The Impact of Domain-Specific Terminology on Machine Translation for Finance in European Languages

Arturo Oncevay **Charese H. Smiley** **Xiaomo Liu**
JPMorgan AI Research
arturo.oncevay@jpmorgan.com,{charese.h.smiley,xiaomo.liu}@jpmchase.com

Abstract

Domain-specific machine translation (MT) poses significant challenges due to specialized terminology, particularly when translating across multiple languages with scarce resources. In this study, we present the first impact analysis of domain-specific terminology on multilingual MT for finance, focusing on European languages within the subdomain of macroeconomics. To this end, we construct a multi-parallel corpus from the European Central Bank, aligned across 22 languages. Using this resource, we compare open-source multilingual MT systems with large language models (LLMs) that possess multilingual capabilities. Furthermore, by developing and curating an English financial glossary, we propose a methodology to analyze the relationship between translation performance (into English) and the accuracy of financial term matching, obtaining significant correlation results. Finally, using the multi-parallel corpus and the English glossary, we automatically align a multilingual financial terminology, validating the English-Spanish alignments and incorporating them into our discussion. Our findings provide valuable insights into the current state of financial MT for European languages and offer resources for future research and system improvements.[1]

1 Introduction

In the age of globalization, MT plays a crucial role in facilitating cross-linguistic communication in different domains such as finance. Accurate translation of financial documents is essential for informed decision-making, regulatory compliance, and international collaboration (Gagnon et al., 2018). However, financial language is characterized by specialized terminology, complex sentence structures, and domain-specific conventions, which present significant challenges for MT systems (Nunziatini, 2019).

Despite the growing relevance of MT in finance, publicly available datasets for evaluating MT systems in this domain remain limited. Existing resources often lack linguistic diversity or domain-specific materials (Volk et al., 2016; Ghaddar and Langlais, 2020), such as terminologies, needed to comprehensively assess translation performance in financial contexts.

In this work, we introduce a new multi-parallel corpus from the European Central Bank (ECB)[2], consisting of financial (macroeconomics) articles published exclusively in 2024, translated and aligned across 22 European languages. Using this resource, we can evaluate how MT systems handle the complexities of financial translation, particularly domain-specific terminology. Given the recent advancements of robust LLMs for generation tasks, such as Llama 3 (Llama Team, 2024), or models with strong multilingual capabilities like AYA23 (Aryabumi et al., 2024), it is essential to assess their performance against large-scale multilingual MT systems, such as NLLB-200 (NLLB Team et al., 2022) and MADLAD-400 (Kudugunta et al., 2024).

To address this, we propose a methodology for analyzing the impact of domain-specific financial terminology on MT performance. By developing and curating an English-language glossary of financial terms, we assess how terminology accuracy influences translation performance at both the language and segment-levels. Additionally, we automatically align a multilingual financial terminology, which we use to complement our discussion.

Our contributions are threefold:

- We introduce a novel *financial and multilingual translation benchmark* from the ECB, covering 22 European languages.

[1] Please contact the author(s) if you want to get access to the resources.

[2] https://www.ecb.europa.eu/

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2758–2775

- We propose a methodology for *terminology analysis*, evaluating how MT systems and LLMs handle domain-specific financial terminology and examining its effect on translation performance.
- We present valuable *resources* for future research and system improvements, including a curated English financial glossary and an aligned multilingual financial terminology.

2 Related Work

MT and Multilinguality in Finance. Several studies have developed parallel corpora in the financial domain, but these are mostly limited to bitext. Ghaddar and Langlais (2020) extracted data from the Canadian government for an English-French corpus, while Fu et al. (2024) and Turenne et al. (2022) mined financial news sites for English-Chinese data, and Luo et al. (2018) created a corpus of financial listings between English and traditional Chinese. However, none of these resources are publicly available. The only exception is Volk et al. (2016), who released a parallel corpus for four languages (English, French, German, and Italian) from a banking magazine. Additionally, in related tasks, Läubli et al. (2019) investigated post-editing productivity for automatic translations in the financial domain, and Zhang et al. (2017) examined sentiment preservation in financial translation between English and German. Finally, Nunziatini (2019) discussed challenges in financial MT, including terminology, though their work is an overview of an MT service implementation for industry use.

Regarding multilingual approaches, Jørgensen et al. (2023) introduced `MultiFin`, a dataset of financial headlines in 15 languages, although it is focused on classification rather than translation.

Terminology in MT. Terminology presents challenges across various specialized domains in MT, as highlighted by two editions of the Shared Task on MT with Terminologies at the Conference on MT since 2021 (Alam et al., 2021). These tasks have focused on domains like medical (Alam et al., 2021) and scientific (Semenov et al., 2023), but not finance. While the shared task primarily explores techniques for leveraging terminology to enhance MT performance, other lines of work have focused on developing more terminology-aware MT models in generic domains, such as in Bogoychev and Chen (2023). To the best of our knowledge, this is the first study to systematically assess the impact of financial terminology on translation performance.

3 Multi-Parallel Financial Dataset

In this section, we describe the creation of a new multi-parallel financial dataset sourced from the ECB, covering macroeconomic topics. We refer to it as multi-parallel because it is aligned at the segment-level[3] across 22 languages from the Indo-European and Uralic language families (see the full list in Table 1). This is the first dataset of its kind in the financial domain, allowing for a fair and consistent evaluation of translation performance across multiple languages.

3.1 ECB Articles Overview

The ECB is the central bank responsible for managing the euro and formulating monetary policy within the Eurozone. We use two main sources from their website that are translated into multiple languages: the "Annual Report 2023", which provides an overview of the ECB's activities related to monetary policy and other key topics, and the "Macroeconomic Projections", published quarterly (we use the first three from 2024). The list of links are provided in Table 7 in the Appendix.

It is worth noting that OPUS (Tiedemann, 2012)—a well-known open parallel corpus repository—includes ECB data, offering bitexts for various languages. However, the most recent release of the OPUS' ECB corpus dates back to 2018.[4] OPUS is widely used in MT research, meaning that this corpus has likely been used to train a wide range of MT systems and LLMs, both open and proprietary. By exclusively using data published in 2024, we ensure that none of the models evaluated in this study have been trained on this specific information, severely limiting data contamination.

3.2 Data Processing

Alignment. We align the downloaded articles for each language using Vecalign (Thompson and Koehn, 2019), a linear sentence aligner that leverages multilingual embeddings. For the sentence encoder, we use the latest version of LASER (Heffernan et al., 2022), which covers over 200 languages.

In our alignment process, we concatenate segments in windows of 3 and align the English text

[3]We use the term segment instead of sentence as many entries consist of multiple sentences within longer paragraphs.

[4]https://github.com/Helsinki-NLP/OPUS/tree/main/corpus/ECB/v1

with the other 21 languages. We retain only the indices where alignment is consistent across all language pairs. After alignment, we retain 80% of the content, indicating that the provided translations were consistently produced across all languages.

Alignment validation. In previous work (Bañón et al., 2020), segment alignment is validated using either a gold standard or a downstream MT task. We adopt the first approach, extracting smaller articles from the same site ("Monetary Policies" articles. See details in Table 7 in the Appendix). These articles consist of 59 perfectly aligned segments across all languages. We permuted and concatenated these sentences and found that our alignment method reliably extracted the same original indices.

Cleaning. After alignment, we applied a cleaning process that removed entries with more non-alphanumeric or numeric words than alphabetic ones. Segments with fewer than 5 words (in English) and duplicate entries were also discarded. After cleaning, we retained 19% of the aligned corpus. Most discarded entries consisted of numerical data, short headers, or footnotes. The final dataset contains 531 aligned segments.[5]

3.3 Dataset Overview

Table 1 summarizes the 531 aligned segments across all languages, with segment lengths reaching up to 400 words in English.[6] Importantly, all languages contain the same number of entries, and each entry is consistently translated across languages, making this resource valuable not only for financial domain applications but also for cross-linguistic comparisons.

For instance, agglutinative languages like Finnish and Estonian show higher Type-Token Ratios (TTR) due to their morphological complexity, resulting in a wider range of unique word forms. In contrast, analytic languages such as English exhibit lower TTR values, reflecting less inflectional diversity. Romance languages tend to have longer average segment lengths (W/S), likely due to syntactic structures that include more function words. Uralic languages, on the other hand, have shorter W/S values, as they convey more information within individual word forms through agglutination. Similar patterns are expected in the translated terminology across these languages.

Group	Language		#T	V	TTR	W/S
Baltic	Lithuanian	lt	33.3k	6.0k	0.18	62.26
	Latvian	lv	32.6k	5.5k	0.17	61.06
Germanic	Danish	da	35.9k	4.7k	0.13	67.23
	German	de	38.2k	5.5k	0.14	71.57
	Dutch	nl	41.3k	4.4k	0.11	77.24
	Swedish	sv	34.5k	5.0k	0.15	64.52
	English	en	38.4k	3.5k	0.09	71.81
Hellenic	Greek	el	45.3k	5.4k	0.12	84.78
Romance	Portuguese	pt	46.6k	4.1k	0.09	87.60
	Italian	it	46.7k	4.4k	0.09	87.39
	Spanish	es	49.8k	4.3k	0.09	93.17
	French	fr	49.3k	4.3k	0.09	92.24
	Romanian	ro	44.5k	5.0k	0.11	83.27
Slavic (C)	Bulgarian	bg	41.0k	5.2k	0.13	76.78
Slavic (L)	Czech	cs	35.8k	6.0k	0.17	66.96
	Polish	pl	35.8k	6.2k	0.17	66.97
	Slovak	sk	35.1k	6.3k	0.18	65.62
	Slovenian	sl	37.0k	5.8k	0.16	69.27
	Croatian	hr	37.2k	5.7k	0.16	69.58
Uralic	Finnish	fi	28.2k	7.3k	0.26	52.80
	Estonian	et	28.0k	6.8k	0.24	52.47
	Hungarian	hu	33.9k	6.8k	0.20	63.49

Table 1: Statistics of the multi-parallel dataset (#T: number of tokens; V: vocabulary; TTR: type-token ratio; W/S: average number of words per segment). All language groups are from the Indo-European family except Uralic, and we separated Bulgarian from the main Slavic group as it uses Cyrillic instead of the Latin script.

4 Benchmarking Financial MT

After constructing the multi-parallel corpus, we evaluate the translation performance of various open-source multilingual MT systems and LLMs. We limit our evaluation to open-source models with a traceable knowledge cut-off date (2023 or earlier) to limit data contamination. It is important to clarify that our goal is not to identify or fine-tune the best model for this dataset. Instead, we aim to compare the challenges that different models face when translating financial content and to analyze the impact of financial terminology (which will be explored further in the next section).

Models For MT systems, we evaluate two models: NLLB-200 (3.3B params.) (NLLB Team et al., 2022) and MADLAD-400 (3B params.) (Kudugunta et al., 2024). Both models are designed for high-performance multilingual translation tasks and cover a wide range of languages.

For LLMs, we evaluate three models: LLAMA3-INSTRUCT (8B params.) (Llama Team, 2024), which has demonstrated robust performance across

[5]We chose to implement strict filtering criteria, including word length, to focus on segments with extensive contextual information for further terminology analysis.

[6]A length distribution is shown in Fig. 3 in the Appendix.

EN→XX	chrF++					COMET				
	MADLAD	NLLB	AYA23	LLAMA3	TOWER	MADLAD	NLLB	AYA23	LLAMA3	TOWER
Baltic	**64.38**	50.49	29.32	36.06	18.92	**92.10**	85.56	55.50	67.80	40.58
Germanic	**67.71**	58.90	50.80	47.74	54.78	**90.27**	86.36	79.19	78.48	84.31
Hellenic	**64.68**	51.10	55.47	40.53	19.01	**90.66**	85.73	89.53	76.10	44.86
Romance	**69.01**	61.87	63.17	55.89	63.38	**88.82**	85.64	87.53	82.48	86.29
Slavic (C)	**69.48**	55.58	37.55	48.71	43.83	**91.21**	86.71	65.27	80.72	78.68
Slavic (L)	**65.21**	53.76	43.29	44.83	39.02	**91.75**	85.61	74.48	79.19	75.24
Uralic	**62.37**	51.34	22.02	36.02	29.29	**92.79**	87.68	48.59	74.22	61.53
Average	**66.28**	55.97	45.39	45.91	43.79	**90.87**	86.11	73.26	77.97	73.06

Table 2: Average translation scores by language group for EN→XX. (Bold = best, underlined = second best).

XX→EN	chrF++					COMET				
	MADLAD	NLLB	AYA23	LLAMA3	TOWER	MADLAD	NLLB	AYA23	LLAMA3	TOWER
Baltic	**65.38**	59.01	46.98	42.89	35.54	**87.61**	83.50	79.24	72.78	68.13
Germanic	**71.14**	64.34	64.63	55.69	67.74	**89.00**	85.66	87.25	79.89	87.78
Hellenic	**70.59**	60.63	64.65	52.38	46.21	**88.76**	83.54	87.38	78.58	76.58
Romance	**72.48**	64.85	67.44	58.73	70.62	**88.90**	85.10	87.77	81.89	88.10
Slavic (C)	**69.31**	61.82	58.35	52.47	63.41	**87.97**	83.54	83.79	78.05	85.41
Slavic (L)	**68.61**	60.39	59.84	51.56	62.39	**87.97**	82.79	84.72	77.60	85.21
Uralic	**65.03**	58.27	44.41	44.85	48.83	**89.04**	85.36	79.68	75.96	80.06
Average	**69.32**	61.85	59.29	52.35	60.15	**88.54**	84.39	84.77	78.43	83.62

Table 3: Average translation scores by language group for XX→EN

multiple generation tasks; AYA23 (8B params.) (Üstün et al., 2024), a model trained on a large multilingual corpus covering diverse tasks[7]; and TOWERINSTRUCT (7B params.) (Alves et al., 2024), an LLM fine-tuned on top of Llama 2, with a focus on translation-related tasks.[8]

Inference We use an NVIDIA A10G GPU for all models. For the MT systems, we set the max new token up to 512, and 1024 for the LLMs, as we have long segments up to 400 words in English. LLMs use 0.3 as temperature, and the translation prompt is shown in Table 6 in the Appendix.

Metrics We employ CHRF++ (Popović, 2017), a character-level metric based on n-gram overlap between system output and reference translations, and COMET (Rei et al., 2020), a semantic-based metric that leverages pretrained multilingual representations to score translation quality.[9]

[7]AYA models are built using an instruction mixture on top of mT5 (Xue et al., 2021), which covers more than 100 languages.

[8]Although TOWER's fine-tuning predominantly focuses on Romance and Germanic languages, it also incorporates a few more distant languages.

[9]We scaled both metrics to a 0-100 range for readability.

4.1 Results and Discussion

The results in Tables 2 and 3 reveal several trends regarding model performance and language group challenges when translating financial documents. Overall, task-specific MT models consistently outperform the LLMs across all language groups and translation directions (EN→XX and XX→EN) based on both evaluation metrics. The superior performance of MT models is expected due to their specialized training for translation tasks, even with fewer parameters compared to LLMs.[10]

MADLAD consistently achieves the highest scores across all language groups and translation directions, likely benefiting from its extensive and diverse training data. NLLB follows closely, particularly excelling in the EN→XX direction, outperforming the LLMs in most cases. In the XX→EN direction, the LLMs, particularly TOWER, show some competitiveness, possibly due to the LLMs' heavy exposure to English text during training.

Regarding metric differences, both CHRF++ and COMET scores generally align in their evaluation

[10]We note that we are exclusively analyzing the baseline or zero-shot translation performance of LLMs. Performance may improve through in-context learning (Zhang et al., 2023) or fine-tuning (Alves et al., 2024), but identifying the best possible model is beyond the scope of this study.

of model performance, capturing consistent differences across models. However, language groups exhibit varying performance depending on the metric and the translation direction.

For instance, for CHRF++, Uralic and Baltic languages score lower in both translation directions, likely due to their morphological complexity and the difficulty of achieving high lexical overlap. However, their higher COMET scores indicate that these translations still retain semantic accuracy, underscoring the importance of balancing lexical and semantic evaluation in financial translation, where precise meaning is critical. Romance languages, on the other hand, perform well on CHRF++ but score relatively lower on COMET in the EN→XX direction. This suggests that while lexical accuracy is high, semantic nuances—such as verb conjugations and gender/number agreements—may not be fully captured. This is particularly relevant in the financial domain, where small semantic errors can impact regulatory compliance and reporting.

4.2 Qualitative analysis.

We conduct a qualitative analysis of translations into English for a high scoring source language, Spanish, and a low scoring language, Finnish and note several key findings. Tables 10 and 11, in the Appendix, present examples of translation outputs for Spanish→English and Finnish→English.

Literal translation of idioms. As expected with many MT systems we see a literal translation from the original texts. For example, the phrase translated more naturally in English as "a leap of faith" is rendered as "a leap into the void" by MADLAD and "a leap into the unknown" by AYA23. For the Finnish translations, we see an even greater divergence from the English reference text with "a courageous venture" (MADLAD) and "a lot of fear of failure" (AYA23). In the financial context, such drastic differences in word choice such as "leap of faith" vs. "a lot of fear of failure" may signal deeper pessimism than intended in the original text.

Common financial phrases and jargon. In Example B, we see minor translation differences such as spelling out % as "per cent" in the AYA23 translation for ES→EN that could harm CHRF++ scores while not impacting COMET. Similarly, in Example A, the entity "EKP" is preserved and not translated to "ECB" in AYA23's FI→EN system output.

More notably, we see differences in the use of financial phrasing across translation pairs. In Example B, we see the financial term *inflation* correctly matched across translations. However, other common financial terms are matched less frequently. For example, we find *input costs* in 3 of 4 texts, *shocks* in 2/4, *easing* and *commodity* in 1/4, and *move sideways* in none of the translations pointing to an overall lack of fluency with financial writing.

Semantic Accuracy. Finally, in comparing the English reference to the translations we see several differences. In Example A, the ES→EN MADLAD translation misses the last sentence altogether. Such an omission would greatly impact COMET scores and could represent a critical information loss in a financial report. For instance, almost at the end of Example A, the frequency of *speech* writing is reported as *occasional* in the reference text but "daily" in ES→EN MADLAD, "usual" in ES→EN AYA23, "sometimes contributed" in FI→EN MADLAD and *speeches* becomes *reports* in FI→EN AYA23 changing the meaning entirely. Overall, in these examples, we see more rigid and stilted translations from the Finnish translations as well as more semantic translation errors than from the Spanish translations possibly due to greater linguistic closeness between the languages.

5 The Challenge of Terminology in Financial MT

Building on our previous analysis of general translation performance, we now focus on the challenge of accurately translating financial terminology. Our methodology is as follows: First, we extract a financial terminology in English (see §5.1), and in the XX→EN translation direction, we can compute the term match accuracy for the predicted English outputs (see §5.2). Then, we introduce our analysis for assessing the relationship between term match accuracy and automatic evaluation metrics for translation, which is conducted at the corpus and segment-level (see §5.3). Finally, using the multi-parallel dataset and English glossary, we align a multilingual financial terminology (see §5.4).

5.1 English Terminology Construction

Our main source is the ECB's glossary[11], from where we extracted and manually curated a terminology of 1,135 unique financial terms in English. Of these, 176 terms have at least one occurrence in the English side of the corpus, resulting in a total of

[11]Only available in English: https://www.ecb.europa.eu/services/glossary/html/glossa.en.html

1,910 matches. The most frequent term is *inflation* with 271 matches.[12] In the Appendix, additional frequent terms can be found in Table 12, and Figure 3 shows the distributions and positive correlation between financial term count and segment length.[13]

5.2 Term Match Accuracy, Prec. and Recall

To evaluate the accuracy of financial terminology translation in the XX→EN direction, we calculate the term match accuracy (T_{acc}) using the English financial terminology extracted earlier. This involves checking whether the expected English terms appear in the MT outputs, focusing only on segments where these terms are present in the reference translations. T_{acc} is calculated as follows:

$$T_{\text{acc}} = \frac{\text{Number of correctly matched terms}}{\text{Total number of terms in reference}}$$

We use regular expressions to identify and match these terms, without considering the position.[14] We also compute precision and recall to provide a more comprehensive evaluation

Table 4 shows the results, averaged by language, where we observe that MADLAD outperformed others with the highest accuracy, precision, and recall, indicating its strong ability to both correctly detect and translate financial terms. NLLB also demonstrated strong performance, though its recall was relatively low, suggesting occasional term omissions in translation. In contrast, AYA23 and TOWER yielded more balanced, though lower, performance across all metrics. LLAMA3 showed the weakest performance, indicating potential challenges in correctly handling financial terminology.

5.3 Relationship between Term Match Accuracy and Translation Quality

With the aim of exploring further the T_{acc}'s correlation with overall translation quality at corpus and segment-levels, we compute the following weighted metrics:

$$\text{Corpus-weighted } T_{\text{acc}} = \frac{\sum(T_{\text{acc},i} \times N_{\text{terms},i})}{\sum N_{\text{terms},i}}$$

[12]We carefully matched the terminology to avoid overlapping of terms such as *debt* and *debt service-to-income ratio*, prioritizing the longer terms, and we also consider the casing to check for acronyms.

[13]While building the terminology, we distinguish between Named Entities and Acronyms (e.g., *European System Risk Board*, *PSPP*) and general terms (e.g., *price stability*, *debt ratio*). However, since our analysis yielded similar results for both types, we do not present any distinctions in this study.

[14]Position is not considered in our calculation, as a stricter analysis is not required for the correlation with MT performance. This could be explored in future work.

Model	Accuracy	Precision	Recall
MADLAD	0.8620 ± 0.22	0.9262 ± 0.19	0.8850 ± 0.21
NLLB	0.7690 ± 0.30	0.8796 ± 0.26	0.8044 ± 0.29
AYA23	0.7072 ± 0.34	0.7945 ± 0.34	0.7220 ± 0.34
TOWER	0.7033 ± 0.35	0.7809 ± 0.35	0.7181 ± 0.35
LLAMA3	0.6194 ± 0.37	0.7271 ± 0.39	0.6342 ± 0.37

Table 4: Financial Term Accuracy, Precision and Recall for XX→EN translations, averaged by language-pair, and sorted (desc.) by Accuracy.

$$\text{Segment-weighted } T_{\text{acc}} = \frac{T_{\text{acc},i} \times N_{\text{terms},i}}{L_{\text{segment},i}}$$

where Corpus-weighted T_{acc} normalizes term accuracy by the total number of terms across all segments, providing a corpus-level evaluation, while Segment-weighted T_{acc} normalizes term accuracy by the length of each segment, offering a more granular, segment-level evaluation.

For the corpus-level analysis, we use the corpus-weighted metric to assess its impact on translation performance across different language pairs in the XX→EN direction. Similarly, for the segment-level analysis, we use the segment-weighted metric to evaluate its impact on translation performance at the segment level for each language-pair/model combination.

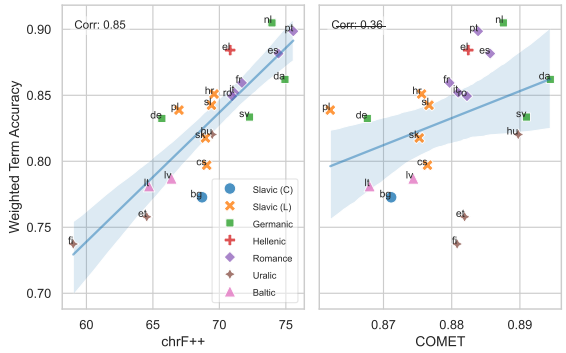
5.3.1 Corpus or Language-level Analysis

Our approach begins by calculating the corpus-weighted T_{acc} per language-pair (XX→EN). This serves as the predictor, while the target variables are the translation evaluation metrics. We then compute Pearson correlation coefficients between the predictor and the target metrics, focusing on statistically significant results ($p < 0.05$).

Results and Discussion. Figure 1 shows the correlation results for one MT system (MADLAD) and one LLM (AYA23). Both models show strong positive correlations between weighted T_{acc} and CHRF++, indicating that terminology accuracy is closely tied to lexical overlap and overall translation performance, especially for financial texts. While for COMET, AYA23 shows a strong correlation, suggesting that LLMs benefit significantly from accurate terminology to improve semantic translation quality. However, MADLAD shows a non-significant correlation with COMET, implying that the room for improvement in semantic accuracy for the MT system may be limited, even if terminology accuracy improves for the corpus.

Regarding the language groups, we note that Baltic and Uralic languages face more challenges

(a) MADLAD



(b) AYA23

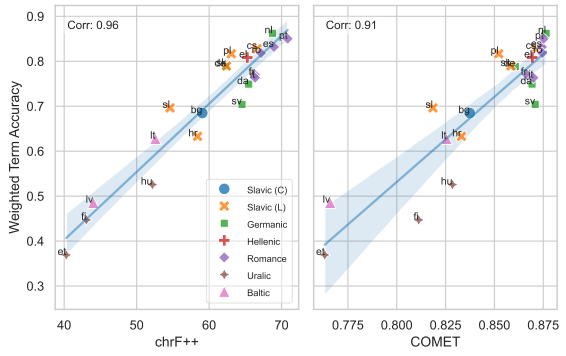


Figure 1: Correlation analysis between Financial Term Accuracy and two translation metrics (CHRF++ (left) and COMET (right)) for the MADLAD (top) and AYA23 (bottom) in the XX→EN direction.

in terminology translation, likely due to their extensive agglutination or affixation processes. In contrast, Romance and Germanic languages consistently perform better, likely due to their closer linguistic ties to English and the relatively easier cross-lingual mapping of terms.

In summary, while CHRF++ strongly correlates with term accuracy for both models, COMET shows varying sensitivity depending on the model type. The patterns for each model type are consistent across the other models (NLLB, LLAMA3 and TOWER), with their correlation results presented in Figure 4 in the Appendix.

5.3.2 Segment-level Analysis

Our methodology is as follows:

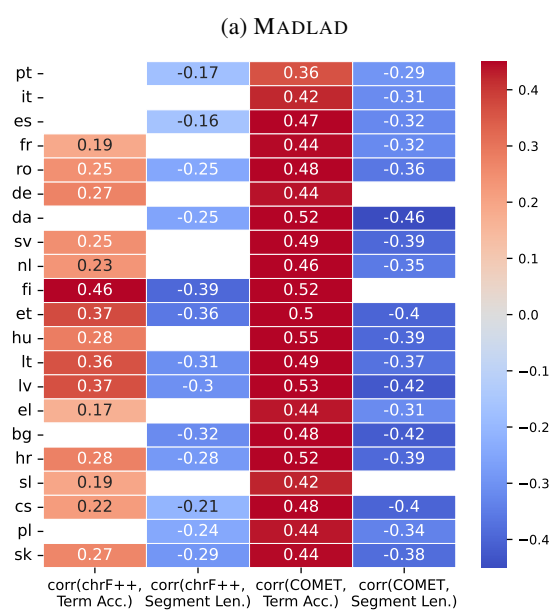
1. For each language-pair/model combination, we compute the segment-weighted T_{acc} for all outputs, which serves as the main predictor.
2. Segment length (number of words) is included as a confounder to control for its known impact on translation performance and to avoid multicollinearity with T_{acc}.
3. We calculate the Variance Inflation Factor (VIF) for T_{acc} and segment length. If VIF > 5, indicating high multicollinearity, we exclude that language-pair/model combination from the analysis.
4. For each target metric (CHRF++ and COMET), we construct Generalized Linear Models (GLM) with T_{acc} and segment length as predictors, assuming a Gaussian distribution.
5. We compute Pearson correlation coefficients between predictors and target metrics for all settings, focusing on statistically significant results ($p < 0.05$).

Results and Discussion. Figure 2 shows the correlation analysis for one MT system (MADLAD) and one LLM (AYA23).[15] MADLAD exhibits stronger correlations between T_{acc} and translation quality, especially for COMET, where correlations reach 0.55 (Hungarian) and remain positive across most languages. This highlights the importance of T_{acc} at the segment-level in improving translation performance for MADLAD, particularly in finance, where accuracy in term translation is critical.

AYA23, however, shows more varied correlations. In some languages (e.g., Romanian, Croatian), moderate positive correlations appear, but others (e.g., Finnish, Hungarian) show weak or near-zero correlations. This suggests that while AYA23 may correctly translate some specialized terms, its overall translation quality is still limited. Even when financial terms are accurate, broader issues like fluency or semantic adequacy reduce the impact of T_{acc} on overall performance at the segment-level.

Regarding the segment-length confounder, LLMs like AYA23, with their larger context window (1024+ tokens), handle longer segments more effectively than MT systems like MADLAD (512 tokens). This likely explains why segment-length has a stronger negative impact on MADLAD, particularly in COMET, while AYA23 shows weaker or near-zero correlations with the confounder. The longer context window in LLMs helps mitigate challenges in translating longer financial texts, making segment length a less critical factor.

[15] The observed patterns are consistent across other LLMs (LLAMA3, TOWER) and MT systems (NLLB), as shown in Fig. 5 in the Appendix.



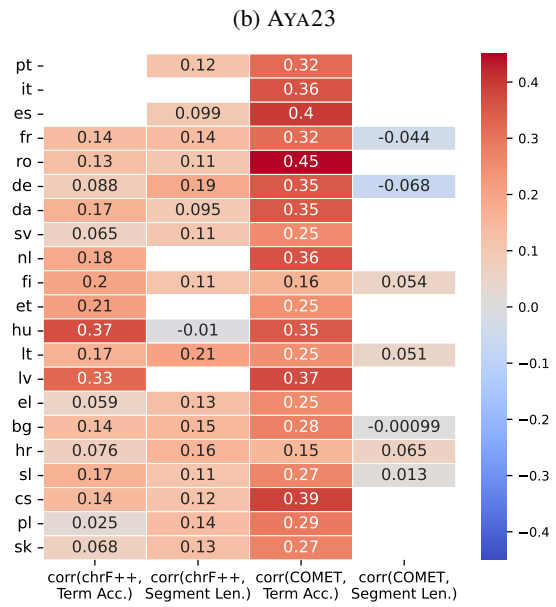


Figure 2: Correlation analysis at the segment-level between the financial term accuracy, weighted by the segment length, and translation performance metrics for one MT system (MADLAD) and one LLM (AYA23).

5.4 Multilingual Financial Terminology

As highlighted in the previous analyses, improving domain-specific translation performance may require leveraging terminology match accuracy. One potential approach is to use additional resources, such as bilingual or multilingual terminologies that provide accurate translations of specialized terms. Building on our multi-parallel financial dataset and the English terminology extracted earlier, we now take the first steps toward constructing a multilingual financial terminology.

Lang.	#TwC	#C	#MWT	Entropy
pt	128	249	92	1.61
es	125	265	105	1.49
it	122	273	107	1.41
fr	120	262	105	1.45
da	119	355	72	1.07
nl	119	337	85	1.12
sv	118	382	86	0.99
sk	118	339	100	1.12
hu	116	421	126	0.87
cs	116	309	106	1.21
bg	117	314	116	1.20
pl	117	311	109	1.19
ro	114	317	117	1.14
hr	114	297	120	1.24
lv	113	302	119	1.22
de	111	378	93	0.94
lt	108	297	105	1.19
sl	108	314	100	1.13
fi	105	400	68	0.86
et	105	350	67	1.00
el	101	277	74	1.19

Table 5: Language statistics for the Multilingual Financial Terminology (aligned from 176 English terms). #TwC: Number of terms with at least one candidate; #C: total number of candidates; #MWT: Number of multi-word candidate terms; Entropy: a measure of diversity or variability in the translations provided by a language across different terms.

Automatic Term Alignment. We utilize SimAlign (Jalili Sabet et al., 2020), a state-of-the-art word alignment tool that uses pretrained multilingual representations. For encoding, we employ XLM-RoBERTa-base (Conneau et al., 2020).

Given that our terminology includes multi-word terms, we process the alignment outputs through several steps. First, we collect all target indices that align to any part of the source term and identify continuous spans in the target text to preserve phrase integrity. This handles cases where terms may be realized differently across languages: one-to-many (when a single English word aligns to multiple target words), many-to-one (when multiple English words align to a single target word), and many-to-many mappings. Additionally, a single English term may have multiple mappings in other languages due to structural differences, such as fusion in Romance languages like Spanish (e.g., "credit card" → "tarjeta" and "tarjeta de crédito") or agglutination in Uralic languages like Finnish (where multiple English words might correspond to a single morphologically complex word).

Table 5 presents statistics of the aligned multi-

lingual terminology, including the number of candidates per language and entropy, which measures variability in translation diversity across terms. Consistently, Romance languages, scoring higher in translation quality and T_{acc}, have more terms with candidates, more total candidates, and higher entropy, suggesting a better chance of retrieving good term candidates. In contrast, Uralic languages, which underperformed, have fewer terms with candidates, fewer total candidates, and lower entropy, indicating less recall potential and less diverse translations. Additionally, the number of multi-word terms may reflect the complexity of terminology, potentially impacting translation performance, as languages with more multi-word terms might face greater challenges in achieving accurate translations, such as Uralic or Slavic ones.

Annotation and Evaluation. Fully curating multilingual alignments is resource-intensive, so in this study, we focus on a single language pair: English-Spanish. After aligning and deduplicating the terms, we manually annotated each alignment as True or False (275 candidates for 176 terms).

Results. The alignment of English and Spanish terminology yielded a precision of 0.57 and a recall of 0.95. This indicates that while the aligner successfully identified most of the correct translations (high recall), it also aligned many incorrect ones (lower precision). Several factors may contribute to this, including long source and target segments that challenge the alignment model and repeated terms within the same segment, which can cause disambiguation issues. Nevertheless, this is a promising result, as it shows that we are highly likely to find the correct terminology translations. However, for this new resource to grow into a large-scale, high-quality multilingual financial terminology, it will be essential to improve the word alignment method or implement strategies to reduce false positives.

6 Diachronic Resources for Financial MT

Our resources are designed to be dynamic and continuously evolving. The **multi-parallel financial corpus** can be updated annually using the provided code, enabling the creation of new benchmarks each year. This will allow researchers to evaluate the latest multilingual MT systems or LLMs against up-to-date financial content and compare them with earlier models, ensuring the benchmarks remain relevant. Besides, the **English financial glossary** and the **multilingual financial terminology** can also be expanded as new terms emerge, further enhancing the evaluation of domain-specific translation accuracy. Additionally, the framework supports the inclusion of new corpora and languages, allowing the benchmark to grow and cover a broader range of languages and financial terms over time.

7 Conclusion

To the best of our knowledge, this is the first systematic study to analyze the impact of domain-specific terminology on translation performance across 22 European languages in the financial domain, where we observed significant correlations between term accuracy and overall translation quality at both the corpus and segment levels. This work involved creating a multi-parallel financial corpus aligned at the segment-level, evaluating the translation performance of diverse MT systems and LLMs, and curating an English financial terminology. We also laid the groundwork for building a multilingual financial terminology, which will be a valuable resource for advancing financial MT.

8 Ethical Considerations

As the text used in this study is taken from the ECB, a publicly available resource, we do not anticipate any ethical concerns with the sourcing of these texts. For each instance, we provide both the original source and target translations. However, as with all MT work, the accuracy of the translated texts is not guaranteed and human oversight still remains needed for use in critical financial applications and we release this data for research purposes only.

9 Limitations

This study has several limitations that present opportunities for future research. Firstly, the dataset is derived from a snapshot of ECB articles from 2024. Due to continuous updates, this data may be included in future iterations of ParaCrawl (Bañón et al., 2020). However, future releases of this resource can also be made using the methodology outlined in this work.

Secondly, while the dataset focuses on macroeconomic and public policy content, the financial domain is broad. Expanding coverage to include topics such as personal finance, corporate finance, risk management, and investing is a limitation and an area for future research.

Additionally, the study evaluates LLMs in a zero-shot setting and focuses on open-source MT systems, intentionally avoiding fine-tuning, in-context learning, and commercial models like Google Translate or DeepL. This approach was chosen to minimize the risk of data contamination and ensure the integrity of the evaluation process. While this design choice limits the exploration of potentially higher-performing models and techniques, it provides a controlled environment for assessing the baseline capabilities of the models.

Thirdly, the study does not include human evaluation of translation quality, which could provide insights into financial clarity, accuracy of terms, and adherence to regulatory norms. This is a limitation and an area for future investigation.

Finally, although the dataset created in this study covers a wide variety of languages, especially within the European context, it does not represent the full variety of dialects and low resource languages present in Europe. Moreover, the languages in this corpus are not largely typologically diverse (more distant language families) or extremely low-resource ones, like languages spoken in Africa or in the Americas.

Acknowledgments

We are thankful to Simerjot Kaur, Elena Kochkina, Samuel Mensah, Joy Sain, and other members of the JPMorgan AI Research team for their insightful feedback since early stages of this work. We also appreciate the feedback of the anonymous reviewers and meta-reviewer.

References

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, Online. Association for Computational Linguistics.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

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.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yuxin Fu, Shijing Si, Leyi Mai, and Xi-ang Li. 2024. Ffn: a fine-grained chinese-english financial domain parallel corpus. In *2024 International Conference on Asian Language Processing (IALP)*, pages 127–132. IEEE.

Chantal Gagnon, Pier-Pascale Boulanger, and Esmaeil Kalantari. 2018. How to approach translation in a

financial news corpus? *Across Languages and Cultures*, 19(2):221 – 240.

Abbas Ghaddar and Phillippe Langlais. 2020. SEDAR: a large scale French-English financial domain parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3595–3602, Marseille, France. European Language Resources Association.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. 2023. MultiFin: A dataset for multilingual financial NLP. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 894–909, Dubrovnik, Croatia. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Samuel Läubli, Chantal Amrhein, Patrick Düggelin, Beatriz Gonzalez, Alena Zwahlen, and Martin Volk. 2019. Post-editing productivity with neural machine translation: An empirical assessment of speed and quality in the banking and finance domain. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 267–272, Dublin, Ireland. European Association for Machine Translation.

Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Linkai Luo, Haiqin Yang, Sai Cheong Siu, and Francis Yuk Lun Chin. 2018. Neural machine translation for financial listing documents. In *Neural Information Processing*, pages 232–243, Cham. Springer International Publishing.

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. *Preprint*, arXiv:2207.04672.

Mara Nunziatini. 2019. Machine translation in the financial services industry: A case study. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 57–63, Dublin, Ireland. European Association for Machine Translation.

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

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.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the WMT 2023 shared task on machine translation with terminologies. In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. 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).

Nicolas Turenne, Ziwei Chen, Guitao Fan, Jianlong Li, Yiwen Li, Siyuan Wang, and Jiaqi Zhou. 2022. Mining an english-chinese parallel dataset of financial news. *Journal of Open Humanities Data*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Martin Volk, Chantal Amrhein, Noëmi Aepli, Mathias Müller, and Phillip Ströbel. 2016. Building a parallel corpus on the world's oldest banking magazine. In *KONVENS*. s.n.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: a case study. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Chong Zhang, Matteo Capelletti, Alexandros Poulis, Thorben Stemann, and Jane Nemcova. 2017. A case study of machine translation in financial sentiment analysis. In *Proceedings of Machine Translation Summit XVI: Commercial MT Users and Translators Track*, pages 49–58, Nagoya Japan.

A Appendix

LLM	Prompt
AYA23, LLAMA3	{"role": "system", "content": "You are a professional translator in the banking and finance domain."}, {"role": "user", "content": "Translate the following text from SRC-LANG into TGT-LANG.\n SRC-LANG: TEXT.\n TGT-LANG: "}
TOWER	{"role": "user", "content": "Translate the following text from SRC-LANG into TGT-LANG.\n SRC-LANG: TEXT.\n TGT-LANG: "}

Table 6: Prompts used for different LLMs. We do not explore other types of complex prompts for translation, as previous work has consistently shown that simple templates achieve robust overall performance (Zhang et al., 2023). Additionally, pushing the capabilities of a model is beyond the scope of our study. TOWER does not include a system role.

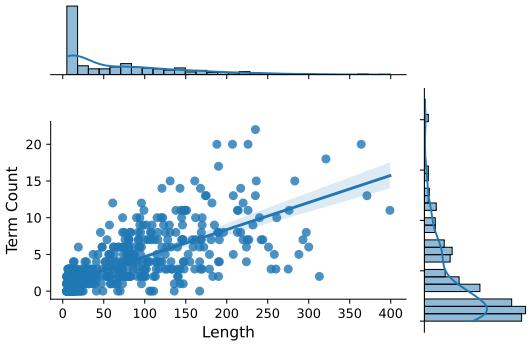


Figure 3: English terminology count per segment versus segment-length (Pearson corr.$= 0.71$, p-val< 0.005).

Title	URL (English)
Annual Report 2023	https://www.ecb.europa.eu/press/annual-reports-financial-statements/annual/html/ecb.ar2023~d033c21ac2.en.html
Macroeconomic Projections June 2024	https://www.ecb.europa.eu/press/projections/html/ecb.projections202406_eurosystemstaff~ee3c69d1c5.en.html
Macroeconomic Projections March 2024	https://www.ecb.europa.eu/press/projections/html/ecb.projections202403_ecbstaff~f2f2d34d5a.en.html
Macroeconomic Projections September 2024	https://www.ecb.europa.eu/press/projections/html/ecb.projections202409_ecbstaff~9c88364c57.en.html
Monetary Policy Decision 25-Jan-2024	https://www.ecb.europa.eu/press/pr/date/2024/html/ecb.mp240125~f738889bde.en.html
Monetary Policy Decision 07-Mar-2024	https://www.ecb.europa.eu/press/pr/date/2024/html/ecb.mp240307~a5fa52b82b.en.html
Monetary Policy Decision 11-Apr-2024	https://www.ecb.europa.eu/press/pr/date/2024/html/ecb.mp240411~1345644915.en.html
Monetary Policy Decision 06-Jun-2024	https://www.ecb.europa.eu/press/pr/date/2024/html/ecb.mp240606~2148ecdb3c.en.html
Monetary Policy Decision 18-Jul-2024	https://www.ecb.europa.eu/press/pr/date/2024/html/ecb.mp240718~b9e0ddd9d5.en.html
All glossary entries	https://www.ecb.europa.eu/services/glossary/html/glossa.en.html

Table 7: List of ECB articles used in this study. To access the articles in a language other than English, replace "en" at the end of the link with the appropriate language code. The "Monetary Policy Decision" articles were used solely to validate the sentence alignment process, whereas "All glossary entries" was used to build the English terminology.

EN→XX	chrF++					COMET				
	MADLAD	NLLB	AYA23	LLAMA3	TOWER	MADLAD	NLLB	AYA23	LLAMA3	TOWER
Bulgarian	69.48	55.58	37.55	48.71	43.83	91.21	86.71	65.27	80.72	78.68
Croatian	64.66	45.28	33.33	45.29	40.68	91.66	80.85	63.17	81.27	74.13
Czech	67.03	58.29	57.59	47.40	42.27	92.52	88.63	90.52	82.00	78.95
Danish	72.15	63.00	43.46	51.16	48.82	91.87	88.57	69.30	80.90	80.87
Dutch	68.42	57.76	60.11	45.61	65.51	89.62	85.02	88.03	75.40	89.46
Estonian	64.23	51.53	20.66	31.16	20.66	93.00	87.82	45.04	66.26	45.10
Finnish	57.36	46.97	20.94	37.02	33.34	93.00	88.14	49.27	78.06	70.56
French	71.12	63.13	65.75	57.06	68.50	87.87	85.08	87.11	80.84	87.33
German	60.29	53.88	53.52	43.65	53.83	87.53	83.60	85.85	76.79	85.93
Greek	64.68	51.10	55.47	40.53	19.01	90.66	85.73	89.53	76.10	44.86
Hungarian	65.53	55.53	24.47	39.90	33.87	92.36	87.07	51.46	78.34	68.94
Italian	65.47	58.90	59.98	52.26	63.25	88.96	85.95	87.73	82.27	88.54
Latvian	64.57	46.70	21.57	36.37	18.41	91.99	83.78	44.46	68.65	39.21
Lithuanian	64.19	54.28	37.07	35.75	19.43	92.21	87.33	66.55	66.96	41.96
Polish	61.24	51.75	54.65	46.45	43.39	90.84	84.63	89.01	83.04	80.18
Portuguese	70.10	63.43	62.63	56.93	67.50	88.19	85.22	86.17	80.99	88.05
Romanian	66.20	55.25	59.92	53.12	48.95	91.26	86.45	89.86	85.49	80.42
Slovak	67.45	59.77	42.61	42.99	31.58	92.28	88.08	76.13	75.65	74.66
Slovenian	65.65	53.70	28.27	42.02	37.17	91.44	85.87	53.57	73.98	68.30
Spanish	72.16	68.64	67.56	60.09	68.71	87.83	85.48	86.79	82.79	87.13
Swedish	69.99	60.97	46.13	50.55	50.97	92.04	88.25	73.59	80.81	80.99
Average	66.28	55.97	45.39	45.91	43.79	90.87	86.11	73.26	77.97	73.06

Table 8: Translation scores per language pair in the EN→XX direction. COMET scores are rescaled to a 0-100 range for readability.

XX→EN	chrF++					COMET				
	MADLAD	NLLB	AYA23	LLAMA3	TOWER	MADLAD	NLLB	AYA23	LLAMA3	TOWER
Bulgarian	69.31	61.82	58.35	52.47	63.41	87.97	83.54	83.79	78.05	85.41
Croatian	69.55	53.08	57.06	51.60	62.38	88.23	77.70	83.51	77.62	85.01
Czech	68.24	62.42	65.17	54.21	64.98	88.17	84.59	87.37	79.54	86.69
Danish	74.02	65.97	64.68	56.26	69.40	89.85	86.03	87.17	79.58	88.44
Dutch	73.13	65.74	67.87	55.65	71.86	89.32	86.08	88.00	78.77	88.74
Estonian	65.00	58.58	39.94	42.68	33.62	88.78	85.18	76.07	74.67	68.73
Finnish	60.02	53.93	42.14	43.03	51.82	88.73	85.25	80.66	76.68	85.16
French	71.61	65.03	65.98	58.97	70.93	88.70	85.43	87.28	82.12	88.21
German	64.94	59.79	61.64	53.17	62.28	87.12	84.05	86.39	80.03	85.84
Greek	70.59	60.63	64.65	52.38	46.21	88.76	83.54	87.38	78.58	76.58
Hungarian	70.06	62.28	51.16	48.85	61.05	89.61	85.65	82.30	76.52	86.30
Italian	70.99	64.19	66.21	58.33	69.78	88.68	85.57	87.61	82.21	88.06
Latvian	66.23	59.74	42.62	45.15	35.73	87.93	83.80	76.22	74.68	67.79
Lithuanian	64.53	58.28	51.34	40.63	35.35	87.29	83.20	82.26	70.88	68.47
Polish	66.75	61.14	61.95	51.60	62.63	86.95	82.97	85.29	77.54	84.87
Portuguese	74.61	66.24	69.97	59.17	73.49	88.93	84.96	88.05	80.79	88.54
Romanian	71.66	62.72	66.73	57.25	66.36	89.08	84.10	87.97	82.30	87.13
Slovak	69.11	63.39	62.03	51.75	60.00	88.19	84.62	86.12	77.67	84.22
Slovenian	69.38	61.91	53.02	48.63	61.95	88.32	84.05	81.30	75.60	85.27
Spanish	73.53	66.04	68.33	59.95	72.51	89.12	85.43	87.94	82.04	88.56
Swedish	72.45	65.86	64.33	57.65	67.41	89.72	86.48	87.43	81.16	88.09
Average	69.32	61.85	59.29	52.35	60.15	88.54	84.39	84.77	78.43	83.62

Table 9: Translation scores per language pair in the XX→EN direction. COMET scores are rescaled to a 0-100 range for readability.

Entry	Text
Spanish	Mi trayectoria profesional en el BCE ha sido extraordinariamente formativa. Comenzó con un salto al vacío, con un nuevo empleo y en un país nuevo. Me incorporé a un BCE en ciernes, como research analyst en la función estadística, que aún se estaba estableciendo. Con solo unos cientos de empleados, el BCE necesitaba que «todos nos pusiéramos manos a la obra», así que compaginábamos múltiples funciones. Una jornada típica era muy diversa, y en el mismo día discutía disposiciones con las áreas de política, diseñaba marcos estadísticos, programaba códigos y redactaba el discurso de turno. Como las políticas institucionales aún estaban dando sus primeros pasos, también sentí gran satisfacción al ver que, a menudo, mis soluciones creativas se implantaban con rapidez.
Finnish	Urapolkuni EKP:ssä on opettanut minulle paljon. Se sai alkunsa rohkeasta heittäytymisestä, kun muutin uuteen maahan ja aloitin uudessa työssä. Tulin EKP:hen tutkimusanalyytikoksi ja työskentelin vielä kehitysvaiheessa olevassa tilastoyksikössä. Työntekijöitä oli silloin vain muutama sata, ja tekemistä oli paljon, joten jokaisen toimenkuvaan kuului monia tehtäviä. Tyypillinen työpäivä oli vaihteleva. Keskustelin eri toimialojen kanssa niiden tarpeista, suunnittelin tilastoinnin rakenteita, kirjoitin ohjelmakoodia ja välillä osallistuin puheiden laatimiseen. Nuoressa organisaatiossa menettelytavat hakivat vielä muotoaan, ja oli innostavaa nähdä omien luovien ratkaisujen toteutuvan usein hyvin nopeasti.
English	My career pathway at the ECB has been remarkably formative. It began with a leap of faith, with a new job in a new country. I joined the ECB at its cusp, as a young Research Analyst in the still-developing Statistics Function. With just a few hundred employees, the ECB needed "all hands on deck", so we juggled multiple roles. A typical day was diverse: I discussed requirements with policy areas, designed statistical frameworks, programmed code and drafted the occasional speech. With institutional policies still in their infancy, I was also thrilled to see my creative solutions often rapidly deployed.
MADLAD's output: Spanish→ English	My career at the ECB has been extremely formative. It started with a leap into the void, a new job and a new country. I joined a budding ECB as a research analyst in the statistical function, which was still in its infancy. With only a few hundred employees, the ECB needed "all of us to get to work", so we were working in multiple roles. A typical day was very diverse, with the same day discussing arrangements with policy areas, designing statistical frameworks, programming codes and writing the speech of the day.
AYA23's output: Spanish→ English	My career path at the ECB has been extraordinarily formative. It started with a leap into the unknown, with a new job and a new country. I joined a fledgling ECB, as a research analyst in the still-evolving statistical function. With only a few hundred employees, the ECB needed everyone to 'pull their weight', so we juggled multiple roles. A typical day was very diverse, and in the same day I would be discussing provisions with the policy areas, designing statistical frameworks, programming code and drafting the usual speeches. As institutional policies were still in their infancy, I also felt great satisfaction in seeing that often my creative solutions were quickly implemented.
MADLAD's output: Finnish→ English	My career path at the ECB has taught me a lot. It began with a courageous venture to move to a new country and start a new job. I joined the ECB as a research analyst and worked in a statistical unit that was still in its infancy. There were only a few hundred people at the time and there was a lot of work to do, so each role involved a variety of tasks. A typical day at work was varied. I discussed with different sectors their needs, designed statistical structures, wrote code and sometimes contributed to the preparation of speeches. In a young organisation, policies were still taking shape and it was exciting to see my own creative solutions being implemented, often very quickly.
AYA23's output: Finnish→ English	The time I spent at the EKP was very instructive. It started with a lot of fear of failure, when I moved to a new country and started a new job. I became a data analyst at the EKP and worked in a still developing department of statistics. At that time there were only a few hundred employees and a lot had to be done, so everyone had many tasks. A typical workday was varied. I discussed the needs of different departments, designed statistical structures, wrote computer programs and occasionally participated in the editing of reports. In the young organization, procedures were still being developed, and it was exciting to see one's own solutions being implemented very quickly.

Table 10: Example A: Reference texts aligned in Spanish, Finnish and English (top), and system outputs for Spanish→English (middle) and Finnish→English (bottom) using two models.

Entry	Text
Spanish	Se prevé que la inflación de los alimentos permanezca prácticamente sin variación a corto plazo y que después disminuya moderadamente como consecuencia de la evolución moderada de los costes de los insumos (panel b del gráfico 7). La tasa de variación de los precios de los alimentos descendió progresivamente hasta situarse en el 2,3 % en julio, debido, en gran medida, a la relajación de las presiones latentes a medida que iban desapareciendo los efectos de las anteriores perturbaciones de los precios de la energía y de las materias primas alimenticias. Se espera que la inflación de los alimentos aumente ligeramente a finales de 2024. Debería mantenerse prácticamente estable en los tres primeros trimestres de 2025, con tasas próximas al 2,5 % sostenidas por la inflación de los alimentos elaborados. Después se prevé que disminuya hasta una tasa media del 2,1 % en 2026, en parte como reflejo del supuesto de una evolución moderada de los precios de las materias primas alimenticias.
Finnish	Elintarvikehintainflaation odotetaan pysyvän suurin piirtein ennallaan lyhyellä aikavälillä ja hidastuvan sen jälkeen maltillisesti tuotantokustannusten vaimean kehityksen vuoksi (ks. kuvion 7 kohta b). Elintarvike-inflaatio hidastui asteittain 2,3 prosenttiin heinäkuussa lähinnä siksi, että hintapaineet laskivat energia- ja elintarvikeraaka-aineiden aiempien hintasokkien vaimetessa. Vuoden 2024 loppua kohden elintarvikeinflaation odotetaan nopeutuvan jonkin verran. Vuoden 2025 kolmella ensimmäisellä vuosineljänneksellä sen odotetaan pysyvän 2,5 prosentin tuntumassa eli suurin piirtein ennallaan jalostettujen elintarvikkeiden hintainflaation ylläpitämänä. Sen jälkeen energiahintainflaation arvioidaan hidastuvan keskimäärin 2,1 prosenttiin vuonna 2026 osittain siksi, että elintarvikeraaka-aineiden hintojen oletetaan kehittyvän maltillisesti.
English	Food inflation is expected to remain broadly unchanged in the near term and subsequently decline moderately on account of moderate developments in input costs (Chart 7, panel b). Food inflation progressively declined to 2.3% in July, owing largely to easing pipeline pressures as the impacts of earlier energy and food commodity price shocks faded. Towards the end of 2024, food inflation is expected increase somewhat. It is expected to move broadly sideways in the first three quarters of 2025, with rates of around 2.5% sustained by inflation in processed food. It is then projected to decline to an average of 2.1% in 2026, partly reflecting the assumption of moderate developments in food commodity prices.
MADLAD's output: Spanish→ English	Food inflation is expected to remain virtually unchanged in the short term and then decline moderately as a result of moderate input cost developments (Figure 7, box b). The rate of change in food prices gradually decreased to 2.3 per cent in July, largely due to the relaxation of latent pressures as the effects of previous shocks in energy and food raw materials prices were disappearing. Food inflation is expected to rise slightly by the end of 2024. It should remain virtually stable in the first three quarters of 2025, with rates close to 2.5 per cent supported by processed food inflation.
AYA23's output: Spanish→ English	Food inflation is expected to remain virtually unchanged in the short term and then decline moderately as a result of the moderate evolution of input costs (graph 7, panel b). The rate of variation of food prices declined gradually to reach 2.3% in July, largely due to the easing of latent pressures as the effects of previous energy and food commodity price disruptions faded. Food inflation is expected to increase slightly at the end of 2024. It should remain virtually stable in the first three quarters of 2025, with rates close to 2.5% sustained by food manufactured inflation. It is then expected to decline to an average rate of 2.1% in 2026, partly reflecting the assumption of a moderate evolution of food commodity prices.
MADLAD's output: Finnish→ English	Food price inflation is expected to remain broadly unchanged in the short term and then slow down moderately owing to the weak developments in production costs (see Figure 7.b). Food price inflation gradually slowed to 2.3% in July, mainly as price pressures on energy and food raw materials subsided from previous shocks. Food price inflation is expected to accelerate somewhat towards the end of 2024. It is expected to remain around 2.5% over the first three quarters of 2025, supported by broadly unchanged processed food price inflation.
AYA23's output: Finnish→ English	The inflation of food prices is expected to remain positive for the time being, with a slight decline in the near future, and to slow down further in the medium term. In particular, food price inflation is expected to remain at around 2.3% in July, mainly due to rising input costs for energy and food production. Food price inflation is expected to decline somewhat by the end of 2024. By the end of 2025, food price inflation is expected to remain at around 2.5% on average, or the highest level in recent years. After that, food price inflation is expected to slow down to an average of 2.1% in 2026, partly due to the expected increase in the cost of food production inputs.

Table 11: Example B: Reference texts aligned in Spanish, Finnish and English (top), and system outputs for Spanish→English (middle) and Finnish→English (bottom) using two models.

Term	Freq	Term	Freq
inflation	271	ECB	241
euro area	162	projections	157
GDP	105	monetary policy	77
HICP	61	euro	55
Eurosystem	53	services	44
Eurostat	34	households	26
fiscal stance	24	cash	23
debt	23	goods	19
unit labour costs	18	Governing Council	17
compensation per employee	15	interest rate	15
liquidity	15	financial stability	14
IMF	13	European Commission	12
collateral	11	APP	10
unemployment rate	10	deposit facility	9
labour productivity	9	ESCB	8
ESRB	8	TIPS	8
budget balance	8	governance	8
payment	8	transactions	8
DLT	7	T2	7
central bank	7	option	7
settlement	7	European Parliament	6
European System of Central Banks	6	asset	6
price stability	6	primary balance	6
CJEU	5	CPI	5
acceptance	5	central bank money	5
debt ratio	5	delivery	5
equity	5	exposure	5
EURIBOR	4	European Systemic Risk Board	4
OECD	4	TLTRO	4
current account	4	financial markets	4
fiscal policy stance	4	labour force	4
member	4	value added	4
CCP	3	International Monetary Fund	3
PSPP	3	T2S	3
TARGET	3	consumer price index	3
credit risk	3	discretionary fiscal policy	3
financial accounts	3	intra-euro area trade	3
limit	3	unwind	3

Table 12: Terminology with frequencies equal or higher than 3 in the English corpus.

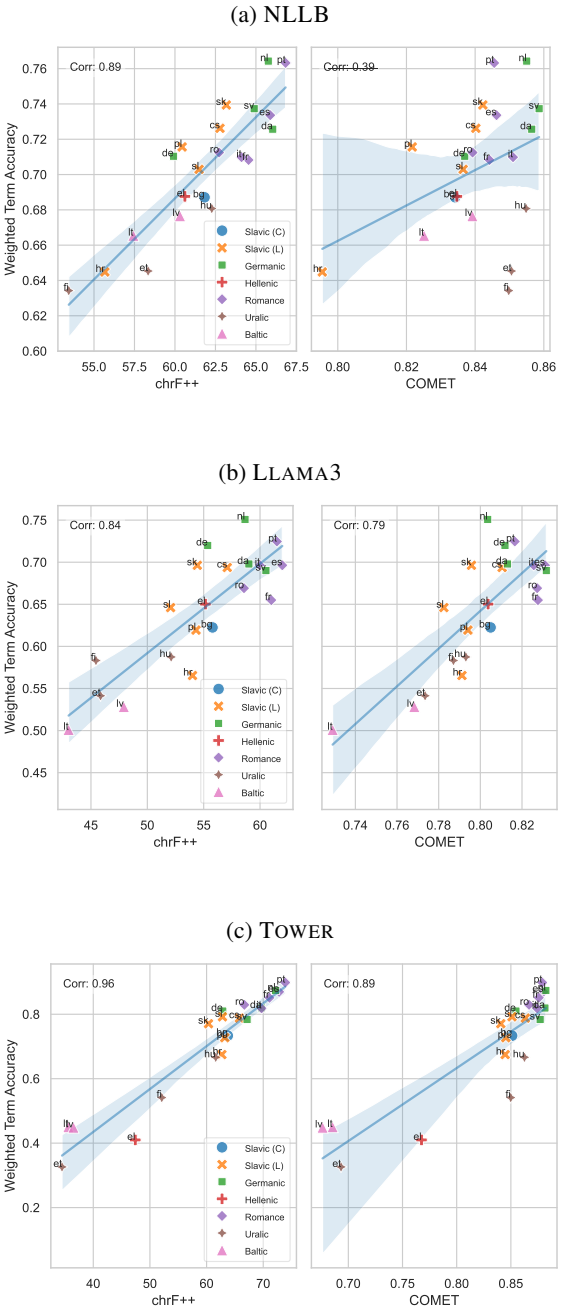
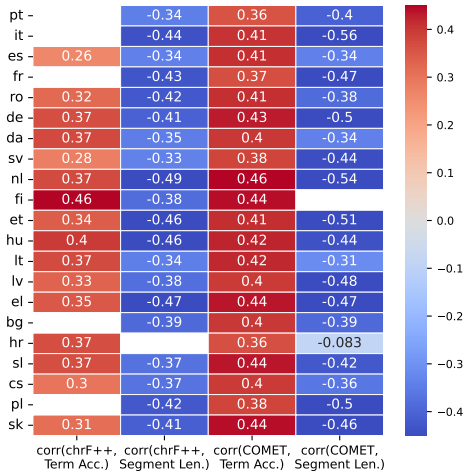
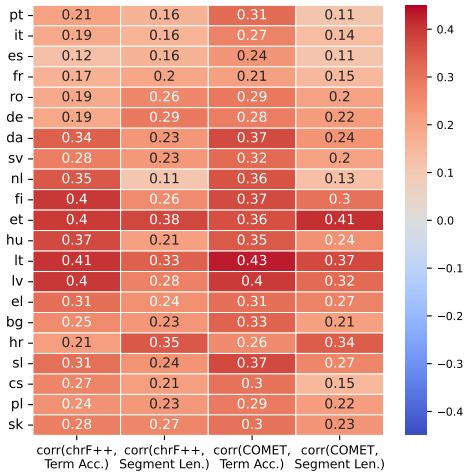


Figure 4: Correlation analysis between the Weighted Financial Term Accuracy and two translation metrics (CHRF++ (left) and COMET (right)) for the NLLB, LLAMA3 and TOWER in the XX→EN direction.

(a) NLLB



(b) LLAMA3



(c) TOWER

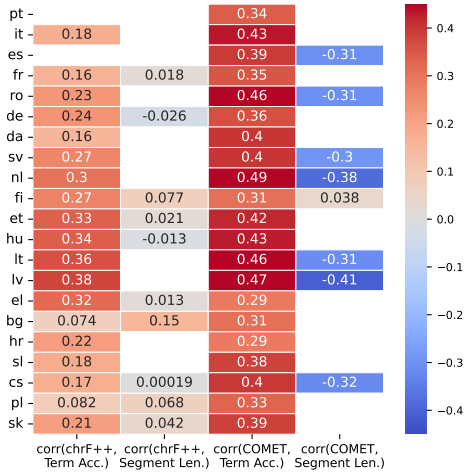


Figure 5: Correlation analysis at the segment-level between the financial term accuracy, weighted by the segment length, and translation performance metrics for one MT system (NLLB) and two LLMs (LLAMA3 and TOWER).