Scoring the Translation: On Target Automatic Keyword-Based Evaluation of Machine Translation in the Sports Domain

Steinþór Steingrímsson, Einar Freyr Sigurðsson
The Árni Magnússon Institute for Icelandic Studies
Edda, Arngrímsgata 5, 107 Reykjavík
{steinthor.steingrimsson, einar.freyr.sigurdsson}@arnastofnun.is

Abstract

We take a closer look at the results of a recent translation shared task at WMT 2025 (the Conference on Machine Translation) and analyse the errors in the output of the four highest-scoring systems. We revise the automatic evaluation method used in Sigurðsson et al. (2025) and compare it to manual evaluation of six machine-translation systems. We find that our results are in line with the manual evaluation, indicating that the test suite can be well suited for evaluating machine translation in this domain. Finally, we publish a list of domain-specific sports terms, namely, in the domains of basketball, chess, football, golf and gymnastics.

Keywords: machine translation, evaluation, terminology, sports, Icelandic

1. Introduction

SportsEval (Sigurðsson et al., 2025), a submission to the test suite subtask of the WMT 2025 General Translation Shared Task (Kocmi et al., 2025) focused on sports-vocabulary translations from English to Icelandic. The submission was a test suite, consisting of 300 segments covering five sports popular in Iceland. The test suite contained 971 term instances, some recurring across segments in order to allow for evaluation of translation consistency. All systems participating in the English→Icelandic language pair in the general translation task were evaluated on this test suite. The findings claim that current MT systems face considerable challenges in the sports domain, as measured by an automatic keyword-based evaluation, supported by manual evaluation of a subset of the translations. While the order of the best-scoring systems was similar using both approaches, there was a large difference in the ratio of terms evaluated as correct depending on whether the evaluation was automatic or manual, with the top four systems translating 47%–55% of terms correctly as measured automatically, as opposed to 71%–78% when evaluated manually.

This begs the question: *Why does the keyword-based approach so commonly reject correct translations?* Furthermore, looking at systems that score highly in the general translation task, what kind of errors do they make in terminology translation?

In this paper, we analyse the translations of six systems, the four highest-scoring ones as reported by Sigurðsson et al. (2025), and two lower-scoring ones. We revise the automatic evaluation approach and reconsider which terms to evaluate. We publish our term lists in English and Icelandic and an updated test suite reflecting the changes made.

Our main contributions are as follows:

- A list of domain-specific terms for five sports popular in Iceland
- A revised test suite, reflecting the list of terms
- A robust approach for keyword-based MT evaluation for terminology translations to Icelandic
- A study of errors made by modern MT systems when translating domain-specific terminology

2. Related Work

While there are numerous studies of the use of sports-related texts in artificial intelligence (see, e.g., Yang et al. 2025 and Xia et al. 2024 for question answering studies) and machine translation (see, e.g., Zhiliang et al. 2024 who address the issue of real-time translation of sports events) there has not been much work on evaluating the quality of automatic translations in this particular domain. We extend the work of Sigurðsson et al. (2025), a submission of sports evaluation test suite for a WMT 2025 shared task.

When measuring machine translation with terminologies, keyword-based approaches are commonly used. In the shared task on machine translation with terminologies at WMT 2023 (Semenov et al., 2023), *term success rate* was calculated by comparing machine-translated terms with their dictionary equivalents and *term consistency* was measured to investigate whether technical terms were translated uniformly. Furthermore, three evaluation approaches were used: Standard metrics to evaluate general translation quality, COMET (Rei et al., 2020) and ChrF (Popović, 2015).

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8862–8871
11–16 May 2026.

Sport	Seg.	Words/Seg.	Keywords	Unique	Repeated	1x	2x	3x	4x	5x
Football	100	26.8	288	197	61	136	32	28	1	0
Basketball	100	31.9	241	153	61	92	37	22	1	1
Chess	50	18.7	71	67	4	63	4	0	0	0
Golf	25	34.6	61	47	11	36	8	3	0	0
Gymnastics	25	20.6	56	46	8	38	6	2	0	0
Total	**300**	**27.3**	**717**	**510**	**145**	**365**	**87**	**55**	**2**	**1**

Table 1: Number of segments, average length of segments, total term checks, unique terms, repeated terms, and exact frequency distributions per domain. 145 unique terms are repeated across the entire test suite.

At WMT 2024, two submissions specifically looked at translations between English and Icelandic. The GenderQueer test suite (Friðriksdóttir, 2024) is built to investigate gender-inclusive translation and whether such translations are appropriate. Ármannsson et al. (2024) used a keyword-based evaluation to examine how adept MT systems are at translating idiomatic expressions and proper names. At WMT 2025, Hauksdóttir and Steingrímsson (2025) also used a keyword-based evaluation to check how skillful MT systems are at translating terminology in official document translations for the European Economic Area.

In this paper, we examine further the keyword-based approach, compare it to manual evaluation and the automatic metrics CometKiwi (Rei et al., 2022) and GEMBA-MQM (Kocmi and Federmann, 2023).

3. A Review of SportsEval

The version of SportsEval submitted to WMT25 and evaluated in Sigurðsson et al. (2025) contains 300 segments divided into five groups, each covering a sport popular in Iceland. The number of segments for each sport differs: Football and basketball have 100 segments each, chess 50 and golf and gymnastics have 25 segments each.

In total, these segments contained 971 term instances in the original version of SportsEval, some recurring across segments in order to evaluate translation consistency. Most of the segments are derived from naturally occurring examples, sometimes with minor changes, but a few are specifically written for the purposes of the evaluation suite.

We review all segments with the aim of removing testing of terms that are not domain related, i.e., used when discussing the particular sport the segment relates to. In the review process we often add allowed translations to the keyword list, as in some cases more than one word can be used for that particular term in Icelandic. The manual evaluation process contributed additions to the Icelandic list of translated terms. Sometimes we remove keywords we deem not to be domain-related terms as well as reduce the number of repetitions in the evaluation set to keep possible over-representation of some terms in check, because while repetitions are important in the test suite for measuring the consistency of the translations, it should be enough for such purposes to allow the same term to appear no more than three times in the evaluation set. When performing a simple keyword count, having the same word appear too often may skew the results.

When reviewing the list of allowed translations of keywords, we looked at the output of the four highest-scoring models. In order to see if this gives the manually evaluated systems an advantage, for comparison we also evaluate two lower-scoring models manually, but without adding terms to the list used for automatic evaluation.

Our amended test suite contains the same segments as before, but the number of keywords evaluated has been reduced to 717, due to removal of some of the repetitions as well as words that we deemed not to be domain-specific terms. Out of the total, 365 keywords occur only once and 145 different keywords occur more than once. Some of the same keywords are used in two or more of the evaluated subdomains. The same keywords never appear more than three times in one and the same subdomain, with the exception that when the same keyword can be either a noun or a verb. Examples of this are the word *foul*, which appears twice as a noun and twice as a verb in basketball texts, and *win*, appearing three times as a noun and once as a verb in football texts. Table 1 shows the number of keywords evaluated for individual sports and the number of times recurring keywords appear.

4. Evaluating MT Systems

We evaluate the MT outputs using four approaches: human evaluation where each term translation is considered, the automatic metrics CometKiwi and GEMBA-MQM, and a keyword-based approach.

System	Correct Terms (string match)	Correct Terms (word match)	Human Evaluation	CometKiwi	GEMBA-MQM
Gemini-2.5-Pro	90.0%	88.1%	90.7%	0.7314	-3.59
Shy	82.1%	78.7%	80.5%	0.7272	-4.78
Erlendur	81.6%	79.8%	80.3%	0.7354	-4.37
GPT-4.1	79.4%	78.2%	79.4%	0.7336	-3.73
TranssionTranslate	73.6%	71.4%		0.7427	-7.58
★ ONLINE-B	72.4%	70.2%		0.7398	-6.71
TowerPlus-9B	70.9%	68.1%		0.7383	-5.05
hybrid	70.2%	66.5%		0.7256	-5.68
ONLINE-G	62.2%	60.3%		0.664	-13.81
Claude-4	61.9%	58.4%		0.7247	-5.80
AMI	60.9%	58.7%	60.5%	0.7486	-10.55
Gemma-3-27B	55.0%	51.5%		0.7081	-10.40
★ Llama-4-Maverick	54.1%	50.6%		0.7144	-8.83
NLLB	47.6%	44.5%		0.6824	-14.19
Mistral-Medium	45.6%	40.7%		0.6997	-11.00
■ GemTrans	44.4%	41.1%		0.7116	-10.97
SalamandraTA	41.1%	35.8%		0.7613	-11.13
Gemma-3-12B	39.7%	37.5%	37.8%	0.6626	-16.33
■ IRB-MT	39.5%	37.0%		0.6515	-15.58
CommandA-MT	30.7%	25.5%		0.7107	-12.17
■ DeepSeek-V3	28.6%	27.9%		0.5693	-16.17
■ Qwen3-235B	24.0%	21.9%		0.5749	-17.19
■ CommandA	23.0%	19.8%		0.622	-20.39
Llama-3.1-8B	20.8%	18.1%		0.5724	-23.71
■ TowerPlus-72B	18.0%	16.7%		0.5019	-18.58
■ UvA-MT	9.5%	8.2%		0.4156	-21.30
Mistral-7B	6.0%	3.5%		0.4216	-24.94
Qwen2.5-7B	6.0%	4.2%		0.426	-24.88
■ AyaExpanse-32B	5.3%	3.5%		0.4687	-24.50
■ EuroLLM-22B	3.8%	3.1%		0.4803	-24.83
■ CommandR7B	2.4%	1.4%		0.4139	-25.00
■ AyaExpanse-8B	2.2%	1.7%		0.3622	-24.97
★ IR-MultiagentMT	2.0%	2.0%		0.3564	-23.43
■ EuroLLM-9B	1.8%	1.3%		0.3831	-24.42

Table 2: Results for the different evaluation approaches. For the human evaluation and keyword-based evaluation the results are reported in terms of accuracy, how many of the total number of terms were correctly translated. We also report on CometKiwi scores (higher is better) and GEMBA-MQM (closer to zero is better).

4.1. Human Evaluation

After revising SportsEval we need to compare manually evaluated translations to automatically evaluated ones in order to understand how accurate our automatic evaluation approach is in comparison to human evaluation. A single human evaluator manually evaluated all sentences translated by six systems, the top four ones as reported by Sigurðsson et al. (2025), and two lower-scoring ones. This gives us a comparison that should be able to confirm or refute that our keyword-based approach is valid for evaluating translations of the segments in the evaluation suite.

4.2. Keyword-Based Evaluation

The most straightforward way to carry out a keyword-based evaluation is simply to collect translations of the terms being evaluated and count how many of them are found in the system translations. In the case of Icelandic, an inflected language where active word formation of compound words is very common, it has to be a bit more complicated. A translation of a term in a sentence may appear in form different from the lemma, so a register of possible word forms has to be consulted. In some cases this is easy for Icelandic as an extensive database of Icelandic inflections, DIM (Bjarnadóttir et al., 2019), is openly available for language technology practitioners. While we find the majority of

the Icelandic terms in our test suite in DIM, some are not there. For these we have to list all possible word forms manually in order to be able to confirm whether the term is correctly translated to the target language, Icelandic. In cases of multiword terms, we almost always have to list possible forms manually. When reviewing the automatic evaluation process we thoroughly went through the word-form registry with a focus on manually adding missing word forms.

Word compounding poses another problem. The correct terms can be compounded with another word and still be admissible. In order to account for such cases we have two options: To try to list all admissible word compounds as possible translations of the term, or have our evaluation script not only look at word tokens, but also strings within words. Doing so has the advantage of accounting for most if not all cases of acceptable translations, given that we have included an exhaustive list of Icelandic words for the term in question. The disadvantage is that in some cases the compounds are inadmissible, and thus a method that accepts translations even if they are only a part of another word gives us some false positives. We report on both approaches in order to explore which one is closer to the human evaluation.

4.3. Other Automatic Metrics

When selecting the best MT system for translating specific domains, it is of utmost importance that the domain-specific vocabulary is correctly translated. But it is also important that the translated text is generally good and if many systems are similarly effective in translating the terminology, using other more general approaches for evaluating translation quality can be helpful.

We do not have human-translated references for all our segments and thus we use two reference-free approaches, CometKiwi and GEMBA-MQM. CometKiwi is a trained neural metric, which has performed well as compared to other metrics (Freitag et al., 2022; Zerva et al., 2022), while GEMBA-MQM is an LLM metric, which prompts a GPT-model to produce MQM error annotations for translated segments. For our evaluation we use GPT-4.1 mini from OpenAI, queried through their API.

CometKiwi runs quite fast on a desktop computer, but GEMBA-MQM requires using a large language model. Evaluating 300 segments for all 34 systems using GPT-4.1 mini cost $6.08, when the price is $0.80 and $3.20 per 1M input and output tokens, respectively.

5. Results

In Sigurðsson et al. (2025), there was quite a divide between the human evaluation and the keyword-based one, with the highest human evaluation being 77.9% while the keyword-based evaluation for the same model was 48.5%. In the work presented here, we fully evaluate six systems manually, the four highest-scoring ones and in order to gain insight into how accurate our approach is for less accurate translations, we also evaluated two systems with lower scores: AMI, which was 11th, and Gemma-3-12B, which was the 18th highest-scoring system in the automatic evaluation. After revising the keywords and updating the keyword-based evaluation approach we find that the keyword-based approach is very much in line with the human evaluation. The results are reported in Table 2 and show that the difference between the two is less than 2 percentage points in all six cases and the order of systems is the same as when we do string matching in our keyword-based evaluation. We thus order all the systems in descending order by the accuracy as measured by the keyword-based string-matching approach.

5.1. Validation of the Automatic Evaluation Metric

To ensure the reliability of our automatic terminology evaluation script, we measured the agreement between the human evaluator and the automatic evaluation (using string matching) both at the granular term level and the strict segment level (where a segment is only marked correct if all terms within it are accurately translated). To quantify inter-rater reliability, we calculated Cohen's κ (Cohen, 1960),

System	Acc.	Kappa (κ)	p-value	FP	FN
Gemini-2.5-Pro	0.903	0.717	0.4576	12	17
Shy	0.907	0.797	0.0890	19	9
Erlendur	0.943	0.880	0.0523	13	4
GPT-4.1	0.960	0.918	0.3865	4	8
AMI	0.943	0.872	0.6276	7	10
Gemma-3-12B	0.983	0.933	0.3711	4	1

Table 3: Per-system segment-level agreement between manual and automatic evaluation. p-values are calculated using McNemar's test with Yates' continuity correction.

which measures the agreement between two raters assessing the same thing. At first we measure the agreement of all the evaluations for the six systems as a whole. For that, at the term level, the automatic metric demonstrates an accuracy of 95.5% and a Cohen's κ of 0.89, and at the segment level, the accuracy is 94.0% with a κ of 0.88, both indicating very strong agreement with human judgment (for interpretation of Cohen's κ, see, e.g., McHugh, 2012).

To evaluate the metric's consistency across different systems, we perform a per-system segment-level agreement analysis. The results in Table 3 show that Cohen's κ remained high across all models (mean κ=0.853), ranging from 0.717 for Gemini-2.5-Pro (substantial agreement) to 0.933 for Gemma-3-12B (almost perfect agreement).

To determine if the automatic script exhibited any systematic directional bias when disagreeing with the human evaluator, we apply McNemar's test for paired nominal data (McNemar, 1947), with Yates' continuity correction (Yates, 1934) to account for the discrete nature of our evaluation, McNemar's test yielded no statistically significant bias for any individual system ($p > 0.05$ in all cases), indicating that while the precision of our approach varies slightly by model, it does not systematically over- or under-penalize any specific system.

Overall, these statistics validate that the automatic script is a highly reliable proxy for human evaluation, allowing us to scale the terminology evaluation across the entire dataset with a known, quantified, and minimal margin of error.

System	Chess		Basketball		Golf		Gymnastics		Football	
	S (%)	T (%)	S (%)	T (%)	S (%)	T (%)	S (%)	T (%)	S (%)	T (%)
Gemini-2.5-Pro	78.9	73.2	91.7	90.0	91.8	88.5	80.4	78.6	92.7	92.0
Shy	74.6	67.6	83.0	78.4	82.0	77.0	41.1	37.5	91.3	89.9
Erlendur	50.7	49.3	86.3	85.5	86.9	85.2	67.9	62.5	86.8	84.7
GPT-4.1	53.5	53.5	84.6	83.8	82.0	80.3	50.0	42.9	86.5	86.1
TranssionTranslate	26.8	25.4	77.6	75.1	88.5	85.2	39.3	35.7	85.4	83.7
ONLINE-B	32.4	31.0	73.4	71.0	83.6	83.6	42.9	35.7	84.7	83.0
TowerPlus-9B	23.9	22.5	74.3	72.6	83.6	80.3	41.1	32.1	82.6	79.9
hybrid	31.0	28.2	83.0	78.4	36.1	31.1	48.2	42.9	80.6	78.1
ONLINE-G	23.9	22.5	62.2	61.0	73.8	72.1	23.2	23.2	76.7	73.6
Claude-4	23.9	19.7	68.1	64.7	55.7	54.1	37.5	32.1	72.2	68.8
AMI	19.7	18.3	69.7	68.1	70.5	68.9	28.6	25.0	68.1	65.3
Gemma-3-27B	22.5	21.1	61.8	59.3	55.7	54.1	19.6	16.1	63.9	58.7
Llama-4-Maverick	38.0	36.6	63.9	61.8	44.3	42.6	12.5	7.1	60.1	54.9
NLLB	14.1	14.1	52.3	48.1	31.1	27.9	14.3	14.3	61.8	58.3
Mistral-Medium	23.9	23.9	50.2	44.4	42.6	32.8	14.3	8.9	53.8	49.7
GemTrans	12.7	12.7	51.5	47.3	39.3	39.3	14.3	12.5	53.1	49.0
SalamandraTA	11.3	9.9	47.3	42.7	41.0	32.8	17.9	12.5	47.9	41.7
Gemma-3-12B	9.9	9.9	48.5	46.5	34.4	31.1	8.9	7.1	46.9	44.1
IRB-MT	9.9	9.9	47.7	45.2	36.1	31.1	10.7	8.9	46.2	43.4
CommandA-MT	8.5	7.0	36.1	32.8	31.1	27.9	17.9	16.1	34.0	25.3
DeepSeek-V3	28.2	28.2	14.9	14.1	41.0	41.0	16.1	14.3	39.9	39.2
Qwen3-235B	14.1	14.1	5.4	5.4	23.0	21.3	5.4	3.6	45.8	41.3
CommandA	8.5	7.0	24.9	22.0	18.0	18.0	7.1	3.6	29.2	24.7
Llama-3.1-8B	4.2	4.2	23.2	21.6	13.1	9.8	0.0	0.0	28.5	24.0
TowerPlus-72B	29.6	29.6	5.4	4.6	59.0	57.4	21.4	21.4	16.3	14.2
UvA-MT	7.0	7.0	8.3	7.1	26.2	23.0	7.1	5.4	8.0	6.9
Mistral-7B	0.0	0.0	7.9	5.4	8.2	6.6	0.0	0.0	6.6	2.8
Qwen2.5-7B	0.0	0.0	7.5	5.8	4.9	4.9	5.4	3.6	6.6	3.8
AyaExpanse-32B	4.2	4.2	8.3	5.0	9.8	9.8	1.8	1.8	2.8	1.0
EuroLLM-22B	0.0	0.0	9.5	7.5	6.6	6.6	0.0	0.0	0.0	0.0
CommandR7B	0.0	0.0	0.4	0.0	6.6	6.6	0.0	0.0	4.2	2.1
AyaExpanse-8B	0.0	0.0	5.0	4.1	3.3	3.3	0.0	0.0	0.7	0.0
IR-MultiagentMT	12.7	12.7	1.2	1.2	0.0	0.0	0.0	0.0	0.7	0.7
EuroLLM-9B	0.0	0.0	4.6	2.9	3.3	3.3	0.0	0.0	0.0	0.0

Table 4: Scores for the five subdomains. S-columns show the accuracy as calculated by string matching and T-columns by token matching.

System	All	Some	Cons.
Gemini-2.5-Pro	127	16	88.8%
Shy	117	18	86.7%
Erlendur	123	13	90.4%
GPT-4.1	124	11	91.9%
TranssionTranslate	120	13	90.2%
ONLINE-B	112	20	84.8%
TowerPlus-9B	108	26	80.6%
hybrid	105	24	81.4%
ONLINE-G	93	27	77.5%
Claude-4	104	7	93.7%
AMI	79	48	62.2%
Gemma-3-27B	90	17	84.1%
Llama-4-Maverick	87	19	82.1%
NLLB	70	38	64.8%
Mistral-Medium	75	20	78.9%
GemTrans	66	23	74.2%
SalamandraTA	63	27	70%
Gemma-3-12B	60	27	69%
IRB-MT	57	30	65.5%
CommandA-MT	45	14	76.3%
DeepSeek-V3	19	58	24.7%
Qwen3-235B	34	20	63%
CommandA	31	17	64.6%
Llama-3.1-8B	22	33	40%
TowerPlus-72B	15	29	34.1%
UvA-MT	7	30	18.9%
Mistral-7B	5	16	23.8%
Qwen2.5-7B	3	15	16.7%
AyaExpanse-32B	5	11	31.3%
EuroLLM-22B	3	7	30%
CommandR7B	0	10	0%
AyaExpanse-8B	1	6	14.3%
IR-MultiagentMT	2	5	28.6%
EuroLLM-9B	1	4	20%

Table 5: Consistency in term translation. 145 unique terms are seen more than once in the test suite, as evident in Table 1. This table shows how many of them are always translated correctly, every time they occur, and how many af them are sometimes translated correctly. We call the sum of these two Recognized Re-occurring Terms (RRT). Consistency is given as terms consistently translated correctly as a percentage of RRT.

5.2. Comparison of Different Evaluation Approaches

We see that in four of the six cases, the automatic approach gives higher accuracy than the human evaluation. In some cases, this is likely due to word compounding done by the translation systems, generating words that the human annotator does not accept as valid Icelandic words. When the keyword-based approach demands that the whole term matches with our keywords, the results are a bit lower than the human evaluation, which can be expected as some valid keywords may be missing. A further inspection into that is done in Section 5.4.

We find that CometKiwi does not give us very useful results in terms of measuring term translations as compared to our manual or automatic approaches. There is not much difference between the highest-scoring systems and the order of the systems differs substantially from other evaluation metrics, with SalamandraTA scoring highest, followed by AMI and TranssionTranslate, which are 17th, 11th and 5th, respectively, in the order of correct terms. GEMBA-MQM on the other hand gives us results that are generally in line with the manual evaluation and keyword-based evaluation, with a few exceptions. GPT-4.1 scores very close to the best system, Gemini-2.5-Pro, which is not surprising as we used the mini version of GPT-4.1 as the backbone for GEMBA-MQM. Apart from a few deviations, GEMBA-MQM seems to give us results that may be useful. Using a more powerful LLM as the backbone and/or running it multiple times and aggregating the scores, as done by Junczys-Dowmunt (2025), may give us even more accurate scores.

As is done in Sigurðsson et al. (2025), in the results table we tag the systems that may be at a disadvantage due to other reasons than just their capability in translating from English into Icelandic: ■ for translations that had missing lines or repetitions, and ★ for translations where sentence splitting was not as expected and the the sentence alignment tool SentEval (Steingrímsson et al., 2023) was used to fix it.

5.3. Comparison of Individual Subdomains

Table 4 exhibits our automatic evaluation results for the subdomains, different sports. We find that quite a few translation systems fare well when translating texts on football, golf and basketball, while few manage to handle chess and gymnastics texts as well. While these two latter sports are among the most popular in terms of practitioners, they are probably not as well represented in media, and therefore less likely to be well represented in training data for language models, which may be at least a part of the explanation as to why most systems fail dismally when translating in these domains, with only four systems translating half of the chess terms correctly and three reaching 50% for gymnastics, as measured by the string-based approach. While defining what is good enough to be acceptable is not trivial, it is fairly safe to say that only one system can possibly reach that, Gemini-2.5-Pro which obtains 78.9%–92.7% accuracy for the subdomains. All the other highest-scoring systems go down to around 50% or lower accuracy for at least one subdomain.

We thus believe that the conclusion of Sigurðsson et al. (2025) still holds, even though our revised evaluation test suite paints a rather better picture than before.

For each domain, a number of terms occur twice or more in the test suite. We inspected whether the ones that were translated correctly at least once were consistently translated correctly. Table 5 shows that many of the highest-scoring systems were quite consistent, with four of the top ten systems obtaining over 90% consistency and another five of them translating more than 80% of re-occurring terms consistently. While some of the lower-scoring systems have rather high consistency numbers, the results may not be very meaningful in their cases as they fail more often than not in translating most of the terms.

5.4. Error Analysis

In the sports dataset, various keywords are a head of a compound. Sigurðsson et al. (2025) mention the English term *supporter*. In one of the football segments, *supporters* is, in fact, the head of the compound *Liverpool supporters*. To make it more likely that the automatic evaluation will catch correct translation of the term in this segment, both *stuðningsmaður* and *Liverpool-stuðningsmaður* are found in the Icelandic list as a translation of *supporter*. However, when the top models are evaluated, it becomes evident that it is difficult to catch all such compounds. An example is *shirt*, which is translated as *treyja* in the Icelandic term list. However, in the segment in question, *shirt* is in fact the head of the compound *Burnley shirt* but *Burnley-treyja* was nonetheless not included in the original list of Icelandic terms, whereas it is found in the updated list. There are various other examples like that which can easily be missed.

The keywords can also be part of a compound without being the head. In the chess dataset, *blindfold chess* is in fact a part of a compound whose head is *exercise*, i.e., *blindfold chess exercises*. The Icelandic term list contains *blindskákæfing* but when a system translates *blindfold chess exercises* as *blindskák**s**æfingar*, with a genitive *-s* following *blindskák* (the genitive of *blindskák* is not *blindskáks* but *blindskákar*) both the automatic token-matching evaluation and the manual evaluation report it as a failed translation (the string-matching approach should report it as correct, on the other hand). That may be a questionable approach, as the system translates the *blindfold chess* part of the compound correctly, if we disregard the genitive ending. In this way, however, we treat the translation the same as when we encounter incorrect word formation or word forms, such as when a system incorrectly translates *trampoline* as a feminine noun, *trampolína*, rather than a neuter noun, *trampólín*, or when a system translates *the mats* in the dative plural as the incorrect word form *dýnurnum* rather than *dýnunum*.

In various cases, the manual evaluation reveals that the list of Icelandic terms is limited and that some translations are missing. For example, *shot from long range* was originally translated only as *skot af löngu færi* (the Icelandic version is, word for word, fairly close to the English version). By checking manually the output of the MT systems, it was discovered that *langskot* (literally 'longshot') was missing and was therefore added to the updated list.

When using a keyword-based approach, we are looking out for whether keywords are translated correctly. However, in some cases, the keyword is not translated as such but the translation as a whole is nevertheless correct. This leads to the assessment being correct in the manual evaluation without us necessarily updating the list for the automatic evaluation. When we had the MT systems translate *Haaland, City's usual penalty taker...* the focus was on seeing whether they would translate *penalty taker* as the noun *vítaskytta*. Not all the models translated this into a noun but rather into a verb phrase, something like *Haaland, who usually **takes penalties** for City...* The list was not updated to accomodate this use, resulting in the manual evaluation marking this as a correct translation but the automatic evaluation marking it as incorrect.

When the Icelandic terms consist of more than one word, the automatic evaluation looks for a specific text string that should contain both, or all, of the words. If another word splits up the string, however, the translation inevitably will be deemed incorrect. By including a chess segment that contained the string *by flagging him*, we were focusing on *flag*, which was, according to the dataset, supposed to be translated as *fella á tíma* (literally 'fail/fell on time'). However, when *fella á tíma* is used in a translation of the segment, the pronoun *hann* (equivalent of English *he*) is most naturally positioned between *fella* and *á tíma*, i.e., *með því að fella hann á tíma*. To make sure that the automatic evaluation catches this, the dataset must contain the string *fella hann á tíma* and not only *fella á tíma*.

It should be mentioned that while the drawbacks of the automatic evaluation method described here lie mainly in false negatives — where the automatic evaluation fails to catch correct translations — there are some cases of false positives. For the chess segment *OMG you're a chess player! No way! So what's your favorite chess piece?* the translation of *chess player* and *chess piece* were being observed. *Chess player* is *skákmaður* in Icelandic whereas *chess piece* is *taflmaður*. One system translated the former incorrectly as *skákleikmaður* and the latter as *skákmaður*. *Skákmaður* is not the correct

term for *chess piece* but would have been correct for *chess player*. The human evaluator deemed these as two incorrect translations whereas the automatic evaluation marks one incorrect and one correct as it does not look at the order that the translations appear in; it only checks if the terms are found somewhere in the segment.

6. Sports-Terminology Lists

As discussed by Sigurðsson et al. (2025), there is a long tradition of terminology work in Iceland (see, e.g., Christensen et al. 2025). Nonetheless, none of the sports that are the basis of SportsEval have a terminology or glossary in the Icelandic terminology bank (Íðorðabankinn)[1] at the Árni Magnússon Institute for Icelandic Studies, which is the center of terminology work in Iceland.

We therefore use SportsEval as a groundwork for sports-terminology lists that we publish in the CLARIN-IS repository (see Section 7 below).

The terminology lists are published in TBX-format, containing the English term, with Icelandic translations, part-of-speech and the subdomain as the subject field for the term.

While terminology lists such as the one published here can be useful for machine translation evaluation, as we have shown, they are also important for translators, journalists or other writers aiming for correct use and consistency in their word choices.

7. License and Availability

The English and Icelandic terms, together with the English segments, evaluation data and all code are found in our GitHub repository.[2]

The terminology lists for each sport are published in the CLARIN repository under a CC BY 4.0 license.[3] We encourage anyone who wants to work with sports terminology to incorporate our work into theirs. These lists are derived from the work published on GitHub and simply include English terms and their Icelandic equivalents. Whereas the lists on GitHub include Icelandic translations like *stuðningsmaður* and *Liverpool-stuðningsmaður* for *supporter* (to facilitate automatic evaluation for the SportsEval test segments, as explained above), the terminologies on the CLARIN repository, which do not contain any context found in the evaluation dataset on GitHub, only list *stuðningsmaður* as a term.

[1] idord.arnastofnun.is

[2] https://github.com/stofnun-arna-magnussonar/SportsEval

[3] http://hdl.handle.net/20.500.12537/380

8. Conclusions and Future Work

We have described a revision of a test suite for sports-related machine translation, where we both revised the list of terms selected for evaluation, possible translations listed and the approach to keyword-based automatic evaluation. We find that our automatic approach using keywords and matching all allowed word forms to the translated segment as a string to allow for unforeseen word compounding, rather than only matching tokens, obtains results very close to a manual evaluation. We also publish all the revised data, terminology lists, evaluation data as well as our code, under open licenses.

As mentioned in Section 3, the manual evaluation process resulted in the addition of some translations that were not found in the original version of the lists. This only applies to the four top-scoring models, as their outputs were inspected while updating the terminology lists, but not those of any other models. While we believe that this contributes to a better overall automatic evaluation assessment it may favor the manually evaluated models. Having said that, when validating our approach in Section 5.1 we did not find any effect of this in the evaluation of our two lowest systems. Still, there may be term translations in the output of some other models that should be considered correct but are not included in our keyword lists for automatic evaluation.

Our evaluation results show that only a handful of the best systems are able to translate sports vocabulary correctly in the majority of cases. While 13 out of 34 systems submitted to the English-Icelandic translation task at WMT 2025 manage to translate terms correctly more than half the time, only three systems do that for all five subdomains, with the translation systems commonly failing in translating chess and gymnastics related vocabulary. These three systems are all among those that were manually evaluated. While a manual evaluation of some of the systems close to these four top-scoring ones may find that a few translations which the automatic approach does not accept are indeed acceptable, they lag so much behind, especially in the chess and gymnastics domains that it is safe to say that no more than three of all the systems evaluated can translate more than half of the terms correctly in all subdomains.

Extending the list of segments to evaluate may give us more accurate results in future evaluations, but as seen by the difference between sports, adding more subdomains might give us more useful information. Experimenting with LLM-based approaches such as GEMBA-MQM, focusing specifically on the domain-specific keywords may also be an interesting direction of study for domain-specific machine translation evaluation.

Extending the evaluation to more translation di-

rections than only English→Icelandic would give better information on the state-of-the-art in sports-related translations in general. As we publish our terminology translations in the context of the Sports-Eval segments and a separate sports-related term list, these can form a basis of evaluation in other languages.

9. Limitations

While the statistical significance tests showed very strong agreement between the manual and automatic evaluations, our manual evaluation was performed by a single human evaluator. This may introduce subjectivity into the assessment and while high Cohen's κ scores suggest the automatic metric is consistent with this evaluator, these findings might be strengthened by employing multiple human annotators and calculating inter-annotator agreement to mitigate potential individual bias.

As our approach and dataset is restricted to a single language pair and a specific domain, it is difficult to generalize about our findings for other conditions. It should also be noted that our approach is of course heavily influenced by the language, using an inflectional database and accounting for active word formation.

A limitation of using a set terminology list for evaluation of domain-specific MT is the problem of new terms of translation variants, requiring ongoing maintenance of the list. If the list is not maintained, its usefulness for this task may slowly diminish.

Finally, we use GEMBA-MQM and suggest adapting it for future work. Currently, we base our GEMBA evaluation on GPT-4.1 mini, which incurs evaluation costs that can quickly become excessive, hindering its application at a large scale. Using open models could make this option more attractive, if such models would be able to carry out a reasonable GEMBA evaluation.

10. Acknowledgments

We thank our co-authors of our previous paper on sports-terminology evaluation, Atli Jasonarson and Magnús Már Magnússon, who are also our collaborators on the terminology lists discussed above. We would also like to thank three anonymous reviewers for valuable feedback on the paper.

11. Bibliographical References

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 (WMT)*, pages 451–458. Association for Computational Linguistics.

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. Linköping University Electronic Press.

Lise Lotte Weilgaard Christensen, Hanne Erdman Thomsen, Bodil Nistrup Madsen, Anna-Lena Bucher, Henrik Nilsson, Claudia Dobrina, Håvard Hjulstad, Åsa Holmér, Johan Myking, Anita Nuopponen, Sirpa Suhonen, Anu Ylisalmi, and Ágústa Þorbergsdóttir. 2025. The Nordic Terminology Community. Research and practice. In *Terminology throughout History. A discipline in the making*, pages 327–364. John Benjamins.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68. Association for Computational Linguistics.

Steinunn Rut Friðriksdóttir. 2024. The Gender-Queer Test Suite. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, pages 327–340. Association for Computational Linguistics.

Selma Dís Hauksdóttir and Steinþór Steingrímsson. 2025. Automated Evaluation for Terminology Translation Related to the EEA Agreement. In *Proceedings of the Tenth Conference on Machine Translation (WMT)*, pages 850–855. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2025. GEMBA-MQM V2: Ten Judgments Are Better Than One. In *Proceedings of the Tenth Conference on Machine Translation (WMT)*, pages 926–933. Association for Computational Linguistics.

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. 2025. 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 (WMT)*, pages 355–413. Association for Computational Linguistics.

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 (WMT)*, pages 768–775. Association for Computational Linguistics.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645. 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 (WMT)*, pages 663–671. Association for Computational Linguistics.

Einar Freyr Sigurðsson, Magnús Már Magnússon, Atli Jasonarson, and Steinþór Steingrímsson. 2025. Up to Par? MT Systems Take a Shot at Sports Terminology. In *Proceedings of the Tenth Conference on Machine Translation (WMT)*, pages 856–865. Association for Computational Linguistics.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023. SentAlign: Accurate and Scalable Sentence Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263. Association for Computational Linguistics.

Haotian Xia, Zhengbang Yang, Yuqing Wang, Rhys Tracy, Yun Zhao, Dongdong Huang, Zezhi Chen, Yan Zhu, Yuan-fang Wang, and Weining Shen. 2024. SportQA: A Benchmark for Sports Understanding in Large Language Models.

Zhengbang Yang, Haotian Xia, Jingxi Li, Zezhi Chen, Zhuangdi Zhu, and Weining Shen. 2025. Sports Intelligence: Assessing the Sports Understanding Capabilities of Language Models Through Question Answering from Text to Video. *Electronics*, 14(3):461.

F. Yates. 1934. Contingency Tables Involving Small Numbers and the χ^2 Test. *Supplement to the Journal of the Royal Statistical Society*, 1(2):217–235.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 Shared Task on Quality Estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99. Association for Computational Linguistics.

Zeng Zhiliang, Wang Lei, and Liu Qiang. 2024. A method for real-time translation of online video subtitles in sports events. *Signal, Image and Video Processing*, 19:146.