

Findings of the WMT26 Terminology Translation Task: The Hard Part is Finding the Terms, Not Using Them

Adrian Charkiewicz^{1,2} Pinzhen Chen³ Thierry Etchegoyhen⁴ Harritxu Gete⁴
Kamil Guttman^{1,2} Xu Huang⁵ David Ponce⁴ Artur Nowakowski^{1,2}
Frédéric Odermatt⁶ Arturo Oncevay Dawei Zhu⁷ Vilém Zouhar⁸ Kirill Semenov⁹
¹Lanigo ²Adam Mickiewicz University ³Queen’s University Belfast ⁴Vicomtech
⁵Nanjing University ⁶44ai ⁷Amazon ⁸ETH Zurich ⁹University of Zurich

Abstract

The WMT26 Terminology Translation Task aims to evaluate machine translation in high-stakes, term-heavy domains (technology, finance, medicine). This year, we focus solely on document-level translation and run two tasks: (1) MT with explicit document-level dictionaries, (2) MT with bitext samples that contain the specific terms. Participants are presented with the texts in three translation directions, two of which feature mid-to-low-resourced morphologically rich languages: Spanish→Basque, English→Polish, and Traditional Chinese→English. This year, the main metrics were multiple variants of overall translation quality and terminology success rate; in line with previous shared tasks, we also compared systems with no terminology, proper dictionaries, and random dictionaries to causally analyze terminology utility. 17 teams participated in our task, submitting 21 systems to Track 1 and 18 systems to Track 2. The results show that document-level translation with explicit dictionaries is close to saturation, and the best systems nearly reach the reference texts. In contrast, for translation with sample bitexts, the spread of the systems is bigger, and the best scores are lower, highlighting the need to concentrate on terminology extraction rather than its use from an explicit source. We also evaluate the grammaticality of the generated texts in Basque and Polish and observe slight trends toward using dictionary forms of the terms and assigning the most frequent grammemes.

★ Authors (except KS) are ordered alphabetically by surname. AO contributed as an independent researcher, and DZ’s work was done outside of Amazon. The views and opinions expressed in this paper are strictly those of the authors. Correspondence to kirill.semenov@uzh.ch.

Sections

- 1 Introduction
- 2 Task description
- 3 Data
- 4 Evaluations and Metrics
- 5 Participants and System Descriptions
- 6 Shared Task Results
- 7 Analysis
- 8 Related Work
- 9 Conclusions

Data & Code

github.com/wmt-conference/wmt26-terminology

Website

www2.statmt.org/wmt26/terminology.html

1 Introduction

In the AI era, one calendar year often feels like many more. The terminology machine translation (MT) field is no exception: we can see papers on terminology-aware translation at EMNLP 2025 (Shen et al., 2025; Berger et al., 2025; Oncevay et al., 2025a; Menis Mastromichalakis et al., 2025), the more recent EACL (Abela et al., 2026) and ACL (Hasan et al., 2026; Di Rosa, 2026), as well as a specialized terminology MT shared task at the Workshop on Natural Language Processing for Turkic Languages (Gebeşçe et al., 2026). These facts together highlight the main urgencies of terminology-aware MT nowadays: the need to benchmark the abundance of the emerging solu-

tions on multiple languages, especially when it comes to less resourced and typologically diverse languages.

Contrary to [Waters et al. \(1973, 2:55\)](#), we can't afford to say that there is time to kill today. Therefore, we altered the pattern of the WMT Terminology Shared Task from every other year ([Alam et al., 2021](#); [Semenov et al., 2023, 2025](#)) to a yearly event. This time, in addition to English, we have included Basque and Polish as the target languages, representing different linguistic families and various levels of available data resources.

Building on last edition's findings showing that sentence-level benchmarks are approaching saturation, this edition shifts entirely to document-level translation. Additionally, we investigate whether translation systems can maintain terminology compliance without having access to an explicit bilingual dictionary. To contrast explicit term injection against implicit term discovery, the shared task features two tracks:

- **Track 1 (Document MT with Explicit Dictionaries):** Systems translate full documents given domain terminology glossaries, requiring contextual term selection and target inflection.
- **Track 2 (Document MT with Sentence Samples):** Systems receive no explicit term dictionaries; instead, they are provided with a minimal sample (~10%) of reference bitexts from the same collection to extract and enforce domain terminology on the fly.

To isolate the causal impact of terminology resources from generic reference leakage, systems in Track 1 are evaluated across three terminology modes (**No Term**, **Proper Term**, and **Random Term**) ([Zouhar, 2023](#); [Semenov et al., 2023](#)), while Track 2 compares performance under **No Term** and **Sample** modes. The shared task attracted broad participation, receiving 27 submissions from 17 teams (21 systems in Track 1 and 18 in Track 2). Analysis of these submissions yields five key findings:

- **LLMs and Agents Lead:** Top-performing systems on the Pareto frontier feature LLM-based pipelines and multi-agent workflows.
- **Explicit Terms vs. Raw Bitexts:** Explicit dictionaries (Track 1) increase term success by 26.4 points on average with a 1.57 chrF++ gain in quality. Sample bitexts (Track 2) yield larger quality gains (+2.95 chrF++) by transferring style and domain phrasing, but yield smaller term

```
Test:
{
  "es": "Caja de cambios\nEn los vehículos, la caja de cambios o caja de velocidades (también llamada simplemente caja) es el elemento encargado de obtener [...]",
  "eu": "Abiadura-kaxa\nIbilgailuetan, abiadura-kaxak edo engranaje-kaxak (kaxa ere esaten zaio, besterik gabe) lortzen du [...]"
}

Terms:
{
  "proper": {"caja de cambios": ["abiadura-kaxa"], "vehículo": ["ibilgailu"], "caja de velocidades": ["abiadura-kaxa"], "..."},
  "random": {"elemento": ["elementu"], "caja": ["kaxa"], "..."}
}
```

Listing 1: A Spanish→Basque example from Track 1.

success gains (+14.6 points).

- **Term Discovery Remains Unsolved:** While top systems in Track 1 reach near-saturation in term success rate, Track 2 exhibits roughly twice the performance spread across systems, showing that discovering terms without explicit glossaries still proves challenging.
- **Metric Blind Spots and Judge Instability:** Neural MT metrics penalize incorrect dictionary usage in Random mode but remain mostly insensitive to proper terminology adoption. Furthermore, reference-free LLM judges exhibit high false-positive error rates (40.3% overall) on human references due to output-discipline failures, frequently emitting error spans for terms they acknowledge in their explanations to be correct.
- **Grammatical yet Statistically Different Term Inflection:** In Basque and Polish, the systems produce grammatically correct and coherent sentences; however, the distributions of the grammemes slightly differ from the ones in the reference texts, concentrating around the dictionary forms or the constructions from the higher-resourced languages.

2 Task Description

The last two shared task editions, despite covering multiple languages and domains, lacked inter-comparability: [Semenov et al. \(2023\)](#) had fiction texts for one of the translation pairs; while [Semenov et al. \(2025\)](#) featured non-overlapping language and domain pairs for two competition tracks. This year, we tried to fix this by including the same domain areas for two out of the three translation pairs. Moreover, we finally moved into the area of less-resourced, morphologically rich languages: Basque and Polish. Finally, we stuck to the multi-track competition from last year, but reconsidered the content of the two tracks, changing them to

```

Test:
{
  "es": "[...] Por ejemplo, los cepillos en que el movimiento  
recíprocamente de la corredera se produce por medio de  
manivela y perno, principalmente en el engranaje, en el  
movimiento de la máquina, reciben el nombre de  
cepillos de manivela. [...]"

  "eu": "[...] Adibidez, irristailuaren joan-etorriko  
mugimendua biraderaren eta bernoaren bidez egiten duten  
arrabotei, nagusiki engranajea eta makinaren  
mugimenduan, biradera-arrabota esaten zaie. [...]"
}

Reference samples:
{
  "es": "Las mortajadoras se utilizan principalmente para  
mecanizar ranuras, pero también se emplean para  
contornear levas, placas, palancas, tallar engranajes,  
etc.",
  "eu": "Mortasatzeko makinak artekak mekanizatzeke erabiltzen  
dira nagusiki, baina erabiltzen dira espekak, plakak  
edo palankak inguratzeko, engranajeak taillatzeko, etab  
. ere."
}

```

Listing 2: A Spanish→Basque example from Track 2.

document-level translation with or without explicit terminologies.

2.1 Tracks

As was stated in last year’s findings (Semenov et al., 2025), the sentence-level translation with explicitly given sentence-level term dictionaries seemed a relatively simple task for most systems, while document-level translation scores showed a considerable divergence and were far from saturation even in the best submissions. Therefore, this year’s shared task features document-level translation in both tracks. The difference lies in the way the systems are exposed to the terms:

- **Track 1: Document MT with Explicit Dictionaries:** The participants are given lists of full documents (up to 2000 words each, 10 documents on average), together with a single terminology dictionary that features terms applied for the whole domain (400 to 1100 terms). Therefore, the system is expected to select the relevant terms for a particular document and use the corresponding translations in context.
- **Track 2: Document MT with Sentence Samples:** The participants are not provided with explicit dictionaries. Instead, they receive small samples of sentences and their reference translations drawn from the same source as the test inputs (comprising roughly 10% of the dataset on average). We consider this scenario more challenging for the participants, since the system needs to identify the terms and align them with their correspondences before the translation.

2.2 Languages and Domains

Spanish→Basque (es-eu)

- **Tracks:** 1 and 2.
- **Domains:** Four sub-domains of Engineering and Technology: Energy and Automotive for Track 1; Machine Tools and Railways for Track 2.
- **Language Overview:** While Spanish is a relatively well-represented language, Basque is a low-resource language spoken mostly in the north of Spain and less in the neighboring regions of France (commonly known as the Basque Country). Additionally, Basque is genealogically and structurally very distant from Spanish and other Indo-European languages, since it is a language isolate and features grammatical categories that are rare for Europe (such as ergative case marking or polypersonal agreement).

English→Polish (en-pl)

- **Tracks:** 1 and 2.
- **Domains:** Medical texts, Mechanical Engineering (both domains equally represented in Track 1 and 2).
- **Language Overview:** Polish can be described as a mid-resourced language (for example, [CommonCrawl](#) shows its resourcedness as approximately equal to Italian and Portuguese). It is one of the most widely spoken Slavic languages (approx. 40M speakers). Its challenges to the MT task include relatively complex morphology (seven nominal cases and dozens of verbal forms, multiple nominal and verbal classes). Additionally, its elaborate system of diacritization and digraphs may present an additional challenge because of over-tokenization.

Traditional Chinese→English (zh-Hant-en)

- **Track:** 2.
- **Domain:** Finance, publications from the Hong Kong Monetary Authority
- **Language Overview:** Although Chinese is one of the most-resourced languages, its data in the traditional writing script is significantly less represented on the internet and in NLP resources. Additionally, this is the only language in our sample that uses a non-Latin script and is distinct typologically and genealogically from the European languages.

2.3 Terminology Modes

In Track 1, we follow the three-way evaluation (Zouhar, 2023; Semenov et al., 2023, 2025), to

enable a causal analysis of the utility of the terminology:

- **No Terminology (No Term):** The system translates the input text without a terminology dictionary.
- **Proper Terminology (Proper Term):** The system is provided with a dictionary of domain-specific terms relevant to the input.
- **Random Terminology (Random Term):** The system receives a dictionary of randomly selected source words (excluding the proper terms) and their aligned translations from the references.

The use of Random Term mode allows us to measure whether improvements in translation quality stem from incorporating terminologies or simply seeing part of the correct translation (Zouhar, 2023; Semenov et al., 2023).

Since Track 2 does not feature explicit term dictionaries, we do not apply a three-way evaluation protocol there. Instead, we only distinguish between the **No Term** (only input texts provided) and **Sample** (input texts provided with a few sample translations) modes.

3 Data

3.1 Track 1: Document Translation with Explicit Dictionaries

3.1.1 Spanish-Basque

Data source. The data for Spanish to Basque translation were sampled from the ADAPTIA-MT suite of resources (Etchegoyhen et al., 2025), which covers four different industrial domains: Automotive, Energy, Railways and Machine Tool. The data were based on original technical documentation in Spanish, provided by experts from the Basque Industry Clusters in each domain and professionally translated into Basque at the document level. For Track 1, we selected data from the Energy and Automotive domains.

Terminology. Four specialised terminologies were manually built for the domains of ADAPTIA-MT (*op. cit.*), by the Basque Centre for Terminology and Lexicography UZEI. The process leveraged both the collected data described above, and existing dictionaries in each respective field, when available. The relevant Spanish terms were extracted semi-automatically from both types of resources and the resulting lists were combined and manually reviewed. Equivalent Basque terms were then manually established, following terminolog-

ical ISO standards, in particular ISO/TC 37. For Track 1, there were 683 and 761 terms for the Energy and Automotive domains, respectively.

Test sets. For Track 1, we used the complete documents from the ADAPTIA-MT test sets for the two selected domains. In the Energy domain, the data consisted of 33 documents, which contained 683 terms. For the Automotive domain, there were 41 documents containing 761 terms in total. The test sets were manually revised, with minor cases of terms not captured in the original test sets due to morphological variation. Additionally, target term annotations were added to support the evaluation protocols of the shared task.

Listing 1 provides a simplified example of the test data for Spanish to Basque translation for Track 1, including sample documents in both languages and terminology samples, both proper and random. The other languages in this track exhibit the same structure.

3.1.2 English-Polish

Data source. The data for English to Polish translation were constructed using Laniqo’s in-house medical and mechanical engineering glossaries. Domain terms were categorized into 12 thematic groups for the medical domain and 15 groups for mechanical engineering. A team of linguists leveraged Gemini 3.1 Pro (DeepMind, 2026) to draft synthetic English source documents for each group, requiring each term to appear 3 to 5 times across the document to measure terminology translation consistency. These English texts were subsequently verified and enriched for higher lexical diversity of the terms. Finally, the documents were automatically translated into Polish using the same model and post-edited by professional translators.

Terminology. Terminology lists were derived from Laniqo’s glossaries, which contain 1,055 medical and 1,264 mechanical engineering terminology pairs mapping English terms to their Polish equivalents.

Test sets. The Track 1 test sets consist of the complete documents selected for this track, comprising 6 documents for the medical domain and 7 for mechanical engineering.

3.2 Track 2: Document Translation with Sample Texts

3.2.1 Spanish-Basque

Data source. For Track 2, we selected data from the Railways and Machine Tool domains in the original ADAPTIA-MT suite, described above.

Terminology. Terminological databases for Track 2 were also based on the ADAPTIA-MT resources for the two selected domains. Terminological data consisted of 761 terms for the Railways domain, and 838 for Machine Tool.

Test sets. For Track 2, we followed the same steps as for Track 1 in terms of manually revising and fixing minor issues in the original test data for Railways and Machine Tool, as needed, along with target term annotation. Since for this track participants were not provided with the terminology, but with texts from which to identify relevant terms, we sampled the minimal set of source sentences that contained all terms to be addressed to properly translate the source documents. Listing 2 provides a simplified example of the test data for Spanish to Basque translation for Track 2, including sample documents and reference sentences from which terms were to be extracted. The other languages in this track exhibit the same structure.

3.2.2 English-Polish

Data source. For Track 2, we allocated the remaining dataset, comprising 6 documents for the medical domain and 8 for mechanical engineering.

Terminology. The underlying terminological resources for Track 2 were drawn from the same Laniqo glossaries used across the shared task.

Test sets. For Track 2, participants were required to identify relevant terminology independently rather than receiving pre-defined dictionaries. To enable this, we applied a selection algorithm to identify the minimal subset of source sentences and their translations covering all unique terms present in the dataset. This sample of context sentences was supplied to facilitate term discovery, while the full documents served as the test set.

3.2.3 Traditional Chinese-English

Data source. We converted the Hong Kong Monetary Authority (HKMA) financial publications into document-level translation data, a data source we used in our 2025 shared task, and we used the same pipeline, PDF to markdown conversion and

manual inspection (Semenov et al., 2025). Different from 2025, where we processed Annual Reports, this year we processed Half-Yearly Monetary & Financial Stability Reports (2017-2025, March) and Sustainability Reports (2022-2024). Each report contains multiple chapters, and we treat each chapter as a *document*, truncate it to 2000 English words by space delimiter, and then truncate the Chinese side based on the number of English paragraphs remaining. We then removed explicit paragraph boundaries “\n\n”, along with notes, tables, figures, and their captions.

Terminology. We created a terminology dictionary by merging two sources: 1) we used the terminology dictionary we developed last year, and 2) we mined and aligned the English and Traditional Chinese terms separately published by HKMA, which “provides brief definitions of terms relevant to Hong Kong’s monetary, banking and financial systems”. We then filtered out entries that appear fewer than three times in this year’s data, based on exact matches on the Chinese side, since Chinese words do not undergo morphological inflections. This results in a 465-item dictionary used for evaluation only.

Test set. We split our data into sample and test as per Track 2 design. To prevent cross-chapter information leakage, we release all documents (chapters) from 2023–2025 half-yearly reports as samples, and use the remaining half-yearly reports and all sustainability reports as the test set. In total, there are 15 sample document translations and 53 documents as test input. This also implies that the sustainability reports form a domain that is “unseen” in samples.

3.3 Unified Data Model

The datasets introduced in Sections 2, 3.1 and 3.2 operate at [three levels](#).

- **Documents:** a multi-paragraph run.
- **Paragraphs:** single to multi-sentence paragraphs that can be recovered from the documents by applying the per-document paragraph delimiter.
- **Segments:** sentence pairs at the granularity they were translated by the human professionals for the shared task. This data level was not made available to the competitors, but is now available at the [task’s evaluation repository](#).

The competitors received the data at the document level. At submission time we enforced a crude

alignment at the paragraph level: The number of paragraphs within each document must match the expected number of paragraphs according to the source. Paragraph i of a submission is expected to translate paragraph i of the source. We accept that systems that have moved content across delimiter lines are penalized by most of our metrics.

Terminologies: The human professionals annotate on the segment level: they mark the actual inflected term and they denote the base lemma of the term. For Track 1 the competitors received the full glossary bag of all terms per language pair. For English→Polish it is mined from all segments, for Spanish→Basque it is an ADAPTIA-MT resource. Since the human professionals were selective in picking the most relevant terminologies per segment, some source side matches were not annotated for that specific segment. For English→Polish 647 / 2586 (25%), for Spanish→Basque 1487 / 4628 (32.1%) of all possible source side matches from the glossary bag are not annotated by the human professionals. These are not considered during evaluation. Among those, for English→Polish 46.8% (303), for Spanish→Basque 56.1% (834) of the non-annotated terms have no match on the reference side. For some of these examples the correct glossary term would be a more narrow definition of the same concept. Systems that more strictly apply terminologies as matched can have degraded performance on general translation quality for these matches.

4 Evaluation and Metrics

4.1 Upper-Bound Scoring

We verify our metric implementation by scoring the references against themselves, which we denote as *gold* submission in Table 6. This then represents an upper bound, since even a fluent human translation does not achieve a perfect score across all metrics.

4.2 Term Success Rate

We consider these points while implementing two versions of term success rates:

1. **Inflection Exact Match:** While a system producing the exact same inflected form as our human professionals is a good sign, this does not consider other inflected usages of the same term that are also acceptable.
2. **Lemmatization:** Can tell us whether a given terminology was used, without giving any insight whether a correct inflection was applied.

tier	space	accepts
annotated	surface	the annotated lemma, verbatim
inflected	surface	the gold inflected form, verbatim
glossary	surface	any glossary alternative, verbatim
annotated-lemma	lemma	lemma of the annotated or of the inflected form
glossary-lemma	lemma	lemma of any glossary alternative
exclusive	both	one disjoint span per required occurrence, over all tiers

Table 1: Matching tiers for term success. Lemma tiers add to the surface tiers rather than replacing them.

Algorithm 1 Term success over one paragraph. MAXDISJOINT returns the largest set of requirements satisfiable by pairwise disjoint spans, so one span can satisfy at most one required occurrence.

Input: paragraph hypothesis h , annotations A

- 1: $S, L \leftarrow$ surface and lemma renderings of h , with a character-span map between them
- 2: **for** $a \in A$ **do**
- 3: $F_s \leftarrow \{a.target, a.inflected\} \cup a.glossary$
- 4: $F_\ell \leftarrow \text{LEMMATIZE}(F_s)$
- 5: $M_a \leftarrow$ word-bounded, case-insensitive matches of F_s in S , and of F_ℓ in L , each mapped into both spaces
- 6: **return** per-tier rate $|\{a : M_a \neq \emptyset\}|/|A|$, and exclusive rate $\text{MAXDISJOINT}(\{M_a\})/|A|$

Lemmatization uses Stanza (Qi et al., 2020), and the candidate term forms are lemmatized the same way as the text, so the two renderings agree even where the lemmatizer is wrong.

3. **Locality:** As documents are often multi-page sized, we require terms to appear within the same paragraph their segment lies in.
4. **Word Boundaries:** Inflected/lemmatized terms should not be a substring of a larger word, therefore we apply [Unicode-aware word boundaries](#).
5. **Span Overlap:** The human annotations do not contain overlapping terminologies (such as annotating both “Covid” and “Covid-19” if the target only contains “Covid-19”). We disallow span overlap during terminology scoring following the algorithm described in Algorithm 1.

Each terminology annotation contains the term as it occurs in the source, the annotated target lemma, the surface form the reference actually uses (inflected), and every alternative the released dictionary allows. This allows fair terminology matching across language pairs with different setups: in Spanish→Basque the annotated lemma appears ver-

batim in the reference for 98–100% of occurrences, but in English→Polish for only 34–41%.

Exact Term Success Rate measures verbatim surface matches of the inflected form from our human-annotated references, of the annotated lemma, or of any glossary alternative.

Lemmatized Term Success Rate is the most lenient terminology matching metric, considering all alternatives and all their lemmatizations, as well as the annotated version and its lemmatized version as a valid usage as described in Table 1 while enforcing no span overlap. We keep track of consumed spans in both the surface and lemma space and disallow further matches in any of them.

4.3 Overall Translation Quality

As an indication of overall quality, we run one string-matching metric alongside several trained metrics: reference-based and reference-free. We report trained metrics as a complement, since neural metrics can be less robust on out-of-domain terminology and on the long, document-length inputs (Zouhar et al., 2024).¹

chrF2++ We compute chrF2++ (Popović, 2017) with sacrebleu (Post, 2018) using default settings.² chrF2++ is tokenisation-insensitive, allowing consistent evaluation across the target languages. We report it at two granularities: a *document* score, treating each entire translation and reference document as a single string (O’Brien et al., 2025), and a *paragraph* score, averaged over the aligned paragraph units.

COMET, COMET-Kiwi and XCOMET. We use COMET (Rei et al., 2022a,b) in its reference-based (wmt22-comet-da) and reference-free QE (wmt22-cometkiwi-da) variants; together with XCOMET-XXL (Guerreiro et al., 2024), which additionally predicts error spans. All three encode the source, hypothesis and reference in separate passes, each capped at 510 tokens. Document scores are the plain mean of the segment scores.

MetricX (reference-based and QE). We use MetricX-24-Hybrid-XL-v2p6 (Juraska et al., 2024), an error score on a 0–25 scale for which lower is better, in its reference-based and reference-free (QE) settings. MetricX concatenates the source, candidate and reference into a single input sequence capped at 1536 tokens, which makes

it sensitive to input length. As for COMET, document scores are aggregated as a plain mean.

All outputs are re-segmented onto the reference segments prior to scoring, which for Track 2 zh-en means cutting whole documents; this process, along with the segment-level aggregation, is detailed in the following section.

4.4 Segment-Level Scoring

Models of the COMET family use the XLM-RoBERTa SentencePiece tokenizer with a maximal sequence length of 512. As source, hypothesis and reference are encoded separately, they should each fit within the budget. Likewise, the MetricX family uses the mT5 tokenizer, which concatenates source, hypothesis and reference and should fit within 1536 tokens of that tokenizer. Since our smallest unit is segments and they usually fit well within the limits, we re-segment hypothesis strings onto the reference segments using mwerAlign (Matusov et al., 2005; Post and Hoang, 2025). Re-segmentation runs strictly within paragraph or document boundaries. Section A provides the token budget census, empirical length effects, and a validation of segment-level alignment against paragraph-level scoring.

5 Participants and System Descriptions

This year, we see 21 systems in Track 1 and 18 systems in Track 2 (12 of which were submitted to both tracks). Their descriptions are provided below. Some teams submitted several systems; we group them by their team names (these are always the prefixes of the model names, e.g., “COZY” and “COZY-flash” both represent the same team). For easier navigation across the variety of approaches, we label them with the main features and components of particular submissions, namely:

- **Type of models**

 **NMT** NMT model (encoder-decoder)

 **LLM** LLM (decoder-only model)

 **multiple models or agentic systems** (agents, preprocessing+postprocessing, etc.)

- **Data**

 **SYNTH** data synthesis (augmentation, etc.)

 **CURATE** data curation (filtering big corpora, enriching training data with annotation, etc.)

- **Model update techniques**

 ***FT** continuous pre-training, supervised fine-tuning, etc.

¹We also probe LLM judges on the gold reference (§7.4).

²chrF2++|#:1|c:mixed|e:yes|nc:6|nw:2|s:no|v:2.6.0

***PO** various types of preference optimization: GRPO, PPO, DPO, etc.

- **Inference-time strategies**

 using additional models to refine translation outputs

 subselecting the vocabulary relevant for the current segment

ICL in-context learning, few-shot prompts, etc.

 multi-metric decoding (using both general quality and term accuracy for sequence choice)

 term injection (for NMT models)

- **Track 2 terminology usage from the seed bitexts**

 selecting the most similar/relevant sentences from the pool of samples

 creating the terminology dictionary (and reducing the task to Track 1)

TARL       The Amazon Hyderabad team (Mhaskar et al., 2026) presented a set of agentic systems that consist of four role-specialized LLM agents: translation, terminology verification, error detection and post-editing modules. The input consists of the source language sentence and relevant terminology pairs obtained by n -gram matching over the complete bilingual dictionary. For Track 2, the term dictionary is first mined from the sample bitext sentences. The Translation Agent (first step) is explicitly prompted to not simply copy the dictionary form of the term, but to inflect it according to the target language grammar; the Terminology Verification Agent (second step) adds the terms that were missed in the first step; after that, the agents treat the terms as locked and can only change the context of the sentence, not the terms themselves. The backbone LLM is Claude-Opus-4-8; no post-training is applied to it.

Cohere (Cohere CAT+) (Kocmi et al., 2026)

 ***FT** ***PO** The model is built on top of the Cohere **Command A+ architecture**. It is a mixture-of-experts (MoE) model that has been supervised fine-tuned (SFT) on both MT- and non-MT tasks (such as post-editing or quality estimation), followed by DPO and Reinforcement Learning with Verifiable Rewards (RLVR). The model does not include any special inference techniques, and the system submission is the direct performance of the model when prompted with the shared task instructions.

SalamandraTA (Liao and Melero, 2026) 

SYNTH **CURATE** ***FT** The team submitted their 7B-INSTRUCT v3.0 open-weight model, fine-tuned from its 7B base version. The training data were created using a “two-way” synthetic pipeline using Gemma-4-31B: one direction generates synthetic documents from medical glossaries, and the other extracts terminology pairs from existing parallel texts. The core of their approach is a data curation rule that filters those texts using their previous model release, v2.0, keeping only the “hard examples”, which are sentences where v2.0 fails to output the required glossary term by default. A controlled study demonstrated that fine-tuning the base model exclusively on these hard examples significantly improves term accuracy compared to using easy or unfiltered data.

COZY (systems: COZY, COZY-flash) 

  The system translates the whole documents, with post-check of the paragraph consistency and re-translation once on mismatch. The lemmatized terms from the source passages are filtered from the global dictionary, and are added to the prompt, which requires the model to inflect them according to the target language grammar. For Track 2, the glossary is extracted with the help of an LLM and provided to the prompt together with the seed bitext examples. The submitted systems differ in the backbone LLM: while **COZY** uses GPT-5.6-Sol (OpenAI, 2026), **COZY-flash** runs the pipeline on Deepseek-V4-Flash (DeepSeek-AI et al., 2026).

MASC-Dec  This system is based on the publicly available **HiTZ/mt-hitz-es-eu** Spanish-Basque NMT model with no additional fine-tuning. It uses a document-scoped terminology memory for consistent term lemma usage throughout the document. To ensure that the Basque terms are inflected correctly, the model derives candidate lemmas by using the Apertium (Forcada and Tyers, 2016) and Stanza (Qi et al., 2020) morphological analyzers, and applies character-aware soft constraints at the decoding step to promote token paths that reflect the term lemmas, while relying on the contextually appropriate suffix based on the parametric knowledge. Exact constraints are reserved for invariant items such as numbers, standards, URLs, and fixed identifiers.

CUNI-ÚFAL (systems: CUNI-UFAL, CUNI-UFAL_{dict-LLM}, CUNI-UFAL_{E2E})  

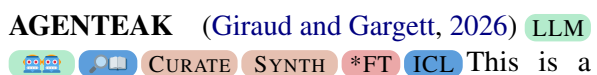
 Papáček and Sourada (2026) developed a model-agnostic framework. It takes both the source sentence and relevant terminology mappings; the team experimented with the placement of terminologies and merging terminologies from random and proper dictionaries. After the initial translation with attempts to preserve the input document structure, the framework runs a terminology revision step to check for incorrect or missing terminologies. For Track 2, where bitext samples are provided instead of terminology dictionaries, they explore 1) multi-turn prompting with examples, 2) dictionary extraction with fallback to Track 1, and 3) a hybrid of both. They used Gemini 2.5 Flash for development and compared it to Gemma 4 31B and GPT-5.6-Sol, with the latter as submitted translations.

HW-TSC  Huang et al. (2026) built a pipeline centered on Qwen3.7-Max, featuring a Track 2 bilingual glossary induction module and a shared document translation module used across both tracks. For both tracks, the shared translator performs local glossary matching against each document or chunk, capping the prompt-injected terminology block at 200 locally relevant term pairs and instructing the model to adapt target morphology and agreement grammatically.

For Track 2, the glossary induction module identifies candidate source terms through two channels: open source-term extraction using Qwen3.7-Max and expression matching against the source side of seed bitexts. The extraction prompt for the open-extraction channel is optimized offline using SkillOpt on the WMT25 Terminology Shared Task (Track 2) English–Chinese finance data and rewritten by Qwen3.7-Max for other language directions. Finally, candidates from both channels are merged, filtered, and aligned to seed target expressions with Qwen3.7-Max before being passed into the shared translation module.

RWS (TQM)  RWS team (Guinea and Alfieri, 2026) submitted Translation-Queries-Memory (TQM), a prompt-based, retrieval-augmented translation system using TranslateGemma-27B, Qwen3.5-35B-A3B-Instruct, and the in-house LW Research Pro model, without task-specific fine-tuning. For Track 1, an in-house terminology detector identifies approved terms and their inflected forms in each source segment and injects the corresponding source–target

pairs into the prompt; TranslateGemma-27B was selected for Spanish–Basque and LW Research Pro for English–Polish based on term accuracy. For Track 2, TQM retrieves semantically similar translation pairs from the provided sample bitext and supplies them as few-shot examples. Model and retrieval-depth selection relied on reference-free quality estimation, resulting in Qwen3.5-35B for Chinese–English and Spanish–Basque, and TranslateGemma-27B for English–Polish, with a different number of samples per language pair.

AGENTEAK (Giraud and Gargett, 2026)  This is a multi-agent pipeline built with LangGraph for Spanish–Basque terminology translation in the automotive and energy domains. Running on a single GPU with open-weight models spanning 4–8B parameters, the system splits translation into five stages: deterministic source-term matching, context-aware terminology selection using Qwen3-4B (reasoning mode), domain-adapted translation via Latxa-Llama-3.1-8B-Instruct (Etxaniz et al., 2024) with soft-enforced target terms, verification and targeted rewriting of non-compliant segments by a reviewer agent, and deterministic formatting fixes.

To compensate for missing domain bitext, Spanish–Basque parallel data was mined from Wikipedia using LaBSE (Feng et al., 2022) embeddings with Bicleaner AI filtering, and augmented with synthetic parallel paragraphs generated by Latxa-Llama-3.1-8B-Instruct conditioned on term bases. The translator was fine-tuned using LoRA with explicit domain tags on a mixture of mined, synthetic, and general-domain corpora.

GRIAL-TA  The GRIAL-TA team (López-Sánchez et al., 2026) submitted systems for both tracks of the English–Polish language pair, using instruction-tuned Gemma 4 models fine-tuned on curated general and domain-specific data, including parallel corpora, terminological databases and bilingual dictionaries. They trained separate models for medicine and mechanical engineering domains, experimenting with different model sizes, prompting configurations, and fine-tuning strategies. For both tracks, translation is performed at the segment level, with the terminology explicitly associated with each segment injected into the prompt. In Track 2, the team first extracts terms from the provided sample bitext using TBXTools,

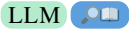
combining a BERT-based extractor with linguistic POS-pattern extraction. The extracted terms are then aligned across English and Polish using LaBSE embeddings, and the resulting terminology pairs are used in the same translation procedure as in Track 1.

UniOr MT (systems: UniOR-ZSL+Term, UniOR-GenPE, UniOR-RulePE, UniOR-Mixed)

LLM                     

allowing even long documents to be translated without explicit sentence or paragraph splitting. In the Revision step, language-specific models are employed to revise the initial translation. Latxa-70B (Etxaniz et al., 2024) is used for Basque revision, and PLLuM-12B (Kocoń et al., 2025) is used for Polish revision. English revision is omitted as it does not improve quality.

VNFusion  The system features terminology-aware candidate generation with source-grounded selection and automatic post-editing, preserving glossary terms with appropriate target-language morphology rather than forcing exact surface forms. Additionally, the authors claim that the system outputs are reviewed with MQM-style review for accuracy, fluency, professionalism, and cross-segment coherence.

Tilde (systems: TO-TM_{top5}, TO-TM_{top15}, TO-G4terms, TO-fuzzy, TO-fuzzy_{stem})     The Tilde team (Pinnis et al., 2026) fine-tuned TildeOpen-30B using data they compiled and augmented for general translation, glossary-guided translation, and translation with translation-memory examples. They used the resulting model for English–Polish translation in both tracks. For Track 1, the system translates documents paragraph by paragraph, narrowing down the provided dictionary to potentially relevant term pairs for each paragraph and including them in the translation prompt. The team compared fuzzy matching, stemming with exact matching, and stemming with fuzzy matching. These methods prioritize recall, leaving contextual disambiguation to the translation model. For Track 2, the team uses the provided sample bitexts either to build a glossary or to supply translation examples. The first approach prompts Gemma 4 31B to extract corresponding domain-specific terms from the source texts and their translations, then uses the resulting glossary as in Track 1. The second retrieves similar bilingual sentence pairs for each source sentence in a paragraph and selects examples in turn from each sentence’s ranked list until a fixed total is reached. These examples guide translation of the entire paragraph.

6 Shared Task Results

Figures 1 and 2 evaluate system capability along two main axes: **terminology success (lemmatized)** on the y-axis and **MetricX-24, reference-**

based,³ as the general quality metric on the x-axis. MetricX-24 is preferred over chrF++ here because chrF++ reacts directly to dictionary injection; a term×chrF++ trade-off would artificially double-count terminology adherence (Semenov et al., 2025). To ensure fair comparisons, these Pareto figures only include non-ablated submissions that covered every direction in a track. Systems lying on the Pareto frontier represent the leading systems for each track. Following §4.2, performance is not evaluated against a nominal maximum of 100%; rather, we contextualize each system’s score against the human (gold) reference scored against itself. Full direction-wise results (including chrF++ and segment-aligned neural metrics) are tabulated in Tables 6 and 7 in the Appendix.

Track 1 is close to its ceiling on terminology. The best five systems fall within 1.2 points of each other, at 94.5–95.7 term success, and the leader sits 0.7 points below the gold ceiling of 96.4. With an explicit dictionary and no constraint on model size, little headroom remains on this measure: what remains is the residue the metric charges to the reference as well, with repeated mentions rendered by a pronoun or a shared head. General quality is not saturated in the same way. The best system reaches 3.55 MetricX against a gold value (human reference) of 2.04, a gap of one and a half points that no system closes, and the field spans 3.55–8.39.

Track 2 is not saturated, and the spread roughly doubles. Without an explicit dictionary, the leader reaches 89.2 term success, 4.2 points below the gold ceiling of 93.4, and the top five systems spread over 2.1 points against 1.2 in Track 1. Twelve systems entered both tracks. On en–pl, where the two tracks draw on the same two domains, term success falls for every one of them when the dictionary is replaced by sample bitexts, by between 1.2 and 53.0 points, with a median of 11.0. The size of the drop is a property of the system rather than of the track: CUNI-UFAL loses 1.2 points, and STRA-MT loses 53.0, and STRA-MT’s own ablation attributes this to the coverage of its mined inventory, which its word aligner populates mostly with single-word terms. Identifying which terms need controlling is where the difficulty of the task now sits.

³We selected MetricX-24 over COMET-22 due to documented underperformance of the latter on Spanish–Basque (Falcão et al., 2024).

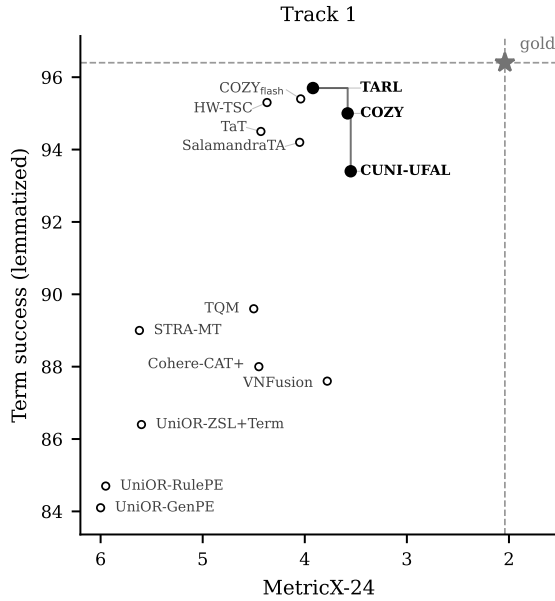


Figure 1: Terminology against general quality in Track 1. x is MetricX-24, an error score, so the axis is inverted. Filled markers are the systems on the Pareto frontier; the dashed lines mark the gold reference scored against itself on each axis. Only systems covering every direction of the track are plotted (6 further systems are omitted on that ground), and of those, 14 of the 15 contenders are shown (term success ≥ 75 , MetricX ≤ 6). Appendix Table 6 reports every submission.

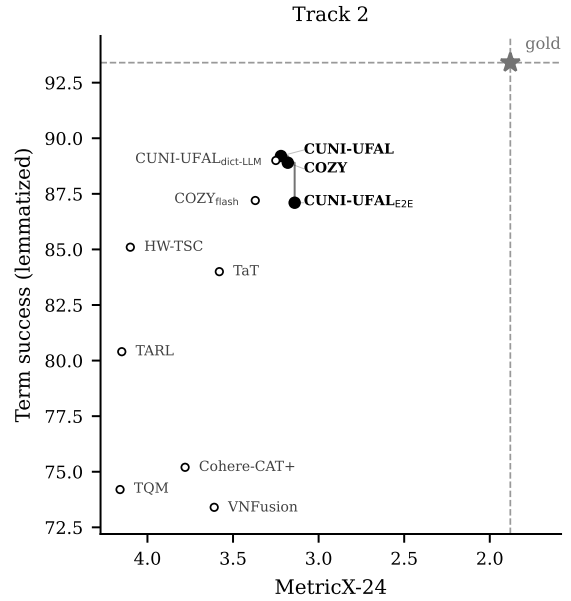


Figure 2: Terminology against general quality in Track 2. x is MetricX-24, an error score, so the axis is inverted. Filled markers are the systems on the Pareto frontier; the dashed lines mark the gold reference scored against itself on each axis. Only systems covering every direction of the track are plotted (5 further systems are omitted on that ground), and of those 11 of the 13 contenders are shown (term success ≥ 70 , MetricX ≤ 5). Appendix Table 7 reports every submission.

The frontier systems pair revision with a large backbone. The frontier systems are TARL, COZY and CUNI-UFAL in Track 1 and CUNI-UFAL, COZY and CUNI-UFAL_{E2E} in Track 2. No system runs a single translation pass: each generates, checks the output against the terms it selected, and revises where the check fails. Also, the teams on both frontiers use a strong general-purpose LLM behind the pipeline rather than a model trained specifically for the task: COZY and CUNI-UFAL both submit GPT-5.6-Sol, and TARL runs Claude Opus-4.8. A revision stage is not by itself what separates them, though—AGENTEAK, STRA-MT, STaR-MT, VNFusion and UniOr’s post-editing paths all verify their output and rewrite it, several of them against the dictionary specifically, without coming close to the frontier. However, where those teams report a backbone, it is an open-weight model between 1.3B and 70B parameters, or GPT-4o-mini. What the leaders add is the capacity of the model for terminology revision. This is consistent with concurrent work on multi-agent MT for financial long documents, where backbone capacity matters more than workflow complexity

for terminology accuracy (Oncevay et al., 2026).

Where the terminology success metric runs out. In zh-en, the terminology gold ceiling is 86.3, lower than in the other directions because the HKMA documents repeat their terms most, and six systems score above it, up to 89.0. They are not more accurate than the reference: a system that repeats the dictionary target at every mention satisfies the metric where a professional translator pronominalises. This is the clearest case for reading the leaderboard against the ceiling, and it bounds how far term success alone should be pushed as an objective.

7 Analysis

7.1 Effect of Terminology Incorporation

We investigate the impact of incorporating terminology on translation quality. As noted in §2.3, Track 1 comprises No Term, Proper Term and Random Term; Track 2 provides no explicit dictionary, so it distinguishes only No Term and Sample. Figure 3 draws one vertical bar per system, spanning that system’s modes, with the tracks side by side

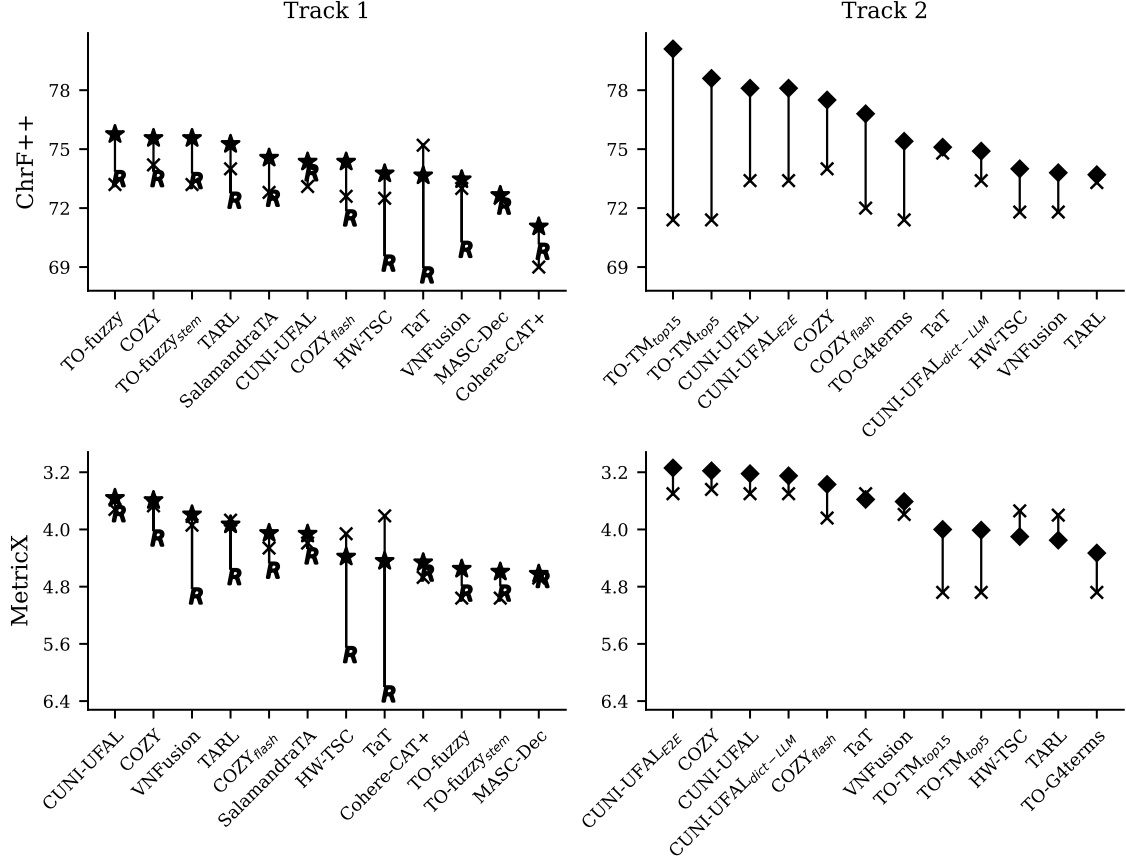


Figure 3: Effect of terminology mode, by track (columns) and metric (rows). Each panel shows the top 12 systems of its track by its own metric in the proper (Track 1) or sample (Track 2) mode, best on the left, so the two panels of a column need not show the same systems in the same order. Legend: $\times$ no term and R random term; $\star$ proper term (Track 1), $\blacklozenge$ sample term (Track 2). MetricX axis is inverted. Both tracks are drawn to the same scale in each row.

and the two metrics in rows. The length of a bar is how much the terminology guidance is worth to that system, and because a row is drawn to one scale, bars are comparable across the two tracks. Each panel shows the twelve highest-scoring systems for readability; the averages quoted below are over all submissions in the track (see Appendix Tables 8 and 9 for the full scores).

In Track 1 the star sits above the $\times$ for almost every system: the dictionary is worth 26.4 points of term success and 1.58 chrF++ on average for all systems. The R sits at the other end of the bar, below the $\times$. A dictionary of random terms leaves a system worse off than no dictionary at all, by 1.09 chrF++ and 2.19 points of term success.

This is not because the random entries name content the source does not contain. They are ordinary source words, a median of three to five per segment, each paired with a single reference-aligned target that is usually a base form. Systems adopt them. For every random entry whose source word occurs

in a document, we checked whether one of its targets appears in the translation: the rate is 42.9% under No Term and 52.1% under Random Term.⁴ The systems that adopt the most also lose the most, at $r = -0.54$ for chrF++ and $r = +0.67$ for MetricX. The star-to- R distance therefore measures how far a system defers to whatever dictionary it is given, and it is widest for TaT and HW-TSC.

Track 2 is the novelty this year, and its bars are the longer ones in the top row. Sample bitexts are worth more than an explicit dictionary in general quality, 2.99 chrF++ against 1.58, and less in terminology, 14.5 points of term success against 26.4. A sample bitext carries register, domain phrasing and sentence structure alongside its terminology, so it improves the translation by more than the terms alone would explain, while being a weaker instru-

⁴A deterministic surface match over the released random dictionary, case-insensitive and word-bounded, on the same documents in both modes; morphology makes it a lower bound in both.

ment for terminology than a dictionary that names the required target outright. The bars also differ in length far more than in Track 1, which suggests that **how a system exploits the samples matters more than whether it receives them**.

The bottom row is the same comparison under a trained metric, and it is nearly flat. MetricX improves by 0.16 in Track 1 and 0.08 in Track 2, against more than twenty-six points of term success. The one place its bars open up is the *R*, which costs 0.57 and drops furthest for the two systems that inserted the random targets most. A trained quality metric is thus sensitive to terminology being *wrong* and close to blind to terminology being *right*, which is the behaviour a terminology metric is needed to cover and which we return to in the following section.

7.2 Ranking Correlation

Terminology-aware translation is a multifaceted problem, and the metrics we report are built to look at different things: a string-matching metric, trained quality metrics, and a terminology metric driven by the dictionary. It is therefore worth asking how differently these metrics rank the participating systems, both to understand what each one contributes and to inform which of them a future evaluation actually needs. Following last year’s shared task (Semenov et al., 2025), we compare the rankings with Kendall’s τ (Kendall, 1938). We correlate at the system level, using each metric’s direction-averaged score, and we negate MetricX-24 so that higher is better for every metric and a positive τ means two metrics agree.

The clearest structure is not in any single cell but in the contrast between two blocks. The overall quality metrics agree with one another more than any of them agrees with Term Success: among chrF++, COMET-22 and MetricX-24 the coefficients average 0.63 in Track 1 and 0.70 in Track 2, whereas the Term Success row averages 0.47 and 0.55. The pattern holds across both tracks, which differ in their systems, their language directions and the form the terminology guidance takes. Every coefficient is nevertheless positive and significant, so the metrics are not measuring unrelated things; better systems are better on all axes. Terminology adherence is simply the axis on which the quality metrics agree least, and it is not recoverable from them.

Some of the metrics we compute are, by contrast, close to redundant with one another.

	Track 1 (21 systems)			Track 2 (18 systems)		
	ChrF++	COMET 22	MetricX 24	ChrF++	COMET 22	MetricX 24
COMET-22	0.61*	–		0.86*	–	
MetricX-24	0.50*	0.79*	–	0.59*	0.65*	–
Term Success	0.36*	0.45*	0.59*	0.55*	0.48*	0.63*

Figure 4: Kendall’s τ correlation between the automatic metrics, on direction-averaged system scores. MetricX is negated so that higher is better for every metric. Comparisons with $p < 0.05$ are marked with *.

XCOMET-XXL ranks systems almost identically to COMET-22 ($\tau = 0.92$ in Track 1, 0.94 in Track 2), and the reference-free variants stay close to their reference-based counterparts: MetricX-24-QE against MetricX-24 with 0.86 and 0.72, and CometKiwi-22 against COMET-22 with 0.70 and 0.79. Reporting all of them would spend space on a distinction the rankings do not make, so Figure 4 shows one representative per family.

The practical conclusion mirrors the one drawn at WMT25, on a larger and more varied set of systems. An evaluation of terminology-aware translation **needs a term-specific metric alongside a quality metric**, and, given how little the reference-free variants change the ordering, one trained metric from each family is enough.

7.3 Grammaticality Estimation

As mentioned in Section 2.2, Basque and Polish have considerable morphological complexity, featuring multiple morphological categories (gender, case, number, definitiveness, etc.) while being relatively under-represented. We therefore ask: when terms are appropriately chosen at the lemma level, how fluent and grammatically coherent are their occurrences in sentence contexts? To answer this, we split the outputs (and aligned references) into sentences and, for each system, selected only passages where the term lemmas were correctly used, based on the lemmatized terminology success rate from Section 4.2.

Grammatical error correction (GEC) and identification (GEI) are well-established fields in NLP (Bryant et al., 2023), but existing solutions and datasets (such as Masciolini et al. 2025) largely focus on spelling errors. Sentences composed of orthographically correct but grammatically incoherent words (e.g., “The dogs *is* chasing the

cat”) remain more difficult to detect, particularly for under-resourced languages⁵. Estimating the log-probability of system outputs with language-specific LLMs is a possible alternative, but this approach has shown inconsistent results (Semenov and Sennrich, 2025). Since there is no straightforward metric targeting grammatical coherence, we therefore conducted several exploratory estimations of the grammatical and syntactic reasonableness of terms in the system outputs, complemented by a short qualitative analysis.

The simplest sanity check of the grammatical coherence of the terms is to analyze their **relative position in the sentence**. An adversarial system could naively add relevant terms retrieved from the dictionary to the end of the sentence, thereby preserving overall fluency while maximizing the terminology success rate; but in natural sentences, we expect terms to be distributed relatively evenly throughout the sentence. We therefore compared the distributions of term positions (normalized by sentence lengths) in the system outputs and corresponding references. The results (see Section B.1) showed that the relative term position patterns in the reference texts are language-dependent: in Basque, terms gravitate towards the beginning of the sentence, while in Polish the opposite is true. All system outputs follow these patterns, with only rare statistically significant differences (up to 2% later than the reference for some systems).

Next, we focused on the inflected forms of the terms. We were interested in understanding **how diverse the grammatical forms of the terms are** in the generated sentences. To measure this, we used the mean size of paradigm (MSP) concept introduced in (Xanthos and Gillis, 2010): we counted the number of unique inflected forms of the terms in the texts and divided it by the number of unique term lemmas (or dictionary values for multi-word terms). A higher MSP value indicates that a system can generate a greater variety of grammatical forms of the same term. The results (shown in Figure 8 in Section B.2) are generally predictable: the references show a slightly higher MSP than the hypotheses, while systems that retrieved more terms show higher MSP than those that retrieved fewer. A notable observation concerns the best-performing systems for Basque: they show a large discrepancy between the numbers of unique inflected forms in

the hypotheses and references, with *considerably more* unique forms in the system outputs. Qualitative analysis showed that, in many cases, this is an artifact of white spaces either on the reference side (annotation artifacts), or on the system output side (system used the correct suffix, but separated it with an excessive white space). Examples of both phenomena and another example of over-regularization are shown in Table 5 in Section B.2. Another notable example is the irregular word “plater”, which should retain a single “r” after a suffix is appended (unlike most words in Basque, where the “r”-final words after suffixation duplicate the consonants). Sometimes it is incorrectly inflected with the doubled “r” (e.g. presio-plattera instead of presio-platera). This is an example of over-regularization or over-correction, a frequent phenomenon in children’s or L2 speech, where speakers who are not fully fluent in a language apply regular morphological rules to exceptions or less frequent patterns (for example, “bayed” instead of “bought”). In recent years, there has been growing evidence of over-regularization patterns in cognitively plausible LMs (Haga et al., 2024).

Another way to address grammaticality is to analyze the **distribution of the grammemes** (particular values of a grammatical category, for example, the counts of the nominative, genitive, and locative cases, which are all grammemes of the “case” category). We did this by selecting only single-word terms and analyzing them with Stanza (Qi et al., 2020): it integrates morphological annotation of features from Universal Dependencies (de Marneffe et al., 2021), where the feats value shows all recognized grammemes for each word. Since the majority of terms are nouns (and, less frequently, adjectives), we focused on the grammatical categories of **case** and **number**.

The aggregated visualization of the case grammeme distributions is shown in Figure 5. Although there are no dramatic changes compared to the reference translations (as reflected in relatively low Jensen-Shannon Distance values), there are several notable observations. First, in Basque, we observe decreasing usage of the Ergative case (ERG). Basque has an Absolutive-Ergative case structure (as opposed to Nominative-Accusative systems, which are predominant among the world’s lan-

⁵To the best of our knowledge, there are no robust grammatical error checkers for Polish and Basque, except for classical NLP approaches, such as (Gojenola and Oronoz, 2000).

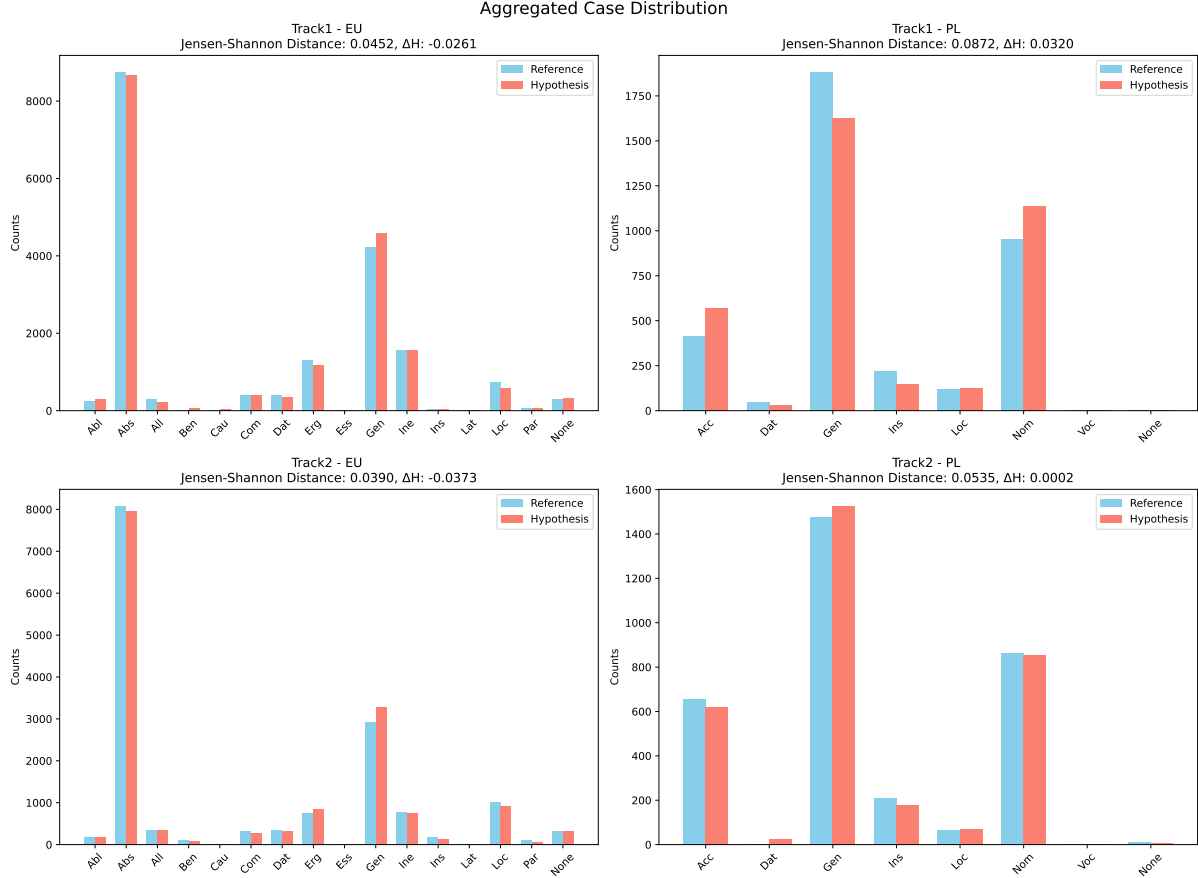


Figure 5: Distribution of the cases in Basque and Polish single-word terms, aggregated from all systems. Each subgraph shows a distribution for a particular language (columns) and track (rows). Within each graph, two colors represent the occurrences for the reference and for the submitted outputs (“hypothesis”). On the top of each subgraph, we provide Jensen-Shannon Distance (ranging from 0 to 1, where 0 means identical distributions), and ΔH shows the difference in entropies of the distributions, which is calculated as $\Delta H = H(ref) - H(hyp)$. The cases are sorted alphabetically.

guages⁶), and we can expect that LLMs, trained predominantly on higher-resourced Nominative-Accusative languages, impose a structural bias on inflection when presented with dictionary forms of words and do not favour the use of ergative constructions. This is indirectly supported by the less pronounced trend in Track 2, since many systems directly leverage sentences mined from the bitexts, increasing the chance of copying nouns in the ergative case.

(1)

- a. *aerosorgailu-a eraiki*
wind.turbine-ABS build
‘build a wind turbine’ [reference]

⁶Dryer and Haspelmath, 2013, c. 98A, show that there are more languages in the world with nominative-accusative structure (over 50), than with Absolutive-Ergative (over 30). More importantly, the most resourced languages are almost only the nominative-accusative languages (German, Russian, Spanish, Japanese, etc.).

- b. *aerosorgailu-aren eraikuntza*
wind.turbine-GEN construction
‘construction of the wind turbine’ [MASC-Dec]
- c. *construcción del aerogenerador*
construction of.the wind.turbine
‘construction of the wind turbine’ [input]

Another notable shift is the increase of the Genitive case (GEN). The qualitative analysis showed that this often happens in the contexts that can be formulated as verbal or as noun phrases. In the above example, we see that the same idea (building a wind turbine) can be expressed as a verb + direct object (in Absolutive case, ABS, 1a), or as a deverbal noun + possessive noun (in GEN, 1b). Although both “verb structures” and “noun structures” are fully grammatical and acceptable in Basque, the latter can be considered “excessively heavy”, and is sometimes perceived as a calque from Spanish

(which extensively uses noun structures, see the input in 1c).

(2)

- a. *w faze telogenu*
in phase.ACC telogen.GEN
‘in the telogen phase’ [reference]
- b. *w telogen*
in telogen.ACC
‘in the telogen [phase]’ [CUNI-UFAL]

Third, the case distribution in Polish shows the pronounced drift towards the nominative and the accusative cases. The Nominative case (NOM, a dictionary form) may reflect that the models tend to use the terms “as they are” when prompted with their dictionary forms. Contrary to our initial assumption (ACC is often homonymous to NOM, so it may be morphological analyzer errors), the qualitative analysis showed that often ACC appears from rephrasing the noun phrases with the nested terms in GEN (e.g. 2a) towards simple phrases with terms in ACC (e.g. 2b).

The distribution of the number grammemes is shown in Figure 7 (Section B.3). Notably, Polish data show a significant drop in the usage of the plural forms (which is reflected in ΔH decrease by 0.12).

The observations in the grammeme distributions, together with qualitative analysis, demonstrate that the models are generally capable not only of selecting appropriate term lemmas but also of arranging them naturally in the sentence, both in terms of their position and coherence with the surrounding context. However, there is a slight but steady drift towards using the dictionary forms of the terms and skewing the variety of grammemes towards several most frequent cases. This is consistent with other evidence of degeneration in morphosyntactic diversity noted by Luo et al. (2024) and Guo et al. (2025), which is one example of the so-called “mode collapse” (Anschel et al., 2025; Qi and Murat, 2026): LLMs’ tendency to generate more “averaged” sequences due to maximum likelihood-based training objectives. Another interpretation of this phenomenon is that, since the models do not fail to arrange the rest of the sentence according to grammatical rules, they seem to favor copying the constrained inputs and then building a grammatically coherent sentence “around” them. This is consistent with the well-established body of research on LLM inductive copying abilities through induction

heads and in-context learning (Olsson et al., 2022; Lv et al., 2025; Jelassi et al., 2024). Such behavior is explicitly enforced by at least one system, TARL.

7.4 Terminology Drift in LLM Judges

LLM judges are now common in MT evaluation, and terminology is one of the standard MQM categories (Freitag et al., 2021; Kocmi and Federmann, 2023). We did not use them to rank systems, but the shared task offers a clean probe: the human reference is what the terminology metric is defined against, so a judge that flags a terminology error on the reference is, with few exceptions, wrong.

We follow the WMT25 Terminology judge setup, in the GEMBA-MQM line (Kocmi and Federmann, 2023): the model returns error spans with an MQM category and severity, weighted 1, 5 and 10 for minor, major and critical. We prompt zero-shot, and use Focus Sentence Prompting (Domhan and Zhu, 2025): the model receives the full source document and its translation as context, and is then asked for errors in one segment at a time, with the relevant dictionary entries repeated alongside each segment. The judge is reference-free, which is what makes the reference usable as a probe. We ran five backbones available in our environment: GPT-5.6-Luna, MiniMax-M2.5, GLM-5, Qwen3-235B, and DeepSeek-V3.2.

These judges emitted 6,507 terminology spans on the gold references. In **23.8%** of them the judge’s own explanation undermines the span: it states outright that the translation is correct (16.2%), or concedes that the required term is present and objects only to its inflected form (7.6%). The rate ranges from 0.3% (GPT-5.6-Luna) to 43.0% (DeepSeek-V3.2). The backbones also differ by an order of magnitude in how many spans they emit at all, so the pooled figure describes the judges that flag most rather than a typical judge, and the range is the safer summary.⁷

Table 2 shows what this looks like. The judges are not ignorant of morphology: they identify the inflected form, check it against the dictionary, conclude that it complies, and emit the span regardless. Several go further and announce the retraction in the explanation (“Removing this span”) while the span still reaches the output. This is an output-

⁷Counted by pattern match over the explanation field, requiring an unnegated sentence that gives an explicit positive verdict on the translation; of 25 flagged spans we read by hand, 23 were genuine concessions. Recall is the weaker side, since a bare “Correct.” is easily missed, so the rate is a lower bound.

stroke → dictionary *udar mózgu*, reference *udar mózgu*
 “The translation uses ‘udar mózgu’, which matches the dictionary entry. This is correct...” (*span emitted as a terminology error*)

aerogenerator → dictionary *haize-sorgailu*, reference *haize-sorgailuetan*
 “...is grammatically correct (with locative suffix *-tan*), but the base form...” (*span emitted as a terminology error*)

turismo → dictionary *turismo*, reference *turismoak*
 “...an inflected plural form of ‘turismo’, which is an accepted target. Therefore, this is correct and should not be flagged. I must retract this as it’s erroneous.” (*span emitted as a terminology error*)

Table 2: Terminology errors reported against the *human reference*, with the judge’s own explanation. The reference uses an accepted target in every case.

discipline failure rather than a knowledge failure. The effect concentrates where inflection is richest, at 29.4% of the spans in en-pl and 25.8% in es-eu against 9.0% in zh-en.

Two things follow for evaluation practice: (i) a result obtained from a single backbone says little about what another would have found, so judge-based terminology findings need more than one model behind them;⁸ and (ii) a judge’s terminology output should not be consumed without a consistency check against its own explanations, which is cheap and, on this evidence, removes a large fraction of the noise. Why these judges behave this way, and whether the terminology annotations that a task like this one provides can be used to correct them, we leave to future work.

8 Related Work

Building terminology-aware MT systems is critical to achieve quality translation that properly covers domain-specific lexical knowledge such as brand names, technical jargon, or client-specific translation constraints. The task is highly relevant for industrial deployment in particular, where accuracy on specialised terminology is one of the key determinants of translation quality (Di Rosa, 2026; Oncevay et al., 2025b). Two main families of approaches have been explored: hard constraints, which alter the decoding process to enforce term generation, and soft constraints, where models are steered towards generating the expected terms without altering the decoding mechanism.

Among hard-constraint approaches, Hokamp

⁸Another mitigation is single-backbone multi-run aggregation (Junczys-Dowmunt, 2025), though its effect on terminology errors specifically has not been investigated.

and Liu (2017) demonstrated the enforcement of lexical constraints via Grid Beam Search (GBS) for terminological adaptation. Although their approach could satisfy lexical constraints, it incurred a significant computational cost. Post and Vilar (2018) introduced an alternative to reduce GBS computational complexity, but hard constraints typically still suffer from increased translation latency, while also facing difficulties to generate proper morphological derivations for the translated terms (Jon et al., 2021; Bergmanis and Pinnis, 2021). As an alternative, Michon et al. (2020) proposed the use of term placeholders, augmented with morphosyntactic annotation, while Hasler et al. (2018) explored finite-state machines and multi-stack decoding for the task.

To address the challenges of lexically constrained decoding, Dinu et al. (2019) introduced a soft-constraint approach based on inline term annotation for Neural Machine Translation (NMT), training on data augmented with term injection and deriving the application of terminology constraints without any modification of the decoding process. Encouraging the models to follow terminological constraints without incurring additional decoding cost has been a fruitful line of research in recent years (Ailem et al., 2022; Xu and Carpuat, 2021; Alam et al., 2021). Although they provide a straightforward alternative at no additional computation cost, soft constraint approaches do not however guarantee a strict adherence to terminological constraints, as they rely on the probabilistic nature of standard NMT decoding mechanisms. The trade-off between accuracy and fluency in terminology-aware NMT has notably been studied by Abela et al. (2026), who proposed a hybrid approach mixing soft data augmentation and constrained inference.

Recently, the rise of LLMs has introduced new paradigms for terminology-aware translation, exploiting their instruction-following capabilities to enforce terminological constraints, via in-context learning to condition or refine translation, or by fine-tuning means (Bogoychev and Chen, 2023; Moslem et al., 2023; Zheng et al., 2024; Wang et al., 2023; Kim et al., 2024). This approach has notably been explored on a variety of languages, including low-resource ones (Oncevay et al., 2025a). For the languages of this shared task, Etchegoyhen et al. (2025) compared soft-constraint NMT and few-shot LLM-based MT in Basque-Spanish, showing higher terminological accuracy with the LLM approach, although with an overall lower

translation quality at the sentence level.

Among shared tasks on terminology-aware MT, [Alam et al. \(2021\)](#) organised the first of its kind in the WMT series, which focused on the medical domain and addressed five language pairs: English to French, Chinese, Russian and Korean, and Czech to German. Their work helped establish a solid foundation for the evaluation of terminology in MT research, with a novel benchmark that was mined semi-automatically from parallel corpora. The second iteration of the task expanded the range of domains, including scientific abstracts and web novels in addition to medical texts, and addressed three translation pairs: Chinese-English, English-Czech, and German-English ([Semenov et al., 2023](#)). They contrasted dictionary usage with information from reference translations, showing that both types of data yielded similar results. This result challenged standard assumptions on terminologies being the crux of meaning in translation, while also opening the debate on limitations of evaluation metrics that are not terminology-centric.

In 2025, the WMT shared task focused on the information technology domain for English to German, Russian and Spanish, at the sentence level, and the finance domain for English-Chinese at the document level. The task supported an evaluation of terminology accuracy and consistency, as well as overall translation quality, by means of different terminology modes: no terminology, proper terminology and random terminology. One of the main conclusions of the task was that document-level translation lagged behind sentence level, showcasing the need for long-form evaluation and more robust metrics overall ([Semenov et al., 2025](#)). In parallel, [Conia et al. \(2025\)](#) organised a task on Entity-Aware MT within SemEval, where they introduced the XC-Translate benchmark, containing over 50K manual translations with entities requiring the use of specialised knowledge for translation.

This year was also marked by a separate terminology-aware English-Turkish MT shared task, which was organized within the SIGTURK workshop ([Gebeşçe et al., 2026](#)). The authors created a new human-annotated dataset based on Wikipedia and university students’ theses, provided the participants with 5 baselines, and analyzed the submissions (all of which were LLM-based, with additional optimization techniques or RAG modules). The results show positive scaling effects from the backbone model size, and a discrepancy between the human annotators and the

models in adhering to the explicit dictionary constraints (models tend to rely on their parametric terminological knowledge rather than on explicit instructions). It is worth noting that the aim of the shared task was different from ours: while we test the models’ abilities in end-to-end MT, the organizers of the English-Turkish MT shared task focused on particular steps of the terminology in machine translation (automated term detection in the texts, and different forms of post-editing). These results help us enrich our understanding of the elements of the terminology-aware MT systems (e.g., terminology extraction in Track 2), which our task did not causally pinpoint.

The present shared task builds upon previous work by including different translation modes, including random terms, while expanding to new language pairs and domains, generalising the benchmarks to document-level translation, providing a novel track where participants are tasked to extract the terminology from raw data, and expanding the set of evaluation metrics.

9 Conclusions

The Terminology Translation Task at WMT-2026 was aimed at estimating the progress in the subfield of terminology-aware document-level translation, especially for lower-resourced morphologically-rich languages. We arranged two tracks (one with explicit dictionaries, another with sample bitext seeds) across three translation pairs (Spanish-Basque, English-Polish, Traditional Chinese-English), covering different fields of medicine, science, and finance. Apart from curating the data, we also focused on the extensive implementation of automated metrics (refining the term success rate metric for the document level and comparing it with multiple surface-based and neural metrics of general MT quality), as well as on qualitative and quantitative analysis of the fluency and grammaticality of the system outputs.

We received approximately 20 submissions for each task, with most teams submitting to all language pairs. This is more balanced compared to last year’s competition, where most systems were concentrating on sentence-level translation (which appeared to have saturated). The variety of approaches suggested by the participants reflects global trends in NLP: most of the systems were using LLMs, often combined with agentic or multi-model approaches.

The results of the competition are encouraging: according to our metrics, the task of MT with explicit dictionaries, even at the document level, is close to saturation in both general quality and term accuracy. This is a significant jump compared to last year, when the systems, tested on considerably higher-resourced languages, were lagging behind on both axes. One of the reasons may be the explicit techniques used by many participants, who filtered the document-level dictionaries and chunked the sentences, thereby reducing this task to the saturated problem of sentence-level MT.

On the other hand, the systems tend to struggle when not provided with explicit terminology: the spread of the scores is significantly larger, and the terminology success rate never reaches 90%. Interestingly, the models that explicitly use terminology mining modules do not seem to be the top-scoring systems.

The comparison of the different metrics shows that we can group most of them by their backbone systems, and that none of the general quality metrics, be it surface-based or neural-based, can substitute for the term success rate. When it comes to the terminology modes, we see that metrics react differently to random and proper term incorporation (neural metrics being more sensitive to the quality of the dictionaries than the surface-based ones). At the same time, the exclusion of the bitext seeds from Track 2 resulted in a significant drop in translation quality.

When addressing under-resourced and morphologically rich languages such as Basque and Polish, we first came across the scarcity of systems available for the automatic analysis of these languages (moreover, the Stanza morphological analyzer that we used in our work was also extensively used by the participants). The analysis of the LLMs-as-Judges that we ran as an experiment against the reference texts, also showed us unstable results depending on the model, as well as inconsistencies between the generated error spans and explanations. On the positive side, the automated checks on the positional and grammatical distributions in the system outputs, together with the manual checks of the most suspicious cases, did not reveal any worrying trends in the grammaticality of the term-aware MT outputs; yet we do observe slight but steady degeneration in the variety of the grammatical forms and grammemes that the systems generate.

Taken together, this raises the following **perspectives** for the future shared tasks:

Data Formats and Languages. The comparison between Tracks 1 and 2 shows that systems seem to saturate on new tasks quickly. This highlights the necessity of focusing on non-trivial and more naturalistic ways through which terminology can be acquired (for example, continuous chapters from books, cross-modal materials, etc.). Additionally, it would be valuable to extend the pool of languages used in our evaluation towards more typologically diverse and less-represented ones, since this year’s results are quite promising.

Metrics. While being more robust for high-resourced languages, neural and LLM-based metrics showed shortcomings and inconsistent performance on our languages, which is supported by previous research. Additionally, although we did not find any serious cases of grammar rule violations in our text, we still think that grammaticality evaluation should be included in the evaluation of submissions in the coming years. Finally, this year’s implementation of the term success rate metric opens the horizon for taking into account more subtle aspects of terminology usage, such as coreference, pronominalization, and other patterns in the relation between terms in the discourse.

Human Evaluation. This year’s input data were manually annotated by domain experts and native speakers of the target languages, marking a new step in the development of our shared task. However, we still lack systematic human evaluation of the system outputs, which should give us a better understanding not only of the submissions, but also of the metrics described above.

Outreach and Interactivity. This year, we experimented for the first time with opening an [interactive leaderboard](#) for participants after the end of the official competition. We think that this will be more helpful for the community, as it creates space for more continuous feedback about the data, metrics, and solutions for the task. We will consider how to integrate this into future shared tasks on a larger scale.

We are open to any collaboration and, as always, especially look forward to ideas about resources that can be used for the test sets and human evaluation. Let’s keep in touch!

Limitations

Neural metrics like COMET and MetricX have shown limited reliability when evaluating transla-

tions into Basque, reducing our confidence in the automated quality scores for those outputs. Our terminology evaluation pipeline and several participating systems rely heavily on Stanza for lemmatization. Imperfect morphological parsing can lead to false positives or negatives in term matching.

Terminology dictionaries were released at the document level rather than the sentence level. While this deliberately mimics professional translation workflows, it introduces alignment ambiguity for systems optimized for sentence-level term injection. In Track 2, the sample sentences provided to systems are exact subsets extracted directly from the target test texts, simplifying the term discovery process.

There is no solid guarantee of data non-contamination. We consider this quite unlikely, as the Polish dataset was created specifically for this task and the Basque dataset was published but removed from public repositories. The HKMA reports are posted online, but they cover a specific, recent time period.

Acknowledgments

We thank all participants for their submissions.

Vilém Zouhar gratefully acknowledges the support of a Google PhD Fellowship.

The Traditional Chinese-English data in Track 2 are derived from publications on the Hong Kong Monetary Authority website. We acknowledge the HKMA’s ownership of the intellectual property rights in the information and in all its copies. We are grateful for the availability of these materials for research purposes.

We also thank Bittor Alkain (EHU) for the fruitful discussion on the Basque case structure.

References

- Kurt Abela, Marc Tanti, and Claudia Borg. 2026. [Balancing fluency and adherence: Hybrid fallback term injection in low-resource terminology translation](#). In *Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026)*.
- Melissa Ailem, Jingshu Liu, and Raheel Qader. 2022. [Encouraging neural machine translation to satisfy terminology constraints](#). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*.
- 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*.
- Oron Anshel, Alon Shoshan, Adam Botach, Shunit Haviv Hakimi, Asaf Gendler, Emanuel Ben Baruch, Nadav Bhonker, Igor Kviatkovsky, Manoj Aggarwal, and Gerard Medioni. 2025. [Group-aware reinforcement learning for output diversity in large language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.
- Nathaniel Berger, Johannes Eschbach-Dymanus, Miriam Exel, Matthias Huck, and Stefan Riezler. 2025. [Learning to translate ambiguous terminology by preference optimization on post-edits](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*.
- 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*.
- 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*.
- Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. [Grammatical error correction: A survey of the state of the art](#). *Computational Linguistics*, 49(3):643–701.
- Antonio Castaldo, Giulia Speranza, Maria Carmen Staiano, Petra Giommarelli, Johanna Monti, and Maria Pia Di Buono. 2026. UniOR MT: An Agentic Pipeline for Terminology Translation. In *To appear in Proceedings of the Eleventh Conference on Machine Translation*.
- Simone Conia, Min Li, Roberto Navigli, and Saloni Potdar. 2025. [SemEval-2025 task 2: Entity-aware machine translation](#). In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, and others. 2024. [Aya expand: Combining research breakthroughs for a new multilingual frontier](#). *arXiv preprint arXiv:2412.04261*.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. [Universal Dependencies](#). *Computational Linguistics*, 47(2):255–308.
- Google DeepMind. 2026. Gemini 3.1 pro: Model card. <https://deepmind.google/models/model-cards/gemini-3-1-pro/>. Accessed: 2026-09-09.

- DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and others. 2026. [Deepseek-v4: Towards highly efficient million-token context intelligence](#). *Preprint*, arXiv:2606.19348.
- Emanuele Di Rosa. 2026. [Multi-agent orchestration for terminology-constrained machine translation in industrial localization](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*.
- 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*.
- Tobias Domhan and Dawei Zhu. 2025. [Same evaluation, more tokens: On the effect of input length for machine translation evaluation using large language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.
- Matthew S. Dryer and Martin Haspelmath, editors. 2013. [The World Atlas of Language Structures Online](#). Max Planck Institute for Evolutionary Anthropology, Leipzig.
- Thierry Etchegoyhen, Harritxu Gete, Joxean Zapirain, Victor Ruiz, and others. 2025. Machine translation in industrial domains: Resources and evaluations. *Procesamiento del Lenguaje Natural*, 75:117–135.
- Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. [Latxa: An open language model and evaluation suite for Basque](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. [COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ariavazhagan, and Wei Wang. 2022. [Language-agnostic BERT sentence embedding](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, and others. 2026. [Translatagemma technical report](#). *Preprint*, arXiv:2601.09012.
- Mikel L. Forcada and Francis M. Tyers. 2016. [Aperium: a free/open source platform for machine translation and basic language technology](#). In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*.
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. [Experts, errors, and context: A large-scale study of human evaluation for machine translation](#). *Transactions of the Association for Computational Linguistics*, 9:1460–1474.
- Miguel Angel Rios Gaona. 2026. [UniVie-HAITrans at WMT26 Terminology Translation Task: Terminology-Guided Data Filtering and Few-Shot Fine-Tuning for Lightweight LLMs](#). In *To appear in Proceedings of the Eleventh Conference on Machine Translation*.
- Ali Gebeşçe, Abdulfattah Safa, Ege Uğur Amasya, and Gözde Gül Şahin. 2026. [Overview of the SIGTURK 2026 shared task: Terminology-aware machine translation for English–Turkish scientific texts](#). In *Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026)*.
- Jurgi Giraud and Andrew Gargett. 2026. [Agenteak: A Multi-Agent Pipeline for Terminology-Constrained Spanish-Basque Translation](#). In *To appear in Proceedings of the Eleventh Conference on Machine Translation*.
- Koldo Gojenola and Maite Oronoz. 2000. [Corpus-based syntactic error detection using syntactic patterns](#). In *Proceedings of the ANLP-NAACL 2000 Student Research Workshop*.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xCOMET: Transparent machine translation evaluation through fine-grained error detection](#). *Transactions of the Association for Computational Linguistics*, 12:979–995.
- Jon Cambra Guinea and Andrea Alfieri. 2026. [RWS at WMT26 Terminology Translation Task](#). In *To appear in Proceedings of the Eleventh Conference on Machine Translation*.
- Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025. [Benchmarking linguistic diversity of large language models](#). *Transactions of the Association for Computational Linguistics*, 13:1507–1526.
- Akari Haga, Saku Sugawara, Akiyo Fukatsu, Miyu Oba, Hiroki Ouchi, Taro Watanabe, and Yohei Oseki. 2024. [Modeling overregularization in children with small language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*.
- Kazi Reyazul Hasan, A. B. M. Alim Al Islam, and Muhammad Abdullah Adnan. 2026. [BanglaSTEM: A parallel corpus and term-weighted evaluation for technical Bangla-English translation](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*.
- 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: Human Language Technologies, Volume 2 (Short Papers).
- 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)*.
- Boqi Huang, Daimeng Wei, Jiaxin GUO, Yuanchang Luo, Hengchao Shang, Zongyao Li, Jinlong Yang, Zhanglin Wu, Yu He, and Xiaoqing Lan. 2026. HW-TSC’s Submission to the WMT26 Terminology Translation Shared Task. In *To appear in Proceedings of the Eleventh Conference on Machine Translation*.
- 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*.
- Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. 2024. [Repeat after me: Transformers are better than state space models at copying](#). In *Proceedings of the 41st International Conference on Machine Learning*.
- Josef Jon, João Paulo Aires, Dusan Varis, and Ondřej Bojar. 2021. [End-to-end lexically constrained machine translation for morphologically rich languages](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
- Marcin Junczys-Dowmunt. 2025. [GEMBA v2: Ten judgments are better than one](#). In *Proceedings of the Tenth Conference on Machine Translation*.
- Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. [MetricX-24: The Google submission to the WMT 2024 metrics shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*.
- Maurice G. Kendall. 1938. [A new measure of rank correlation](#). *Biometrika*, 30:81–93.
- 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*.
- Tom Kocmi, Alexandre Berard, Phil Blunsom, Samuel Cahyawijaya, Shaun Rafael Cassini, Ona de Gibert, Aidan Gomez, Nithya Govindarajan, Shun Kiyono, Olivia Lasche, and others. 2026. Cohere CAT+: Submission to the General MT Shared Task. In *To appear in Proceedings of the Eleventh Conference on Machine Translation*.
- Tom Kocmi and Christian Federmann. 2023. [GEMBA-MQM: Detecting translation quality error spans with GPT-4](#). In *Proceedings of the Eighth Conference on Machine Translation*.
- Jan Kocoń, Maciej Piasecki, Arkadiusz Janz, Teddy Ferdinan, Łukasz Radliński, Bartłomiej Koptyra, Marcin Oleksy, Stanisław Woźniak, Paweł Walkowiak, Konrad Wojtasik, and others. 2025. [Pllum: A family of polish large language models](#). *Preprint*, arXiv:2511.03823.
- Xixian Liao and Maite Melero. 2026. SalamandraTA at WMT 2026 Terminology Shared Task: Hard Examples Are Better Teachers. In *To appear in Proceedings of the Eleventh Conference on Machine Translation*.
- Gonzalo López-Sánchez, Antoni Oliver, marta cozzini, Mercè Vázquez, Patricia Morales-Hurtado, and Sergi Alvarez-Vidal. 2026. GRIAL-TA participation in the WMT26 Terminology Shared Task. In *To appear in Proceedings of the Eleventh Conference on Machine Translation*.
- Jiaming Luo, Colin Cherry, and George Foster. 2024. [To diverge or not to diverge: A morphosyntactic perspective on machine translation vs human translation](#). *Transactions of the Association for Computational Linguistics*, 12:355–371.
- Ang Lv, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Rui Yan. 2025. [Language models “grok” to copy](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*.
- Guilherme Aren Marchetti, Gil Rocha, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2026. STaR-MT: Select, Translate, and Revise Pipeline for Terminology Aware Machine Translation. In *To appear in Proceedings of the Eleventh Conference on Machine Translation*.
- Arianna Masciolini, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfah, Ricardo Muñoz Sánchez, Elena Volodina, and Robert Östling. 2025. [The MultiGEC-2025 shared task on multilingual grammatical error correction in NLP4CALL](#). In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*.
- Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. [Evaluating machine translation output with automatic sentence segmentation](#). In *Proceedings of the Second International Workshop on Spoken Language Translation*.
- Orfeas Menis Mastromichalakis, Giorgos Filandrianos, Maria Symeonaki, and Giorgos Stamou. 2025. [Assumed identities: Quantifying gender bias in machine translation of gender-ambiguous occupational terms](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

- Shivam Ratnakant Mhaskar, Vrunda Nileshkumar Sukhadia, Anurag Deshmukh, Manan Sharma, and Pawan Kumar Rajpoot. 2026. Language-agnostic, prompt-optimized terminology translation: Genetic-pareto prompt evolution and an agentic repair loop for the wmt26 terminology shared task. In *To appear in Proceedings of the Eleventh Conference on Machine Translation*.
- 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*.
- Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. [Adaptive machine translation with large language models](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*.
- Dayyán O’Brien, Bhavitvya Malik, Ona de Gibert, Pinzhen Chen, Barry Haddow, and Jörg Tiedemann. 2025. [DocHPLT: A massively multilingual document-level translation dataset](#). In *Proceedings of the Tenth Conference on Machine Translation*.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.
- Arturo Oncevay, Zhijin Guo, Elena Kochkina, Damiano Fassina, Simerjot Kaur, Charese Smiley, Xiaomo Liu, and Manuela Veloso. 2026. Simple beats spend: Cost-performance trade-offs for multi-agent machine translation of financial long documents. In *Findings of the Association for Computational Linguistics: EMNLP 2026*.
- Arturo Oncevay, Elena Kochkina, Keshav Ramani, Toyin Aguda, Simerjot Kaur, and Charese Smiley. 2025a. [Translating domain-specific terminology in typologically-diverse languages: A study in tax and financial education](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.
- Arturo Oncevay, Charese Smiley, and Xiaomo Liu. 2025b. [The impact of domain-specific terminology on machine translation for finance in European languages](#). In *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)*.
- OpenAI. 2026. [Gpt-5.6 system card](#).
- Aleš Manuel Manuel Papáček and Tomáš Sourada. 2026. CUNI-ÚFAL at the WMT26 Terminology Translation Task. In *To appear in Proceedings of the Eleventh Conference on Machine Translation*.
- Marcis Pinnis, Martins Kronis, Emils Grims, Ervins Jakovelis, Roberts Rozis, and Toms Bergmanis. 2026. Tilde’s Submission for the WMT2026 Shared Task on Terminology Translation. In *To appear in Proceedings of the Eleventh Conference on Machine Translation*.
- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*.
- Matt Post and Hieu Hoang. 2025. [Effects of automatic alignment on speech translation metrics](#). In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*.
- 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)*.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
- Yuan Qi and Alim Murat. 2026. Structure-guided generation and verification: constructing high-quality parallel corpora for low-resource agglutinative languages. *J. King Saud Univ. - Comput. Inf. Sci.*, 38(5).
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, and others. 2022b. [CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*.
- D. Rummel and S. Ball. 2001. [The IATE project - towards a single terminology database for the European Union](#). In *Proceedings of Translating and the Computer 23*.
- Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. [CCMatrix: Mining billions of high-quality parallel sentences on the web](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

- 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*.
- Kirill Semenov and Rico Sennrich. 2025. [Measuring the effect of disfluency in multilingual knowledge probing benchmarks](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.
- 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*.
- Sherrie Shen, Weixuan Wang, and Alexandra Birch. 2025. [Liaozhai through the looking-glass: On paratextual explicitation of culture-bound terms in machine translation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.
- Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, and others. 2026. [Gemma 4 technical report](#). Preprint, arXiv:2607.02770.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, and others. 2022. [No language left behind: Scaling human-centered machine translation](#). Preprint, arXiv:2207.04672.
- Vania Velda, Yigyu Hwang, and Yongho Kim. 2026. STRA-MT: Deterministic Terminology Control for Document-Level Machine Translation. In *To appear in Proceedings of the Eleventh Conference on Machine Translation*.
- Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. [Document-level machine translation with large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
- Roger Waters, Nick Mason, David Gilmour, and Richard Wright. 1973. [Time](#).
- Aris Xanthos and Steven Gillis. 2010. Quantifying the development of inflectional diversity. *First Lang.*, 30(2):175–198.
- 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.
- Fan Zhang, Mei Tu, Mengchao Zhang, Jinting Wu, and Bowen Zhang. 2026. TaT at WMT26 Terminology Translation Shared Task. In *To appear in Proceedings of the Eleventh Conference on Machine Translation*.
- Jiawei Zheng, Hanghai Hong, Feiyan Liu, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation. *arXiv preprint arXiv:2402.15061*.
- Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. [Pitfalls and outlooks in using COMET](#). In *Proceedings of the Ninth Conference on Machine Translation*.
- Vilém Zouhar. 2023. [Machine translation that peeks at the reference](#). Report.

unit	over 510	longest
en-pl paragraphs (gold)	83 of 208	951
en-pl submission paragraphs	43%	1570
zh-en documents (gold)	53 of 53	3056
zh-en submission documents	100%	3430
es-eu paragraphs	5 of 5897	1015
en-pl gold segments	0 of 2317	129
es-eu gold segments	0 of 9231	328
zh-en sentence pairs	0 of 1417	471

Table 3: XLM-R tokens per unit against the 510-token cap. Segments fit everywhere; paragraphs and documents do not.

A Segment-Level Scoring in Detail

This appendix expands Section 4.4: the input budgets that force segment-level scoring, the length effect we measured, the policy for pieces that still do not fit, and a validation of the alignment against paragraph-level scoring.

Every encoder has a budget. The trained metrics truncate silently: the model scores whatever fits, and MetricX discards the reference first and then the candidate. A unit is within budget when each of its three sides stays under the XLM-R cap of 510 tokens, taking the longest of source, hypothesis and reference, and when the assembled MetricX template, source: . . . candidate: . . . reference: . . . , stays under 1536. Table 3 shows what the released units meet. Paragraph-level scoring is comfortable only for es-eu; it is impossible for en-pl and zh-en. The providers’ segments fit everywhere with room to spare.

The soft cap, measured. Beyond the hard cap there is a soft one, since XCOMET-XXL was trained on sentences and drifts long before 510 tokens. Scoring the en-pl gold reference against itself makes this concrete. At paragraph level, mean XCOMET falls monotonically with reference length: 0.80 below 200 tokens, 0.69 at 200–299, 0.41 at 300–399, 0.19 at 400–499 and 0.22 beyond. The same references scored as segments average 0.964 with a median of 0.990 over 2317 segments, with no length effect beyond a mild dip for the longest sentences. A pipeline in which the reference does not sit at its ceiling is wrong before any system is scored, which is the argument for scoring at segment level rather than a matter of convenience.

Aligning the hypothesis. Only one side lacks boundaries. Source and reference segments are

paired sentence by sentence in the providers’ data, whereas a submitted paragraph is a single string. We therefore re-segment the hypothesis onto the reference segmentation with mwerAlign (Matusov et al., 2005; Post and Hoang, 2025), which finds by dynamic programming the cut points minimising the word error rate against the reference segments; it is deterministic, runs on CPU, and takes seconds for the whole task. The alignment runs strictly inside one paragraph and never across a delimiter, so the paragraph contract established by the release stays intact and piece i pairs with reference segment i , and hence with source segment i .⁹

Units that do not fit. Every unit carries a status, and a scorer never calls a model on a unit that is not ok. An empty piece, an omission or an alignment slip, takes the worst score of the metric; both are charged rather than excused. A piece still over the cap after alignment also takes the worst score, with the reason stored on the unit, because truncating it would reward padding. A piece that is long but under the cap is scored normally, since the metric already penalises padding by itself, and no piece is split further: below the sentence there is no gold fragment to align a sub-sentence cut to. These cases are rare. Over every submitted file and mode, 28 units exceed the cap (13 in es-eu, 15 in en-pl, none in zh-en), and every one is an output three to seven times the length of its reference; 2726 units are empty, out of roughly 530k in total.

Validating the alignment. Because es-eu paragraphs are only one or two sentences long, paragraph-level COMET-22 is a sound reference point there, which lets us compare the two granularities where both are trustworthy. Across 176 files scored at both levels, file means correlate at Pearson $r = 0.989$; segment-level scores sit 1.2 points lower on average, uniformly, as one-word title segments and boundary slips each cost a little. The ten largest gaps all belong to one weak system family, whose outputs align less cleanly, and system ranks are preserved.

⁹mwerAlign is indifferent about which of several non-matching words a cut lands on, and has no preference for sentence-final punctuation. This is immaterial to the trained metrics, which grade such a boundary by a small margin. Snapping cuts to sentence punctuation is the natural refinement.

System Name	Track1	Track2
TARL	1406546	1226360*
Cohere-CAT+	1141509	739788
SalamandraTA	1268804	—
COZY	5390576	4619400
COZY _{flash}	1302994	1064397
TQM	756978	633106
HW-TSC	1348542	1252334
TO-TM _{top5}	—	20048
TO-TM _{top15}	—	17280
TO-G4terms	—	63435
TO-fuzzy	90319	—
TO-fuzzy _{stem}	87078	—
STRA-MT	1122966	406796*
GRIAL-TA	63470*	7494
AGENTEAK	607251*	—
CUNI-UFAL _{dict-LLM}	—	1518986
CUNI-UFAL _{E2E}	—	1010062*
CUNI-UFAL	1371870	10571034*
UniOR-ZSL+Term	1037314	—
UniOR-GenPE	941874	—
UniOR-RulePE	960528	—
UniOR-Mixed	803832	—
TaT	1354118*	1253770*
STaR-MT	30780	351178
VNFusion	1286217	963835*

Table 4: Paired one-sided Wilcoxon signed-rank test for relative position shifts, calculated for every system and every track separately. “—” values show that the system was not submitting for a particular track, asterisks show the statistics for which p-value is smaller than 0.05.

B Grammaticality Estimation

B.1 Relative Term Position Distribution

The aggregated distributions of the relative term positions are shown in the Figure 6. We can see that the differences between the reference and the hypothesis distributions are not substantial.

A more aligned comparison can be made by calculating the difference between the relative positions in the aligned pairs of sentences. We measured the shifts in relative positions within each sentence, and computed the paired Wilcoxon signed-rank test (one-sided since the visualization was showing that the systems tend to locate the terms later, i.e., only to the right). The results are shown in Table 4: most values are statistically insignificant.

B.2 Mean Size of Paradigm and Its Components

The mean size of paradigm is calculated by the formula:

$$MSP = \frac{\#inflected_{unique}}{\#lemmas_{unique}}, \text{ where } \#inflected_{unique} \text{ stands for the number of}$$

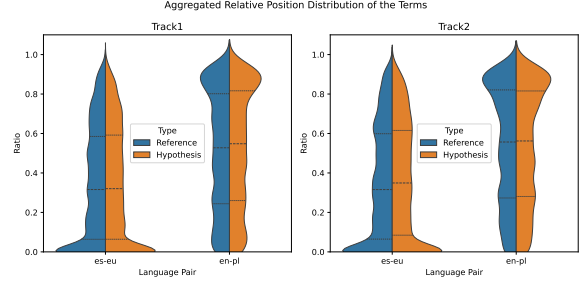


Figure 6: Aggregated relative term position distributions for two tracks and two target languages. Each violin plot shows the distribution in the reference texts (on the left) and in the system outputs (“hypothesis”, on the right).

the unique inflected forms observed in the corpus, and $\#lemmas_{unique}$ is the number of the unique lemmas in the corpus. Higher MSP demonstrates a larger variety of the inflected forms, therefore, we generally expect the references to have higher MSP compared to the systems, and better systems to have higher MSP compared to the worse ones. For our study, we concentrate only on the term inflection; therefore we only analyze the expected dictionary forms of the target terms and their real occurrences; on the other hand, we include into the sample the multi-word terms.

The MSP ratio consists of the two values, which we can plot in two dimensions: by putting $\#lemmas_{unique}$ over x-axis, and $\#inflected_{unique}$ over y-axis. This is what we do in Figure 8. First, we see that our assumptions are supported by data: worse systems, i.e., the ones with a lower $\#lemmas_{unique}$, have a larger discrepancy between the reference and the hypothesis $\#inflected_{unique}$ values; references in Polish have a higher $\#inflected_{unique}$ than their corresponding hypothesis values. On the other hand, we see a notable difference in the best-performing systems in Basque: the better the performance, the bigger the $\#lemmas_{unique}$, the higher $\#inflected_{unique}$ in the system hypotheses compared to the references.

The qualitative analysis showed that many compound nouns in the reference texts were incorrectly spelled with an additional white space, while the system outputs contained orthographically correct compounds. There were also opposite cases, where the system over-tokenized the terms: although the correct suffix was used, it was separated by an excessive white space. We show the examples of it in Table 5.

Reference	System output	Error in
itsasontsi- gupil	itsasontsi-gupil	Reference
potenziat- fluxua	potenziat-fluxua	Reference
Balazta-sistema k	Balazta-sistemak	MT output
kondentsadore aren	kondentsadorearen	MT output

Table 5: Examples of incorrect spacing in the references and system outputs.

B.3 Distribution of Number Grammmemes

Figure 7 shows the difference in distribution of the number grammemes between the references and system outputs.

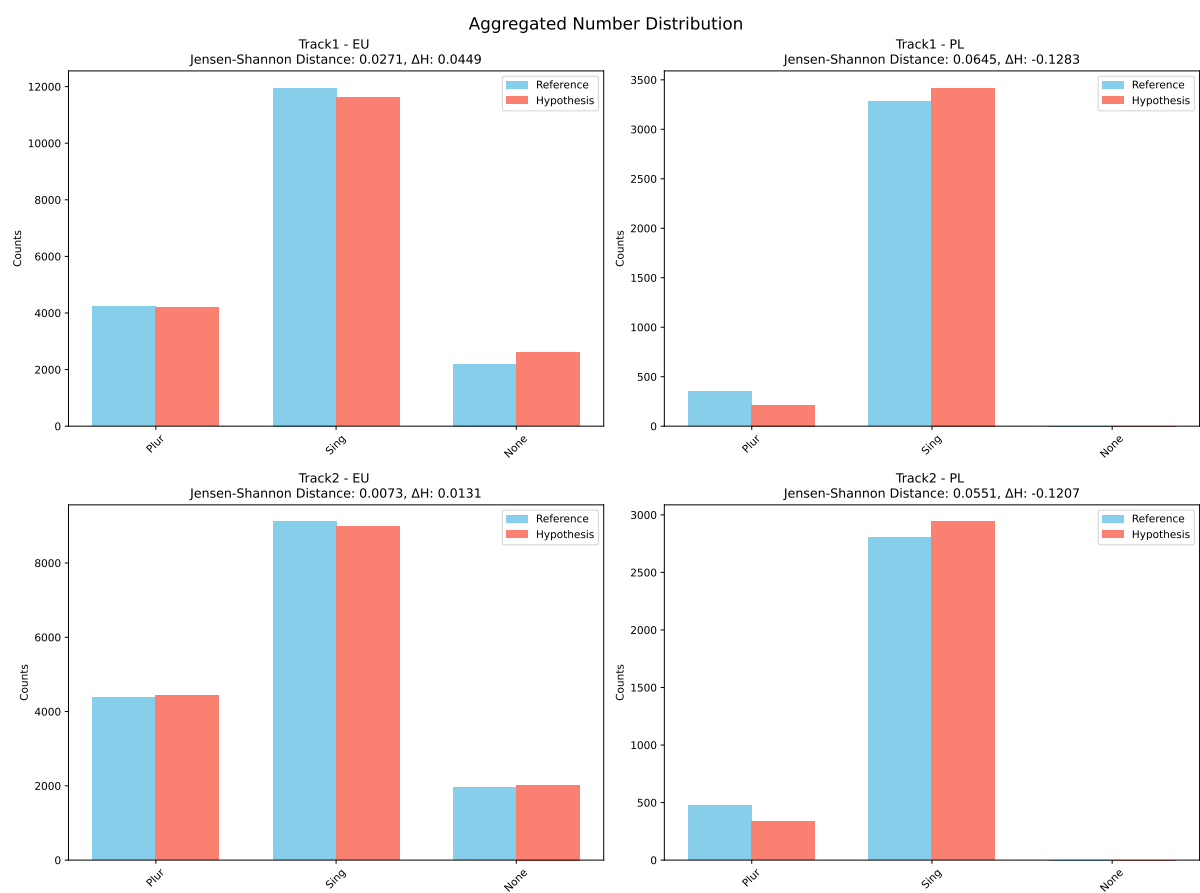


Figure 7: Difference in distributions of the number grammemes between the references and system outputs. The legend is analogous to the one in Figure 5. The “None” value stands for the cases where the word was not assigned a number category at all.

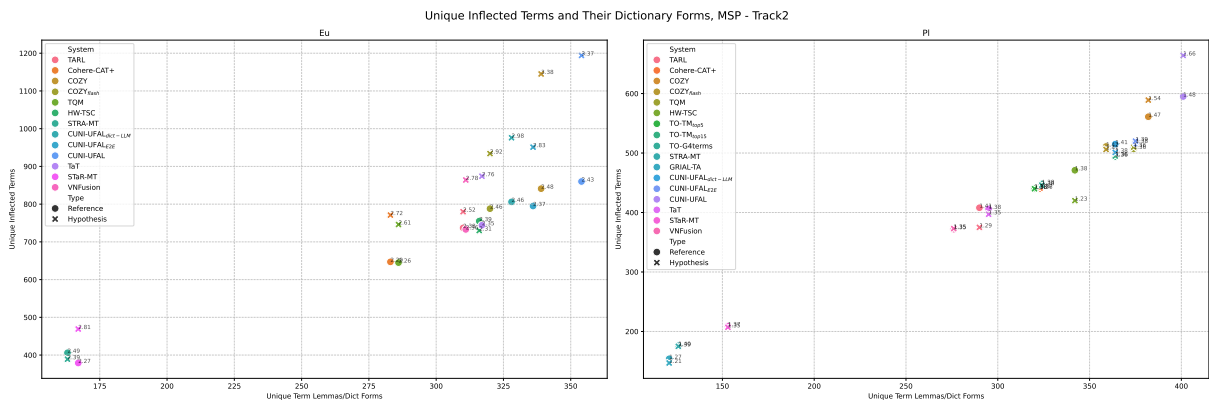
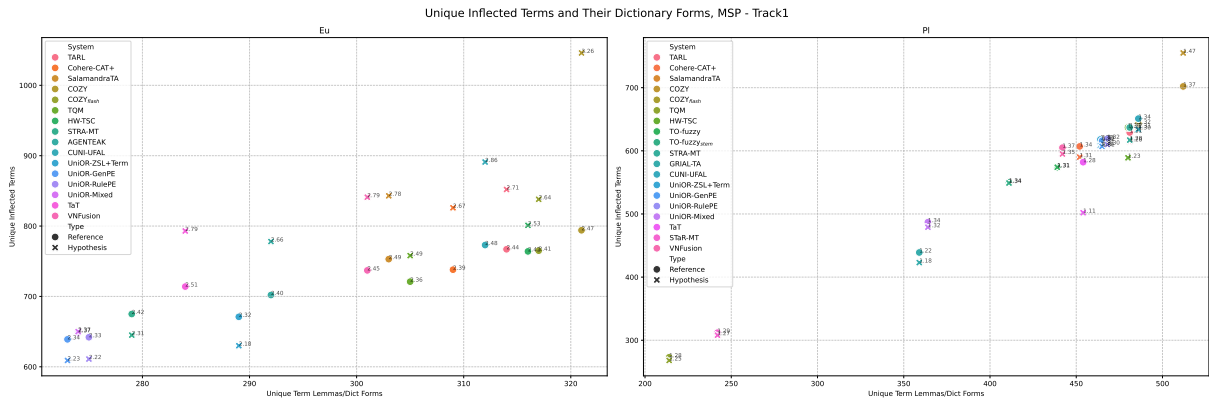


Figure 8: Distribution of the unique term lemmas (or dictionary forms for multi-word expressions, $\#lemmas_{unique}$), x-axis, and the unique inflected forms ($\#inflected_{unique}$), y-axis. The wide spread over the x-axis is explained by the fact that we only analyze the occurrences of the successful term retrieval for a particular system, so some systems could retrieve more unique terms. Generally, we expect a better system to have higher values over both axes. The color of the points reflects the MT system, and the shape shows whether the calculations were made on the system outputs (“hypothesis”) or on the aligned reference sentences. Next to each data point, we show the *MSP* values (higher is better).

System	chrF++			Term success (lemma)			COMET-22			CometKiwi-22			XCOMET-XXL			MetricX-24 XL ↓			MetricX-24 XL QE ↓		
	Avg			es-eu			Avg			Avg			Avg			Avg			Avg		
	en-pl	es-eu	100.0	en-pl	Avg	93.6	en-pl	es-eu	96.4	en-pl	es-eu	96.3	en-pl	es-eu	96.5	en-pl	es-eu	96.5	en-pl	es-eu	96.5
gold	100.0	100.0	100.0	96.4	93.6	93.6	96.3	96.5	96.5	78.5	81.5	75.4	93.2	96.5	96.5	89.8	2.04	2.26	1.82	3.45	3.47
TARL	75.3	77.1	73.5	95.7	95.4	95.4	88.3	89.0	87.6	77.2	79.2	75.1	83.6	85.8	81.4	3.92	4.31	3.53	4.16	4.49	3.84
COZY <i>flash</i>	74.4	<u>76.5</u>	72.2	95.4	94.4	94.4	88.0	88.6	87.4	76.9	79.3	74.4	83.0	85.7	80.4	4.04	4.33	3.76	4.33	4.54	4.12
HW-TSC	73.8	75.6	72.1	95.3	95.8	95.8	87.4	88.0	86.7	76.2	78.6	73.8	82.2	85.1	79.3	4.37	4.75	3.98	4.63	4.88	4.39
COZY	75.6	75.9	75.3	95.0	93.6	93.6	89.2	89.6	88.8	77.3	79.7	75.0	85.5	87.9	83.2	3.58	3.93	3.24	3.95	4.13	3.77
TaT	73.7	73.6	73.8	94.5	97.2	97.2	91.7	86.8	86.3	76.0	77.1	74.9	80.7	80.2	81.2	4.43	5.19	3.68	4.66	5.26	4.06
SalamandraTA	74.6	75.9	73.3	94.2	93.2	93.2	95.2	87.8	88.6	77.0	79.4	74.6	83.0	85.7	80.2	4.05	4.45	3.66	4.30	4.58	4.01
CUNI-UFAL	74.4	74.2	74.5	93.4	91.6	91.6	89.2	89.6	88.7	77.3	79.5	75.1	85.9	88.4	83.5	3.55	3.93	3.17	3.88	4.08	3.68
TQM	70.0	73.8	66.1	89.6	89.8	89.8	89.4	86.5	89.5	83.4	76.0	80.0	71.9	80.4	88.4	72.3	4.50	3.85	5.15	4.49	4.00
STRA-MT	69.3	72.9	65.8	89.0	91.7	91.7	86.3	84.1	87.0	81.3	74.1	78.3	69.8	74.8	82.4	67.3	5.62	5.10	6.14	5.44	5.03
AGENTEAK	-	-	68.8	-	-	-	-	-	-	84.6	72.5	-	72.5	-	-	75.7	-	-	4.71	4.67	-
Cohere-CAT+	71.1	74.9	67.3	88.0	84.9	84.9	91.1	87.0	89.2	84.9	76.3	79.6	72.9	80.7	87.2	74.1	4.45	4.14	4.75	4.46	4.21
VNFusion	73.5	72.8	74.1	87.6	84.6	84.6	90.7	88.0	88.9	87.1	77.6	79.3	75.8	84.0	86.7	81.3	3.78	4.11	3.45	3.81	4.09
MASC-Dec	-	-	72.7	-	-	-	-	-	-	84.5	72.2	-	72.2	-	-	75.2	-	-	4.61	4.71	-
UniOR-ZSL+Term	68.8	73.4	64.1	86.4	87.9	87.9	84.9	83.9	87.9	79.9	74.3	79.1	69.6	75.8	84.0	67.6	5.60	4.72	6.48	5.44	4.75
UniOR-RulePE	68.9	73.5	64.2	84.7	88.4	88.4	81.0	82.6	87.9	77.3	73.2	79.1	67.2	74.2	84.1	64.3	5.95	4.69	7.21	5.82	4.74
UniOR-GenPE	68.8	73.4	64.2	84.1	87.7	87.7	80.5	82.5	88.0	77.0	73.0	79.2	66.9	73.9	84.2	63.7	6.00	4.66	7.33	5.85	4.70
TO-fuzzy	-	75.8	-	-	80.8	80.8	-	-	88.1	-	79.6	79.6	-	-	84.6	-	-	4.54	-	4.59	-
TO-fuzzy <i>stem</i>	-	75.6	-	-	78.0	78.0	-	-	88.0	-	79.6	79.6	-	-	84.4	-	-	4.58	-	4.60	-
UniOR-Mixed	70.4	74.5	66.3	76.4	70.8	70.8	81.9	76.7	74.6	78.8	67.0	65.8	68.3	65.8	64.5	67.0	7.79	8.97	6.61	7.42	8.42
GRIAL-TA	-	71.0	-	-	57.3	57.3	-	-	76.9	-	68.2	68.2	-	-	64.9	-	-	8.39	-	7.62	7.62
STaR-MT	-	68.6	-	-	44.9	44.9	-	-	85.2	-	<u>78.5</u>	78.5	-	-	78.5	-	-	5.71	-	5.25	5.25

Table 6: Full results for Track 1: document-level chrF++, lemmatized exclusive term success (%) and the trained metrics on aligned segments, averaged over the directions a system entered and per direction. Reference-free metrics are given beside their reference-based counterparts. ↓ marks an error score, for which lower is better. Best in bold, runner-up underlined; gold is the reference scored against itself.

System	Proper				Random				Noterm							
	Term	chrF++	COMET	MetricX	MetricX QE	Term	chrF++	COMET	MetricX	MetricX QE	Term	chrF++	COMET	MetricX	MetricX QE	
gold	96.4	100.0	96.4	2.04	3.45	—	—	—	—	—	—	—	—	—	—	—
TARL	95.7	75.3	88.3	3.92	4.16	66.8	72.4	86.5	4.66	4.60	68.6	74.0	88.1	3.87	3.93	3.93
COZY _{flash}	95.4	74.4	88.0	4.04	4.33	65.0	71.5	86.6	4.57	4.56	65.9	72.6	87.3	4.26	4.32	4.32
HW-TSC	95.3	73.8	87.4	4.37	4.63	64.0	69.2	83.8	5.75	5.56	64.7	72.5	87.9	4.06	4.15	4.15
COZY	95.0	75.6	89.2	3.58	3.95	69.5	73.5	87.9	4.12	4.23	73.0	74.2	88.8	3.67	3.86	3.86
TaT	94.5	73.7	86.8	4.43	4.66	64.4	68.6	82.3	6.30	6.12	71.2	75.2	88.4	3.81	4.00	4.00
SalamandraTA	94.2	74.6	87.8	4.05	4.30	64.1	72.5	86.8	4.37	4.33	65.8	72.8	87.2	4.19	4.16	4.16
CUNI-UFAL	93.4	74.4	89.2	3.55	3.88	69.3	73.8	88.6	3.78	3.96	71.3	73.1	88.6	3.72	3.88	3.88
TQM	89.6	70.0	86.5	4.50	4.49	50.1	67.1	84.5	5.43	5.02	52.3	67.6	85.5	4.81	4.49	4.49
STRA-MT	89.0	69.3	84.1	5.62	5.44	45.0	64.6	80.0	7.34	6.65	47.0	66.3	82.3	6.22	5.68	5.68
AGENTEAK	88.7	68.8	84.6	4.71	4.67	75.2	68.3	84.1	4.99	4.87	75.7	68.3	84.0	4.94	4.81	4.81
Cohere-CAT+	88.0	71.1	87.0	4.45	4.46	55.9	69.8	86.4	4.61	4.45	55.5	69.0	86.1	4.67	4.42	4.42
VNFusion	87.6	73.5	88.0	3.78	3.81	61.0	69.9	85.4	4.93	4.72	65.4	73.0	87.7	3.94	3.85	3.85
MASC-Dec	86.8	72.7	84.5	4.61	4.71	76.0	72.1	84.4	4.70	4.81	77.4	72.5	84.3	4.68	4.78	4.78
UniOR-ZSL+Term	86.4	68.8	83.9	5.60	5.44	45.5	65.4	81.5	6.33	5.86	48.7	66.0	83.0	5.92	5.46	5.46
UniOR-RulePE	84.7	68.9	82.6	5.95	5.82	45.0	65.8	80.5	6.63	6.19	47.4	66.4	81.2	6.46	6.04	6.04
UniOR-GenPE	84.1	68.8	82.5	6.00	5.85	44.5	66.0	80.2	6.75	6.29	47.1	66.4	81.4	6.43	6.00	6.00
TO-fuzzy	80.8	75.8	88.1	4.54	4.59	51.7	73.5	86.9	4.89	4.64	52.1	73.2	86.7	4.96	4.68	4.68
TO-fuzzy ^{stem}	78.0	75.6	88.0	4.58	4.60	51.3	73.4	87.0	4.89	4.64	52.1	73.2	86.7	4.96	4.68	4.68
UniOR-Mixed	76.4	70.4	76.7	7.79	7.42	41.6	67.7	75.0	8.31	7.75	47.1	67.5	78.3	7.34	6.86	6.86
GRIAL-TA	57.3	71.0	76.9	8.39	7.62	40.0	69.3	75.3	9.00	8.09	38.8	69.2	75.6	9.03	7.94	7.94
STaR-MT	44.9	68.6	85.2	5.71	5.25	39.9	64.7	79.7	7.14	6.56	44.6	68.9	84.9	5.67	5.19	5.19

Table 8: Track 1: lemmatized term success (%), document chrF++, COMET-22, MetricX-24 XL and MetricX-24 XL QE (both lower is better) under the released dictionary, a random dictionary and no dictionary. Random and noterm are scored against the proper dictionary.

System	Sample					Noterm				
	Term	chrF++	COMET	MetricX	MetricX QE	Term	chrF++	COMET	MetricX	MetricX QE
gold	93.4	100.0	95.2	1.88	3.36	—	—	—	—	—
CUNI-UFAL	89.2	78.1	88.7	3.22	3.64	<u>73.7</u>	73.4	87.5	<u>3.50</u>	3.60
CUNI-UFAL _{dict-LLM}	89.0	74.9	88.5	3.25	3.51	<u>73.7</u>	73.4	87.5	<u>3.50</u>	3.60
COZY	88.9	77.5	88.9	3.18	<u>3.51</u>	74.7	74.0	87.6	3.44	3.55
COZY _{flash}	87.2	76.8	88.4	3.37	3.66	65.8	72.0	86.3	3.84	3.76
CUNI-UFAL _{E2E}	87.1	78.1	88.9	3.14	3.54	73.7	73.4	87.5	3.50	3.60
HW-TSC	85.1	74.0	86.5	4.10	4.23	63.4	71.8	86.5	<u>3.74</u>	3.69
TaT	84.0	75.1	87.6	3.58	3.75	72.5	74.8	87.7	<u>3.50</u>	3.62
TARL	80.4	73.7	85.8	4.15	4.16	68.7	73.3	86.4	3.80	3.76
TO-G4terms	79.8	75.4	88.5	4.33	4.34	51.1	71.4	86.8	4.88	4.44
TO-TM _{top15}	78.7	80.1	89.7	4.00	4.35	51.1	71.4	86.8	4.88	4.44
Cohere-CAT+	75.2	73.4	86.8	3.78	3.75	56.2	69.7	85.3	4.32	3.94
TQM	74.2	73.0	86.2	4.16	4.06	52.2	67.8	83.4	5.10	4.48
TO-TM _{top5}	73.7	<u>78.6</u>	89.9	4.01	4.18	51.1	71.4	86.8	4.88	4.44
VNFusion	73.4	73.8	87.0	3.61	3.47	63.3	71.8	86.2	3.79	3.51
UniVie-HAITrans	57.8	70.3	82.1	6.59	6.13	40.4	64.9	79.1	7.65	6.78
STaR-MT	53.1	67.3	83.9	4.87	4.33	51.8	65.8	82.4	5.36	4.77
STRA-MT	48.0	65.9	80.9	6.19	5.35	48.8	65.9	80.9	6.17	5.34
GRIAL-TA	32.1	62.2	64.4	13.12	11.92	43.3	68.1	76.6	8.28	7.51

Table 9: Track 2: lemmatized term success (%), document chrF++, COMET-22, MetricX-24 XL and MetricX-24 XL QE (both lower is better) under with and without the sample bitexts. Random and noterm are scored against the proper dictionary.