Up to Par? MT Systems Take a Shot at Sports Terminology

Einar Freyr Sigurðsson, Magnús Már Magnússon, Atli Jasonarson, Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies
Reykjavík, Iceland
einar.freyr.sigurdsson,magnus.mar.magnusson,atli.jasonarson,
steinthor.steingrimsson@arnastofnun.is

Abstract

We present a submission to the WMT25 test suite subtask, focusing on the capabilities of MT systems to translate sports-related language. Although many sports attract extensive media attention and feature a rich, polysemous language, often shaped by active neologism and community-driven translations, the sports domain has received relatively little focus in MT research. In English-Icelandic automatic translations, sports-specific vocabulary often appears to be mistranslated. Our test suite is designed to test whether this observation holds merit. We evaluate 34 systems, both automatically and manually, and find that sports language poses challenges to a varying degree for all the systems.

1 Introduction

With the advent of large language models (LLMs), significant advances have been made in machine translation (MT) (Kocmi et al., 2024). While general translation capabilities are impressive for many high-resource and even some less-resourced languages, many MT systems commonly fail when dealing with specialized vocabulary or other rare peculiarities. We have noticed that some commonly used systems seem more prone to making errors when the topic is sports than for other common topics in the media. To investigate this further we built a test suite and submitted it to the WMT 2025 Test Suite subtask (Kocmi et al., 2025a). We compile a list of segments discussing five different sports: Basketball, football, golf, gymnastics and chess. These are international sports that are popular and have been played in Iceland for decades. As a result, these sports often have a well-established and diverse vocabulary that many speakers need to agree upon. It is therefore important that different translation systems apply vocabulary that is known and in actual use by speakers in order to translate sports texts successfully. Additionally, while "new" sports known from abroad start being played in Iceland, the vocabulary may consist of a somewhat high proportion of loanwords adapted to Icelandic (using, e.g., English-oriented stems with Icelandic inflection). Vocabulary for sports that have a long history in Iceland, however, strikes a balance between using loanwords and new (and old) words applying other methods. Our test suite is made available in plain text format with term annotations as well as evaluation code on GitHub.[1]

2 Related Work

There is a long tradition of terminology work in Iceland (see, e.g., Christensen et al. 2025). The Icelandic terminology bank (Íðorðabankinn)[2] at the Árni Magnússon Institute for Icelandic Studies is the center of terminology work in Iceland. As of August 2025, it hosts more than 70 terminologies and glossaries. Most of these give translations of the terms in one or more languages, and English is usually one of them. However, only one of said terminologies and glossaries deal specifically with sports — that is the terminology on climbing.

Furthermore, even if a given translation system may contain, e.g., material from published dictionaries in its training data, its usefulness may be limited by the fact that some well-known words from sports may be missing. Also, various words generally known by speakers of English are used in a very specific sense in sports, that may differ from the general use. Such terms may have a different translation in Icelandic altogether. An example is the noun *paint*, as in *The paint is wet*. In basketball, this refers to a specific place on a basketball court. Whereas *paint* in a general sense could be translated as 'málning' in Icelandic, in the basketball sense it can be translated as 'teigur, vítateigur', meaning something like 'free-throw lane', or even

[1] github.com/steinst/WMT25_Sports_Test_Suite
[2] idord.arnastofnun.is

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 856–865
November 8–9, 2025.

Sport	Segments	Words/Segment	Unique Terms	Repeated Terms
Football	100	26.7	196	94
Basketball	100	31.8	154	89
Chess	50	18.7	67	4
Gymnastics	25	20.5	46	10
Golf	25	34.5	48	14

Table 1: Number of segments, average length of segments, number of terms and how many terms are repeated in the segments. 142 unique terms are repeated in the test suite, some more than once.

as 'undir körfunni', meaning 'below/under the basket'.

We therefore find it interesting — and challenging — to look into vocabulary that may not be very well documented in lexicographic works. However, at least some of the vocabulary can be found in texts online, e.g., in corpora such as Tímarit.is[3] (e.g., Hrafnkelsson and Sævarsson 2014) and the Icelandic Gigaword Corpus[4] (Steingrímsson et al., 2018; Barkarson et al., 2022).

Knowles et al. (2023) study the performance of NMT systems at handling terminology and find that the systems are twice as likely as humans to commit terminology errors in the Hansard corpus. They measure the accuracy by counting the occurrences and translations of unique terms. In the shared task on machine translation with terminologies at WMT 2023, three evaluation approaches were used (Semenov et al., 2023): Standard metrics to evaluate general translation quality, COMET (Rei et al., 2020) and ChrF (Popović, 2015); Term success rate, or accuracy, was calculated by comparing machine-translated terms with their dictionary equivalents; Term consistency was measured to investigate whether technical terms were translated uniformly across the entire corpus.

At WMT 2024, two submissions specifically looked at problems in translating 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) uses a keyword-based evaluation to check how adept MT systems are at translating idiomatic expressions and proper names.

[3] timarit.is

[4] igc.arnastofnun.is

English Segment	Icelandic Terms
She remembers when she started learning a **kip** during her first week of **bars**.	[kippur, langkippur] tvíslá

Table 2: An English segment containing sport terms, and their Icelandic counterparts.

3 Test Suite and Translations

The test suite consists of 300 segments in total, divided into five documents. The segments can be sentences or short paragraphs containing a few sentences. An example of a segment from gymnastics is shown in Table 2.

3.1 The Dataset

We decided to select popular and well-established sports for the test suite. Football (soccer in the USA) has a more than 100-year history in Iceland with the first rule book being published in 1907 (*Knattspyrnulög* 1907, Jasonarson 2025). Football is hugely popular in Iceland with a lot of coverage in the news, online and in various podcasts. Football games are broadcast live with Icelandic commentary, including games in the Icelandic divisions, games played in the biggest leagues in the world and games played by national teams. Written live commentary is also available, especially from the Icelandic leagues. There is definitely no shortage of texts if one wants to study the grammar and vocabulary of football.

We also include 100 segments of texts on basketball — another very popular sport with a lot of media coverage — whereas we have fewer segments for chess (50), golf (25) and gymnastics (25). This could have been done differently but instead of focusing on, e.g., two sports, we went for covering five, with three of them having fewer segments.

By having more text from football and basketball, we can evaluate consistency in recurring terms but there are fewer terms that occur more than once in the chess, golf and gymnastics set. Table 1 shows the number of segments, length of segments and number of terms covered for each sport represented in the test suite.

We selected sentences containing sports terminology from news reports and other sports-related coverage to ensure a varied distribution of vocabulary. Certain terms were intentionally repeated to allow for examination of model consistency. To preserve the original meaning when extracting sentences from their context, minimal edits were applied, such as shortening or rephrasing.

We mark the terms in each segment and register the Icelandic equivalent. This information is used for automatic evaluation of MT systems when applied to the test suite. Following Ármannsson et al. (2024), we carry out a keyword-based evaluation and compare the results to a manually evaluated sample to inspect whether the automatic approach judges the systems in a way close to what humans would do.

3.2 The Work Process

When creating the test suite and evaluating the different translation systems, the work process was as follows. First of all, we gathered the actual English data, with the goal of creating 300 segments. These range from single sentences to short paragraphs consisting of a few sentences. The majority of the data are real examples found through various online sources, although we have in some cases simplified them somewhat or added a bit of context to make it clear what sport is being discussed. A few of the sentences are purely synthetic.

When we were collecting the data, we picked out term candidates and assigned them translations. When finalizing the list of terms, one of the authors read through all the segments and inspected the term candidates with the goal of limiting internal discrepancy in the translations of a term that occurs more than one time in each sport. However, in order to reduce the impact of a single term on the overall outcome, he tried to limit the occurrences of a single term within a single sport to three. Even if a term together with its translation occurs three times in one sport, that term may also be found in the list within a different sport. An example is *penalty* which is a term in football and golf. While it can be translated as 'víti' in both sports, we also translate it as 'vítaspyrna' in football and 'refsing' in golf. 'Refsing' would not work as a translation for *penalty* in football and likewise, e.g., 'vítaspyrna' would not work as a translation of the word in golf (the head noun of the compound *vítaspyrna* is *spyrna* 'kick'). Furthermore, some terms occur as a noun and a verb within a single sport. An example is *foul*. We decided to count the occurrences of it as a noun separately from when it occurs as a verb, meaning that it may be found three times as a noun and three times as a verb in one and the same sport. Also, we treat the noun *foul* in basketball separately from, e.g., the term *technical foul*.

When the list of terms had been finalized, another author ran the data through automatic evaluation and prepared a document for two of the authors, different from the two already mentioned, for manual evaluation. The evaluation process is described and discussed in Section 4.

3.3 Problematic Translations

We evaluate translations generated by 34 MT systems. Out of the 34 systems, we received the original test suite segmentation from 15 systems, i.e. 300 segments with a 1-to-1 mapping to our original segments. The other 19 systems delivered more or fewer segments. For two systems, IR-MultiagentMT and ONLINE-B, all text for each sport was translated as one document and we recieved the translations as 5 lines, one for each document. For these documents we segmented the text using NLTK (Bird and Loper, 2004) and then used SentAlign (Steingrímsson et al., 2023) to match the translations to the correct segments. We fixed the other outputs mostly manually in order to be able to carry out our automatic evaluation approach. Table 3 lists the systems we fixed the output for, explains what was out of order and how we fixed it. In many cases the root of the problem seems to be the inability of the systems to translate long documents so they fail on multiple translations due to that.

4 Evaluation Methodology

4.1 Automatic Evaluation

We employ a rather simplistic approach to automatic evaluation. For each segment in our test suite, the English terms have been identified and their Icelandic counterparts registered.

System Name	Lines	Problem	Fix
Erlendur	300	Incorrect order.	Manually
IR-MultiagentMT	5	Documents returned as one line.	SentAlign
Mistral-Medium	302	Split segments.	Manually
ONLINE-B	5	Documents returned as one line.	SentAlign
UvA-MT	100	Many lines missing. Translated as documents but seems to stop after approx. 15 sentences in each document.	Manually
TowerPlus-72B	264	Many lines missing. Fails on long documents, gets stuck in a loop and starts repeating previous translations.	Manually
Qwen2-235B	221	Many lines missing for basketball.	Manually
Llama-4-Maverick	329	Many football segments split.	SentAlign
IRB-MT	593	Football starts repeating in a loop when 12 sentences are left, thus only contains 288 valid translations in total.	Manually
GemTrans	343	Football starts repeating itself when two lines are left, we remove all text produced after repetitions start.	Manually
Gemma-3-27B	301	Split segments.	Manually
Gemma-3-12B	303	Split segments.	Manually
EuroLLM-22B	199	All football in one line. (Translations in Swedish, not Icelandic.)	Manually
EuroLLM-9B	176	Gymnastics in one line. Football missing. (Translations in Swedish, not Icelandic.)	Manually
DeepSeek-V3	197	Basketball and football fails because the document is too long. Translations missing for these two sports.	Manually
CommandR7B	277	Multiple errors. Gymnastics don't finish but start repeating previous translations. Basketball translations phrased like headlines (and in German, not Icelandic).	Manually
CommandA	293	Some basketball translations missing.	Manually
AyaExpanse-32B	299	Missing translations.	Manually
AyaExpanse-8B	241	Many missing translations. Some translations repeated.	Manually

Table 3: The submissions by some of the MT systems did not contain the same number of lines as our test suite did. Section 3.3 describes how we tried to align the source to the correct translated segments.

System	Correct Terms	Chess	Basketball	Golf	Gymnastics	Football
Gemini-2.5-Pro	**54.57%**	**64.79%**	**52.26%**	51.61%	**58.93%**	**53.79%**
Shy	48.48%	61.97%	46.09%	50.00%	26.79%	51.03%
Erlendur	47.92%	39.44%	48.56%	46.77%	48.21%	49.66%
GPT-4.1	46.95%	43.66%	48.15%	46.77%	35.71%	48.97%
TranssionTranslate	43.35%	21.13%	45.27%	**54.84%**	25.00%	48.28%
★ ONLINE-B	43.07%	26.76%	42.39%	50.00%	25.00%	49.66%
TowerPlus-9B	42.94%	21.13%	45.27%	48.39%	28.57%	47.93%
hybrid	40.03%	26.76%	46.09%	19.35%	30.36%	44.48%
Claude-4	36.98%	18.31%	42.39%	35.48%	19.64%	40.69%
ONLINE-G	36.98%	22.54%	36.21%	40.32%	10.71%	45.52%
AMI	35.32%	14.08%	42.80%	43.55%	16.07%	36.21%
Gemma-3-27B	34.21%	19.72%	37.86%	37.10%	12.50%	38.28%
★ Llama-4-Maverick	32.41%	29.58%	38.68%	25.81%	7.14%	34.14%
NLLB	30.75%	12.68%	35.39%	19.35%	10.71%	37.59%
Mistral-Medium	30.06%	18.31%	31.28%	37.10%	8.93%	34.48%
■ GemTrans	27.01%	9.86%	33.33%	25.81%	7.14%	30.00%
SalamandraTA	24.38%	9.86%	28.40%	22.58%	14.29%	26.90%
Gemma-3-12B	23.41%	8.45%	27.98%	24.19%	7.14%	26.21%
■ IRB-MT	22.99%	8.45%	27.57%	25.81%	8.93%	24.83%
■ DeepSeek-V3	18.28%	22.54%	9.05%	25.81%	10.71%	24.83%
CommandA-MT	18.14%	4.23%	24.69%	16.13%	8.93%	18.28%
Llama-3.1-8B	14.40%	4.23%	17.28%	11.29%	0.00%	17.93%
■ CommandA	13.85%	5.63%	15.64%	12.90%	5.36%	16.21%
■ Qwen3-235B	12.88%	9.86%	3.70%	12.90%	5.36%	22.76%
■ TowerPlus-72B	9.14%	19.72%	1.65%	38.71%	12.50%	5.86%
■ UvA-MT	5.82%	7.04%	5.76%	16.13%	7.14%	3.10%
■ AyaExpanse-32B	3.74%	2.82%	5.35%	9.68%	0.00%	2.07%
Qwen2.5-7B	3.74%	0.00%	3.70%	4.84%	1.79%	4.83%
Mistral-7B	3.32%	0.00%	4.53%	6.45%	0.00%	3.10%
■ EuroLLM-22B	2.63%	0.00%	6.17%	6.45%	0.00%	0.00%
■ CommandR7B	1.66%	0.00%	0.41%	6.45%	0.00%	2.41%
★ IR-MultiagentMT	1.66%	11.27%	1.23%	0.00%	0.00%	0.34%
■ AyaExpanse-8B	1.39%	0.00%	2.47%	3.23%	0.00%	0.69%
■ EuroLLM-9B	1.25%	0.00%	2.88%	3.23%	0.00%	0.00%

Table 4: Automatic evaluation of each model across categories. The systems are in order of overall accuracy, with the highest scoring system being the only one that translated more than 50% of terms correctly. Accuracy is the ratio of correct translations to total term occurrences. The highest score for each domain is in bold letters.

We use these keywords when evaluating, by inspecting each translated segment and checking whether it contains the expected Icelandic term or terms. As Icelandic is an inflected language, we must consider all possible forms of the terms. For terms consisting only of one word, we look up all possible forms of the word in DIM, the Database of Icelandic Morphology (Bjarnadóttir et al., 2019), and if it is not found there, we manually list the forms. For multiword terms we manually create possible forms and list them. If any form of the term is found in the translation we count that as correct. We expect this approach to give us a close approximation of how correctly the MT systems translate the terminology. In some cases though, correct forms of terms may be missing, or a form is used that might make the translation ambiguous or wrong although our approach marks it as correct. There may also be variations of some terms, that we do not register but would be considered correct by a human judge. In order to inspect how close the automatic evaluation is to human judgments, we carry out a manual evaluation for comparison.

4.2 Manual Evaluation

For manual evaluation, we randomly sample segments from all five subdomains, depending on how many sentences they have in the test suite. For basketball and football we select 10 segments, 7 segments for chess and 5 for golf and gymnastics. We then collect translations for these segments from all evaluated systems. Human evaluators judge the translations and check if the term is correctly translated, without regard to the registered Icelandic term translations. Scores for each system are given as a percentage of correctly translated terms.

5 Results

We report on the results of our automatic evaluation approach and manual evaluation and compare the outcomes. In Section 3.3 we discuss how the outputs of some systems were problematic. In the results tables, we tag the systems that may be at a disadvantage due to other reasons than just their capability in translating from English into Icelandic. We use two tags: ■ for translations that had missing lines or repetitions, possibly because all segments for a given sport were being translated as one document, but the model failed to handle such long documents, and ★ for translations where we had to run sentence alignment to match translations to the original segments. These tagged systems may score lower than expected, when compared to the results in the general translation task (Kocmi et al., 2025b,a), in most cases likely due to inability to handle long documents. Two of the three systems that we realigned using SentAlign, ONLINE-B and Llama-4-Maverick, seem to score similarly to the systems in the general translation task, so splitting up and realigning may have had minimal effects in these cases.

5.1 Automatic Evaluation

We find that according to our automatic evaluation, see Table 4, only one system manages to translate over 50% of the terms correctly. This system, Gemini-2.5-Pro, also scores highest in all domains except for golf, where it has the second highest score. There is a markedly large difference between the highest scoring system and the next, but the systems in second to fourth place are not very far from each other.

It is also noteworthy that there can be a large difference between systems within domains, while the difference on average is not substantial. Only six systems translate more than 25% of the chess terms correctly, while half or more reach that threshold for basketball, football and golf. In spite of that two systems translate almost over 60% of the chess terms correctly.

We also looked at terms that occur twice or more in the test suite and inspected whether the ones that were translated correctly at least once were consistently translated correctly. Table 5 shows that this was rarely the case, with only two systems consistently translating over 50% of these terms correctly. This is a disappointment, as it indicates that even when a system translates the terminology correctly, it may be incidental.

5.2 Manual Evaluation

The results of the manual evaluation are given in Table 6. While there is a general consensus between the automatic and manual evaluation with respect to the order of best systems, there are some variations. The highest scoring system in the automatic evaluation, by a good margin, switches seat with the second best system in the manual evaluation. This being said, the sample was much smaller in the manual evaluation, with only 77 terms being checked in 37 segments. The difference between the top systems is thus only one correct translation.

System	Some	All	Cons.
Gemini-2.5-Pro	108	56	51.85%
Shy	99	44	44.44%
Erlendur	96	44	45.83%
GPT-4.1	100	45	45.00%
TranssionTranslate	96	43	44.79%
★ ONLINE-B	96	42	43.75%
TowerPlus-9B	99	40	40.40%
hybrid	96	32	33.33%
Claude-4	81	42	51.85%
ONLINE-G	87	38	43.68%
AMI	90	31	34.44%
Gemma-3-27B	78	36	46.15%
★ Llama-4-Maverick	75	33	44.00%
NLLB	77	30	38.96%
Mistral-Medium	70	29	41.43%
■ GemTrans	66	20	30.30%
SalamandraTA	60	24	40.00%
Gemma-3-12B	56	19	33.93%
■ IRB-MT	58	18	31.03%
■ DeepSeek-V3	58	5	8.62%
CommandA-MT	42	17	40.48%
Llama-3.1-8B	41	10	24.39%
■ CommandA	33	11	33.33%
■ Qwen3-235B	38	10	26.32%
■ TowerPlus-72B	26	5	19.23%
■ UvA-MT	22	3	13.64%
■ AyaExpanse-32B	12	3	25.00%
Qwen2.5-7B	11	2	18.18%
Mistral-7B	8	2	25.00%
■ EuroLLM-22B	8	1	12.50%
■ CommandR7B	7	0	0.00%
★ IR-MultiagentMT	6	1	16.67%
■ AyaExpanse-8B	5	0	0.00%
■ EuroLLM-9B	3	1	33.33%

Table 5: Consistency in term translation. 142 unique terms are seen more than once in the test suite. The table shows how many of them are sometimes translated correctly, and how many of them are consistently translated correctly, every time they occur. Consistency is given as a percentage of terms consistently translated correctly of reoccurring terms that are translated correctly at least once.

System	Accuracy
Shy	77.92%
Gemini-2.5-Pro	76.62%
Erlendur	72.73%
GPT-4.1	71.43%
TowerPlus-9B	68.83%
TranssionTranslate	66.23%
hybrid	62.34%
★ ONLINE-B	62.34%
ONLINE-G	61.04%
Claude-4	55.84%
AMI	54.55%
Gemma-3-27B	42.86%
★ Llama-4-Maverick	40.26%
■ GemTrans	37.66%
Mistral-Medium	35.06%
NLLB	29.87%
■ IRB-MT	28.57%
■ DeepSeek-V3	28.57%
Gemma-3-12B	27.27%
SalamandraTA	23.38%
■ CommandA-MT	19.48%
■ TowerPlus-72B	18.18%
Llama-3.1-8B	11.69%
■ Qwen3-235B	11.69%
CommandA	11.69%
■ UvA-MT	6.49%
■ AyaExpanse-32B	3.90%
★ IR-MultiagentMT	2.60%
Mistral-7B	1.30%
■ EuroLLM-22B	1.30%
Qwen2.5-7B	0.00%
■ AyaExpanse-8B	0.00%
■ CommandR7B	0.00%
■ EuroLLM-9B	0.00%

Table 6: Manual evaluation of a sample of translations from each system.

The accuracy in the manual evaluation is higher than in the automatic one. This could be expected as it is likely that some word forms or variations of the terms are not registered in our lists even though a human would consider them correct and mark them as such when evaluating them. Even though the manual evaluation paints a prettier picture in this regard, it shows, like the automatic evaluation, that even the best systems are still failing quite often when translating this specialized, but common, vocabulary.

6 Limitations

There are various limitations to the current work, some of which are discussed below.

In various segments, the English terms are a part of a larger compound. We have tried to avoid including such compounds as listed terms in some places but where we think it is appropriate, we try

to make the translated terms capture this fact. An example is the term *point guard*, which is translated as 'leikstjórnandi, ás'. However, in one segment, *point guard* is part of the compound ***star** point guard*.[5] In addition to the translations 'leikstjórnandi, ás', we include '**stjörnu**leikstjórnandi, **stjörnu**ás'. When translating to Icelandic, English compounds can sometimes be reworded as a phrase rather than a single compound. An example is *Liverpool supporter* which appears in one of the segments in our football set. *Supporter* is a term in the set and one of the translations we give for it is 'stuðningsmaður'. By rephrasing we could translate *Liverpool supporter* as 'stuðningsmaður Liverpool', meaning 'a supporter of Liverpool', but if we would insist on translating it as a compound, we could translate it as 'Liverpool-stuðningsmaður' — and we give that as a possibility in our test suite in that particular segment. In some cases, however, we may have failed to notice that a term is a part of a compound but when a translation system translates the whole term as "one" word (without spaces or without rephrasing) the automatic evaluation would mark it as incorrect. A dataset with more synthetic data could more easily avoid such compounds.

The manual evaluation makes it clear that we miss out various translation possibilities we did not think of, showing that the automatic evaluation is limited. Furthermore, only one annotator evaluated each translation. If multiple annotators would have evaluated the same translations they would possibly have had some disagreements. That would also have allowed us to calculate inter-annotator agreement, which could give us an indication of whether a manual evaluation such as the one we carried out is straightforward or whether it demands multiple annotators for each segment.

We try to reflect actual use in our translations. That is not an easy task, especially when term use in written texts, such as in fairly formal news coverage, differs from spoken language use. An example might be *layup* in basketball which we only translate as 'sniðskot'. However, searching for a string starting with "layup" in the Icelandic Gigaword Corpus gives 219 results. 'Sniðskot' has a formal feel to it but it captures the English term. Sometimes, however, we were not sure whether the Icelandic translation captures it completely or is well-known enough. An example is the basketball term *screen*, which occurs as a noun in the test suite and is translated as 'hindrun' in the Icelandic version of the FIBA basketball rule book;[6] we made the decision to give an Icelandic spelling version of *screen*, i.e., 'skrín', in addition to 'hindrun' which goes back centuries in the language and can be translated as 'hindrance, obstacle', even though we translated *layup* as 'sniðskot' only (a relatively new compound in the language whose head is *skot*, meaning 'shot'). Further work that takes a closer look at the actual usage would be interesting and working with professionals in each sport on the translations of the terms would be ideal.

Which words should be given as terms in a work like this can be debated. We have the verb and the noun *win* in various places listed as terms but as winning is not specifically found in one sport but not the other we might want to exclude some such words when we are exploring a translation system's ability to translate the vocabulary in a certain branch of sport. The same goes for, e.g., parts of the body: A word like *shin* is probably used more in a sport like football than golf but, nevertheless, one could certainly disagree with our decision of leaving that in as a football term in our dataset.

As natural data are the bulk of segments in the test suite, it is not always clear from the segment's context alone what the sport in question is. In a few places, we added information in the segments. Instead of using the original unchanged in the segment *A screen is the legal action of a player who, without causing undue contact, delays or prevents an opponent from reaching a desired position* we added "In basketball" to make it clear what the sport is: ***In basketball**, a screen is the legal action of a player who, without causing undue contact, delays or prevents an opponent from reaching a desired position*. However, we did not generally focus on this when finalizing the dataset and what impact this has on the translation scores merits further study. Moreover, a synthetic dataset could control for this.

[5] Note that although the term *point guard* is a compound on its own, we focus here on *star* + *point guard*, where only *point guard* is a term according to our dataset and not the whole compound *star point guard*.

[6] See, e.g., the 2017 version here: https://www.kki.is/library/Skrar/Leikreglur_i_Korfuknattleik_2018.pdf.

7 Conclusions and Future Work

Overall, the results indicate that more attention needs to be paid to the language of sports in LLMs and MT and the generally low scores confirmed our suspicion that MT systems have a hard time at translating sports texts adequately. Even the highest scoring systems do not do a very good job at translating the language of sports, and in all cases consistency is lacking.

Categorizing the errors could be useful for analyzing comparative differences between system types. Such categorization could also be useful for building MT systems better suited for translating sports language.

We intend to use our test suite to automatically evaluate new systems and keep track of their competence in translating sports terminology. While the sports we selected are some of the most popular ones that have a specialized vocabulary in Icelandic, expanding the test suite by adding more sports could be useful. Covering the terminology of each sport more thoroughly could help us get more accurate results and having at least two occurrences of each term could make our consistency evaluation more precise. It would also be interesting to translate the keyword list into more languages than Icelandic to investigate whether this problem is specific to Icelandic.

Acknowledgments

This project was funded by the Language Technology Programme for Icelandic 2024–2026. The programme, which is managed and coordinated by Almannarómur, is funded by the Icelandic Ministry of Culture, Innovation and Higher Education.

We thank Árni Jóhannsson, Eiríkur Stefán Ásgeirsson and Jóhannes Gísli Jónsson for answering questions on specific terms relating to basketball, golf and chess, respectively. Thank you to Ágústa Þorbergsdóttir for discussions on terminology work in Iceland. We would also like to thank two anonymous reviewers for valuable feedback on the paper.

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*, pages 451–458. Association for Computational Linguistics.

Starkaður Barkarson, Steinþór Steingrímsson, and Hildur Hafsteinsdóttir. 2022. Evolving Large Text Corpora: Four Versions of the Icelandic Gigaword Corpus. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2371–2381. European Language Resources Association.

Steven Bird and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217. 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.

Knattspyrnulög. 1907. Íþróttafjelag Reykjavíkur.

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

Örn Hrafnkelsson and Jökull Sævarsson. 2014. Digital libraries of historical Icelandic newspapers, periodicals, magazines and old printed books. In *Language Resources and Technologies for Processing and Linking Historical Documents and Archives Deploying Linked Open Data in Cultural Heritage – LRT4HDA, LREC 2014*. European Language Resources Association.

Atli Jasonarson. 2025. Áhliða óvinamegin: Um orðaforða Knattspyrnulaga frá 1907. Paper presented at the 38th Rask-ráðstefna um íslenskt mál og almenna málfræði, January 24th, University of Iceland.

Rebecca Knowles, Samuel Larkin, Marc Tessier, and Michel Simard. 2023. Terminology in Neural Machine Translation: A Case Study of the Canadian Hansard. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 481–488. European Association for Machine Translation.

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. 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.

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.

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. Association for Computational Linguistics.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.