Adapting Multilingual Embedding Models to Historical Luxembourgish

Andrianos Michail
University of Zurich
andrianos.michail@cl.uzh.ch

Corina Julia Raclé
University of Zurich
corinajulia.racle@uzh.ch

Juri Opitz
University of Zurich
jurialexander.opitz@cl.uzh.ch

Simon Clematide
University of Zurich
simon.clematide@cl.uzh.ch

Abstract

The growing volume of digitized historical texts requires effective semantic search using text embeddings. However, pre-trained multilingual models face challenges with historical content due to OCR noise and outdated spellings. This study examines multilingual embeddings for cross-lingual semantic search in historical Luxembourgish (LB), a low-resource language. We collect historical Luxembourgish news articles from various periods and use GPT-4o for sentence segmentation and translation, generating 20,000 parallel training sentences per language pair. Additionally, we create a semantic search (Historical LB Bitext Mining) evaluation set and find that existing models perform poorly on cross-lingual search for historical Luxembourgish. Using our historical and additional modern parallel training data, we adapt several multilingual embedding models through contrastive learning or knowledge distillation and increase accuracy significantly for all models. We release our adapted models and historical Luxembourgish-German/French/English bitexts to support further research. [1]

1 Introduction

Exploration possibilities of historical texts, such as newspapers, have advanced rapidly due to digitization efforts by libraries and archives (Ehrmann et al., 2023a). Traditionally, tools relied on keyword-based searches, often enhanced with semantic enrichment techniques such as named entity recognition (Ehrmann et al., 2023b).

Recent embedding benchmarks (Muennighoff et al., 2023; Enevoldsen et al., 2025) show that massively multilingual embedding models, trained on diverse multilingual corpora, perform well in both multilingual and cross-lingual semantic search. These models have also become integral in Retrieval-Augmented Generation (RAG), where they help retrieve more relevant and contextually appropriate documents, thereby improving the faithfulness of generated responses.

However, for low-resource languages like Luxembourgish (LB), where multilingual models have limited exposure, their performance remains uncertain. Applying these models to semantic search in imperfectly digitized historical collections introduces additional challenges, as they must handle OCR errors and historical spelling variations. The disparity between these noisy, historical texts and the clean, modern digital-born data used to train multilingual models, combined with their limited support for Luxembourgish, complicates the development of effective exploration tools for historical Luxembourgish newspaper archives.

To address this issue, we compile 2,338 historical Luxembourgish news articles from different time periods and use GPT-4o to segment and translate them into modern French (FR), English (EN) and German (DE). The resulting parallel sentences serve as fine-tuning data to adapt existing multilingual embedding models for imperfectly digitized historical Luxembourgish.

Our main contributions:
(1) We adapt multilingual embeddings for digitized historical Luxembourgish by generating training data through a prompt-based translation approach with GPT-4o.
(2) We define a historical bitext mining task and create a high-quality cross-lingual semantic search test set with 233 source news articles (LB-DE: 2,127; LB-FR: 2,157; LB-EN: 2,105 sentences).
(3) We fine-tune and evaluate off-the-shelf models – *M-MPNet* (Reimers and Gurevych, 2020), *LaBSE* (Feng et al., 2022), *M-GTE* (Zhang et al., 2024), and *LuxEmbedder* (Philippy et al., 2025) – to as-

[1] See https://github.com/impresso/histlux_emb for our released models, data and source code.

Proceedings of LaTeCH-CLfL 2025, pages 291–298

sess our adaptation methods.
(4) We propose and evaluate a 1:1 data mixing strategy that balances noisy historical texts with clean modern texts to minimize performance degradation on modern Luxembourgish benchmarks.

2 Related Work

This section reviews relevant embedding models that support Luxembourgish semantic search, including monolingual Luxembourgish models and multilingual embeddings.

Reimers and Gurevych (2020) use knowledge distillation through a strong paraphrase-trained English embedding model and parallel data to create cross-lingually aligned models. Multiple instances of such models have been open sourced and a particularly powerful and popular one is *paraphrase-multilingual-mpnet-base-v2* (**M-MPNet**) which was trained on over 50 languages. Later within this work, we will explain how we extend this model to also support Luxembourgish.

The multilingual bitext mining model **LaBSE** (Feng et al., 2022) is trained with translation ranking loss and negative samples. It has been trained roughly on less than 100 Luxembourgish-English sentence pairs and specializes in zero-shot bitext mining.

A recent model is GTE Multilingual (**M-GTE**) (Zhang et al., 2024), a multilingual embedding model designed for long context text representation and reranking. *M-GTE* has been trained with hard negatives and has included 50,000 Luxembourgish pairs within its contrastive pre-training.

Specific model adaptations to Luxembourgish have also been developed. One example is **LuxemBERT** (Lothritz et al., 2022), a monolingual BERT model pre-trained for Luxembourgish using augmented data, partially generated by translating texts from closely related languages and incorporating relevant text sources.

Closely related to our work, **LuxEmbedder** (Philippy et al., 2025) used OpenAI's *text-embedding-3-small* and *LaBSE* to mine a set of parallel sentences for each pair of languages between Luxembourgish, English, and French. These parallel sentences (up to 20,000 per pair) were then used to further fine-tune *LaBSE*, improving performance on modern Luxembourgish evaluation sets. However, its ability to handle Luxembourgish texts from different historical periods—potentially affected by digitization errors common in large-scale historical text collections, remains unclear.

Our work aims to extend existing embedding models to better perform cross-lingual semantic search within a collection of historical, OCR-noisy Luxembourgish texts. The conditions of these texts combined with the different spelling variations[2] poses an interesting generalization challenge to the models.

3 Method

To adapt and evaluate embedding models for digitized historical Luxembourgish news articles, we create parallel texts by translating them into modern German, French, and English. This allows the models to learn cross-lingual representations and improves their ability to align historical Luxembourgish with contemporary languages for semantic search.

3.1 Parallel Historical Luxembourgish

We build our translated parallel data sets LB-DE, LB-FR and LB-EN from monolingual Luxembourgish texts sourced from the publicly available BNL newspaper archive.[3] Our data consists of articles from newspapers published between 1841 and 1948. To select diverse samples for translation, we first cluster the articles into 2,000 groups by K-Mean on a 100-topics LDA model output[4] and keep the 605 clusters with more than 20 articles.

We select articles through a two-step process, resulting in a total of 2,340 articles, as shown in Figure 1. First, we retrieve the most representative article from each cluster, ensuring it contains between 5 and 20 sentences (cutting of the remaining sentences). In a second round, we randomly sample three additional articles per cluster under the same length conditions.

We prompt GPT-4o to segment historical Luxembourgish articles and generate sentence-level translation pairs separately for German, French and English (see Prompt 2). The model is instructed to preserve the original meaning and structure as closely as possible while reconstructing sentences affected by OCR errors that could hinder translation. This process yields approximately 22,500 sentence pairs for the LB-DE, LB-FR and LB-EN pair. Notably, GPT-4o appears to perform sentence

[2]Luxembourgish had no standardized spelling until 1946 and underwent multiple further reformations (eg. in 1999)

[3]https://data.bnl.lu/data/historical-newspapers

[4]Taken from the impresso-project.ch (Ehrmann et al., 2020).

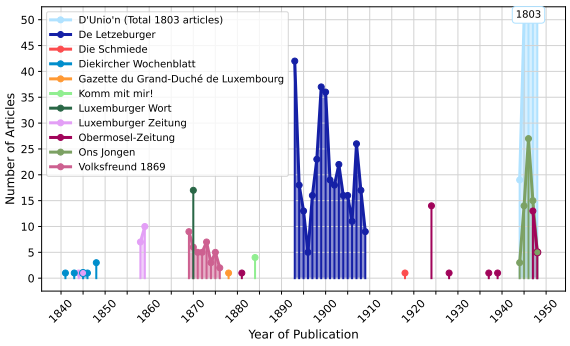

Figure 1: Source LB articles per newspaper per year.

segmentation consistently resulting in 65.0% of sentences forming exact quadruplets (4-way parallel) across the four languages.

To ensure fidelity to the original articles, we calculate the percentage of regenerated Luxembourgish sentences that do not exactly match their source texts. These account for 1.4% of all historical Luxembourgish sentences per language pair, which we manually correct. Most mismatches result from missing or added punctuation, modernized spelling, and, in rare cases, errors caused by the LLM not adhering to the instructed format.

To assess translation quality, a quadrilingual native speaker of Luxembourgish (LB) annotated 100 randomly selected sentence quadruplets after removing 15 samples with severe OCR problems. Of the 100 LB sentences presented without context, 88 were judged to be comprehensible or at least confidently guessable (23). The remaining 12 were considered incomprehensible due to OCR errors and archaic spellings, and their translations were not evaluated.

For the comprehensible and confidently guessable sentences, the German translations were rated as adequate in 78 cases (88.6%), with missing minor details in 9 cases and 1 case of inadequate translation. The French translations showed a similar pattern: 78 were adequate (88.6%), 9 were missing minor details, and 1 was inadequate. The English translations were also adequate in 78 cases (88.6%), with 10 missing minor details. A sample of the annotated dataset is available in the appendix (Table 2).

3.2 Framing an Evaluation Task: Historical LB Bitext Mining

From our parallel dataset, we set aside a held-out test set of 233 articles (2,127 sentences) to establish a historical semantic search benchmark for Luxembourgish-to-German, French, and English bitext mining (LB<->DE/FR/EN). A prediction is considered a true positive if the embedding model assigns a higher similarity to the correct parallel sentence than to any of the 2k alternative candidates. We report the bidirectional average accuracy. To minimize false negatives caused by near-identical sentences, we exclude candidate sentences with a Levenshtein similarity score above 0.85 to the source sentence, after removing non-alphanumeric characters from both. This filtering affects 57 source-candidate pairs (2.7%) in German, 65 (3%) in French, and 76 (3.6%) in English. A human review at different thresholds confirms the appropriateness of the filtering process and the chosen threshold.

3.3 Modern LB Evaluation Tasks

We replicate two evaluation tasks on modern Luxembourgish from Philippy et al. (2025).

ParaLux is a monolingual paraphrase detection test set designed to evaluate embedding models. Performance is measured by the proportion of cases (a total of 312 triplets) in which an embedding model assigns a higher similarity to an anchor-positive pair than to an anchor-negative pair. The negative sentences are adversarially generated to maintain high lexical similarity and manually verified to ensure they are true negatives.

SIB 200 (LB) is a repurposed subset of the 'Flores' dataset (NLLB Team et al., 2022; Adelani et al., 2024), used for monolingual zero-shot topic classification. In this task, texts are assigned to template sentences representing candidate topics based on embedding similarity.

3.4 Adapting Multilingual Embedding Models to Historical LB

3.4.1 Datasets

*Hist*orical: We use 2,105 historical LB newspaper articles (excluding held-out articles) with their sentence-level translations to create a parallel training set for the following language pairs: LB-DE (20,092), LB-FR (20,010), and LB-EN (19,054) sentences.

Modern: Philippy et al. (2025) extracted 89,405 LB-FR and 28,172 LB-EN parallel sentence pairs from RTL.lu, a trilingual news platform. This dataset was used to fine-tune the *LuxEmbedder* model.

Model	Training Data	Historical LB Bitext Mining				Modern LB		
		LB↔FR	LB↔EN	LB↔DE	AVG	SIB 200 (LB)	ParaLux	AVG
Random Baseline	–	00.00	00.00	00.00	45.97	14.28	50.00	32.14
text-embedding-3-small	–	78.36	75.08	82.33	78.59	40.20	15.71	27.96
text-embedding-3-large	–	86.18	83.63	88.15	85.99	58.82	26.28	42.55
M-MPNet	–	46.32	45.04	46.55	45.97	24.71	26.60	25.66
M-MPNet(+LB Distilled)	LB↔EN (Hist)	87.23	87.53	89.14	87.97	42.65	56.09	49.37
	LB↔EN (Modern)	75.55	77.03	78.09	76.89	**59.51**	80.13	69.82
	LB↔EN (Mixed)	**89.32**	**89.55**	**91.44**	**88.79**	59.41	**80.45**	**70.48**
LaBSE	–	93.12	95.27	94.01	94.13	43.24	**38.14**	40.69
LaBSE (Hist)	LB↔FR	**97.73**	97.22	98.10	97.68	39.61	25.00	32.31
	LB↔EN	97.24	97.44	97.96	97.54	41.76	22.44	32.10
	LB↔DE	97.08	97.01	**98.52**	97.54	34.02	14.74	24.38
LaBSE (Mixed)	LB↔FR	97.40	**97.55**	98.22	97.35	**45.69**	31.73	47.66
	LB↔EN	96.80	97.34	97.82	**97.75**	45.59	36.86	**50.23**
LuxEmbedder	–	84.49	85.09	85.48	85.02	**65.59**	**52.24**	**58.92**
LuxEmbedder (Hist)	LB↔FR	**97.47**	97.51	98.24	97.74	50.39	32.37	41.38
	LB↔EN	97.18	97.29	98.26	97.58	54.12	28.85	41.49
	LB↔DE	97.25	**97.72**	**98.43**	**97.80**	46.76	26.60	36.68
LuxEmbedder (Mixed)	LB↔FR	96.97	97.32	97.77	97.72	56.86	38.46	38.71
	LB↔EN	97.41	97.58	98.26	97.32	56.86	43.59	41.23
M-GTE	–	83.68	80.12	87.55	83.78	55.78	**70.51**	63.20
M-GTE (Hist)	LB↔FR	95.18	94.23	96.05	95.15	59.12	57.05	58.09
	LB↔EN	**95.81**	95.56	96.52	**95.96**	54.71	55.77	55.24
	LB↔DE	95.23	94.61	**97.65**	95.83	45.29	42.31	43.80
M-GTE (Mixed)	LB↔FR	95.53	95.11	96.78	95.80	60.98	60.26	60.62
	LB↔EN	95.48	**95.58**	96.55	95.87	**67.84**	64.10	**65.70**
M-GTE (Hist, Modern: 120k)	LB↔DE/FR/EN	**96.83**	**97.15**	**97.93**	**97.30**	62.16	62.82	62.75

Table 1: Performance (accuracy) of the examined models and our adapted variants within the Historical and Modern Luxembourgish evaluation sets. The last row shows an adapted model trained on the maximum available data, with details found at the end of Section 4.

Training Data Configurations We investigate three data mixing strategies for model training:
(1) *Hist*orical: 20,000 translated sentence pairs (LB↔FR, LB↔EN, or LB↔DE) from historical texts.
(2) *Modern*: 20,000 bitext-mined sentence pairs (LB↔FR or LB↔EN) from modern Luxembourgish news.
(3) *Mixed*: 20,000 *Hist* sentence pairs with 20,000 *Modern* sentence pairs in mixed batches.

3.4.2 LB Knowledge Distillation

We adapt *M-MPNet* for historical LB using multilingual knowledge distillation (Reimers and Gurevych, 2020). The original English model *paraphrase-mpnet-base-v2* serves as a frozen teacher, while LB-EN parallel sentences are used to train the *M-MPNet* student model to embedded LB sentences similar to their English translations. We fine-tune *M-MPNet* for five epochs using each of the three data mixing strategies: (1) *Hist*orical, (2) *Modern*, and (3) *Mixed*.

3.4.3 Contrastive Loss

We adapt *LaBSE*, *LuxEmbedder*, and *M-GTE* to historical LB using contrastive learning. Specifically, we fine-tune these embedding models using *MultipleNegativesRankingLoss* (Henderson et al., 2017), with a batch size of 8 for one epoch. For fine-tuning, we apply two of the previously defined data mixing strategies: (1) *Hist*orical and (3) *Mixed*.

4 Results

Table 1 shows the performance of the off-the-shelf and adapted models on the historical Luxembourgish bitext mining and the modern LB evaluation tasks.

Among the off-the-shelf(in cursive) models, *LaBSE* is the strongest model in all three languages. Surprisingly, *LuxEmbedder*, a LB-tuned version of *LaBSE*, shows an average performance drop of 9pp across language pairs in our bitext mining task, despite improved performance on the modern LB tasks. Similarly, *M-GTE* underperforms *LaBSE*

by 10.4pp. Both OpenAI embedding models (*text-embedding-3-small/large*) show moderate performance.

Among the models contrastively adapted using the *Hist* pairs, the performance in historical bitext mining improves significantly, reaching up to 97.8% accuracy. Notably, after domain adaptation, *LuxEmbedder* matches the adapted *LaBSE*, reaching over 97.8% accuracy and closing the performance gap observed in the standard models. Meanwhile, the customized *M-GTE* models lag behind by about 2pp. Interestingly, across all model architectures, training on any language pair improves performance similarly across all pairs, showing a positive cross-lingual transfer.

These models experience significant performance drops in modern LB evaluations, particularly in *ParaLux*. However, adapting these models with mixed batches of *Hist* and *Modern* sentence pairs partially mitigated performance loss on the Modern LB evaluation tasks. Within *LaBSE* and *M-GTE*, this adaptation even improved the performance on SIB-200 topic classification, while sacrificing only up to 1% of the performance on historical bitext mining. These mixed-data adapted models provide an overall stronger general backbone for cross-lingual semantic searching within collections containing both historical and modern Luxembourgish.

The *M-MPNet* model, before distilling EN-LB knowledge, performs poorly in all LB evaluations, despite its proven exact matching capabilities in other languages, confirming its lack of support for the language. After distilling LB with any dataset, the model performs magnitudes better across the board. When distilled with a single dataset, the model performs best on historical semantic search tasks with the *Hist* sentence pairs. In contrast, when distilled with *Modern* sentence pairs, the model excels on modern LB tasks, achieving 80% accuracy on ParaLux[5] and outperforming the second-best *M-GTE*, which achieves 70%. Finally, distilling with the mixed data set yields the best results in all evaluations, demonstrating the synergy between the two sources.

However, even the *Mixed*-data distilled *M-MPNet* model only achieves an average accuracy across pairs of 90% in historical bitext mining, trailing the contrastive domain-adapted *LaBSE* and *LuxEmbedder* by 8pp and the off-the-shelf *LaBSE* model by 4pp.

The Final Model: Mix it All For a final all-purpose model covering both historical and modern LB, we contrastively adapt *M-GTE* to all language pairs of *Hist* while preserving an equal number of *Modern* sentence pairs, regardless of language. The adaptation dataset consists of 20,000 LB-FR/EN/DE (*Hist*), 20,000 LB-EN (*Modern*), and 40,000 LB-FR (*Modern*), for a total of 120,000 sentence pairs.

It is the best-performing historical semantic search *M-GTE* model, achieving an average accuracy of 97.5% across all language pairs. This model outperforms the adapted *LaBSE* and *LuxEmbedder* models on SIB-200 (+6pp) and ParaLux (+20pp), while performing similarly to them in the historical bitext mining evaluations.

5 Conclusions

In this work, we explore the adaptation of multilingual embedding models to digitized historical LB texts, a task where off-the-shelf models struggle due to limited exposure and their reliance on clean modern data. To address this issue, we generate parallel sentence-segmented documents by translating historical Luxembourgish newspaper articles into French, English, and German using GPT-4o.

To evaluate the effectiveness of adaptation, we design a historical bitext mining task with a held-out test set of 233 articles. Our results show that adaptation to parallel historical data improves retrieval accuracy by up to 13pp. However, this adaptation introduces trade-offs, particularly reducing performance on modern LB tasks that require high semantic precision, such as adversarial paraphrase detection. We mitigate this problem through a balanced data mixing strategy that helps preserve modern LB performance while improving historical text semantic search capabilities.

These results demonstrate the effectiveness of domain adaptation for historical text processing and suggest that such approaches could benefit low-resource languages facing digitization challenges. Such improvements are particularly relevant for libraries and archives, where effective cross-lingual semantic search can improve the discoverability of historical documents and support digital exploration.

[5]As shown in recent work (Michail et al., 2025), results on adversarial paraphrase discrimination test sets might not accurately reflect performance on semantic search in general. Therefore, this result should be interpreted with caution.

Limitations

Our findings paves the way for better semantic search systems within Luxembourgish archives. On the one hand, our method demonstrates clear benefits for the targeted use case, effectively embedding heterogeneous digitized historical texts and revealing shortcomings in off-the-shelf models. Through our exploration of adaptation methodologies, we have produced practical embeddings for semantic search while mostly preserving modern LB performance. On the other hand, we have not strictly reached a single best model across all evaluation sets. For example, in all of our adapted models, performance on *ParaLux* declines, possibly indicating interference with modern LB understanding and reduced sensitivity to semantic nuances.

Overall, we have applied a single adaptation method for each model type across all available data mixes, ensuring alignment with the models' initial training methods. Exploring alternative adaptation approaches may reveal additional patterns.

One problem with our evaluation is that they are all at the sentence level, whereas applications of such models would often be at the paragraph, article, or document level. The hypothesis that our improved performance would be reflected when embedding longer segments of text is possible, but not guaranteed. Lastly, while our research focuses on historical Luxembourgish, our methodology may also be useful for developing semantic search models in other underrepresented languages, which we do not examine in this study.

Acknowledgments

We would like to thank Fred Phillipy for helping with the human annotation. This research is conducted under the project *Impresso – Media Monitoring of the Past II Beyond Borders: Connecting Historical Newspapers and Radio*. Impresso is a research project funded by the Swiss National Science Foundation (SNSF 213585) and the Luxembourg National Research Fund (17498891).

References

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Maud Ehrmann, Estelle Bunout, and Frédéric Clavert. 2023a. *Digitised Historical Newspapers: A Changing Research Landscape*, pages 1–22. De Gruyter Oldenbourg, Berlin, Boston.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023b. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, 56(2).

Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströbel, and Raphaël Barman. 2020. Language resources for historical newspapers: the impresso collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 958–968, Marseille, France. European Language Resources Association.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, (...), and Niklas Muennighoff. 2025. MMTEB: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *Preprint*, arXiv:1705.00652.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2022. LuxemBERT: Simple and practical data augmentation in language model pre-training for Luxembourgish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5080–5089, Marseille, France. European Language Resources Association.

Andrianos Michail, Simon Clematide, and Juri Opitz. 2025. PARAPHRASUS: A comprehensive benchmark for evaluating paraphrase detection models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8749–8762, Abu Dhabi, UAE. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Fred Philippy, Siwen Guo, Jacques Klein, and Tegawende Bissyande. 2025. LuxEmbedder: A cross-lingual approach to enhanced Luxembourgish sentence embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11369–11379, Abu Dhabi, UAE. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

A Appendix

System:

You are a professional translator specializing in the translation of historical Luxembourgish newspaper articles into modern Standard {German/French/English}.
Your task is to translate paragraphs from such newspapers, provided to you by the user. These paragraphs may contain old spellings, outdated expressions, and likely a lot of OCR errors, as they are extracted from 19th-century LB newspapers. Please translate each sentence individually into modern Standard {German/French/English}. Prioritize retaining the original meaning, expressions, and any nuanced tone in each translation, even if the result sounds somewhat unconventional or even bad in {German/French/English}. If an expression is ambiguous due to its historical nature or OCR errors, attempt to reconstruct the most probable meaning based on linguistic context. Ensure that all punctuation and whitespace is preserved exactly. Do not add any extra formatting such as backticks, markdown, or additional symbols.

Please return the source sentences and your translations in the following format as JSON:

```
{"translation": [
{"lb": "lb_sent1", "{de}": "{de}_sent1"},
{"lb": "lb_sent2", "{de}": "{de}_sent2"},
{"lb": "lb_sent3", "{de}": "{de}_sent3"},
...]}
```

Figure 2: Zero-shot prompt template given to GPT-4o for the segmentation and translation of historical Luxembourgish newspaper articles to modern French(fr)/English(en)/German(de).

Newspaper	Year	Sentence	LB Compr.	FR Compr.	EN Compr.	DE Compr.
D'Union	1946	**LB**: Si bestét aus 18 000 tuben, weit 30 tonnen a kascht 400 000 dollar. **FR**: Elle est composée de 18 000 tubes, pèse 30 tonnes et coûte 400 000 dollars. **EN**: It consists of 18,000 tubes, weighs 30 tons, and costs 400,000 dollars. **DE**: Sie besteht aus 18.000 Röhren, wiegt 30 Tonnen und kostet 400.000 Dollar.	Comprehensible	Adequate	Adequate	Adequate
Ons Jorgen	1946	**LB**: „Daf net, mais ech hu kc properen Teller me'," **FR**: "Certainement pas, mais je n'ai plus d'assiette propre," **EN**: "Not really, but I don't have a single clean plate anymore." **DE**: „Das nicht, aber ich habe keinen sauberen Teller mehr.	Comprehensible	Adequate	Adequate	Adequate
De Letzeburger	1893	**LB**: De Batti: Elo hätte mer d'Stemmung gut eriwer, hätte mer elo och nach Rén. **FR**: Le Batti : Maintenant, nous aurions bien passé l'ambiance, si seulement nous avions aussi encore Rén. **EN**: Batti: Now we would have a good atmosphere if we also had some rain. **DE**: Der Batti: Jetzt hätten wir die Stimmung gut geschafft, hätten wir jetzt auch noch Regen.	Confidently Guessable	Adequate	Adequate	Adequate
De Letzeburger	1905	**LB**: Op d'Weis: Das ist im Lehen hfisslich eingerichtet. **FR**: À la manière de : Cela est arrangé vilaine dans la vie. **EN**: To the tune: It is poorly arranged in life. **DE**: Zur Melodie: Das ist im Leben hässlich eingerichtet.	Confidently Guessable	Adequate	Adequate	Adequate
De Letzeburger	1893	**LB**: Wann och d'Liss'ché wéss ze feischtren. **FR**: Même si Liss'ché sait lutter. **EN**: Even if Lisette knows how to flirt. **DE**: Wenn auch die Liss'ché weiß zu feilschen.	Incomprehensible	/	/	/
Obermosel-Zeitung	1924	**LB**: In vielen vorkern Bincl alle Krank, 80 cla,BB «lie ?eläer nicdt deBtellt terrien Können. **FR**: Dans de nombreux villages, tous sont malades, si bien que les champs ne peuvent pas être cultivés. **EN**: In many places, all are sick, so that the fields cannot be tended. **DE**: In vielen Dörfern sind alle krank, so dass die Felder nicht bestellt werden können.	Incomprehensible	/	/	/

Table 2: Sample of quadruplets of parallel sentence as used within our human evaluation of the dataset quality.