# Morphology Aware Source Term Masking for Terminology-Constrained NMT

**Ander Corral** and **Xabier Saralegi**
Orai NLP Technologies
{a.corral,x.saralegi}@orai.eus

## Abstract

Terminology-constrained NMT systems facilitate the forced translation of domain-specific vocabulary. A notable method in this context is the *copy-and-inflect* approach, which appends the target term lemmas of constraints to their corresponding source terms in the input sentence. In this work, we propose a novel adaptation of the *copy-and-inflect* method, referred to as *morph-masking*. Our method involves masking the source terms of the constraints from the input sentence while retaining essential grammatical information. Our approach is based on the hypothesis that *copy-and-inflect* systems have access to both source and target terms, allowing them to generate the correct surface form of the constraint by either translating the source term itself or properly inflecting the target term lemma. Through extensive validation of our method in two translation directions with different levels of source morphological complexity, Basque to Spanish and English to German, we have demonstrated that *morph-masking* is capable of providing a harder constraint signal, resulting in a notable improvement over the *copy-and-inflect* method (up to 38% in term accuracy), especially in challenging constraint scenarios.

## 1 Introduction

While Neural Machine Translation (NMT) achieves high quality results in general-purpose translation scenarios, it frequently encounters challenges with precise technical terminology in specialized domains, as noted by Alam et al. (2021). To address this limitation, terminology-constrained NMT facilitates the forced translation of specific terminology, ensuring consistent and reliable translation of domain-specific vocabulary, thus considerably reducing post-editing efforts.

Recent research in terminology-constrained NMT predominantly adopts a data-driven approach. This method involves teaching systems to apply terminology constraints through training with synthetic, task-specific data (Dinu et al., 2019; Michon et al., 2020; Bergmanis and Pinnis, 2021). Specifically, Bergmanis and Pinnis (2021) introduced a *copy-and-inflect* method. This method appends the lemmas of constraints' target terms to their corresponding source terms within the input sentence. The system is then trained to produce translations by appropriately copying and inflecting these target terms based on the context (see annotation example in Table 1).

However, available evidence suggests that *copy-and-inflect* methods do not consistently enforce terminology constraints (Bergmanis and Pinnis, 2021; Zhang et al., 2023). Our hypothesis is that these methods, having access to both the source and target terms of the constraints, only provide a *soft* constraint. In other words, they might generate the correct surface form of the constraint either by translating the source term directly or by properly inflecting the lemma of the target term.

Given this hypothesis, we introduce a novel variation of the *copy-and-inflect* method designed to provide a stronger constraint signal to the system. Specifically, **we propose to mask the source terms of constraints in the input sentence while retaining the crucial grammatical information, such as as gender, number, grammatical cases, etc.**. We contend that maintaining this information is vital, especially for morphologically rich languages like Basque, to prevent any degradation in translation quality due to a loss of grammatical context after masking.

While much of the previous research examining the effects of masking source terms has focused on English as the source language (Dinu et al., 2019; Exel et al., 2020; Michon et al., 2020), we evaluate our approach on two translation directions, each with varying degrees of source morphological complexity: English to German and Basque to Spanish. These language pairs were selected



*Findings of the Association for Computational Linguistics: EACL 2024*, pages 1676–1688

to encompass a wide variety of linguistic features and complexities. Spanish and Basque, belonging to different language families, display significant differences in morphology and syntax. Although English and German are both Germanic languages and share some similarities, German has a much more complex morphology. Consistent with previous research by Bergmanis and Pinnis (2021), we translate to morphologically rich languages to assess the inflection capabilities of the systems.

To the best of our knowledge, the Basque to Spanish translation direction has not been previously explored. Consequently, we have manually created a challenging test set[1] for this translation direction, which we anticipate will be a valuable resource for subsequent research.

## 2 Related Work

Works addressing terminology-constrained NMT mainly fall into two different categories: a) constrained decoding-based approaches and b) data-driven approaches.

**Constrained decoding approaches** modify the decoding algorithm to force the model to apply terminology constraints when predicting the next token (Hokamp and Liu, 2017; Post and Vilar, 2018; Hu et al., 2019). While constrained decoding ensures the presence of the required terminology, it can significantly slow down the decoding process (Dinu et al., 2019) and strict enforcement of the constraints can result in lower quality translations (Bergmanis and Pinnis, 2021).

**Data-driven approaches** train systems with synthetic task-specific data to learn how to apply terminology constraints (Dinu et al., 2019; Michon et al., 2020; Bergmanis and Pinnis, 2021). The main advantage of this approach is that it does not require any changes in the model architecture nor in the decoding algorithm. There is no inference time overhead either. As a result, recent efforts have concentrated on methodologies employing various data generation strategies for this task.

For instance, Bergmanis and Pinnis (2021) proposed a *copy-and-inflect* method which appends constraint's target terms lemmas to their corresponding source terms in the input sentence. With additional source factors (Sennrich and Haddow, 2016) they indicate whether the words in the input sequence belong to the source term of the constraint, to the target term or the word is not part of the constraint. Then, the system is trained to generate translations by properly copying and inflecting those target terms depending on the context. The method is based in the original *copy* method proposed by Dinu et al. (2019) but they use lemmas instead of the final form of the terms. This is specially important when translating to morphologically rich languages where each word has several surface forms depending on the context.

Related to our masking approach, both (Dinu et al., 2019) and (Exel et al., 2020) explore what they refer to it as the *replace* setting, in which the source term is entirely masked. While Dinu et al. (2019) report findings similar to the *append* setting, Exel et al. (2020) find that the *replace* method underperforms. Notably, both studies evaluate the *replace* setting using English as the source language, a language that has fewer surface forms per word compared to morphologically rich languages, such as Basque.

## 3 Our method: morphology aware source term masking

We introduce a novel adaptation of the *copy-and-inflect* method (Bergmanis and Pinnis, 2021) which we call 'morphology aware source term masking', hereinafter referred to as, *morph-masking*. This approach involves masking the source term of the constraints within the input sentence, aiming to deliver a more robust constraint signal to the system. Before masking, grammatical information -such as gender, number, and grammatical cases- is extracted from the masked term. We argue that this information is crucial, especially for languages with complex morphology like Basque, to prevent losing grammatical details after masking that could adversely impact the overall translation quality. The target term lemma and the tokens representing the extracted grammatical information are then inserted in place of the masked source term.

As in Bergmanis and Pinnis (2021), we distinguish constraints from the original source sentence words using additional source factors (Sennrich and Haddow, 2016). We employ BIO tags —abbreviations for "Beginning, Inside, and Outside" tags— which are frequently utilized in Named Entity Recognition (NER) tasks, to annotate target terms. These tags are instrumental in structuring and labeling constraints, especially for multi-word terms. Additionally, we use an extra information

[1]Datasets used in the experiments are available at https://github.com/orai-nlp/terminology-constrained-NMT

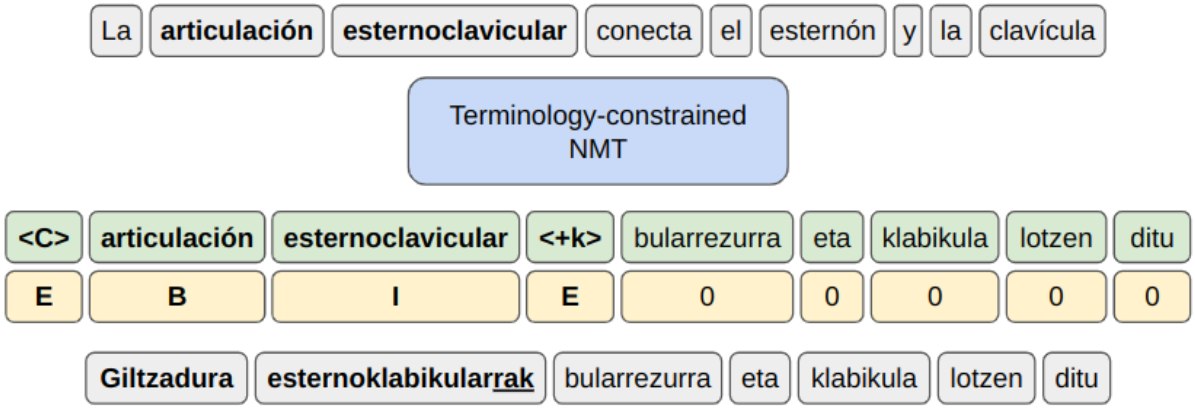


Figure 1: Illustration of the proposed annotation method *morph-masking*. Constraints' source terms in the input sentence are masked and replaced with the target terms while preserving the necessary grammatical information in the source terms such as as the gender, the number, grammatical cases, etc. We differentiate constraints from the original source sentence words using additional source factors (Sennrich and Haddow, 2016). English translation: *The **sternoclavicular joint** connects the sternum and the clavicle.*

tag (E) to differentiate between words and the extracted grammatical information tokens. See Figure 1 for a complete example of the proposed annotation.

We annotate a constraint only when the source term appears in the source sentence and the target term is present in the reference sentence. To identify annotation candidates, both source and reference sentences, as well as dictionary entries, are first lemmatized. This lemmatization step is essential to find words in morphologically complex languages such as Basque. Target terms are annotated in their dictionary form, that is, lemmatized. Our annotations are limited to common nouns, proper nouns, and adjectives.

To compare our *morph-masking* method to the *copy-and-inflect* method proposed by (Bergmanis and Pinnis, 2021), we follow their annotation guidelines to generate the training data. That is, constraints' target terms lemmas are appended to the source terms in the input sentence. In this case, additional source factors are used to differentiate between source terms (1), target terms (2) and other words (0). Table 1 shows an annotation example for both methods. To ensure a fair comparison, the same constraints are employed in generating the examples for both methods.

### 3.1 Extracted grammatical information

Understanding the intricacies of a language is essential when it comes to accurately extracting grammatical information. Each language has its unique set of rules, structures, and nuances controlling how words are inflected and modified. For instance, Basque is an agglutinative language primarily characterized by its rich suffix-based morphology. These inflectional suffixes indicate the grammatical case (absolutive, ergative, dative,...) of words within a sentence. The morphology of these suffixes depends on several grammatical features, such as the number, either singular, plural or undefined (*mugagabea*). In English those grammatical cases are commonly encoded using prepositions leaving the word unaltered. Consequently, each word in Basque has a higher number of variations in comparison to minimally-inflected languages such as English. For example the word *dog* in English can adopt two forms depending on the number, *dog* (singular) and *dogs* (plural). In contrast, the corresponding word *txakur* can have multiple forms depending on the grammatical cases and features, such as, *txakurra*, *txakurrak*, *txakurrarekin*, *txakurrentzat*, *txakurrarena*, etc. If the Basque word *txakurrentzat* (plural benefactive case) is masked, meaning *for the dog*, essential grammatical information from the original source sentence is lost. In this case a plural token (<*pl*>) and the grammatical case token (<+*entzat*>) are extracted and appended to the input sentence.

In this study, we focus on analyzing Basque and compare it with English. Specifically, for Basque we extract the grammatical case suffixes for common nouns, proper nouns and adjectives. The plural number for common nouns is also extracted. For English we only extract the plural number and a comparative/superlative token for common nouns.

| | |
|---|---|
| Glossary entry | *giltzadura esternoklabikular –> articulación esternoclavicular* |
| Source | **Giltzadura esternoklabikularrak** bularrezurra eta klabikula lotzen ditu gorputzean |
| copy-and-inflect | **Giltzadura**$_1$ **esternoklabikularrak**$_1$ **articulación**$_2$ **esternoclavicular**$_2$ bularrezurra$_0$ eta$_0$ klabikula$_0$ lotzen$_0$ ditu$_0$ gorputzean$_0$ |
| morph-masking | **<C>**$_E$ **articulación**$_B$ **esternoclavicular**$_I$ **<+k>**$_E$ bularrezurra$_0$ eta$_0$ klabikula$_0$ lotzen$_0$ ditu$_0$ gorputzean$_0$ |
| Translation | La **articulación esternoclavicular** conecta el esternón y la clavícula en el cuerpo humano |
| English | *The **sternoclavicular joint** connects the sternum and clavicle in the human body* |

Table 1: Comparison of the *copy-and-inflect* and *morph-masking* annotation methods for the Basque to Spanish translation direction. Additional source factors are represented by subscripts. For the *morph-masking* method, the ergative grammatical case of the original Basque term **Gitzadura esternoklabikularrak** is extracted and appended as an extra token **<+k>**. Casing information, **<C>**, is also extracted as an additional token.

For both languages the casing of the source word, either uppercased or cased, is also used as additional information. See Appendix A for more details on the extracted grammatical information and the corresponding tokens.

## 4 Experimentation

All the systems were trained using the default configuration for the Transformer architecture (Vaswani et al., 2017) as implemented in the PyTorch version of the OpenNMT toolkit (Klein et al., 2017). We apply BPE tokenization (Sennrich et al., 2016) learned on 32,000 merge operations on the joint training parallel data. Sentences larger than 100 subwords after tokenization are discarded from the training set.

First, general purpose NMT systems were trained to be used as the baselines. The Basque-Spanish baseline was trained on the Basque-Spanish portion of the Paracrawl corpus (v9) (Bañón et al., 2020). Data was splitted into train, validation and test sets with 3.3M/5K/5K parallel sentences respectively. The total vocabulary size after applying BPE tokenization was 42K for Basque and 36K for Spanish. Similarly, the English-German baseline was trained on the English-German portion of the Paracrawl corpus (v9). In this case, training, validation and test sets consist of 278M/5K/5K parallel sentences respectively. A vocabulary size of 58K tokens is used for both English and German.

We followed an annotation method designed for easy extension across a broad spectrum of language pairs. To achieve this, we decided to leverage the Apertium toolkit (Forcada et al., 2011), an open-source rule-based machine translation toolkit that already covers many language pairs. This toolkit provides essential tools for lemmatization and morphological analysis, both crucial for our annotation process. Additionally, Apertium offers bilingual dictionaries, which we employ as constraints. Although many of the dictionary entries can potentially be commonly used words, we argue that the system must learn how to apply terminology constraints rather than learning the annotated words themselves.

Apertium's Basque-Spanish and English-German bilingual dictionaries were used for the annotation step. These dictionaries were divided into train and test set, with the test set comprising 10% of the entries. For the Basque-Spanish language pair we annotate the entire training parallel data following the annotation procedure described in Section 3. Segment pairs lacking annotations -samples for which no constraint was found- were discarded. For the English-German translation direction, we limited our annotation to 10M sentences from the training data, also skipping samples without annotations. Annotating the full training parallel data, 278M segments, in this case is an expensive task and there should be enough annotated training samples to learn the task. We generate samples with different number of constraints. Specially, 50% of the samples have a single constraint while the remaining samples are annotated with 2 to 5 constraints randomly sampled. Source factors are appropriately transposed from word-level to BPE token level.

Unlike prior works (Bergmanis and Pinnis, 2021; Zhang et al., 2023), the terminology-constrained systems were trained by fine-tuning the baseline

system on the annotated data sets. This avoids training the system from scratch which means already existing strong baselines can be adapted to handle terminology constraints. To avoid catastrophic forgetting, as systems must perform equally well on terminology constrained and unconstrained data, we follow a mixed fine-tuning strategy (Chu et al., 2017). A weighted combination (2:1 ratio[2]) of unconstrained and constrained data is used during training and validation steps. For validation purposes, we concatenate with a 1:1 ratio.

The baseline and the fine-tuned systems were trained until they converged based on perplexity results from the validation set, using an early stopping criterion of 5 consecutive checkpoints. Validation is performed every 10,000 steps in the case of the baseline system whereas fine-tuning validation is performed every 1,000 steps. All the systems were trained on a single RTX 2080-Ti GPU device.

We evaluate our systems using BLEU and chrF++ scores provided by the sacreBLEU tool (Post, 2018). Additionally, we report COMET (Rei et al., 2020) scores[3], a metric which focuses on the semantic similarity by leveraging the recent breakthroughs in neural language modeling. While BLEU, chrF++ and COMET metrics measure the overall translation quality of the systems, task specific metrics are required. To address this, we determine the accuracy of the correctly translated constraints in terms of term-level constraint accuracy (TCA) as in (Zhang et al., 2023).

## 5 Task oriented challenging test sets

Many publicly available test sets for this task are based on an oversimplified constraint annotation method, as discussed in Bergmanis and Pinnis (2021) and Zhang et al. (2023). The conventional annotation method involves automatically identifying and annotating terms from a term database within a corpus of parallel sentences (Dinu et al., 2019). While seemingly tailored to the task, this approach raises questions about its reflection of real-world scenarios. In most cases, term databases contain highly specialized domain specific terms which are not present in general out-domain parallel corpora. Consequently, many complex and valuable terms are not found and are subsequently discarded, resulting in simple test sets for which the baseline already obtains competitive enough results. Additionally, this approach lacks control over the number and complexity of the constraints annotated.

We posit that terminology-constrained NMT becomes useful in cases where the baseline model fails to produce the correct target term of the constraints. The ideal test set should contain more complex and specialized terminology constraints that align with real-life requirements.

**Basque-Spanish test sets**. To the best of our knowledge, the Basque to Spanish translation direction has not been previously addressed. As a result, we curated two high-quality and challenging test sets for this translation direction. These sets were meticulously crafted to emulate real-life applications of terminology-constrained NMT. In the following lines we describe the handcrafted test sets and Table 2 shows detailed figures about the test sets.

*Euskalterm*. The aim of this test set is to prioritize the incorporation of specific terminology constraints, focusing on the terms rather than on the parallel sentences. Initially, a collection of 300 terms was curated from the publicly accessible Euskalterm term database[4]. The Euskalterm database contains specialized terminology for a diverse set of domains. Terms with a varying number of words were chosen, ranging form one to five words. Instead of relying on parallel corpora to find these terms, we asked a native speaker to craft up to two Spanish sentences for each term. This approach was taken to ensure that the corresponding Basque translations of the terms include a wide variety of complex suffix patterns. Subsequently, these sentences were meticulously translated into Basque, ensuring the inclusion of the constraints.

*Euskalterm multi*. In a similar fashion to Zhang et al. (2023) we designed a test to measure the influence of varying constraint counts within a sentence. For this purpose, we utilized the *Elhuyar* parallel corpus publicly available at OPUS (Tiedemann, 2012). We carefully removed samples already present in the training data and selected a set of 50 parallel sentences. Then, a linguistic expert manually selected 4 terms from each of the extracted parallel sentences. These terms comprised noun phrases and proper names of varying word lengths.

[2]Initial experiments showed that 2:1 ratio for unconstrained and constrained data respectively works well.

[3]The recommended model *wmt20-comet-da* was employed and it already covers both Spanish and German.

[4]https://opendata.euskadi.eus/katalogoa/-/euskalterm-hiztegi-terminologikoak/

| Test set | Language pair | Sents. | Terms | Avg. Terms | Avg. Words |
|---|---|---|---|---|---|
| Paracrawl | EU-ES | 710 | 836 | 1.2 | 1.0 |
| Euskalterm | EU-ES | 550 | 550 | 1.0 | 2.6 |
| Euskalterm multi | EU-ES | 50 | 205 | 4.1 | 2.7 |
| IATE | EN-DE | 414 | 452 | 1.1 | 1.0 |
| Automotive Test Suite | EN-DE | 766 | 986 | 1.3 | 0.7 |

Table 2: Statistics for the created Basque to Spanish test sets. *Avg. Terms* indicates the average number of terms annotated per sentence. *Avg. Words* means the average number of words for each target term.

Furthermore, as in Dinu et al. (2019), we automatically annotated the test portion of the Paracrawl data set using the test subset of the bilingual dictionary extracted from Apertium (Referred to as *Paracrawl*). As mentioned earlier, this test set does not mimic a challenging real-life scenario. Instead it is used for comparison purposes against the more complex and challenging *Euskalterm* test set.

**English-German test sets**. For the English to German translation direction we utilized two publicly available test sets: *Automotive Test Suite* test set introduced in Bergmanis and Pinnis (2021) and *IATE* from Dinu et al. (2019).

The *Automotive Test Suite* test set consists of parallel sentences in English, Estonian, German, Latvian, and Lithuanian, with terminology constraints derived from a glossary constructed by professional translators. The *IATE* test set was created by automatically annotating *IATE* terms in the out-domain WMT newstest 2017 test set. Consequently, many common nouns, such as *sport*, *bridge*, *trip*, are annotated. We note that some of them appear multiple times. Additionally, terms are annotated in their surface form which means their final form is known beforehand. Therefore, this test set is only used for comparison purposes with prior work.

## 6 Results

This section presents the results of our experimental work, emphasizing a comparative analysis between our proposed *morph-masking* method and the *copy-and-inflect* method (Bergmanis and Pinnis, 2021). We evaluate the performance of the terminology-constrained fine-tuned systems for the Basque to Spanish and English to German translation directions, aiming for comprehensive insights and conclusions[5].

**Overall translation quality**. First, we examine the overall translation quality of the fine-tuned terminology aware systems for a terminology unconstrained setting, as systems are required to perform effectively with and without terminology constraints. We use the *Flores200* benchmark (NLLB Team, 2022) which encompasses both Basque-Spanish and English-German translation directions for the same set of sentences.

| EU-ES | | | |
|---|---|---|---|
| System | BLEU | chrF++ | COMET |
| Baseline | 18.4 | 44.7 | 0.551 |
| copy-and-inflect | 18.2 | 44.7 | 0.543 |
| morph-masking | 18.4 | 44.9 | 0.548 |

| EN-DE | | | |
|---|---|---|---|
| System | BLEU | chrF++ | COMET |
| Baseline | 36.1 | 61.3 | 0.616 |
| copy-and-inflect | 36.1 | 61.0 | 0.614 |
| morph-masking | 36.0 | 61.1 | 0.617 |

Table 3: Results for the Basque-Spanish and English-German overall translation quality evaluation on the Flores200 benchmark. BLEU, chrF++ and COMET scores are reported. Terminology-unconstrained baselines are compared against our proposed *morph-masking* method and the *copy-and-inflect* method.

Table 3 shows the results of the overall translation quality evaluation on the *Flores200* test. For the Basque to Spanish translation direction, the baseline and both of the terminology aware methods, *copy-and-inflect* and *morph-masking*, perform similarly without any statistically significant differences. Similarly, in the English to German translation, both fine-tuned terminology aware systems perform on par with the baseline.

**Terminology accuracy**. Terminology accuracy rates are reported for the task specific test sets described in Section 5. Both the *copy-and-inflect* and *morph-masking* systems are evaluated with and without applying terminology constraints to the test sets.

[5]Additional experiments were conducted on a proprietary test. See Appendix C.

| System | C. | Euskalterm | | | | Paracrawl | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | BLEU | chrF++ | COMET | TCA | BLEU | chrf++ | COMET | TCA |
| Baseline | No | 51.5 | 72.5 | 0.872 | 44.18 | **39.3** | 61.7 | 0.681 | 90.43 |
| copy-and-inflect | No | 51.1 | 72.3 | 0.871 | 45.09 | 39.1 | 61.5 | 0.682 | 90.31 |
| | Yes | 50.8 | 72.4 | 0.844 | 45.64 | **39.3** | **61.8** | 0.683 | 91.27 |
| morph-masking | No | 51.0 | 72.3 | 0.876 | 44.73 | 39.1 | 61.6 | 0.688 | 90.31 |
| | Yes | **57.8*** | **77.4*** | **0.916*** | **83.45** | 39.0 | 61.5 | 0.675 | **93.30** |

Table 4: Basque to Spanish terminology accuracy (TCA) scores in addition to translation quality scores (BLEU, chrF++ and COMET) for the task specific *Euskalterm* and *Paracrawl* test sets. **C.** column means whether terminology constraints are applied or not. * indicates statistically significant (p-value $\leq 0.05$) differences by conducting paired bootstrap resampling with respect to the baseline. Best scoring systems are highlighted in bold.

**Basque-Spanish results** (Table 4). For the *Euskalterm* test set, the baseline struggles to correctly translate terminology constraints, with less than half of the terms being correctly translated. This aligns with our intent to create a challenging test set. While the *copy-and-inflect* method exhibits a slight improvement over the baseline, it too largely falls short in enforcing terminology constraints. Conversely, *morph-masking* notably outperforms the baseline in terms of constraint accuracy. This is also reflected in the translation quality with significantly better results. This discrepancy in performance can be attributed to the constraint signal they impose. The *morph-masking* method enforces a harder constraint signal by entirely eliminating the source term. Under the unconstrained setting, both methods perform at par with the baseline.

Results on the *Paracrawl* test set reaffirm that this test set doesn't effectively emulate challenging real-life scenarios. The baseline system already achieves satisfactory TCA scores. Consequently, both fine-tuned terminology-aware systems show only marginal improvement, with *morph-masking* leading slightly. Many common nouns were annotated for which the system seems to be confident enough to provide its own term translation even though constraints are provided.

**English-German results** (Table 5). On the *IATE* test set, the baseline already achieves a high TCA score. As explained in Section 5, this test set represents a relatively basic benchmark for evaluating terminology-constrained NMT. Both fine-tuned terminology aware systems substantially improve TCA results and *morph-masking* obtains the best results. Higher TCA scores are slightly reflected in the translation quality for the *copy-and-inflect* system, although none of the systems significantly improve the baseline.

On the more challenging *Automotive test suite* test set, the baseline struggles to accurately translate constraints, as evidenced by its TCA score. While substantially surpassing the baseline, the *copy-and-inflect* system underperforms when compared to our method which achieves outstanding results.

**Impact of constraint counts**. Similarly to Zhang et al. (2023), we evaluate the robustness of our proposed method, *morph-masking*, against varying constraint counts per sentence in the Basque to Spanish translation direction. The objective of this evaluation is to determine whether masking multiple source terms leads to a significant loss of essential information. For this purpose, four variations of the *Euskalterm multi* are generated with constraints counts ranging from 1 to 4, $C_i$ where $1 <= i <= 4$. Constraints are randomly selected from the four constraints of each sample. Results with no constraints ($C_0$) are also provided. Sentence-level constraint accuracy (SCA) (Zhang et al., 2023) scores are reported in addition to TCA scores. That is, translations are considered correct only if they meet all the constraints in the sentence. Results are shown in Table 6.

As expected, an increase in the number of constraints typically results in improved translation quality, as translations align more closely with the references. All configurations yield high TCA scores. However, as the number of constraints rises, SCA scores decrease, indicating the increasing difficulty in ensuring that all specified terms appear in the translations. Nevertheless, $C_4$ clearly surpasses the unconstrained $C_0$ setting proving our approach is useful to address challenging multiple constraints settings.

**Grammatical information ablation study**. To highlight the importance of the extracted grammatical information during the masking of source terms, we conducted an ablation study on our method. In

| System | C. | IATE | | | | Automotive test suite | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | BLEU | chrF++ | COMET | TCA | BLEU | chrF++ | COMET | TCA |
| Baseline | No | 32.4 | 57.6 | **0.546** | 86.95 | 31.0 | 56.5 | 0.478 | 72.37 |
| copy-and-inflect | No | 32.6 | 57.7 | 0.535 | 86.95 | 31.0 | 56.4 | 0.473 | 71.37 |
| | Yes | **32.8** | **58.2*** | 0.540 | 94.91 | **32.8*** | 59.0* | 0.553* | 86.76 |
| morph-masking | No | 32.7 | 58.0 | **0.546** | 86.73 | 30.9 | 56.4 | 0.477 | 72.37 |
| | Yes | 32.6 | 57.9 | 0.534 | **96.02** | 32.3* | **59.3*** | **0.589*** | **95.40** |

Table 5: English to German terminology accuracy (TCA) scores in addition to translation quality scores (BLEU, chrF++ and COMET) for the task specific *IATE* and *Automotive Test Suite* test sets. **C.** column means whether terminology constraints are considered or not. * indicates statistically significant (p-value $\leq 0.05$) differences by conducting paired bootstrap resampling with respect to the baseline. Best scoring systems are highlighted in bold.

| $C_i$ | BLEU | chrF++ | TCA | SCA |
|---|---|---|---|---|
| 0* | 40.9 | 63.9 | 70.24 | 22.00 |
| 1 | 41.5 | 64.7 | 90.38 | 90.00 |
| 2 | 42.3 | 65.5 | 90.48 | 86.00 |
| 3 | 43.1 | 66.1 | 90.97 | 78.00 |
| 4 | 43.0 | 66.0 | 89.76 | 66.00 |

Table 6: Basque-Spanish results for the assessment of the impact of different constraint counts per sample on the *Euskalterm multi* test set. BLEU, chrF++ and TCA, as well as, sentence-level constraint accuracy (TCA) are reported. *$C_0$ is evaluated against the 4 constraints in each sample.

this study, we removed all tokens related to grammatical information, resulting in only the source terms being masked from the input sentence. This approach aligns with the *replace* method described in Exel et al. (2020), and hence we will refer to it as *replace*. The results of this ablation study for both the Basque to Spanish and English to German translation directions are presented in Table 7 and Table 8 respectively.

Although both methods perform similarly in terms of term accuracy, the results reveal a substantial drop in the translation quality for the *replace* method when compared to *morph-masking*. The observed differences vary depending on the morphological richness of the source language, being less pronounced for English. For morphologically rich languages like Basque, completely masking the source term leads to a significant loss of essential grammatical information, which adversely impacts the final translation quality. Although the *replace* method underperforms, it still markedly outperforms the baseline. This suggests that the system can compensate for the missing information by leveraging the surrounding context. Please refer to Appendix B for illustrative examples showcasing the outcomes of the methods in the ablation study.

| System | Euskalterm | | |
|---|---|---|---|
| | BLEU | chrF++ | TCA |
| Baseline | 51.5 | 72.5 | 44.18 |
| morph-masking | **57.8*** | **77.4*** | **83.45** |
| replace | 54.5*$^\dagger$ | 75.7*$^\dagger$ | 83.27 |
| **System** | **Paracrawl** | | |
| | BLEU | chrF++ | TCA |
| Baseline | **39.3** | **61.7** | 90.43 |
| morph-masking | **39.0** | **61.5** | **93.30** |
| replace | 38.0$^\dagger$ | 60.9$^\dagger$ | 91.51 |

Table 7: Results of the grammatical information ablation study for the Basque to Spanish translation direction. We report BLEU, chrF++ and TCA scores on the *Euskalterm* and *Paracrawl* test sets. * indicates statistically significant (p-value $\leq 0.05$) differences by conducting paired bootstrap resampling with respect to the baseline, while $^\dagger$ indicates statistically significant differences between *morph-masking* and *replace* methods. Best scoring system is highlighted in bold.

## 7 Conclusions

In this work, we tackle terminology-constrained NMT using a data-driven approach that does not require changes in the system architecture or decoding algorithm. In particular, we introduce a novel variation of the *copy-and-inflect* method introduced by Bergmanis and Pinnis (2021). Our proposed method aims to provide a stronger constraint signal by masking the source terms of the constraints in the input sentence, while retaining essential grammatical information from the source terms, such as gender, number, grammatical cases, and so forth.

By evaluating our approach on two translation directions —Basque to Spanish and English to Ger-

| System | IATE | | |
|---|---|---|---|
| | BLEU | chrF++ | TCA |
| Baseline | **32.4** | **57.6** | 86.95 |
| morph-masking | **32.6** | **57.9** | 96.02 |
| replace | 32.3$^{\dagger}$ | 57.7$^{\dagger}$ | **96.24** |
| **System** | **Automotive test suite** | | |
| | BLEU | chrF++ | TCA |
| Baseline | 31.0 | 56.5 | 72.37 |
| morph-masking | **32.3*** | **59.3*** | **95.40** |
| replace | **32.2*** | 59.0*$^{\dagger}$ | 94.53 |

Table 8: Results of the grammatical information ablation study for the English to German translation direction. We report BLEU, chrF++ and TCA scores on the *IATE* and *Automotive test suite* test sets. * indicates statistically significant (p-value $\leq 0.05$) differences by conducting paired bootstrap resampling with respect to the baseline, while $^{\dagger}$ indicates statistically significant differences between *morph-masking* and *replace* methods. Best scoring system is highlighted in bold.

man, each having varying degrees of source morphological complexity- we demonstrate that our *morph-masking* method offers a harder constraint signal. This leads to performance improvements over the *copy-and-inflect* method in all scenarios. Removing source terms not only maintains the performance but also compels the model to utilize the provided target term in the output translations. This confirms our hypothesis that the *copy-and-inflect* method can sometimes allow the system to disregard the given target term, instead defaulting to its standard translation for the source term. Through an ablation study, we further highlight the importance of preserving essential grammatical information, especially for morphologically rich languages like Basque, to achieve superior translation quality and term accuracy.

Additionally, we show that fine-tuning a general purpose NMT system with synthetically generated data for the terminology-constrained NMT task is sufficient for the system to learn how to apply terminological constraints.

## Limitations

While we validated our *morph-masking* method on two translation directions, each with distinct source morphological complexity (English to German and Basque to Spanish), further exploration is needed to assess its adaptability to other languages, particularly those from diverse language families with unique structures and nuances influencing word inflections and modifications.

## 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.

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.

Toms Bergmanis and Mārcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Miriam Exel, Bianka Buschbeck, Lauritz Brandt, and Simona Doneva. 2020. Terminology-constrained neural machine translation at SAP. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280, Lisboa, Portugal. European Association for Machine Translation.

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

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 (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. Improved lexically constrained decoding for translation and monolingual rewriting. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Elise Michon, Josep Crego, and Jean Senellart. 2020. Integrating domain terminology into neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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 Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. 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.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Huaao Zhang, Qiang Wang, Bo Qin, Zelin Shi, Haibo Wang, and Ming Chen. 2023. Understanding and improving the robustness of terminology constraints in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6029–6042, Toronto, Canada. Association for Computational Linguistics.

## A Extracted grammatical information

This section provides a more detailed and comprehensive explanation of the grammatical information extracted for our *morph-masking* method for each of the source languages involved in the experiments: Basque and English. Additionally, we present a compilation of the unique tokens that were incorporated into the vocabularies of the respective systems.

Specifically, for Basque we extract the grammatical case suffixes for common nouns, proper nouns and adjectives. The plural number for common nouns is also extracted. For English we only extract the plural number and a comparative/superlative token for common nouns. For both languages the letter casing of the source word, either uppercase or cased, is also extracted (see Table 9 and Table 10 respectively). The amount of information extracted

varies with the morphological complexity of the source language, resulting in lesser extraction from morphologically simpler languages like English.

## B Ablation results

Table 11 presents examples from the ablation study, illustrating the performance differences between the *morph-masking* and *replace* methods in the context of Basque to Spanish translation direction. For each method, we provide the input alongside its respective translation, supplemented by the English translation to enhance comprehension of the results.

The first example showcases the importance of the extracted grammatical information as evidenced by the *replace* method's failure to capture the causal grammatical case (<+gatik>). Conversely, the subsequent example demonstrates how the *replace* method can potentially compensate for the missing information (comitative grammatical case, <+ekin>), by effectively utilizing contextual cues, thereby achieving a comparable translation.

## C Additional Basque-Spanish tests results

We also created an additional proprietary test set which comprises specialized terminology from vocational training courses as well as their example usage parallel sentences. Terms and examples are divided into one word constraints and multiple words constraints, that is, the two versions of the test which we call *Laneki single* and *Laneki multi* respectively. Although there is just a single constraint per sample, they provide a useful insight as they consist of real-life examples. They are also much more testing samples than in the other tests, 3738 samples for *Laneki single* and 6864 for *Laneki multi*. Statistics for these tests are shown in Table 12. Results are shown in Table 13.

| **Basque** | |
|---|---|
| **Information** | **Token** |
| **Grammatical case** | |
| Absolutive | <+a> |
| Ergative | <+ak> |
| Comitative | <+ekin> |
| Allative | <+ra>, <+gana> |
| Benefactive | <+entzat> |
| Terminative | <+aino>, <+ganaino> |
| Causal | <+gatik> |
| Instrumental | <+z> |
| Possessive genitive | <+en> |
| Local genitive | <+ko> |
| Directive | <+antz>, <+ganantz> |
| Ablative | <+tik>, <+gandik> |
| Innesive | <+an>, <+gan> |
| Dative | <+i> |
| Partitive | <+ik> |
| Prolative | <+tzat> |
| **Number** | |
| Plural | <+pl> |
| **Letter case** | |
| Cased | <C> |
| Uppercase | <U> |
| **Other declensions** | |
| Excessive | <+egi> |
| Comparative | <+ago> |

Table 9: Extracted grammatical information and the corresponding tokens for Basque language.

| **English** | |
|---|---|
| **Information** | **Token** |
| **Number** | |
| Plural | <+pl> |
| **Letter case** | |
| Cased | <C> |
| Uppercase | <U> |
| **Other** | |
| superlative | <sup> |
| comparative | <comp> |

Table 10: Extracted grammatical information and the corresponding tokens for English language.

| | |
|---|---|
| Glossary entry | *transformagarri –> transformable* |
| Source | Bere izaera **transformagarriarengatik**, sofa hau erraz bihur daiteke ohe. |
| Target | **Por** su naturaleza **transformable**, este sofá puede convertirse fácilmente en una cama. |
| morph-masking | Bere izaera **transformable <+a>**$_E$ **<+gatik>**$_E$, sofa hau erraz bihur daiteke ohe.<br>**Por** su naturaleza **transformable**, este sofá se puede convertir fácilmente en una cama. |
| replace | Bere izaera **transformable**, sofa hau erraz bihur daiteke ohe.<br>Su carácter **transformable**, este sofá se puede convertir fácilmente en una cama. |
| *English* | ***Due to*** *its* ***transformable*** *nature, this sofa can easily be converted into a bed.* |

| | |
|---|---|
| Glossary entry | *mekanismo eragile elektromekaniko –> mecanismo accionador electromecánico* |
| Source | **Mekanismo eragile elektromekanikoarekin**, atea modu eraginkorrago eta isilagoan irekitzen eta ixten da. |
| Target | **Con** el **mecanismo accionador electromecánico**, la puerta se abre y cierra de forma más eficiente y silenciosa. |
| morph-masking | **<C>**$_E$ **mecanismo accionador electromecánico <+a>**$_E$ **<+ekin>**$_E$, atea modu eraginkorrago eta isilagoan irekitzen eta ixten da.<br>**Con** el **mecanismo accionador electromecánico** la puerta se abre y cierra de forma más eficiente y silenciosa. |
| replace | **mecanismo accionador electromecánico**, atea modu eraginkorrago eta isilagoan irekitzen eta ixten da.<br>El **mecanismo accionador electromecánico** abre y cierra la puerta de forma más eficiente y silenciosa. |
| *English* | ***With*** *the* ***electromechanical drive mechanism****, the door opens and closes more efficiently and quietly.* |

Table 11: Comparison of the results for the *morph-masking* and *replace* methods for the Basque to Spanish translation direction. For each method we provide the input and the resulting translation (rows *morph-masking* and *replace*). We also include the English translation for better understanding of the results (rows *English*). The first example showcases the importance of the extracted grammatical information as evidenced by the *replace* method's failure to capture the causal grammatical case (<+gatik>). Conversely, the subsequent example demonstrates how the *replace* method can potentially compensate for the missing information (comitative grammatical case, <+ekin>), by effectively utilizing contextual cues, thereby achieving a comparable translation.

| Test set | Language pair | Sents. | Terms | Avg. Terms | Avg. Words |
|---|---|---|---|---|---|
| Laneki single | EU-ES | 3738 | 3958 | 1.1 | 1.0 |
| Laneki multi | EU-ES | 6864 | 6924 | 1.0 | 2.5 |

Table 12: Statistics for the *Laneki single* and *Laneki multi* test sets for the Basque to Spanish translation direction. *Avg. Terms* indicates the average number of terms annotated per sentence. *Avg. Words* means the average number of words for each target term.

| System | C. | Laneki single | | | Laneki multi | | |
|---|---|---|---|---|---|---|---|
| | | BLEU | chrF++ | TCA | BLEU | chrf++ | TCA |
| Baseline | No | 34.0 | 59.0 | 75.62 | 40.0 | 64.3 | 74.99 |
| copy-and-inflect | No | 34.0 | 59.1 | 75.75 | 40.1 | 64.4 | 74.91 |
| | Yes | 34.2* | 59.3* | 79.61 | 40.2* | 64.5* | 78.15 |
| morph-masking | No | 34.0 | 59.1 | 75.52 | 40.2 | 64.4 | 74.91 |
| | Yes | **34.7*** | **59.9*** | **94.34** | **40.7*** | **65.0*** | **91.46** |

Table 13: Basque to Spanish terminology accuracy (TCA) scores in addition to translation quality (BLEU, chrF++) scores for the task specific *Laneki* test sets. Two versions of the test set are presented, with single word contraints and multi-word constraints respectively. **C.** column means whether terminology constraints are applied or not. * indicates statistically significant (p-value $\leq 0.05$) differences by conducting paired bootstrap resampling with respect to the baseline. Best scoring systems are highlighted in bold.