Automated Evaluation for Terminology Translation Related to the EEA Agreement

Selma Dís Hauksdóttir
The University of Iceland
Reykjavík, Iceland
sdh22@hi.is

Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies
Reykjavík, Iceland
steinthor.steingrimsson@arnastofnun.is

Abstract

This paper presents a submission to the WMT25 test suite subtask, focusing on terminology in official documents related to the EEA Agreement. The test suite evaluates the accuracy of MT systems in translating terminology for the English→Icelandic translation direction when applied to EEA documents. We focus on the use of terminology in four domains of the agreement; science, technology, economics, and society. We find that manual evaluation confirms that our test suite can be helpful in selecting the best MT system for working with these domains. Surprisingly, an online system which does not achieve very high scores on the general translation task, according to preliminary results, is most adept at translating the terminology our test suite evaluates. The test suite and evaluation code are openly available on Github: https://github.com/steinst/WMT25_EEA_test_suite

1 Introduction

Pozzo (2020) argues that the law of the European Community has a multilingual framework, with 24 official languages, which calls for the use of descriptive language to maintain equal distance between each language, with legal terminology being culture-bound. The EEA Agreement is translated into two additional languages; Icelandic and Norwegian. The meaning associated with legal texts is often disputed, even in monolingual texts. The target text should perform at the same function as the source text, since they're equivalent in legal sense (Bago et al., 2022). Term inconsistencies are therefore unacceptable. Until recently, most MT systems translated documents sentence by sentence, which could result in inconsistencies in translation of terminology (see e.g. Semenov and Bojar (2022)). The Translation Centre at the Ministry for Foreign Affairs in Iceland translates around 9000 pages related to the EEA Agreement each year (Steindórsdóttir, 2022). They have been testing an MT system trained on their own corpus but the results show that the system outputs are mediocre at best. Bago et al. (2022) state that one of the main complaints of the translators is the lack of correct terminology and consistency in the MT output.

Current LLM-based systems are capable of considering larger context. In this paper we present a test suite which can help us understand if systems based on this new MT paradigm can translate the terminology correctly. We focus on terminology translations from English to Icelandic in the EEA Agreement in our submission for the WMT25 Test Suite subtask (Kocmi et al., 2025a). We evaluate 34 systems, both automatically and manually, and find that surprisingly, two out of the three highest scoring systems are Online-systems. We release our test suite and evaluation code for others to build on and to allow for further comparison between future models in this domain.

2 Related Work

Semenov and Bojar (2022) measured consistency, unambiguity and adequacy in automatically translated legal texts by counting the correct occurrences of the exact term. Gašpar et al. (2022) found that the Herfindahl-Hirshman Index (HHI) can be used to measure the consistency of terminology in a translated corpus, since the HHI works on a small amount of data.

$$\mathrm{HHI} = \sum_{i=1}^{n} S_i^2 \tag{1}$$

i ranges over n different translations for the specific term translated in the relevant text. S_i is the ratio of the number of times when the term was translated as i to the total number of times it was translated. Therefore the HHI score will be 5.0 if a term has two different translations.

$$\frac{\sum_{j=1}^{p} \sum_{i=1}^{n} \left(\frac{f_i}{k_j} \cdot 100 \right)^2}{p} \tag{2}$$

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 850–855
November 8–9, 2025.

An overall translation consistency index (C_t) for a source term is calculated as follows: p is the number of translation having the source term t, and each frequency share is calculated as the ratio of its frequency f_i to the total translation occurrence within a product k_j. The score ranges between 0 and 10, with 10 being perfect consistency.

Alam et al. (2021) introduced a benchmark for evaluation of quality and consistency of terminology translation in a shared task at WMT21, with creating new evaluation datasets that were annotated by professional translators for their terminology consistency. They found that general translation quality does not have to be sacrificed for terminology compliance. Semenov et al. (2023) evaluated the efficiency of using segment-level terminology dictionaries in a shared task at WMT23, and concluded that an improvement in MT performance when using a terminology dictionary ranged between 0 and 10 ChrF points.

Two participants in the WMT 2024 Test Suite subtask (Kocmi et al., 2024) focused specifically on English→Icelandic translations: Friðriksdóttir (2024) focuses on various aspects of gender-inclusive translation, including LGBTQIA+ terminology and whether translations are current and culturally appropriate, as the terminology in that domain has been updated repeatedly. Ármannsson et al. (2024) focus on idiomatic expressions and proper names in their test suite. Their evaluation is keyword-based, checking if content words in the idioms are translated correctly and whether the proper names have correct translations.

3 Methodology

We built an automated keyword-based evaluation regarding the EEA Agreement. The aim was to test the ability of MT systems when it comes to the translation of terms in the Agreement, as disambiguation is important when it comes to the translation of legal texts. To test this, we ran the automated evaluation and confirmed our method with manual evaluation. We manually collect sentences from EU regulations and directives relevant to the EEA agreement. For each sentence, we tag terms and find the standard Icelandic translation for each term on the official Icelandic website for the EEA Agreement.[1] The Icelandic translations are used for automatic evaluation and a subset of the translations from each MT system is manually evaluated in order to understand whether the automatic evaluation is close to human judgment.

3.1 Test Suite Compilation

The terms were manually extracted from 32 EU Regulations and two EU Directives that were translated into Icelandic and published in 2024 and 2025. The aim was not to test the regulations, but to collect a diverse and descriptive sample of keywords that appear in the EEA Agreement. In some cases, we added a simple verb phrase if necessary, to build a coherent sentence. The terms were divided into four subgroups: science, technology, economics, and society, with as little overlay as possible. The subgroups are based on the groups at the Translation Centre, where the EEA Agreement is divided into said subgroups. Every sentence contained at least one term that we tested, but we did not test every term in each sentence, especially not recurring terms, since we were not testing consistency in this test suite. We gathered every word form of the terms and used it for the automatic evaluation.

The sentences were exported as a txt-file, which was sent to the test suite subtask of WMT25. Once we got the translated sentences from the 34 MT systems, we ran the automatic evaluation, see 3.2. We manually evaluated translation of the terms in 50 sentences for each system to test our automatic evaluation method, see 3.3.

Due to an error in the layout of the txt-file, some of our sentences were split, so we ended up with 256 sentences and 408 keywords. We disregard the erroneous sentences in the input and report only on the error-free ones. We have however published a corrected version on Github along with the test suite and evaluation codes, for others to build on and compare other models.

3.2 Automatic Evaluation

The automatic evaluation is keyword-based. For each MT system we check all 256 complete sentences and disregard the ones that were split up before submission. The check inspects if a given translation contains the Icelandic terms, by comparing the translation to all possible inflectional forms of the term. We look up the word forms in DIM, the Database of Icelandic Morphology (Bjarnadóttir et al., 2019), and if they are not found there, we manually create a list of acceptable forms. If the translation contains the term in any form accepted in our lists we count that as correct.

[1] https://gagnagrunnur.ees.is/

3.3 Manual Evaluation

We manually evaluated 1700 translations, 50 for each system. The sentences were chosen randomly from the complete set of 256 translations and for all systems we evaluated translations of the same 50 sentences. The evaluator is a PhD student in Translation Studies and a former translator at the Translation Centre, with a three year background as a professional translator. The manual evaluation took around 10 hours, since the focus was only on the keywords and not the sentences. One point was given for every term that was correctly translated, and the maximum points available correlated therefore with the number of terms. A point was given for acceptable translations, other than the ones that were included in our keyword list. 117 other translations were accepted, mainly synonyms, and rephrasing of terms consisting of more than one word. Additionally, we inspect the ratio of sentences that have all terms correctly translated.

4 Results

Results of the automatic evaluation are presented in Table 1 and manual evaluation in Table 2. While the main difference between the evaluation approaches is that the manual evaluation paints a picture where many of the MT systems seem to be quite adept at dealing with EEA terminology, achieving up to almost 80% accuracy, the accuracy being the ratio of terms correctly translated according to the human annotator. The automatic evaluation gives substantially lower scores, which may indicate that the keyword and word inflection lists used for the automatic evaluation are sometimes lacking. Even though that is the case, the order of the systems is very similar in the two evaluations, manual and automatic. Our main takeaway from comparison is thus that if we trust our manual evaluation to be reliable and can use that to help us select the best MT system to help us with EEA translations, we can also trust the automatic evaluation using our test suite, as the order of the system is almost identical, with the same systems being in the top 3 seats of both lists. If we compare our results to the preliminary rankings for the WMT25 general translation task (Kocmi et al., 2025b), we find that our order of systems is quite different. The most surprising results are that the system achieving first place on both our lists, ONLINE-G, is actually a low scoring system in the general translation task, ending up in 24th place out of 33 systems. We wonder why this is and speculate whether this might be an encoder-decoder system that actually contains EEA texts in their training data. If an evaluated system is trained on data from the domain being evaluated, possibly containing the same or very similar structures as being evaluated in the test suite, this data leakage can lead to overestimation of the models capabilities, see e.g. Zhu et al. (2024) and Zeng et al. (2024). This could explain why the system is particularly good in this task, but not in the general translation task where LLMs seem to have an advantage. This is not necessarily a far-fetched idea, as a substantial part of ParIce (Barkarson and Steingrímsson, 2019; Steingrímsson and Barkarson, 2021), a parallel English-Icelandic parallel corpus, comprises data from EEA-documents and this corpus is among those distributed on OPUS[2]. Other top scoring systems, on the other hand, are all in the top seats in the preliminary system ranking table.

5 Conclusions and Future Work

We evaluated 34 MT systems using our test suite, automatically and manually. The evaluation shows that while a few of the systems translate the majority of the terms correctly, they are all quite far from perfect. Our automatic evaluation orders the systems in a similar way to the manual evaluation, indicating that an automatic approach such as this one can be useful to help translators find the most useful system for the task.

A larger set of sentences and terminology would improve our test suite, especially if we include terminology from other subdomains. Given the large amount of published documents relating to the EEA Agreement it is almost, if not entirely, impossible to test for every single term that appears in those documents. We plan to look into the frequency of terms in order to reconstruct the test suite in a way that may be more indicative of real-world usage, on the one hand giving terms that appear often in the Agreement more weight, but on the other make sure that a representative part of terms that rarely appear is also included. To build further on the test suite, we also plan to look into results for each subdomain and see if MT systems perform better for a specific domain. Another interesting area is to test the translation of neologism, as acts about new developments, especially scientific and technological ones, often call for new terminology. Finally, looking into the translation of terms that have more

[2] https://opus.nlpl.eu/

System	Term Acc. (%)	Sentence Acc. (%)
ONLINE-G	55.9	42.2
Erlendur	53.3	41.0
Gemini-2.5-Pro	46.5	30.5
ONLINE-B	46.5	33.2
TranssionTranslate	46.2	32.8
hybrid	38.7	26.6
Claude-4	38.5	27.7
SalamandraTA	38.5	25.0
TowerPlus-9B	37.5	23.8
Shy	36.8	22.7
GPT-4.1	34.9	23.8
TowerPlus-72B	32.7	20.7
DeepSeek-V3	30.0	17.2
AMI	27.8	17.2
Llama-4-Maverick	27.8	17.6
NLLB	26.6	14.8
CommandA-MT	24.9	14.8
IR-MultiagentMT	24.2	14.5
Gemma-3-27B	20.1	8.6
Mistral-Medium	18.9	9.4
GemTrans	16.5	8.2
IRB-MT	13.8	6.3
UvA-MT	13.1	5.9
Gemma-3-12B	11.6	4.3
Qwen3-235B	10.9	3.9
CommandA	7.5	2.0
Llama-3.1-8B	3.1	0.8
AyaExpanse-32B	2.7	0.8
Qwen2.5-7B	0.7	0.0
CommandR7B	0.5	0.0
EuroLLM-9B	0.5	0.4
EuroLLM-22B	0.2	0.0
Mistral-7B	0.2	0.0
AyaExpanse-8B	0.0	0.0

Table 1: Automatic evaluation of the systems.

System	Term Acc. (%)	Sentence Acc. (%)
ONLINE-G	79.6	64
Erlendur	76.3	64
Gemini-2.5-Pro	72	62
TranssionTranslate	71	60
ONLINE-B	67.7	54
Claude-4	60.2	46
Shy	60.2	44
hybrid	59.1	48
TowerPlus-9B	59.1	44
GPT-4.1	57	44
SalamandraTA	55.9	42
IR-MultiagentMT	44.1	30
TowerPlus-72B	44.1	26
NLLB	43	26
DeepSeek-V3	40.9	20
AMI	37.6	26
CommandA-MT	37.6	22
Llama-4-Maverick	35.5	18
Gemma-3-27B	31.2	14
Mistral-Medium	30.1	10
GemTrans	26.9	16
Gemma-3-12B	25.8	16
UvA-MT	25.8	14
IRB-MT	24.7	14
Qwen3-235B	16.1	4
CommandA	12.9	2
Llama-3.1-8B	5.4	4
AyaExpanse-32B	2.2	0
AyaExpanse-8B	1.1	0
CommandR7B	1.1	0
EuroLLM-22B	1.1	0
EuroLLM-9B	0	0
Mistral-7B	0	0
Qwen2.5-7B	0	0

Table 2: Manual evaluation of the systems.

than one allowed Icelandic translation, based on context and subgroups, could help us understand problems that translators might miss and special attention has to be paid to.

This test suite can be adapted to other languages with relative ease, which allows further work on other language directions.

6 Limitations

This work did not consider consistency especially, which would be a logical next step, to check whether the terms are consistently translated in the same way, or whether for some MT systems correct translations may be fortuitous incidents.

Our selection of terms was not always systematic and we do not always consider all terms in a given sentence. We were not able to check every category under each subgroup in this test suite due to the size limitations.

Due to time limits we were not able to add the accepted translations, and all the word forms of said translations, from the manual evaluation to the list of accepted terms. The keyword and word inflection lists are therefore lacking for the automatic

evaluation.

As mentioned above, some sentences were split up and we ended therefore with fewer sentences than anticipated, and therefore a smaller test suite. Both the submitted test suite and a fixed one are available in the Github repository.

References

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

Bjarki Ármannsson, Hinrik Hafsteinsson, Atli Jasonarson, and Steinþór Steingrímsson. 2024. Killing Two Flies with One Stone: An Attempt to Break LLMs Using English-Icelandic Idioms and Proper Names. In *Proceedings of the Ninth Conference on Machine Translation*, pages 451–458, Miami, Florida, USA. Association for Computational Linguistics.

Petra Bago, Sheila Castilho, Edoardo Celeste, Jane Dunne, Federico Gaspari, Níels Rúnar Gíslason, Andre Kåsen, Filip Klubička, Gauti Kristmannsson, Helen McHugh, and et al. 2022. Sharing high-quality language resources in the legal domain to develop neural machine translation for under-resourced European languages. *Revista de Llengua i Dret*, (78):9–34.

Starkaður Barkarson and Steinþór Steingrímsson. 2019. Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 140–145, Turku, Finland. Linköping University Electronic Press.

Kristín Bjarnadóttir, Kristín Ingibjörg Hlynsdóttir, and Steinþór Steingrímsson. 2019. DIM: The Database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 146–154, Turku, Finland. Linköping University Electronic Press.

Steinunn Rut Friðriksdóttir. 2024. The GenderQueer Test Suite. In *Proceedings of the Ninth Conference on Machine Translation*, pages 327–340, Miami, Florida, USA. Association for Computational Linguistics.

Angelina Gašpar, Sanja Seljan, and Vlasta Kučiš. 2022. Measuring Terminology Consistency in Translated Corpora: Implementation of the Herfindahl-Hirshman Index. *Information*, 13(2).

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025a. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025b. Preliminary Ranking of WMT25 General Machine Translation Systems.

Barbara Pozzo. 2020. Looking for a Consistent Terminology in European Contract Law. *Lingue Culture Mediazioni - Languages Cultures Mediation*, 7.1:103–126.

Kirill Semenov and Ondřej Bojar. 2022. Automated Evaluation Metric for Terminology Consistency in MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 450–457, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Hanna Kristín Steindórsdóttir. 2022. Þýðingamiðstöð utanríkisráðuneytisins. ESB-textar og sérstaða þeirra í þýðingum.

Steinþór Steingrímsson and Starkaður Barkarson. 2021. ParIce: English-icelandic parallel corpus (21.10). CLARIN-IS.

Xianfeng Zeng, Yijin Liu, Fandong Meng, and Jie Zhou. 2024. Towards multiple references era – addressing data leakage and limited reference diversity in machine translation evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11939–11951, Bangkok, Thailand. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.