Balancing Fluency and Adherence: Hybrid Fallback Term Injection in Low-Resource Terminology Translation

Kurt Abela[1] **Marc Tanti**[2] **Claudia Borg**[1]

[1]Department of Artificial Intelligence, University of Malta
[2]Institute of Linguistics and Language Technology, University of Malta
{kurt.abela, marc.tanti, claudia.borg}@um.edu.mt

Abstract

Integrating domain-specific terminology into Machine Translation systems is a persistent challenge, particularly in low-resource and morphologically-rich scenarios where models lack the robustness to handle imposed constraints. This paper investigates the trade-off between *static* dictionary-based data augmentation and *dynamic* inference constraints (Constrained Beam Search). We evaluate these methods on two high-to-low resource language pairs: English-Maltese (Semitic) and English-Slovak (Slavic). Our experiments reveal a dichotomy: while dynamic constraints achieve near-perfect Terminology Insertion Rates (TIR), they drastically degrade translation quality (BLEU) in low-resource settings, breaking the fragile fluency of the model. Conversely, static augmentation improves terminology adherence on unseen terms in Maltese (4% → 19%), but fails in the context of a highly inflected language like Slovak. To resolve this conflict, we propose **Hybrid Fallback Term Injections**, a strategy that prioritizes the fluency of static models while using dynamic constraints as a safety net. This approach recovers up to 90% of missing terms while mitigating the quality degradation of pure constraint approaches, providing a viable solution for high-fidelity translation in data-scarce environments.

1 Introduction

Specialised translation domains require strict adherence to specific terminology. A primary challenge in low-resource MT is the "data-efficiency" problem: how to integrate a newly provided terminology database into a system when domain-specific parallel sentences are unavailable or extremely scarce. This paper focuses on methods that use dictionaries to improve terminology adherence without requiring massive parallel corpora. In Machine Translation (MT), this remains an open problem, especially in a low-resource scenario. While Neural Machine Translation (NMT) systems have achieved remarkable fluency, they struggle with rare, domain-specific terms, often opting for frequent synonyms or hallucinations (Koehn and Knowles, 2017; Raunak et al., 2021). This issue is exacerbated in low-resource settings, where the model's limited exposure to diverse contexts makes it resistant to learning new vocabulary and fragile when forced to use specific terms (Hasler et al., 2018).

Current approaches to terminology integration generally fall into two categories: *training-side* (static) and *inference-side* (dynamic). Static methods, such as source-side terminology injection or inline annotation, involve embedding terminology directly into the source sentences during training to bias the model's generation (Dinu et al., 2019). Dynamic methods, such as Constrained Beam Search (CBS) or Grid Beam Search, manipulate the decoding algorithm to force the inclusion of specific tokens at runtime (Post and Vilar, 2018; Hokamp and Liu, 2017).

In high-resource scenarios, dynamic constraints are often preferred for their strict enforcement of terminology. However, in low-resource settings, these constraints can be detrimental (Dinu et al., 2019). A low-resource model, when forced to include a constraint that it does not naturally predict, often sacrifices syntactic coherence, resulting in severe disfluency that satisfies the constraint but fails as a translation (Hasler et al., 2018). Conversely, static methods preserve fluency but do not ensure term inclusion (Dinu et al., 2019), particularly when dealing with the complex morphology typical of low-resource languages (e.g., Semitic or Slavic families), where a lemma-based injection may not match the required inflection (Bergmanis and Pinnis, 2021).

In this work, we explore this stability-adherence trade-off through an evaluation of English-Maltese (ENG-MLT) and English-Slovak (ENG-SLK)

Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), pages 78–86
March 28, 2026 ©2026 Association for Computational Linguistics

translation in the fisheries domain.[1] Our contributions are as follows:

1. **Data Curation via Topic Classification:** For both language pairs, we established a specialised domain baseline by aligning documents from the European Parliament's Committee on Fisheries (PECH)[2] and legislative texts from EUR-Lex.[3] For the ENG-MLT pair, we significantly expanded this corpus by using a BERT-based topic classifier (Micallef et al., 2022) to filter domain-specific documents from the generic DGT Translation Memory (Steinberger et al., 2012).

2. **Benchmarking Augmentation vs. Constraints:** We demonstrate that Acontextual Drilling is effective for Maltese (raising TIR from 5% to 55%) but struggles in the Slovak context due to morphological mismatch. Conversely, we show that while CBS achieves 100% TIR, it causes a significant regression in BLEU scores (up to -6 points in some setups), confirming the fragility of low-resource models under hard constraints.

3. **The Hybrid Fallback Solution:** We introduce a pipelined decoding strategy that attempts translation via the augmented model first, falling back to constrained decoding only when specific terminology is missing. This method achieves the "best of both worlds," recovering ∼90% of terminology while largely preserving the fluency of the static model.

The remainder of this paper is organized as follows: Section 2 reviews related work in terminology constraints and data augmentation. Section 3 details our domain adaptation pipeline and the proposed hybrid strategy. Section 4 presents the comparative experimental results and a qualitative analysis of morphological barriers. Finally, Section 5 concludes the study by listing the limitations and future work.

2 Related Work

The integration of specific terminology into NMT outputs has been approached primarily through two distinct paradigms: inference-time constraints (dynamic) and training-time data augmentation (static).

2.1 Dynamic Inference Constraints

The foundational work in forcing specific lexical constraints during decoding is Grid Beam Search (GBS) by Hokamp and Liu (2017), which extends beam search to ensure the inclusion of target tokens. While effective, GBS suffers from high computational costs. Post and Vilar (2018) improved upon this with Constrained Beam Search (CBS), using a finite-state machine to track constraint satisfaction with significantly lower overhead. This is the implementation used in our experiments.

Despite their guarantees, dynamic constraints are known to be fragile. Hasler et al. (2018) observed that applying hard constraints can degrade overall translation quality if the model is not prepared to handle them. Furthermore, they highlighted that in constrained decoding, models often produce "copying" errors or syntactic breaks when the constrained term conflicts with the model's internal language model. Our work extends these observations specifically to the low-resource regime, quantifying the "fragility" of low-resource models when subjected to such constraints.

Recent shared tasks further highlight the limitations of purely hard inference-time constraints. The WMT 2025 Terminology Translation Task (Semenov et al., 2025) reports that systems relying solely on constrained decoding often struggle with fluency and document-level consistency, motivating hybrid approaches that combine constraint-aware training with selective inference-time control. These results align with our findings that strict decoding constraints, while effective for term insertion, can be brittle when models lack sufficient domain support.

2.2 Static Data Augmentation

An alternative approach involves teaching terminology via data augmentation, often referred to as source-side injection or inline annotation. Song et al. (2019); Dinu et al. (2019) demonstrated that exposing models to terminology directly in the source sentence, either by replacing source words with target translations or appending dictionaries,

[1] Code can be found on Github: https://github.com/MLRS/Balancing-Fluency-and-Adherence-Hybrid-Fallback-Term-Injection

[2] Documents retrieved from the European Parliament Public Register: https://www.europarl.europa.eu/committees/en/pech/documents/

[3] Access to European Union law: https://eur-lex.europa.eu/

can bias the model towards correct terminology usage without altering the decoding algorithm.

More recent work has revisited training-side terminology integration in light of large-scale pretrained and instruction-tuned models. Xu and Carpuat (2021) propose soft lexical constraints that are optimized jointly with the translation objective, allowing models to trade off constraint satisfaction and fluency during training rather than enforcing hard decisions at inference time. Similarly, Kim et al. (2024) investigate efficient terminology integration for LLM-based translation systems, showing that even strong generative models benefit from explicit terminology signals when translating specialised content. These findings reinforce the view that soft or fallback-based mechanisms are better aligned with model uncertainty, particularly in low-resource or domain-shifted scenarios.

Other approaches focus on "soft constraints" using special tags for inline annotation. Bergmanis and Pinnis (2021) and Exel et al. (2020) proposed embedding target terminology directly into the source sentence (alongside the corresponding source term) using special tokens (e.g., `<term_start> term <term_end>`). This allows the model to learn a copying mechanism or a specific translation path for marked terms. While successful in high-resource settings, the efficacy of these methods in low-resource, morphologically rich scenarios remains under-explored.

2.3 Low-Resource and Domain Adaptation

Domain adaptation in low-resource NMT is notoriously difficult due to the risk of overfitting or catastrophic forgetting (Koehn and Knowles, 2017). Williams et al. (2023) established baselines for English-Maltese using generic data, while Benkova et al. (2021) investigated similar trade-offs between general and domain-specific systems for English-Slovak. Both highlight the need for specialised data filtering.

To address data scarcity, techniques like back-translation (Sennrich et al., 2016) and transfer learning are standard. However, precise terminology adaptation often requires lexical overlap that back-translation alone cannot guarantee. Recent trends use pre-trained monolingual models to filter parallel corpora for domain specificity (Aulamo et al., 2020; Zhang et al., 2020b). This approach ensures that the limited compute budget of low-resource training is spent on high-quality, relevant samples.

3 Data and Methodology

We investigate the impact of terminology integration strategies on two low-resource language pairs: English-Maltese (ENG-MLT) and English-Slovak (ENG-SLK). Both pairs involve morphologically rich target languages (Semitic and Slavic, respectively) and present distinct challenges for constraint adherence.

In our experiments, we use baseline models for each language pair to evaluate the models' knowledge of the new domain. For our experiments, we chose the **fisheries** domain (EU legislation and reports regarding maritime policy) as the new domain. The choice is made on the basis of the availability of high-quality, domain-specific terminology in the IATE database, combined with the accessibility of parallel European Parliament Committee on Fisheries (PECH) documents, which allows for a controlled evaluation of domain adaptation in a low-resource setting.

We experiment with two key techniques: (i) Static Acontextual Augmentation and (ii) Dynamic Constraint Decoding (CBS). Based on the results, we propose a third strategy, which we call Hybrid Fallback. This uses the output of Static Acontextual Augmentation to verify the presence of the required target term. If the term is missing, it falls back to re-decoding using CBS as a safety net.

To conduct the evaluation, we automatically collate a parallel corpus of European Parliament documents related to fisheries. We split the dataset for both fine-tuning and testing.

3.1 Data Curation and Domain Adaptation

We established a specialised domain dataset for both Maltese and Slovak, with curation methods varying according to resource availability, as described below.

3.1.1 Maltese (ENG-MLT) Data Setup

Generic Baseline: We trained a baseline model using the 2.9M sentence pairs from the English-Maltese parallel corpus curated by Williams et al. (2023). The corpus is heavily skewed towards formal domains, consisting of approximately 55% legal, 32% parliamentary, and 11% health-related texts, with less than 1% representing generic or informal domains. We utilized the raw text versions of the corpus to apply our own preprocessing pipeline. This involved re-tokenizing the target side using the `MLRS/BERTu` vocabulary (Micallef et al., 2022) and filtering sentence pairs to match the sequence length constraints of our Transformer architecture.

In-Domain Data Mining: To curate a domain-specific fine-tuning set, we used a semi-supervised filtering approach. We first fine-tuned a BERT-based Maltese model, **BERTu** (Micallef et al., 2022), on the MultiEURLEX dataset (Chalkidis et al., 2021) to predict top-level domain labels based on EUROVOC descriptors (Publications Office of the European Union, 2023). This classifier, which achieved an F1 score of 77.4 (macro-average over 21 top-level domains), was applied to the generic DGT Translation Memory (Steinberger et al., 2012). Documents classified as *Fisheries* for which both the Maltese and English documents were available were collected to create the Fine-Tuning set.

Fine-Tuning Set: The mined DGT data was combined with manually aligned legislative texts (see below), yielding a robust specialised domain corpus of **96,558** sentence pairs. From this corpus, we randomly held out 2,500 pairs for validation and 2,500 pairs for testing, leaving 91,558 pairs for fine-tuning.

3.1.2 Slovak (ENG-SLK) Data Setup

Generic Baseline: We trained a baseline model on the OPUS-100 corpus (Zhang et al., 2020a), consisting of approximately 1M sentence pairs. While multilingual pre-trained models such as Tiedemann and Thottingal (2020) are publicly available and partially trained on the same dataset, we opted to train a custom bilingual model for two primary reasons. Firstly, we are using the Fairseq toolkit, thus using the same framework throughout allows for a seamless and controlled fine-tuning pipeline. Secondly, it ensures architectural parity with our ENG-MLT setup, maintaining the same Transformer configuration to ensure that observations regarding terminology adherence are not skewed by differences in model capacity. There is a strong domain overlap with this training data and our test set, since OPUS-100 contains a lot of legal data.

In-Domain Manual Curation: Unlike the Maltese setup, where we used a classifier for data selection, we relied on manual curation rather than automated mining. We aligned documents from two primary sources:

1. **Committee Drafts:** Draft reports from the European Parliament's Committee on Fisheries (PECH), retrieved from the public register[4].
2. **Legislative Texts:** Specific fisheries regulations (e.g., CELEX:52023PC0587[5]) scraped from **EUR-Lex**[6].

Fine-Tuning Set: This manual curation resulted in a high-quality but extremely sparse specialised domain dataset of **5,736** sentence pairs. Due to the data scarcity, we reserved a smaller split of 1,000 pairs each for validation and testing, resulting in a remaining training set of 3,736 pairs.

3.1.3 Terminology Dictionaries and POS Filtering

We extract terminology from the Interactive Terminology for Europe (IATE) database[7], filtering specifically for the fisheries domain. For **ENG-MLT**, we find 3,343 unique term pairs, whilst for **ENG-SLK**, we find 1,365 unique term pairs.

To investigate the impact of grammatical category on constraint stability, we processed these dictionaries using language-specific Part-of-Speech (POS) taggers. We used `bert-base-uncased` fine-tuned for POS tagging (Devlin et al., 2018) to identify and isolate nouns within the English source terms. This allowed us to create a **Nouns-Only** subset of the dictionary to assess how results would be affected by the morphologically rich nature of both language pairs. We present the comparative results between the full dictionary and this nouns-only subset in Section 4.

Test Set Annotation (Seen vs. Unseen) To test the models' capacity for generalisation (versus memorisation), we stratified the test set terminology based on exposure during training. We cross-referenced every target term required in the test sets against the respective training corpora. Terms present in the training data were classified as *Seen*, while those appearing exclusively in the test set were classified as *Unseen*. This distinction enables the calculation of **Unseen TIR** in Section 4, to serve as a metric for determining how much a model can handle terminology constraints that it has not previously memorized from the source data.

[4] https://www.europarl.europa.eu/committees/en/pech/home/highlights

[5] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52023PC0587

[6] https://eur-lex.europa.eu/homepage.html

[7] https://iate.europa.eu/

3.2 Model Architecture and Training

All translation models are based on the Transformer architecture (Vaswani et al., 2017) and were trained using the Fairseq toolkit (Ott et al., 2019). We deliberately avoid using massive pre-trained models to ensure that any improvements in terminology adherence are a direct result of our data-efficient injection strategies rather than inherited weights from a high-resource multilingual model.

Hyperparameters We use a standard Transformer configuration suitable for low-resource settings. The models comprise 6 encoder and 6 decoder layers, with an embedding dimension of 512, a feed-forward dimension of 2,048, and 8 attention heads. For training, we use the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.98$) with an inverse square root learning rate schedule and a warmup of 4,000 steps. We apply label smoothing ($\epsilon = 0.1$) and dropout (0.3) to mitigate overfitting.

Tokenization To ensure consistent handling of morphology, we use WordPiece tokenization. For the ENG-MLT pair, we use the `MLRS/BERTu` tokenizer (Micallef et al., 2022) for the target side, as this vocabulary was used in the pre-trained Maltese NMT baseline. For the English-Slovak pair, we use the `gerulata/slovakbert` tokenizer (Pikuliak et al., 2021) for the target side to maintain methodological symmetry. For the English source side, we use the `bert-base-cased` tokenizer.

3.3 Methods for Terminology Integration

We compare two approaches to integrating terminology into the NMT pipeline and propose a third approach referred to as Hybrid Fallback Term Injection.

3.3.1 Method 1: Static Acontextual Augmentation

We adopt a list-based data augmentation strategy, often referred to as "dictionary drilling" (Exel et al., 2020). Given a terminology dictionary D, we append each term pair directly to the fine-tuning data as a distinct training example. To ensure the model attends to these short sequences during training, we apply an oversampling factor of $K = 10$, meaning that each dictionary entry is repeated 10 times and appended to the training corpus.

3.3.2 Method 2: Dynamic Constrained Decoding (CBS)

For dynamic integration, we use CBS (Post and Vilar, 2018). During inference, we identify all source terms present in the input sentence via regex matching against the IATE dictionary. The complete list of corresponding target terms is passed to the decoder as positive constraints, forcing their inclusion in the output.

3.3.3 Method 3: Hybrid Fallback Term Injection Strategy

We propose a pipelined decoding strategy to balance fluency and adherence. The system first generates a translation using the statically augmented model (Method 1). We then automatically verify the presence of the required target terms in the output.

Verification Mechanism: The verification process uses strict string matching against the dictionary term. If the augmented model generates a term that does not exactly match the glossary form, it is flagged as missing.

- If the augmented model successfully produced the terms (exact match), the translation is accepted.
- If terms are marked as missing, the system falls back to re-decoding the specific sentence using CBS (Method 2).

This strategy uses CBS as a "safety net" for stubborn terms that the static model fails to recall.

4 Results and Discussion

We evaluate the proposed methods on the specialised fisheries test sets. We report BLEU and chrF++ for translation quality. For constraint adherence, we report **Terminology Insertion Rate (TIR)** and **Unseen TIR** (performance on terms not present in the training corpus). Statistical significance is calculated using a paired t-test ($p < 0.05$).

4.1 Main Results: The Stability-Adherence Trade-off

Table 1 (ENG-MLT) and Table 2 (ENG-SLK) present the results. When analyzing Static Augmentation, a clear dichotomy emerges between the two language pairs. For Maltese, static injection significantly improves adherence, quadrupling the Unseen TIR from 4.38% to 19.06%. In contrast, for Slovak, the method fails to yield improvements, with TIR remaining stagnant.

Meanwhile, CBS shows a consistent pattern across both languages: they achieve perfect TIR (100%) but at the cost of translation quality (BLEU). The Hybrid method consistently bridges this gap, as it was able to maintain high BLEU with a high TIR.

4.1.1 Static Augmentation

For **ENG-MLT**, the fine-tuned baseline already shows high adherence (TIR 49.73%) due to specialised further fine-tuning. Importantly, the model maintains the high translation quality of the baseline (60.21 BLEU and 76.62 chrF++) while significantly increasing TIR to 55.47% and quadrupling performance in unseen TIR (4.38% → 19.06%)

In contrast, for **ENG-SLK**, Static Augmentation fails to make an impact, with TIR remaining stagnant at $\sim$16%. While the complex morphology of Slovak presents a challenge, we attribute this failure primarily to the scarcity of specialised training data. As detailed in Section 3, the Slovak fine-tuning set consisted of only $\sim$5,700 specialised domain pairs, compared to over 96,000 for Maltese. This limited quantity appears insufficient for the model to effectively generalize the dictionary terms injected via augmentation.

4.1.2 CBS

Dynamic CBS achieves 100% TIR across the board but incurs a significant quality penalty, confirming the fragility of low-resource models under hard constraints. For **ENG-MLT** (Table 1), applying CBS to the Static model precipitates a sharp drop in BLEU from 60.21 to 53.85.

We observe an identical pattern in **ENG-SLK** (Table 2). The Slovak model suffers a comparable degradation, dropping from 61.16 to 55.80 BLEU (-5.36 points). This consistency suggests that the brittleness of constrained decoding is not language-specific but rather a symptom of the low-resource regime. When these models are forced to include tokens they cannot probabilistically support, they sacrifice local syntactic coherence, resulting in disfluent output.

4.1.3 Hybrid Fallback Term Injection Strategy

The **Hybrid Fallback Term Injection** method navigates this trade-off between translation quality and terminology adherence by utilizing CBS only as a secondary decoding pass. This approach consistently yields higher BLEU scores than pure constrained decoding (Method 2) while maintaining high TIR.

In **ENG-MLT**, the full hybrid strategy achieves 54.14 BLEU, representing a modest improvement over the 53.85 score produced by pure CBS. A more significant improvement is observed with the *Nouns-only* hybrid variant, which reaches 59.15 BLEU. Although this configuration results in a slight reduction in adherence (85.71% vs. 90.94% Unseen TIR), it demonstrates that noun-based constraints are less disruptive to the model's target-side fluency. Furthermore, the Nouns-only approach reduces the CBS fallback frequency from 26.76% to 11.16%, indicating that the static model is more likely to generate correct nominal forms naturally. These results suggest that the quality degradation observed in pure CBS is largely driven by forcing the inclusion of non-nominal terms, which require complex Semitic morphological agreement that the low-resource model cannot reliably support.

For **ENG-SLK**, the full hybrid method similarly improves translation quality over pure CBS (57.15 vs. 55.80 BLEU). The most effective results are achieved by the *Nouns-only* hybrid variant, which reaches 63.93 BLEU, notably surpassing the specialized fine-tuning baseline of 61.48. Critically, this variant maintains 100.0% Unseen TIR. The fact that restricting constraints to nouns improves both quality and adherence in Slovak suggests that non-nominal constraints often introduce syntactic conflicts that prevent the decoder from successfully placing even valid terms. By focusing on noun-based constraints, the model achieves perfect adherence with a low fallback rate (12.60%), significantly improving both decoding speed and output quality compared to pure constrained approaches.

4.2 Qualitative Analysis

To understand the source of the BLEU degradation in fully constrained models, we performed a manual error analysis on the ENG-MLT language pair (Table 3).

Firstly, we observe that the baseline model is prone to hallucination in this low-resource setting. As shown in the first example, the model hallucinates domain-specific terms such as "merluzz" (cod) and invents years in the target output even when they are absent from the source. This confirms the instability of the unaugmented low-resource model.

Table 1: ENG-MLT Results. Comparison of Baseline, Static Augmentation, CBS, and Hybrid Fallback Term Injection. Speed is measured in sentences per second (sent./sec). (†) indicates statistical significance ($p < 0.05$) compared to the Fine-tuned Baseline.

Model / Method	**BLEU**	**chrF++**	**TIR**	**Unseen TIR**	**Speed** *(sent./sec)*
Baseline model	32.11	52.21	34.38%	5.00%	43.18
Fine-tuning on specialised corpus	59.96	67.91	49.73%	4.38%	45.80
Method 1: Static Augmentation					
Static Aug. (Acontextual Drill)	**60.21**†	**68.13**†	55.47%	19.06%	**48.99**
Method 2: Dynamic Constraints (CBS)					
Static Aug. + Dynamic CBS	53.85	62.49	**100.00%**	**100.00%**	5.91
Static Aug. + Dynamic CBS (Noun Constraints Only)	58.63	66.81	**100.00%**	**100.00%**	8.28
Method 3: Hybrid Fallback Term Injection					
Hybrid (Static → CBS)	54.14	62.69	89.85%	90.94%	5.27
Hybrid (Static → CBS, Nouns Only)	59.15	66.95	83.62%	85.71%	6.81

Table 2: ENG-SLK Results. Comparison of Baseline, Static Augmentation, CBS, and Hybrid Fallback Term Injection. Speed is measured in sentences per second (sent./sec). (†) indicates statistical significance ($p < 0.05$) compared to the Fine-tuned Baseline.

Model / Method	**BLEU**	**chrF++**	**TIR**	**Unseen TIR**	**Speed** *(sent./sec)*
Baseline model	60.95	62.88	17.28%	0.71%	22.73
Fine-tuning on specialised corpus	**61.48**	**62.99**	16.71%	2.13%	**22.68**
Method 1: Static Augmentation					
Static Aug. (Acontextual Drill)	61.16	62.79	16.43%	2.13%	22.14
Method 2: Dynamic Constraints (CBS)					
Static Aug. + Dynamic CBS	55.80†	58.38†	**100.00%**	**100.00%**	5.47
Static Aug. + Dynamic CBS (Noun Constraints Only)	60.39†	62.45†	**100.00%**	**100.00%**	8.02
Method 3: Hybrid Fallback Term Injection					
Hybrid (Static → CBS)	57.15†	59.45†	83.85%	66.67%	4.42
Hybrid (Static → CBS, Nouns Only)	63.93†	65.50†	92.50%	100.00%	5.32

However, the Dynamic (CBS) approach introduces a different error type: context blindness, where the model ignores the semantic context of the source sentence to satisfy a lexical constraint. For instance, the English term "header" (in a document context) was constrained to the fisheries translation "qtugħ ir-ras" (the physical decapitation of a fish). Similarly, the term "Draft" (as in a document draft) was forced to "Pixka" (a catch of fish). Because CBS forces these terms regardless of the surrounding context, the model sacrifices semantic logic to satisfy the constraint, leading to a drop in BLEU scores.

Finally, the Hybrid Fallback Term Injection approach demonstrates better data integrity. In cases like alphanumeric codes (e.g., "COD/03AN"), where baselines often attempt to translate the substring "COD" into the fish name "Bakkaljaw," the Hybrid Fallback Term Injection approach correctly identifies when to fallback, preserving the code exactly as required.

5 Conclusion

In this paper, we investigate the challenge of integrating domain-specific terminology into low-resource MT. Our results reveal a fundamental trade-off between *stability* (translation fluency) and *adherence* (terminology usage) in data-scarce environments.

We find that standard integration methods exhibit distinct limitations. Static Acontextual Augmentation proved effective for Maltese, where fine-tuning was performed on a concentrated, specialised domain dataset. However, the same method failed for Slovak, potentially due to the smaller specialised domain fine-tuning data setup.

In contrast, Dynamic Constraints (CBS) achieved high TIR but imposed a severe penalty on translation quality. The significant regression in BLEU scores confirms that low-resource models lack the probability mass to accommodate forced tokens without compromising syntactic coherence.

To resolve this, we introduced a Hybrid Fallback Term Injection strategy. By prioritizing the fluent output of the augmented model and using constrained decoding solely as a fallback mechanism, this approach recovered up to 90% of missing terminology without the catastrophic quality loss associated with pure constraints.

Table 3: **Qualitative Error Analysis.** Examples of hallucination and context blindness.

Method	Error Type	Description
Baseline	Hallucination	**Src:** "The Commission proposes..." **Out:** "L-istokk tal-merluzz..." **Tgt:** "Il-Kummissjoni tipproponi" *Analysis:* Hallucinates "merluzz" (cod) and repeats years erroneously.
CBS	Context Blindness	**Src:** "...header of each amendment" **Out:** "...qtugħ ir-ras..." **Tgt:** "l-intestatura ta' kull emenda" *Analysis:* Forces the fisheries term for "heading" (decapitation of fish), resulting in the "decapitation of an amendment."
CBS	Context Blindness	**Src:** "Draft opinion" **Out:** "Pixka ta' opinjoni" **Tgt:** "Abbozz ta' opinjoni" *Analysis:* Forces "Pixka" (fish/catch) instead of "Abbozz" (document draft).
Hybrid	Data Integrity	**Src:** "COD/03AN" **Out:** "COD/03AN" **Tgt:** "COD/03AN" *Analysis:* Correctly preserves alphanumeric codes where baselines often attempt to translate "COD" to "Bakkaljaw" (which is the translation of the fish "Cod").

5.1 Limitations and Future Work

Computational Cost The primary drawback of the Hybrid Fallback Term Injection implementation is inference latency. Because the Static model fails to recall terms frequently (recall $\approx$ 55%), it triggers the expensive CBS fallback for nearly half the test set. As a result, it operates at a lower speed (sent./sec) than the CBS baseline in our experiments. Future work should explore lighter-weight constraint mechanisms to improve this speed-accuracy profile.

Morphological Complexity We emphasize that our current verification mechanism uses strict surface-form matching. While this allows for rigorous testing of adherence, it penalizes valid morphological inflections. Future work proposes exploring *inflection-aware data augmentation* and morphological analyzers in the verification loop to better handle the varied grammatical cases required by the target context.

References

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. Opusfilter: A flexible tool for filtering and combining parallel corpora. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4254–4261.

Lucia Benkova, Dasa Munkova, L'ubomír Benko, and Michal Munk. 2021. Evaluation of english–slovak neural and statistical machine translation. *Applied Sciences*, 11(7):2948.

Toms Bergmanis and Marcis Pinnis. 2021. Context-independent terminology translation with neural machine translation models. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 142–153.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068.

M. Exel, M. Cettolo, and P. Passban. 2020. Terminology-constrained neural machine translation training data generation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 231–240.

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 506–512.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1535–1546. Association for Computational Linguistics.

Sejoon Kim, Mingi Sung, Jeonghwan Lee, Hyunkuk Lim, and Jorge Gimenez Perez. 2024. Efficient terminology integration for LLM-based translation in specialized domains. In *Proceedings of the Ninth Conference on Machine Translation*, pages 636–642, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Kurt Micallef, Albert Gatt, Marc Tanti, Lonneke van der Plas, and Claudia Borg. 2022. Pre-training data quality and quantity for a low-resource language: New corpus and BERT models for Maltese. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 90–101, Hybrid. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Li, Ashish Ghouri, Michael Dauphin, Michael Auli, and David Grangier. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Matúš Pikuliak, Štefan Grivalský, Martin Konôpka, Miroslav Blšták, Martin Tamajka, Viktor Bachratý, Marián Šimko, Pavol Balážik, Michal Trnka, and Filip Uhlárik. 2021. Slovakbert: Slovak masked language model. *Preprint*, arXiv:2109.15254.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1314–1324. Association for Computational Linguistics.

Publications Office of the European Union. 2023. EuroVoc: the eu's multilingual thesaurus.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. Findings of the WMT25 terminology translation task: Terminology is useful especially for good MTs. In *Proceedings of the Tenth Conference on Machine Translation*, pages 554–576, Suzhou, China. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing nmt with pre-specified translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 449–459.

Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Carrasco-Benitez Manuel, Schlüter Patrick, Przybyszewski Marek, and Gilbro Signe. 2012. DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 454–459. European Language Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Aiden Williams, Kurt Abela, Rishu Kumar, Martin Bär, Hannah Billinghurst, Kurt Micallef, Ahnaf Mozib Samin, Andrea DeMarco, Lonneke van der Plas, and Claudia Borg. 2023. UM-DFKI Maltese speech translation. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 433–441, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Weijia Xu and Marine Carpuat. 2021. EDITOR: An edit-based transformer with repositioning for neural machine translation with soft lexical constraints. *Transactions of the Association for Computational Linguistics*, 9:311–328.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639.

Boliang Zhang, Ajay Nagesh, and Kevin Knight. 2020b. Parallel corpus filtering via pre-trained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8545–8554, Online. Association for Computational Linguistics.