. We release a professionally translated Spanish corpus and a provenance dataset documenting the translation process. The quality of the data was validated on the FLORES+ benchmark with English-Spanish neural machine translation models, achieving an average chrF++ score of 34.9.

1 Introduction

In recent years, the NLP community has made significant strides in reducing the data gap for hundreds of languages (Tiedemann, 2012; Bañón et al., 2020; Federmann et al., 2022; NLLB Team et al., 2022). Nonetheless, finding parallel corpora for machine translation and other NLP applications remains challenging for many language pairs (Haddow et al., 2022; Ranathunga et al., 2023). The WMT24 Open Language Data Initiative shared task aims to continue expanding language coverage with contributions from communities of native speakers.

This work describes our data collection method to expand the SEED dataset (NLLB Team et al., 2022; Maillard et al., 2023) with the Spanish language. Specifically, we focus on Latin American Spanish varieties to match the existing coverage of this language in the FLORES+ benchmark (NLLB Team et al., 2022). While the Spanish language benefits from the availability of multiple parallel corpora datasets (Aulamo et al., 2020), the majority of this corpora features other well-resourced languages such as English and French, and translation directions of regional significance to other languages like Asturian and Quechua remains a challenge (Oliver et al., 2023; Ahmed et al., 2023).

The multilingual alignment of the SEED dataset (Doumbouya et al., 2023) allows for the addition of a single corpus to enable dozens of translation directions into low-resource languages. Including Spanish represents an essential step toward incorporating other low-resource languages where finding English translators is challenging, as was the case for Ligurian, where half the data was translated from Italian (NLLB Team et al., 2022).

Considering the impact that high-quality parallel corpora can have on machine translation performance (Maillard et al., 2023), this work aims to facilitate extending the SEED dataset while supporting quality improvements in Spanish machine translation. Our main contributions are:

1. The expansion of the SEED dataset with professional translations of Latin American Spanish, created by native speakers, along with a neural machine translation baseline.
2. The open-source release of SEED-CAT, a web computer-aided translation tool explicitly designed to assist human translators in the translation of SEED files.
3. The automated generation and public release of a provenance dataset documenting the creation of each Spanish translation.

2 Background

Language overview According to a 2022 report from the Cervantes Institute,[1] there are more than 496 million native Spanish speakers in the world. Speakers are mainly concentrated in the Americas and the Iberian Peninsula, with Mexico having the largest population.

Spanish is an Ibero-Romance language that developed from Latin on the Iberian Peninsula. Thanks to its global expansion, this language has evolved into several dialectal variations. An example of this variation is the absence of the informal

[1] https://cvc.cervantes.es

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 624–635
November 15–16, 2024. ©2024 Association for Computational Linguistics

second-person plural ‘vosotros’ in Latin American Spanish, where most varieties use the pronominal form ‘ustedes’ to address speakers in both formal and informal contexts (Hualde et al., 2012).

Although there are social, phonological, and lexical variations, Spanish retains a fundamental cohesiveness (Hualde et al., 2012). The Royal Spanish Academy and the Association of Academies of the Spanish Language collaborate to publish a unified set of orthography, dictionaries, and other language resources. The Spanish writing system is based on the Latin script, with the addition of the character ⟨ñ⟩ forming an alphabet of 27 letters (Hernández Gómez, 2015). This script is represented in our collected data.

Seed dataset The SEED dataset (NLLB Team et al., 2022; Maillard et al., 2023), currently managed by the Open Language Data Initiative (OLDI),[2] contains 6,193 parallel sentences in English along with professional translations into 40 low-resource languages. The English sentences were originally sampled from thousands of Wikipedia articles across various categories such as arts, history, mathematics, people, and technology, offering diverse content from notable topics (NLLB Team et al., 2022). In this work, we use the English corpus (eng_Latn) from the SEED dataset as the source text for the Spanish translations.

FLORES+ benchmark FLORES+ is an evaluation benchmark for machine translation with support for 212 languages based on the initial FLORES-101 dataset (Goyal et al., 2022) and its recent expansions (NLLB Team et al., 2022; Doumbouya et al., 2023). The collection of this data involved a rigorous and iterative quality assurance process with professional translators, pre-defined standards, post-editing, and automatic quality assessments. We rely on this benchmark to assess the quality of the Spanish translations.

Computer-aided translation CAT, or computer-aided translation, refers to software tools, such as word processing, translation memory (TM), and terminology management, used by human translators to assist the translation process (Bowker and Fisher, 2010). Studies have shown that these tools can enhance the productivity and translation quality of human translators (Federico et al., 2012; Koehn, 2009). While machine translation differs from other CAT tools, as it is the primary driver of the translation (Bowker and Fisher, 2010), modern CAT suites often include machine translation as a key feature. According to a user survey study involving 736 translators (Zaretskaya et al., 2017), machine translation ranked as the third most commonly used functionality, following translation memory and terminology management. CAT users from that study and other usability surveys (Alotaibi, 2020; Vargas-Sierra, 2019) also reported dissatisfaction with the ease of use and learnability of these systems, highlighting the importance of user-friendly interfaces for computer-assisted translation.

3 Data Collection

Seed-CAT Various commercial CAT solutions exist, with SDL Trados, Memsource, and Wordfast being recognized as popular options by different research (Alotaibi, 2020; Picton et al., 2017). Apart from requiring purchasing a license, these systems use custom file formats that may not be compatible with other tools, leading to interoperability issues in translation projects. Using general-purpose software can also result in unaligned parallel sentences due to translators re-ordering the files, a problem highlighted by Doumbouya et al. (2023) in their review of the original NLLB-SEED dataset (NLLB Team et al., 2022). Furthermore, commercial CAT systems often integrate machine translation models, such as Google Translate and DeepL, that restrict the use of their outputs for training other models.

Recognizing these challenges, we release SEED-CAT,[3] an open-source web application specifically designed to assist human translators in translating SEED dataset files. This application was at the center of our data collection efforts and was designed with the three core principles.

- The user interface and features are optimized for usability, device compatibility, and seamless integration with the SEED dataset. The list of languages and corpora is fetched at runtime from the dataset’s repository, and metadata is displayed alongside each sentence (Figure 1).

- The system architecture facilitates application deployment, as it does not require configuring databases or user accounts. Data persistency is achieved via IndexedDB,[4] a transactional

[2] https://oldi.org/

[3] https://github.com/josecols/seed-cat

[4] https://www.w3.org/TR/IndexedDB/

database for object storage in web browsers.

- The application data model adheres to the W3C PROV-DM (Missier and Moreau, 2013) recommendation for data provenance, adding an additional layer of transparency to the translation creation process.

SEED-CAT integrates a focused set of features, such as machine translation and terminology consultation. Machine translation is supported for local[5] and remote inference, with the latter being recommended for broader device compatibility. The machine translation feature relies on the `facebook/nllb-200-distilled-600M` model (NLLB Team et al., 2022) to generate outputs. Likewise, terminology consultation in English is enabled by WordNet (Miller, 1995).

Users can also compare translations using text differencing (Myers, 2023), with word-based comparison for Latin-based languages and character-based comparison for other scripts. Additionally, part-of-speech color highlighting for English words can be toggled based on user preference. This feature relies on the Brill tagger (Brill, 1992) and a Treebank tokenizer (Marcus et al., 1993), both implemented using the `natural` library.[6]

Sourcing translators We recruited a team of ten freelance translators who were individually sourced through Fiverr, an online marketplace for digital services. We relied on the platform's reputation system and freelancer profiles to identify potential candidates. The final translators were selected based on specific criteria: native Latin American Spanish speakers, a minimum of two years of translation experience on the platform, at least 500 completed projects, and a brief English conversation assessment. The median translator had nine years of experience, 1,900 completed projects, and 779 reviews. In addition, we sourced an independent freelance translator with a degree in Applied Languages who underwent a similar vetting process. Additional background information on all translators is reported in Table 1.

Compensation Each translator determined their compensation separately based on the number of English words in their assigned task. The tasks were divided into two stages: *translation* and *review*, which had different compensation rates. The median compensation per translated word was 0.017 US dollars, with an average of 0.022, and the median compensation per reviewed word was 0.012 US dollars. Translators were also given a user guide to the SEED-CAT application, along with data samples and translation guidelines, to help them assess the complexity of the task when determining their rates.

Category	Detail	%
Education	Master's degree	9.1
	Bachelor's degree	54.5
	Course or certificate	27.3
	No formal training	9.1
Country	Argentina	9.1
	Chile	9.1
	Colombia	9.1
	Mexico	18.2
	Panama	9.1
	Venezuela	45.5

Table 1: Percentage distribution of participants by educational background in translation and country of origin.

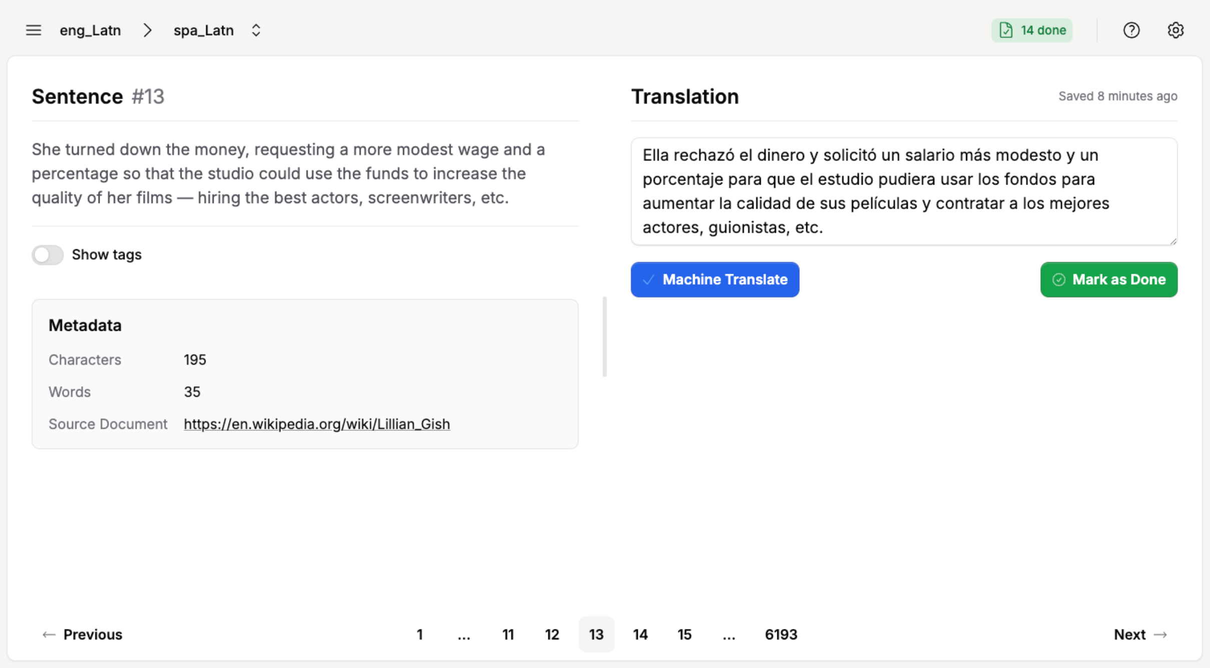
Translation workflow The translation process was divided into two stages: first, translating all sentences from English to Spanish, and second, reviewing every sentence to ensure accuracy and quality. Both phases were carried out by the team of translators using the SEED-CAT application.

A team of ten translators completed the initial translation phase in 16 days. The translators worked on a contiguous set of segments for better contextual reference, with an average task size of 593 sentences. Each translator received a unique URL to access SEED-CAT, which automatically configured their browser with the target language file (`spa_Latn`), sentence range, and user identifier. Translators did not need to handle administrative tasks such as creating user accounts or managing assignments. When they opened the application URL, they were prompted to review and acknowledge the translation guidelines, and then they were directed to their first assigned segment for translation.

The review phase was carried out by three translators. Each reviewed segments that were initially translated by others, finishing the task in five days and proofreading an average of 2,064 sentences. The translators received a specific URL to open SEED-CAT in `review` mode. In this mode, the application automatically loads and deserializes the

[5] https://github.com/xenova/transformers.js

[6] https://github.com/NaturalNode/natural/



She turned down the money, requesting a more modest wage and a percentage so that the studio could use the funds to increase the quality of her films — hiring the best actors, screenwriters, etc.

Ella rechazó el dinero y solicitó un salario más modesto y un porcentaje para que el estudio pudiera usar los fondos para aumentar la calidad de sus películas y contratar a los mejores actores, guionistas, etc.

Metadata

Characters	195
Words	35
Source Document	https://en.wikipedia.org/wiki/Lillian_Gish

Figure 1: SEED-CAT's user interface with two resizable panels to display the original sentence and the translation editor. Users can select different languages, track progress, review guidelines, or access other actions, such as importing/exporting provenance graphs using the top navigation bar. Translators can also open the source document and generate machine translations.

provenance information collected up to that point, enabling the consolidation of the translation and review history of a sentence into a single PROV-JSON file (Huynh et al., 2013). A total of 686 translations were copy-edited, with most corrections involving mistranslations, syntactic and lexical refinements, and grammatical issues such as verb agreement. Additionally, the decimal and thousand separators were standardized following established Spanish orthographic norms (Real Academia Española, 2010).

Dataset sample The final dataset contains 6,193 Spanish sentences (152,664 words) professionally translated by eleven native speakers from six countries in Latin America. Table 2 provides a brief excerpt of the parallel sentences.

Provenance dataset According to the World Wide Web Consortium (W3C) (Groth and Moreau, 2013), "provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness."

During the translation process, the SEED-CAT application automatically recorded provenance information on how activities such as `EditTranslation`, `MachineTranslate`, and `QueryWordNet` were used to generate, invalidate, and revise translations. This information can be serialized into JSON files (Huynh et al., 2013), enabling data sharing with its complete history. Users can also import these files and modify the data entities while maintaining the provenance's integrity.

In the PROV data model (Missier and Moreau, 2013), entities, activities, and agents are linked through relations. These links can be used to create a directed graph to visualize dependencies and data interactions. Appendix E provides examples of these graphs from the Spanish dataset. This dataset containing 6,193 PROV-JSON files is released as part of our SEED contribution.

System usability scale The system usability scale (SUS) (Brooke, 1996) is a standardized 10-item questionnaire for assessing perceived usability. Users rate each statement of the survey on a scale from 1 to 5, enabling the calculation of the SUS score, which ranges from 0 to 100. Substantial research has found that this score is a reliable metric of perceived system usability (Lewis, 2018). We administered the SUS questionnaire to the eleven translators involved in the project to evaluate the SEED-CAT application, obtaining an SUS score of

#	English	Spanish
663	For Gibbon, "The decline of Rome was the natural and inevitable effect of immoderate greatness.	Para Gibbon, "El declive de Roma fue el efecto natural e inevitable de la grandeza excesiva.
2079	By 1843 Richard Hoe developed the rotary press, and in 1844 Samuel Morse sent the first public telegraph message.	En 1843 Richard Hoe inventó la prensa rotativa, y en 1844 Samuel Morse envió el primer mensaje público por telégrafo.
5500	But mental ideas or judgments are true or false, so how then can mental states (ideas or judgments) be natural processes?	Pero las ideas o juicios mentales son verdaderos o falsos, entonces, ¿cómo pueden los estados mentales (ideas o juicios) ser procesos naturales?

Table 2: Sample sentences from the `spa_Latn` dataset with English source text and corresponding translations.

82.95. Appendix C summarizes the participants' responses.

4 Experimental Validation

Following the shared task's recommendation for experimental validation, we trained four bilingual machine translation models on the 6,193 newly collected Spanish sentences and evaluated their performance on the FLORES+ benchmark.

4.1 Data

Italic experiments To validate our model training setup, we reproduced the bilingual results reported by Maillard et al. (2023) for bidirectional translations between English and three Italic languages: Friulan (`fur_Latn`), Venetian (`vec_Latn`), and Ligurian (`lij_Latn`). We focus on these languages due to their linguistic relation to Spanish and their complete data availability on the SEED dataset and the FLORES+ benchmark.

Spanish experiments For our bidirectional English and Spanish (`spa_Latn`) machine translation models, we divided the collected data described in Section 3 into two versions: one before and one after the translation review process. This allowed us to analyze the scoring effect of a more streamlined review process based solely on proofreading and copy-editing in contrast to the iterative quality assurance pipeline implemented in FLORES-200 (NLLB Team et al., 2022). Table 3 summarizes the employed corpora with their corresponding source, size, and split.

Language	Split	Lines	Source

`eng`, `fur`
`lij`, `vec`
	train	6193	Seed

	valid	997	FLORES+
	test	1012	FLORES+
`spa`	**train**	**6193**	**This work**
	valid	997	FLORES+
	test	1012	FLORES+

Table 3: Corpora used in model experiments. Our contribution is highlighted in bold font.

4.2 Tokenization

We trained a SentencePiece model (Kudo and Richardson, 2018) on the `train` split for each language pair using a joined vocabulary of 8k tokens and byte-pair encoding (BPE) (Sennrich et al., 2016) for subword segmentation. In total, we trained three tokenizers for the bilingual Italic experiments and two tokenizers for the Spanish experiments, one for each version of the translated dataset.

4.3 Models

The machine translation models in this work are implemented with the `fairseq` toolkit (Ott et al., 2019) using the `transformer` architecture (Vaswani et al., 2017). Modifications are also made to match the bilingual model configurations in Maillard et al. (2023) for comparison purposes. The resulting architecture consists of 8 attention heads, 6 encoder and decoder layers, each with 4096-dimensional feedforward networks. We trained each model with an inverse square root learning rate of 0.001 and 400 warm-up updates. Training is conducted on a cloud virtual machine with an

NVIDIA L4 24GB GPU and an image preloaded with Debian 11, Python 3.10, PyTorch 1.13, and CUDA 11.3. Data preparation, model training, and evaluation recipes are available.[7]

Italic models We train two models per language pair, one for each direction (eng_Latn ↔ xxx_Latn) between English and the three selected Italic languages from the Seed dataset. We use the dev split of the FLORES+ benchmark for validation and the highest BLEU score (Papineni et al., 2002) on this split as the checkpoint selection criterion. Training is stopped when the validation BLEU score fails to improve after 10,000 gradient updates, and the selected checkpoint is used to calculate the chrF++ scores (Popović, 2015) on the FLORES+ devtest split. These scores provide a baseline for guiding our training parameters until achieving performance on par with the results reported by Maillard et al. (2023). This enables comparing the metrics from our Spanish models and assessing the quality of our spa_Latn data contribution.

Spanish models Using only the newly collected Spanish data, we trained four models, two for each version of the spa_Latn dataset. Model training and architecture parameters were defined during the Italic experiments and remained constant for these models. For validation, we used the dev split from the FLORES+ dataset. Training was conducted for 2,000 epochs (averaging a total runtime of 12h 31m), with the best checkpoint selected based on the highest validation BLEU score.

4.4 Results

Italic experiments We evaluated all model hypotheses using the sacrebleu tool (Post, 2018) against the devtest split of the FLORES+ benchmark. Table 4 compares the original performance of bilingual machine translation models reported by Maillard et al. (2023) with our reproduction attempts, which employed a similar model architecture and training routine. Our reproduced models nearly match the average chrF++ score for the English-to-Italic direction, falling short by 0.2 points while showing an improvement of 2.8 chrF++ points for the Italic-to-English direction.

Spanish experiments We achieved a chrF++ score of 35.0 for English-to-Spanish translation and 34.7 for the reverse direction by training exclusively on the collected spa_Latn data. The average score of 34.9 is comparable to the 35.1 mean obtained by the Italic models trained on existing SEED corpora. This result suggests that the new Spanish training data is representative of the spa_Latn data in the FLORES+ benchmark.

Language	Original		Reproduction	
	eng→	→eng	eng→	→eng
fur_Latn	35.4	35.6	35.7	36.8
lij_Latn	34.1	32.1	33.4	36.0
vec_Latn	33.5	32.3	33.2	35.5
Average	34.3	33.3	34.1	36.1

Table 4: Performance comparison (chrF++) between original (Maillard et al., 2023) and our reproduced bilingual models for three Italic languages (fur_Latn, lij_Latn, vec_Latn).

Analyzing the effect of the translation review process, we observed an average improvement of 0.3 chrF++ points. Specifically, the English-to-Spanish model trained on the reviewed data improved from 34.5 to 35.0, while the reverse direction decreased slightly from 34.8 to 34.7. Figure 2 breaks down the performance of the four bilingual Spanish models.

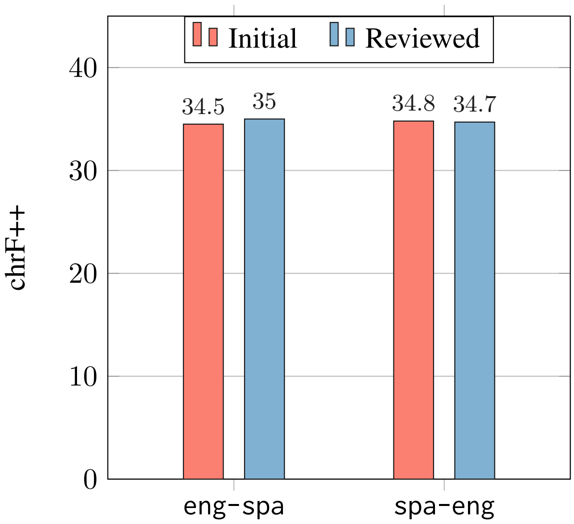


Figure 2: Performance (chrF++) of the eng_Latn↔ spa_Latn bilingual models trained on two versions of Spanish data: before (*Initial*) and after the translation review process (*Reviewed*).

5 Discussions

Seed English corpus Each line in the English corpus is an excerpt from a Wikipedia article,

[7] https://github.com/josecols/seed-cat/tree/main/nmt

which may consist of complete sentences or fragments. Translators identified two primary challenges when working with this data: incomplete sentences and a lack of context due to changes in the original article. For example, segment 5540, “`By way of example, they provide two proofs of the irrationality of .`” is missing an object at the end. Translations of such sentences inevitably reflect the original issues.

Furthermore, Wikipedia articles support versioning,[8] so including the date of compilation in the metadata or augmenting the dataset with provenance information could enable the correct context retrieval at the time of translation.

Seed-CAT The SEED-CAT application facilitated the translation of the English corpus, the review of translations, and the generation of the provenance dataset. Translators rated its usability highly, with an “A” grade based on the Sauro–Lewis curved grading scale (Lewis and Sauro, 2018). Notably, the system’s perceived learnability, identified in Alotaibi (2020) and Zaretskaya et al. (2017) as a key area for improvement in other CAT systems, scored the highest in our study, with an average of 93.2. This result underscores our efforts in user-centered design to make participation by language communities more accessible.

Translation workflow During the review phase, translators proofread and copy-edited the entire `spa_Latn` dataset, modifying 686 sentences and 1,815 words. Although the cost of this phase was lower than the initial translation, they were still comparable. Given the marginal improvement in the post-review model’s metric and the significant impact of high-quality parallel sentences on machine translation performance (Maillard et al., 2023), teams should consider allocating review resources toward generating more translations. In our case, this approach could have generated 3,498 additional translations of similar average length.

6 Conclusions

This work presented the SEED-CAT application and its role in expanding the SEED dataset with a professionally translated Spanish corpus. By integrating a provenance data model and its serialization in SEED-CAT, we automatically obtained a JSON dataset detailing the origin of each translation. Our experimental machine translation validation on the FLORES+ benchmark demonstrates that the collected Spanish data is of high quality, achieving on-par performance with other established language pairs in the SEED dataset. The excellent grade from the system usability scale survey suggests that the SEED-CAT application has the potential to facilitate the inclusion of additional languages in future efforts.

Limitations

To effectively support the expansion of the SEED dataset, localizing the SEED-CAT user interface is essential. Future translation projects may involve translators who work in languages other than English. Identifying these relevant languages and implementing the UI localization requires further work. Similarly, the machine translation feature is constrained by the availability of open models and their supported translation directions.

While the provenance dataset includes timestamps for when activities are performed, this information is not a reliable source for measuring the time taken to translate a sentence or other similar metrics. Users may experience interruptions, and the system does not track user engagement or attention.

Latin American Spanish varieties exhibit dialectal divisions that affect morphosyntactic features such as word order and verb tense (Hualde et al., 2012). Our data collection methodology does not distinguish between these variations. However, the specific variety spoken by each translator who participated in the project is detailed in Appendix D.

Acknowledgements

We thank the anonymous reviewers for their comments and suggestions, as well as the dedicated team of translators for their expertise and dedication in producing translations for this work.

References

Nouman Ahmed, Natalia Flechas Manrique, and Antonije Petrović. 2023. Enhancing Spanish-Quechua machine translation with pre-trained models and diverse data sources: LCT-EHU at AmericasNLP shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 156–162, Toronto, Canada. Association for Computational Linguistics.

[8] `https://en.wikipedia.org/wiki/Help:Page_history`

Hind M. Alotaibi. 2020. Computer-assisted translation tools: An evaluation of their usability among arab translators. *Applied sciences*, 10(18):6295–.

Mikko Aulamo, Umut Sulubacak, Sami Virpioja, and Jörg Tiedemann. 2020. OpusTools and parallel corpus diagnostics. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3782–3789, Marseille, France. European Language Resources Association.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Lynne Bowker and Desmond Fisher. 2010. Computer-aided translation. *Handbook of translation studies*, 1:60–65.

Eric Brill. 1992. A simple rule-based part of speech tagger. In *Third Conference on Applied Natural Language Processing*, pages 152–155, Trento, Italy. Association for Computational Linguistics.

J Brooke. 1996. Sus: A "quick and dirty" usability scale. *Usability Evaluation in INdustry/Taylor and Francis*.

Moussa Doumbouya, Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Abdoulaye Sow, Séré Moussa Doumbouya, Daouda Bangoura, Fodé Moriba Bayo, Ibrahima Sory 2. Condé, Kalo Mory Diané, Chris Piech, and Christopher Manning. 2023. Machine translation for nko: Tools, corpora, and baseline results. In *Proceedings of the Eighth Conference on Machine Translation*, pages 312–343, Singapore. Association for Computational Linguistics.

Marcello Federico, Alessandro Cattelan, and Marco Trombetti. 2012. Measuring user productivity in machine translation enhanced computer assisted translation. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers*, San Diego, California, USA. Association for Machine Translation in the Americas.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10.

Paul Groth and Luc Moreau. 2013. PROV-overview. W3C note, W3C. Https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Elena Hernández Gómez. 2015. Diccionario panhispánico de dudas.

Jose Ignacio Hualde, Antxon. Olarrea, and Erin. O'Rourke. 2012. *The handbook of Hispanic linguistics*, 1st ed. edition. Blackwell handbooks in linguistics. Wiley-Blackwell, Chichester, West Sussex [England] ;.

Trung Dong Huynh, Michael O. Jewell, Amir Sezavar Keshavarz, Danius T. Michaelides, Huanjia Yang, and Luc Moreau. 2013. The prov-json serialization. Project report, W3C.

Philipp Koehn. 2009. A process study of computer-aided translation. *Machine translation*, 23(4):241–263.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

James Lewis. 2018. The system usability scale: Past, present, and future. *International Journal of Human-Computer Interaction*, pages 1–14.

James R. Lewis and Jeff Sauro. 2018. Item benchmarks for the system usability scale. *J. Usability Studies*, 13(3):158–167.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzmán. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

George A Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Paolo Missier and Luc Moreau. 2013. PROV-dm: The PROV data model. W3C recommendation, W3C. Https://www.w3.org/TR/2013/REC-prov-dm-20130430/.

Eugene W. Myers. 2023. Ano(nd) difference algorithm and its variations. *Algorithmica*, 1(1–4):251–266.

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, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Antoni Oliver, Mercè Vàzquez, Marta Coll-Florit, Sergi Álvarez, Víctor Suárez, Claudi Aventín-Boya, Cristina Valdés, Mar Font, and Alejandro Pardos. 2023. TAN-IBE: Neural machine translation for the romance languages of the Iberian peninsula. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 495–496, Tampere, Finland. European Association for Machine Translation.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Aurélie Picton, Emmanuel Planas, and Amélie Josselin-Leray. 2017. Monitoring the use of newly integrated resources into cat tools: A prototype. In *Trends in e-tools and resources for translators and interpreters, Approaches to Translation Studies*, 45, pages 109–136. Brill Editions.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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, Belgium, Brussels. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

Real Academia Española. 2010. *Ortografía de la lengua española*. Espasa, Madrid.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Chelo Vargas-Sierra. 2019. Usability evaluation of a translation memory system. *Quaderns de Filologia - Estudis Lingüístics*, 24:119.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Anna Zaretskaya, Gloria Corpas Pastor, and Míriam Seghiri. 2017. Chapter 2: User perspective on translation tools: Findings of a user survey. In *Trends in E-Tools and Resources for Translators and Interpreters*, pages 37–56. Brill.

A Performance of bilingual models

Table 5 summarizes the performance on the FLORES+ benchmark of the bilingual machine translation models using both BLEU and chrF++ metrics.

B Machine translation examples

Table 6 presents three sample machine translations generated by the `eng-spa` model trained solely on the 6,193 reviewed translations of the `spa_Latn` data.

C System usability scale

Table 7 details the responses of each translator to the system usability scale (SUS) survey. The columns correspond to each numbered statement as they appear in the standard questionnaire (Brooke, 1996), while the rows represent the translators in no particular order. The table also summarizes the average score per participant, the score per question, and the total SUS score.

Model	BLEU		chrF++
	valid	test	test
eng→fur	10.3	10.4	35.7
eng←fur	10.8	10.0	36.8
eng→lij	7.5	8.0	33.4
eng←lij	9.6	9.3	36.0
eng→vec	7.0	6.3	33.2
eng←vec	9.9	9.4	35.5
Average	9.2	8.9	35.1
spa←eng	8.4	8.1	35.0
spa→eng	7.2	7.2	34.7
Average	7.8	7.7	34.9

Table 5: Performance of the bilingual models evaluated using automatic metrics on the `valid` and `test` splits.

D Spanish varieties

Table 8 relates each translator identifier in the provenance dataset with their specific Latin American Spanish variety.

E Provenance graphs

Figures 3 and 4 depict the provenance graphs of two translations. The translation process for each sentence can vary significantly, leading to graphs of different complexity. These graphs were generated using the Python `prov` package.[9]

[9] https://github.com/trungdong/prov

#	English	Spanish
663	This behavior oftentimes results in rifts between the leaders and the rest of the team.	Este comportamiento de los resultadosmes inestables entre los líderes de los líderes y el resto del equipo.
702	European influence and colonialism began in the 15th century, as Portuguese explorer Vasco da Gama found the Cape Route from Europe to India.	La influencia europea y el colonialismo comenzó en el siglo XV, como Portugués Verés Vasco Gama fundó la Cautela de Europa desde Europa.
1009	Japanese work culture is more hierarchical and formal that what Westerners may be used to.	La cultura japonés es más jerárquica y que se puede utilizar en los occidentales.

Table 6: Sample machine translations from the eng-spa bilingual model. The English source sentences are drawn from the devtest split of the FLORES+ benchmark.

User	1	2	3	4	5	6	7	8	9	10	*Score*
1	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	100.00
2	2.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	2.00	70.00
3	3.00	4.00	4.00	4.00	3.00	4.00	4.00	4.00	4.00	4.00	95.00
4	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	100.00
5	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	100.00
6	4.00	1.00	4.00	1.00	4.00	1.00	4.00	4.00	4.00	4.00	77.50
7	2.00	3.00	4.00	4.00	2.00	2.00	4.00	3.00	3.00	4.00	77.50
8	1.00	4.00	4.00	4.00	4.00	4.00	4.00	1.00	4.00	4.00	85.00
9	2.00	3.00	3.00	4.00	2.00	2.00	4.00	4.00	4.00	4.00	80.00
10	2.00	2.00	2.00	3.00	2.00	1.00	3.00	2.00	2.00	4.00	57.50
11	2.00	3.00	4.00	4.00	1.00	4.00	2.00	1.00	3.00	4.00	70.00
Score	68.18	79.55	90.91	88.64	75.00	75.00	90.91	77.27	88.64	95.45	**82.95**

Table 7: System usability scale scores for each translator (normalized).

Translator ID	Variety	Glottocode
14a33724-59b6-45f3-b056-f9d384e48a59	Caribbean Spanish	cari1288
2460a2a5-1a59-4e0a-afff-a83be7af3872	Caribbean Spanish	cari1288
d67b54ab-6325-47be-b578-02f4b7ba942c	Chilean Spanish	chil1286
599ec44e-1b13-4f0c-a71f-296bbf0f2c6a	Mexican Spanish	mexi1248
ef29b2b9-ecc8-4766-95a7-40b794d0053f	Mexican Spanish	mexi1248
548b0e62-71a4-448c-ab47-96f58f81a935	Rioplatense Spanish	riop1234
237fa953-c66e-4d5c-9f5a-919b171766be	Venezuelan Spanish	vene1262
142058e1-0375-4b16-bcc3-655af871ff1c	Venezuelan Spanish	vene1262
8fa01aed-835b-4912-b648-c86ae67e3599	Venezuelan Spanish	vene1262
250663c9-8d8e-43da-a116-840b8cf39cf4	Venezuelan Spanish	vene1262
e730a639-0928-4801-a97b-f070e661dff9	Venezuelan Spanish	vene1262

Table 8: Translator identifiers and their corresponding Latin American Spanish varieties with Glottocodes.

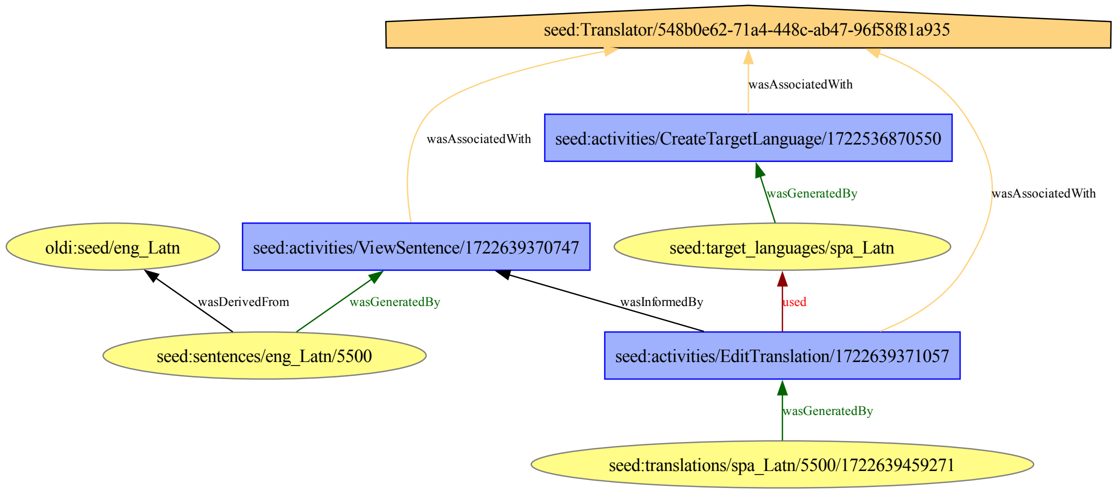

seed:Translator/548b0e62-71a4-448c-ab47-96f58f81a935
wasAssociatedWith
wasAssociatedWith
seed:activities/CreateTargetLanguage/1722536870550
wasGeneratedBy
wasAssociatedWith
oldi:seed/eng_Latn
seed:activities/ViewSentence/1722639370747
seed:target_languages/spa_Latn
wasDerivedFrom
wasGeneratedBy
wasInformedBy
used
seed:sentences/eng_Latn/5500
seed:activities/EditTranslation/1722639371057
wasGeneratedBy
seed:translations/spa_Latn/5500/1722639459271

Figure 3: This provenance graph represents a simple workflow in which the translator consulted the original English text and translated it into Spanish in a single, continuous edit.

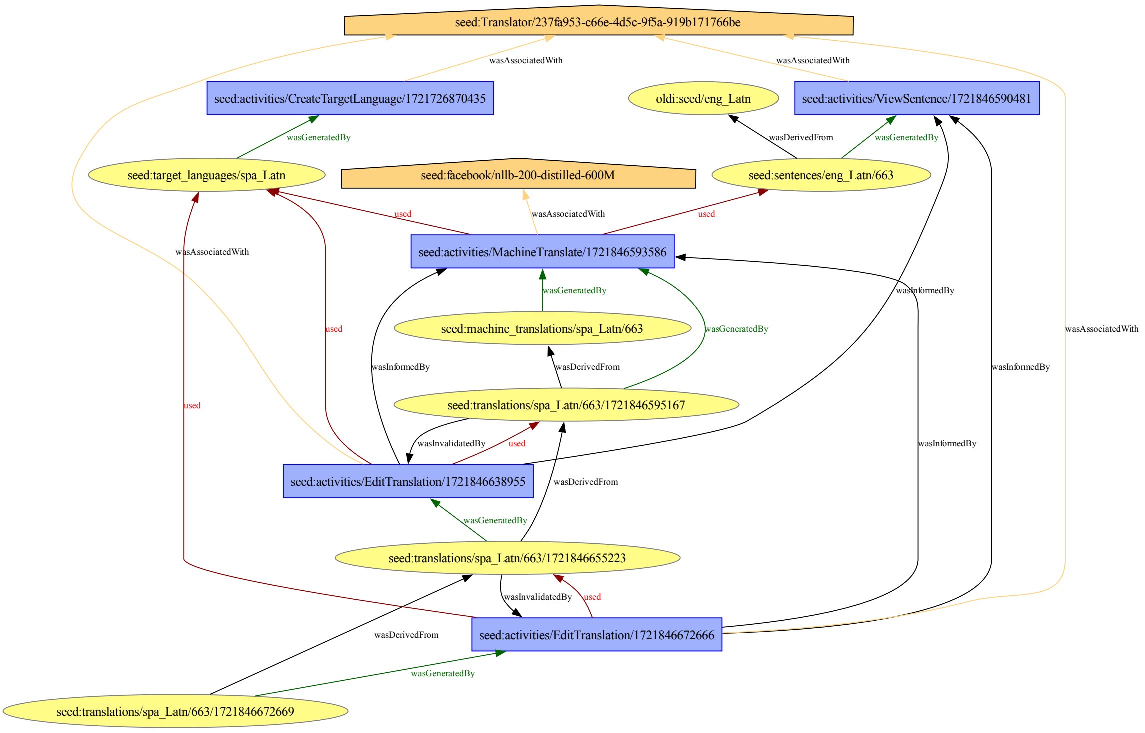

seed:Translator/237fa953-c66e-4d5c-9f5a-919b171766be
wasAssociatedWith
wasAssociatedWith
seed:activities/CreateTargetLanguage/1721726870435
oldi:seed/eng_Latn
seed:activities/ViewSentence/1721846590481
wasGeneratedBy
wasDerivedFrom
wasGeneratedBy
seed:target_languages/spa_Latn
seed:facebook/nllb-200-distilled-600M
seed:sentences/eng_Latn/663
used
wasAssociatedWith
used
wasAssociatedWith
seed:activities/MachineTranslate/1721846593586
wasGeneratedBy
wasInformedBy
used
seed:machine_translations/spa_Latn/663
wasGeneratedBy
wasAssociatedWith
wasInformedBy
wasDerivedFrom
wasInformedBy
used
seed:translations/spa_Latn/663/1721846595167
wasInvalidatedBy
used
wasInformedBy
seed:activities/EditTranslation/1721846638955
wasDerivedFrom
wasGeneratedBy
seed:translations/spa_Latn/663/1721846655223
wasInvalidatedBy
used
wasDerivedFrom
seed:activities/EditTranslation/1721846672666
wasGeneratedBy
seed:translations/spa_Latn/663/1721846672669

Figure 4: This provenance graph represents a workflow that begins with an initial machine translation, followed by two rounds of copy-editing.