

Translating Syntactic Structures: An English→Icelandic Test Suite for Syntactic Divergence

Einar Freyr Sigurðsson,¹ Steinþór Steingrímsson,¹ Bjarki Ármannsson,^{1,2}
Atli Jasonarson,¹ Helga Svala Sigurdardóttir,¹ Jón Símon Markússon²

¹The Árni Magnússon Institute for Icelandic Studies, Reykjavík, Iceland

²The University of Iceland, Reykjavík, Iceland

{einar.freyr.sigurdsson, steinthor.steingrimsson, bjarki.armannsson,
atli.jasonarson, helga.svala.sigurdardottir}@arnastofnun.is
jsm@hi.is

Abstract

We present a test suite intended to study how well different machine translation systems do when translating syntactic constructions found in English that do not have a direct equivalent in Icelandic. We manually evaluate the translations and compare it to an automatic evaluation. Whereas the results demonstrate that the best-performing models achieve high accuracy in various categories, they still struggle with others.

1 Introduction

As modern machine translation (MT) systems edge closer to or, in some cases, outperform human translators (Kocmi et al., 2025), the need for highlighting persistent shortcomings becomes even more prevalent. In this paper, we present an English→Icelandic test suite submitted to the WMT 2026 Test Suites shared task (Kocmi et al., 2026). Our test suite evaluates MT performance on syntactic constructions¹ in English which, for many speakers, cannot be directly mapped onto Icelandic, but instead require sophisticated grammatical reconstruction. We discuss a manual and an automatic process of evaluating model translations and examine our results with respect to these different methods. In our WMT submission we also asked for translations into Faroese. We received 20 system outputs for Faroese as opposed to 27 for Icelandic. An inspection of the output revealed that in some cases the alleged Faroese was not genuine Faroese. We used langdetect² with our own Faroese profile³ to evaluate the language of the outputs and found that six of the 20 systems do

not seem to contain legible Faroese and one of the systems contained full chat responses, with English and Faroese. We decided to leave the evaluation of Faroese to future work.

In constructing our test suite, we are motivated not just by the fact that syntactic constructions that are dissimilar across language pairs should prove challenging, but that the near-ubiquitous use of Large Language Models (LLMs) in state-of-the-art MT systems (Xu et al., 2024) and the overwhelming extent to which those models are trained on English data is likely to impart English-like syntax onto their output (translations or otherwise) in lesser-resourced languages such as Icelandic (see, e.g., Wendler et al. 2024; Guo et al. 2025; Schut et al. 2025 for some evidence of this behaviour in multilingual models).

The extent of English influence on Icelandic grammar has been studied (e.g., Nowenstein and Sigurjónsdóttir 2024), but studies show that native speakers in Iceland tend to resist systematic structural changes. Instead, English influence on syntactic structure mostly shows up in – although it is not limited to – isolated contexts, such as amongst heritage speakers of North-American Icelandic, rather than native speakers living in Iceland (Angantýsson et al., 2025). Given that MT systems are mostly trained on English-centric data, our test suite examines how capable these systems are at resisting structural influence from English.

Our test suite along with all translations and evaluations, manual and automatic, is available on GitHub.⁴

2 Related Work

Previous iterations of WMT have produced English→Icelandic test suites focusing primarily on lexical items such as idiomatic expressions and

¹While the properties we are interested in are first and foremost syntactic in nature, the properties of some constructions that we look at may lie at the syntax-morphology or syntax-semantics interface.

²<https://pypi.org/project/langdetect>

³https://github.com/steinst/langdetect_profiles

⁴https://github.com/stofnun-arna-magnussonar/wmt26_syntactic_test_suite

proper names (Ármannsson et al., 2024), gender-inclusive translation (Friðriksdóttir, 2024), legal terminology (Hauksdóttir and Steingrímsson, 2025) and sports-specific vocabulary (Sigurðsson et al., 2025; Steingrímsson and Sigurðsson, 2026) (with mixed results). Less research has been carried out in investigating translations of syntactic phenomena across the two languages, with notable exceptions all focusing on various features of the Icelandic grammatical gender system (Sólmundsdóttir et al., 2022; Friðriksdóttir, 2024; Ármannsson et al., 2025). This, therefore, is to our knowledge the first test suite specifically highlighting syntactic phenomena in *English* that manifest differently in Icelandic.

For other language pairs, there exist plenty of MT test sets focusing on specific grammatical features that behave differently in the two languages (see, e.g., Macketz et al. 2022; Avelino et al. 2022; Manakhimova et al. 2025). In many cases, automatic or semi-automatic methods are used to try and evaluate model performance on those sets, which we attempt here alongside our manual evaluation.

Inspired by Zouhar et al. 2026, we base our manual evaluation process on the general idea that human annotators perform better when presented with multiple translations of the same text, rather than single outputs in isolation.

3 Test Suite and Translations

3.1 The Work Process

When we chose the topic of the dataset – constructions that do not have direct translation equivalents in Icelandic – we had a few construction types in mind, such as pseudo-passives, to name one. We then looked for more types in various texts, especially works that have a wide coverage of different constructions (e.g., Huddleston and Pullum 2002). The categories of different syntactic phenomena in English were thus chosen by the authors of this paper, with the first author in particular serving as editor. When deciding which constructions to include, we did not use any specific criteria; as we are not aware of previous work that focuses on translations of syntactic constructions from English to Icelandic, we did not know what to expect in terms of which constructions would be hard or easy to translate.

Test items were constructed in various ways – hand-written, found in or based on online searches

Figure 1: A screenshot showing the setup of our manual evaluation tool. Each evaluator was presented with a short description of the syntactic construction at hand (not visible in this image, but see Appendix A for details), a source sentence in English and several different translations into Icelandic, with three criteria to judge for each translation.

and linguistic works, edited from synthesized text – but all were eventually proof-read by a native English speaker and linguist, during the process of which some items were altered or removed.

3.2 The Dataset

The test suite consists of 345 sentences in English over 20 categories, with each category focusing on a specific syntactic phenomenon which behaves differently in English and Icelandic.

The distribution over each category is shown in Table 1, see Appendix A for explanations of the selected phenomena and their different realizations in English and Icelandic.

4 Evaluation Methodology

4.1 Manual Evaluation

In the manual evaluation process, four evaluators in total were asked to judge Icelandic translations from English. This involves both checking whether the translations as such, taking the English original into account, are adequate and judging whether the Icelandic text itself is acceptable (for the constructions in question, this involves looking at the syntactic structure of the Icelandic translations). The evaluators were asked to judge the Icelandic segments according to their native linguistic intuition.⁵

⁵The evaluators’ task, therefore, partly resembles acceptability judgment tasks, which are widely used in syntactic research. It should be noted that the method of using speaker judgments has faced various criticisms and concerns (e.g., Labov 1996; Gibson and Fedorenko 2010, 2013). See, however, at least Sprouse and Almeida (2012, 2013, 2017); Jamieson et al. (2025).

Category	No. of items	Example
Applicatives	20	Adam booked Anna a cab.
Pseudo-passives	17	Nobody likes to be shouted at.
Topicalization and word order	16	Now you can take a nap.
Anaphoric <i>one</i>	20	Not that one – the one beside it.
Exceptional case marking	20	I assumed him to be an artist.
Wh-movement (<i>which</i> +noun)	10	Which horse did she buy?
Wh-movement (<i>whose</i> +noun)	10	Whose son was hired?
Wh-movement (<i>whose</i> +noun’s+noun)	10	Whose boss’s wife was elected?
<i>Tough</i> movement	20	The location is hard to get to.
Pseudo-coordination	20	Can you try and hurry up?
Inalienable body-part possession	20	He washed his hands.
Causative alternation verbs	24	She burned the letters.
Modal verbs	20	You may begin the exam now.
Verb phrase ellipsis	19	Most birds can fly, but penguins cannot.
Subjunctive mood	24	I doubt the lion is telling the truth.
Bare relative clauses	20	The peas I like are not sold here.
Reduced relative clauses	20	The bridge built in 1890 still carries traffic.
Garden path sentences	5	The boat sailed down the river sank.
The understood initiator of passives	20	In Norton, a tree was uprooted by the storm.
Parasitic gaps	10	Here is the melody he sang before hearing.

Table 1: The distribution of items over categories in our test suite, with example sentences shown for each category.

Each sentence to be judged in the manual evaluation was accompanied with a discussion of the construction involved and an example sentence. We thus try to make it clear to the evaluators where their focus should be directed. In addition to judging the translation of the constructions (construction acceptability), we want to know whether other parts of the translation – not taking the construction acceptability into account – are acceptable (translation acceptability). Note that even if construction and/or translation acceptability is inadequate, the meaning of the sentence may still be preserved.

We therefore used three binary-choice criteria in the evaluation process where we tease apart 1) the acceptability of the construction’s translation, 2) the preservation of meaning and 3) the overall acceptability of the translation when not taking the first criterion into account:

1. Construction acceptability
 - Acceptable ◦ Unacceptable
2. Meaning preserved
 - True ◦ False
3. Translation acceptability
 - Acceptable ◦ Unacceptable

For the evaluation process, we designed a tool for the human evaluators where, on a single page (or

a scrollable container), an example from the test suite was shown together with the name of the construction, a description of it and translations of the example from different models, see Figure 1.⁶ Three evaluators were assigned a random sample of 74 items – for each source sentence translations from 20 models were evaluated at a time – with the order randomized and shown to the evaluators without them knowing which models generated the translations. Each of the items given to one evaluator was also given to another evaluator to make sure inter-annotator agreement could be calculated.

The fourth evaluator’s process differed in that he evaluated all translations by seven models. First he was given the output of four models: AMI (the submission of the Árni Magnússon Institute for Icelandic Studies), Google Translate (which we assume to be the most commonly used translation tool for Icelandic), Gemini 3.1 Pro (the best-performing model for Icelandic at WMT25) and Gemma 4 - 31B (a relatively well-performing open-weights model). Subsequently, based on the results of the automatic evaluation, the lowest-scoring system (GLM 4 - 9B) and the two highest-scoring

⁶The evaluation tool, customizable for other binary evaluations, is available on GitHub: <https://github.com/stofnun-arna-magnussonar/AMI-binary-evaluator>

systems (GPT 5.5 and TRIVE) were manually evaluated.⁷ During the manual evaluation of the latter three models, the evaluator was shown the scores he had given to the first four models to check for consistency in his evaluation and was able to reassess and alter his original scores when he thought that was appropriate.

It is not always easy to distinguish between construction acceptability and translation acceptability. As an example, consider a short sentence like *The kittens drowned*. In this case, the focus is on verbs in the causative alternation and if the translation of the sentence is *Kettlingarnir druknaðu*, then *kittens* is correctly translated but *drowned* only partially: it has the intransitive *-n* morpheme but the past plural ending is incorrect and the spelling should have two k-s, i.e., *drukknðu*. Whether this should lead to both construction unacceptability and translation unacceptability can be difficult to decide. Also, if, for example, a translation of a single word is incorrect, it can simultaneously affect the evaluation of whether the meaning is preserved and the construction or translation acceptability. Here, we can take the pseudo-coordination segment *What did you try and do?* as an example. The construction involves *try and do* but when a model does not translate *did* + *try* as *reyndir* ‘tried’ but as *reykti* ‘smoked’, this results in construction/translation unacceptability and the meaning being lost.

4.2 Automatic Evaluation

We wanted to investigate whether it was feasible to use an LLM to automatically evaluate the syntactic accuracy and overall quality of the translations. We want to see whether the evaluations of an LLM are in line with manual evaluations, if given explicit instructions. If the results line up, an automatic assessment might be considered useful to probe the general capabilities of AI models in the area being tested by our test suite. The user nevertheless needs to keep in mind that a simple prompt may not be enough to counter the English-centric bias of the models and interpret automatic results with caution.

We constructed a prompt for each translated sentence and queried Google Gemini Flash Lite 3.5, chosen simply because it is not one of the models being evaluated. We provided the LLM with a detailed linguistic description of a targeted syntactic category (including grammatical

constraints and contrastive source-target examples), the original English source sentence, and the machine-translated Icelandic hypothesis.

The LLM was instructed to evaluate each translation strictly across the same three distinct, binary criteria we evaluated in the manual evaluation: construction acceptability, meaning preservation and general translation acceptability.

To facilitate downstream analysis and ensure standardized outputs, the LLM was constrained to return its evaluation in a strict JSON format, providing both a categorical judgment and a brief textual reasoning for each of the three criteria. A complete example of the prompt is given in Appendix B.

5 Results

We evaluate the translations both manually and automatically and compare the results.

5.1 Manual Evaluation

The overall results of our manual evaluation are shown in Table 2, with GPT 5.5 coming out well on top as the only model to achieve over 90% construction accuracy. The worst performer, Tiny Aya Global, records 14.2% construction accuracy, and, in total, 10 of 27 submitted models achieve only 50% construction accuracy or less.

Interestingly, the best-performing models by construction accuracy record *higher* scores according to our other two metrics, meaning preserved and translation accuracy, but the worst-performing models record *lower* scores according to those two secondary metrics. Intuitively, we might have expected the pattern of the best performers to hold for the meaning-preserved metric throughout, i.e., even if a model fails at accurately remodeling the challenging syntactic construction in question, the overall gist of the sentence might still likely come across in the translation. In fact, it seems as though isolating the construction accuracy as one of our three metrics is actually more forgiving in the case of low-quality translation. That is, an evaluator might reward a translation for construction accuracy in isolation based on a detail such as the position of the verb in the sentence (for the ‘Topicalization and word order’ category) or simply the presence of the relative complementizer *sem* (for the ‘Bare/reduced relative clauses’ categories) – even though the model’s translation is otherwise incomprehensible.

Table 3 shows how the submitted models per-

⁷Even though these were displayed in random order, the evaluator was aware how these three models were selected.

formed overall in each category of our test suite. Construction accuracy was highest in the cases of pseudo-coordination (90.7%), bare relative clauses (85.8%) and reduced relative clauses (83.9%) and lowest in the cases of garden path sentences⁸ (25.7%, although it must be taken into account that this is the smallest category and fewer translations were evaluated compared with other categories), wh-movement with *whose*+noun (12.3%) and wh-movement with *whose*+noun’s+noun (9.5%).

We also highlight the performance by category of the three best-performing models in isolation in Table 4, which shows similar findings: the three best performers score a perfect 100% in construction accuracy on eight categories, seven of which are also in the top eight in Table 3. The bottom six categories are also the same in both tables, with the best-performing models particularly struggling with the two previously mentioned wh-movement categories (overall construction accuracy 38.9% and 37.7% – the only categories where the top three fall below 50% accuracy).

As shown in Table 5, inter-evaluator agreement was generally high for our manual evaluation. Fleiss’s kappa for each of our three criteria is between 0.61 and 0.80, meaning ‘substantial agreement’ according to the definition in Landis and Koch 1977.

5.2 Automatic Evaluation

The results of the automatic evaluation are given in Tables 6 and 7. Table 6 shows all submitted models and their performance, while Table 7 highlights the overall performance on each category of our test suite.

The overall model rankings are broadly similar to those obtained through our manual evaluation – a somewhat encouraging result regarding the reliability of our prompt-based method and chosen model. Four models out of the top five in the manual evaluation also make the top five here and the same goes for the bottom five. The most striking difference in terms of individual models is surely the case of Gemini 3.1 Pro, which finished in third place in our manual evaluation (construction accuracy: 87.2%) but only managed 15th place in the automatic one (construction accuracy: 65.5%). This can be explained by the fact that Gemini 3.1 Pro’s output contained a full chat response with English

scaffolding around the Icelandic translation. The model doing the automatic evaluation considered this additional text as a part of the answer, leading to responses that could be considered false negatives. Note that this translation prompt was removed from the model’s output for our manual evaluation, but left untouched for the automatic evaluation as our intention was to investigate how accurate a completely automatic evaluation could be. This example shows that even for automatic evaluation, it is necessary to manually inspect the data to eliminate factors that can lead to flawed results.

Some more general findings on the differences between our two evaluation methods include the fact that the automatic method seems to score the construction accuracy of lesser models even higher than we do in our manual evaluation. The bottom three models (the same ones according to both evaluations) all finished with a construction accuracy of between 14.2% and 20.2% in our manual evaluation, whereas in the automatic one, the range is 22.9%–32.8%. The differences towards the top of the rankings are noticeably lesser, although we as manual evaluators scored GPT 5.5 as high as 93.4%, while no model reached 90.0% in the automatic evaluation. This means a slightly more limited range in scores overall for the automatic method.

Finally, the scores over the different categories line up with those from our manual evaluation to a certain degree, but there are nevertheless some interesting deviations. For certain constructions, our evaluation model seems to approve of the MT models’ translations to a greater extent than human evaluators do. Construction accuracy on bare and reduced relative clauses is estimated at 93.9% and 92.6%, respectively, compared with 85.8% and 83.9% in the manual evaluation. An even bigger increase is observed for parasitic gaps, where scores from manual evaluators indicate only 67.6% accuracy but our model assigns a total score of 87.4%. This suggests that even though parasitic gaps did not prove to be among the most difficult constructions for the MT models, they may potentially pose difficulties for LLMs (on which most of the MT models are based) when it comes to assessing grammaticality – although this would naturally need to be pursued further, with different models.

⁸It should be noted that even though all of our garden path sentences are in fact reduced relative clauses, we list them as a separate category.

System	Translations	Evaluated	Ratings	Constr. Acc.	Meaning Pres.	Transl. Acc.
GPT 5.5	345	345	533	93.4% $\pm$ 0.0	98.7% $\pm$ 0.0	94.0% $\pm$ 0.0
TRIVE	345	345	536	87.9% $\pm$ 0.0	96.6% $\pm$ 0.0	90.3% $\pm$ 0.0
Gemini 3.1 Pro	345	345	537	87.2% $\pm$ 0.0	96.8% $\pm$ 0.0	92.9% $\pm$ 0.0
Wayfinder	345	74	196	82.7% $\pm$ 7.0	95.9% $\pm$ 2.5	88.8% $\pm$ 4.4
SCIR-TG-MT	345	74	192	82.3% $\pm$ 6.6	91.1% $\pm$ 4.6	83.3% $\pm$ 6.0
UvA-MT	345	74	191	80.6% $\pm$ 7.1	92.7% $\pm$ 3.6	80.1% $\pm$ 6.4
Lumen	345	74	190	80.5% $\pm$ 7.4	91.6% $\pm$ 5.2	84.7% $\pm$ 6.0
Tower 9B	345	74	189	79.4% $\pm$ 7.1	84.7% $\pm$ 6.3	76.7% $\pm$ 7.2
GoogleTranslate	345	345	539	76.4% $\pm$ 0.0	90.2% $\pm$ 0.0	80.5% $\pm$ 0.0
SRPOL	345	74	191	68.6% $\pm$ 8.1	79.6% $\pm$ 7.1	58.6% $\pm$ 8.4
DeepSeek V4 Pro	345	74	189	67.7% $\pm$ 8.9	88.4% $\pm$ 4.9	77.2% $\pm$ 6.3
AMI	345	345	540	66.9% $\pm$ 0.0	56.5% $\pm$ 0.0	51.1% $\pm$ 0.0
Dubformer	345	74	191	65.4% $\pm$ 8.4	81.2% $\pm$ 6.2	57.1% $\pm$ 8.1
SalamandraTA	345	75	194	64.9% $\pm$ 8.6	77.3% $\pm$ 6.9	67.5% $\pm$ 7.2
Qingqiu-MT-9B	345	74	190	62.1% $\pm$ 8.2	66.8% $\pm$ 8.8	53.2% $\pm$ 8.9
Cohere CAT+	345	74	192	62.0% $\pm$ 8.4	65.1% $\pm$ 8.2	46.9% $\pm$ 8.8
Gemma 4 - 31B	345	345	534	54.5% $\pm$ 0.0	41.6% $\pm$ 0.0	30.0% $\pm$ 0.0
Command A+	345	74	192	50.0% $\pm$ 9.1	39.6% $\pm$ 8.9	33.9% $\pm$ 7.7
Qwen 3.6 - 27b	345	74	190	47.9% $\pm$ 8.5	44.7% $\pm$ 8.9	28.9% $\pm$ 7.7
GPT OSS 120B	345	74	192	45.3% $\pm$ 8.5	42.2% $\pm$ 8.7	27.1% $\pm$ 6.9
Mistral Medium 3.5	345	74	193	38.9% $\pm$ 8.8	39.9% $\pm$ 8.6	28.5% $\pm$ 6.7
Ministral 3 - 14B	345	74	194	28.4% $\pm$ 7.2	17.0% $\pm$ 6.3	8.8% $\pm$ 3.9
GPT OSS 20B	345	74	191	28.3% $\pm$ 7.8	18.8% $\pm$ 7.0	12.6% $\pm$ 5.5
Qwen 3.5 - 9b	345	74	190	27.4% $\pm$ 7.4	13.2% $\pm$ 6.2	10.0% $\pm$ 5.2
GLM 4 - 9B	345	345	534	20.2% $\pm$ 0.0	5.4% $\pm$ 0.0	3.0% $\pm$ 0.0
Gemma 4 - 4B	345	74	190	18.4% $\pm$ 5.5	7.4% $\pm$ 3.6	4.2% $\pm$ 3.1
Tiny Aya Global	345	74	190	14.2% $\pm$ 5.0	6.8% $\pm$ 4.9	5.3% $\pm$ 3.3

Table 2: An overview of all models’ performance on our test suite, according to our manual evaluation (note that the output of some models was evaluated in its entirety by one of the evaluators). Models are ranked by their construction accuracy score.

6 Limitations

We discuss some potential limitations in this section.

The research-group profile introduces potential biases. None of the authors who composed the sentences are native English speakers, although the sentences were manually reviewed by a native speaker to mitigate language errors. Furthermore, test items produced by different authors might differ in length or difficulty and possibly other unanticipated factors. When the majority, or even all, of the sentences for a specific construction were created by a single author, they might differ with respect to, e.g., length or vocabulary difficulty from sentences belonging to different constructions not created by that same author.

Additionally, the group responsible for reviewing the translations was homogeneous regarding gender, age and educational background. None

of them is a native speaker of English, increasing the likelihood that they do not catch important nuances in the translations or even accept sentences that should not be accepted or, vice versa, incorrectly reject perfectly fine translations of English sentences that might, e.g., be ambiguous.⁹

The tasks given to the human annotators and the model were quite similar. However, they are not identical, and the model may require more explicit instructions. The human annotators were asked to rely on their native linguistic intuition whereas the model was prompted with the following (see Appendix B): “You are an expert linguist and trans-

⁹An example of an ambiguous sentence in the test suite is *Here is the melody he sang before hearing* which some systems translated in accordance with our interpretation of the sentence (i.e., that someone sang the melody before hearing it). However, at least one model translated the sentence such that the person in question sang the melody before having his hearing. This kind of ambiguity can be difficult to assess and evaluate when focusing on particular syntactic constructions.

Category	Translations	Evaluated	Ratings	Constr. Acc.	Meaning Pres.	Transl. Acc.
Pseudo-coordination	540	180	258	90.7% $\pm$ 3.3	81.0% $\pm$ 4.6	77.1% $\pm$ 4.9
Bare relative clauses	540	221	409	85.8% $\pm$ 3.2	55.0% $\pm$ 4.7	52.6% $\pm$ 4.9
Reduced relative clauses	540	280	634	83.9% $\pm$ 2.6	71.0% $\pm$ 3.4	56.9% $\pm$ 3.7
Subjunctive mood	648	268	507	81.3% $\pm$ 3.1	70.8% $\pm$ 3.9	56.6% $\pm$ 4.2
Topicalization and word order	432	172	336	78.9% $\pm$ 3.6	71.7% $\pm$ 5.1	48.2% $\pm$ 5.6
Wh-movement with <i>which</i> +noun	270	90	134	76.1% $\pm$ 6.7	76.9% $\pm$ 6.3	78.4% $\pm$ 6.6
Modal verbs	540	240	488	73.0% $\pm$ 3.7	63.5% $\pm$ 4.3	54.5% $\pm$ 4.0
Parasitic gaps	270	130	272	67.6% $\pm$ 5.5	56.6% $\pm$ 5.7	41.9% $\pm$ 5.9
Verb phrase ellipsis	513	193	340	64.4% $\pm$ 5.0	67.6% $\pm$ 4.9	55.6% $\pm$ 4.7
Anaphoric <i>one</i>	540	220	446	64.1% $\pm$ 3.6	59.9% $\pm$ 4.5	55.2% $\pm$ 4.4
Causative alternation verbs	648	328	737	63.0% $\pm$ 3.6	62.6% $\pm$ 3.5	62.4% $\pm$ 3.4
Pseudo-passives	459	159	260	60.0% $\pm$ 5.9	54.2% $\pm$ 5.8	42.7% $\pm$ 6.0
Inalienable bodypart possession	540	280	622	59.8% $\pm$ 3.8	63.0% $\pm$ 3.8	58.7% $\pm$ 3.7
Exceptional case marking	540	200	364	56.9% $\pm$ 5.2	62.1% $\pm$ 5.2	54.4% $\pm$ 5.2
Applicatives	540	220	407	49.1% $\pm$ 4.5	59.2% $\pm$ 4.7	58.5% $\pm$ 4.7
<i>Tough</i> movement	540	180	280	45.4% $\pm$ 5.7	71.4% $\pm$ 5.0	62.5% $\pm$ 5.2
The understood initiator of passives	540	240	485	34.4% $\pm$ 3.7	56.9% $\pm$ 4.4	37.1% $\pm$ 3.9
Garden path sentences	135	35	35	25.7% $\pm$ 12.5	42.9% $\pm$ 14.2	45.7% $\pm$ 14.3
Wh-movement with <i>whose</i> +noun	270	130	293	12.3% $\pm$ 3.6	65.2% $\pm$ 5.7	50.5% $\pm$ 5.2
Wh-movement with <i>whose</i> +noun's+noun	270	130	273	9.5% $\pm$ 3.5	38.5% $\pm$ 5.3	44.7% $\pm$ 5.0

Table 3: The performance of all submitted models by category, according to our manual evaluation.

lation evaluator specializing in English to Icelandic translation. Your task is to evaluate a machine-translated Icelandic sentence against its English source.” Since we did not investigate any prompt variations, it remains unclear whether the model’s performance could align more closely with human judgment through tinkering with the prompts.

7 Conclusions and Future Work

In future work, our findings can be built on in several areas.

First, our study focuses on syntactic structures which we considered likely translation challenges. Applying our three-factored evaluation system to standard, everyday sentences could establish a baseline, helping us determine how well translation systems generally perform with respect to syntactic structure. Furthermore, it might shed light on whether such a granular scoring system reveals nuances which simple binary scoring systems miss.

Second, comparing model error patterns to those of L2-speakers could yield interesting insights. Actual errors that L2-speakers tend to make in Icelandic (e.g., Glišić and Ingason 2021; Garðarsdóttir and Þorvaldsdóttir 2020) could be the basis of further evaluation of MT systems in future work, as could grammatical structures in Icelandic that seem to pose particular problems for language models (Ármannsson et al., 2026).

Third, we believe that various constructions in our test suite are such that if translated in a direct

manner from English, they become ungrammatical or unacceptable for most speakers. There are still other constructions, such as *tough* movement, that have more inter-speaker variation in the language community. This variation may or may not be reflected in the inter-annotator agreement to some degree. While syntactic variation in Icelandic has been a fruitful field of study, we hope the current project will inspire even further linguistic variation research with respect to the constructions in our test suite.

Fourth, we want to evaluate the Faroese output discussed in Section 1 to see if our dataset is also applicable to a related language.

Fifth, our dataset could be extended, adding more examples of the current structures, as well as adding new ones. A dataset of such sentence structures could be fed to LLMs, potentially improving their fluency.

Finally, we noticed a trend amongst some models, where they heavily paraphrase tricky structures, such as complex wh-phrases (e.g., *Whose grandmother’s car was destroyed?*). It would be worth researching whether this relates to common AI text traits (see Jaashan and Bin-Hady 2025 for a discussion on these traits in ChatGPT and DeepSeek) or to translation theories about the tendency of translated text becoming more simplified and standardized than the source text (Baker, 1993).

Category	Translations	Evaluated	Ratings	Constr. Acc.	Meaning Pres.	Transl. Acc.
Bare relative clauses	60	60	90	100.0% $\pm$ 0.0	94.4% $\pm$ 0.0	86.7% $\pm$ 0.0
Causative alternation verbs	72	72	136	100.0% $\pm$ 0.0	100.0% $\pm$ 0.0	97.1% $\pm$ 0.0
Modal verbs	60	60	99	100.0% $\pm$ 0.0	99.0% $\pm$ 0.0	88.9% $\pm$ 0.0
Pseudo-coordination	60	60	74	100.0% $\pm$ 0.0	100.0% $\pm$ 0.0	97.3% $\pm$ 0.0
Reduced relative clauses	60	60	115	100.0% $\pm$ 0.0	100.0% $\pm$ 0.0	92.2% $\pm$ 0.0
Subjunctive mood	72	72	108	100.0% $\pm$ 0.0	100.0% $\pm$ 0.0	97.2% $\pm$ 0.0
Topicalization and word order	48	48	72	100.0% $\pm$ 0.0	100.0% $\pm$ 0.0	94.4% $\pm$ 0.0
Wh-movement with <i>which</i> +noun	30	30	37	100.0% $\pm$ 0.0	100.0% $\pm$ 0.0	100.0% $\pm$ 0.0
Pseudo-passives	51	51	66	98.5% $\pm$ 0.0	98.5% $\pm$ 0.0	93.9% $\pm$ 0.0
Verb phrase ellipsis	57	57	79	96.2% $\pm$ 0.0	100.0% $\pm$ 0.0	92.4% $\pm$ 0.0
Inalienable bodypart possession	60	60	114	94.7% $\pm$ 0.0	93.0% $\pm$ 0.0	92.1% $\pm$ 0.0
Parasitic gaps	30	30	52	94.2% $\pm$ 0.0	96.2% $\pm$ 0.0	90.4% $\pm$ 0.0
Anaphoric <i>one</i>	60	60	94	86.2% $\pm$ 0.0	93.6% $\pm$ 0.0	92.6% $\pm$ 0.0
Exceptional case marking	60	60	84	83.3% $\pm$ 0.0	95.2% $\pm$ 0.0	92.9% $\pm$ 0.0
Applicatives	60	60	89	83.1% $\pm$ 0.0	100.0% $\pm$ 0.0	97.8% $\pm$ 0.0
<i>Tough</i> movement	60	60	75	81.3% $\pm$ 0.0	97.3% $\pm$ 0.0	93.3% $\pm$ 0.0
The understood initiator of passives	60	60	100	73.0% $\pm$ 0.0	97.0% $\pm$ 0.0	82.0% $\pm$ 0.0
Garden path sentences	15	15	15	53.3% $\pm$ 0.0	73.3% $\pm$ 0.0	80.0% $\pm$ 0.0
Wh-movement with <i>whose</i> +noun	30	30	54	38.9% $\pm$ 0.0	100.0% $\pm$ 0.0	87.0% $\pm$ 0.0
Wh-movement with <i>whose</i> +noun's+noun	30	30	53	37.7% $\pm$ 0.0	88.7% $\pm$ 0.0	90.6% $\pm$ 0.0

Table 4: The performance of the three best-performing models, according to our manual evaluation, sorted by category.

Task	Fleiss' κ	Mult.-annot. items
Construction acceptability	0.705	2000
Meaning preserved	0.746	2000
Translation acceptability	0.636	2000

Table 5: Overall inter-evaluator agreement, calculated using Fleiss's kappa.

Acknowledgments

We thank an anonymous reviewer for comments, which improved the paper. Thank you to Iris Edda Nowenstein for discussions on acceptability judgments and English influence on Icelandic grammar. This project was funded by the Language Technology Programme for Icelandic 2024–2026. The programme, which is managed and coordinated by Almennarómur, is funded by the Icelandic Ministry of Culture, Innovation and Higher Education.

References

- Ásgrímur Angantýsson, Iris Edda Nowenstein, and Höskuldur Thráinsson. 2025. *V2 violations in different variants of Icelandic: A common denominator?* *Nordic Journal of Linguistics*, 48(3):301–329.
- Mariana Avelino, Vivien Macketanz, Eleftherios Avramidis, and Sebastian Möller. 2022. *A Test Suite for the Evaluation of Portuguese–English Machine Translation*. In *15th International Conference of Computational Processing of the Portuguese Language. Computational Processing of the Portuguese Language (PROPOR-2022)*, March 21–23, Fortaleza,

Brazil, pages 15–25. Springer International Publishing.

- 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.

- Bjarki Ármannsson, Finnur Ágúst Ingimundarson, Iris Edda Nowenstein, and Einar Freyr Sigurðsson. 2026. *Icelandic Gender Agreement Shows LLMs' Failure to Effectively Generalize*. *University of Pennsylvania Working Papers in Linguistics*, 32.1.

- Bjarki Ármannsson, Finnur Ágúst Ingimundarson, and Einar Freyr Sigurðsson. 2025. *An Icelandic Linguistic Benchmark for Large Language Models*. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 37–47, Tallinn, Estonia. University of Tartu Library.

- Mona Baker. 1993. *Corpus Linguistics and Translation Studies — Implications and Applications*. In *Text and technology: In honour of John Sinclair*, pages 233–250. John Benjamins.

- Jóhanna Barðdal. 2001. *Case in Icelandic: A Synchronic, Diachronic, and Comparative Approach*. Doctoral Dissertation, Lund University.

- Noam Chomsky. 1981. *Lectures on Government and Binding*. Foris, Dordrecht.

- Elisabet Engdahl. 1983. Parasitic Gaps. *Linguistics and Philosophy*, 6:5–34.

System	Constr. Acc.	Meaning Pres.	Transl. Acc.
GPT 5.5	89.3%	98.8%	98.0%
TRIVE	89.3%	99.1%	98.5%
SCIR-TG-MT	85.2%	97.1%	96.2%
Wayfinder	85.2%	98.0%	99.1%
Lumen	84.1%	97.1%	96.2%
UvA-MT	83.8%	98.5%	98.8%
Tower 9B	83.2%	94.5%	93.9%
GoogleTranslate	79.4%	94.8%	92.2%
SalamandraTA	74.8%	91.0%	88.4%
SRPOL	72.8%	93.6%	87.5%
Dubformer	72.5%	94.8%	82.9%
Qingqiu-MT-9B	72.5%	89.0%	76.2%
Cohere CAT+	72.2%	88.4%	78.5%
DeepSeek V4 Pro	70.1%	98.0%	93.6%
Gemini 3.1 Pro	65.5%	78.5%	78.3%
AMI	63.5%	73.0%	72.8%
Gemma 4 - 31B	63.5%	80.0%	55.9%
Command A+	59.4%	68.4%	51.0%
Qwen 3.6 - 27b	56.8%	71.0%	53.3%
GPT OSS 120B	54.8%	78.0%	54.8%
Mistral Medium 3.5	51.0%	62.6%	45.8%
Qwen 3.5 - 9b	43.8%	49.6%	25.2%
GPT OSS 20B	38.3%	45.2%	24.4%
Ministral 3 - 14B	37.7%	50.7%	26.7%
Gemma 4 - 4B	32.8%	37.4%	8.1%
Tiny Aya Global	27.2%	31.0%	6.1%
GLM 4 - 9B	22.9%	16.5%	3.2%

Table 6: An overview of all models’ performance on our test suite, according to our automatic evaluation. Models are ranked by their construction accuracy score.

- 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.
- María Garðarsdóttir and Sigríður Þorvaldsdóttir. 2020. [A processability approach to the development of case in L2 Icelandic](#). *Language, Interaction and Acquisition*, 11(1):68–98.
- Edward Gibson and Evelina Fedorenko. 2010. [Weak quantitative standards in linguistics research](#). *Trends in Cognitive Sciences*, 14(6):233–234.
- Edward Gibson and Evelina Fedorenko. 2013. [The need for quantitative methods in syntax and semantics research](#). *Language and Cognitive Processes*, 28(1–2):88–124.
- Isidora Glišić and Anton Karl Ingason. 2021. [The Nature of Icelandic as a Second Language: An Insight from the Learner Error Corpus for Icelandic](#). In *Selected Papers from the CLARIN Annual Conference 2021*, pages 23–33. Linköping Electronic Conference Proceedings.
- Adele E. Goldberg and Laura A. Michaelis. 2017. [One Among Many: Anaphoric One and Its Relationship With Numeral One](#). *Cognitive Science*, 41(S2):233–258.
- Jacqueline Guéron. 2006. Inalienable Possession. *The Blackwell Companion to Syntax*, 2:589–638.
- Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, and Henry Xiao. 2025. [Do Large Language Models have an English Accent? Evaluating and Improving the Naturalness of Multilingual LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3823–3838, Vienna, Austria. 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*,

Category	Constr. Acc.	Meaning Pres.	Transl. Acc.
Bare relative clauses	93.9%	81.7%	73.7%
Reduced relative clauses	92.6%	86.3%	72.8%
Pseudo-coordination	90.7%	86.8%	80.2%
Parasitic gaps	87.4%	75.9%	68.2%
Topicalization and word order	86.8%	80.8%	69.7%
Subjunctive mood	84.8%	85.3%	72.6%
Modal verbs	72.2%	78.2%	69.7%
Wh-movement with <i>which</i> +noun	68.5%	83.0%	77.4%
Exceptional case marking	64.1%	76.3%	69.1%
Causative alternation verbs	61.9%	70.5%	61.7%
Inalienable body-part possession	61.1%	72.8%	61.1%
Anaphoric <i>one</i>	58.5%	77.8%	63.9%
Verb phrase ellipsis	57.5%	86.0%	73.9%
The understood initiator of passives	56.7%	69.4%	54.4%
Applicatives	56.5%	77.4%	63.9%
Pseudo-passives	34.6%	58.6%	45.1%
<i>Tough</i> movement	34.4%	88.3%	64.8%
Garden path sentences	20.0%	28.1%	58.5%
Wh-movement with <i>whose</i> +noun's+noun	17.8%	57.0%	50.7%
Wh-movement with <i>whose</i> +noun	14.1%	68.2%	59.6%

Table 7: The performance of all submitted models by category, according to our automatic evaluation.

- pages 850–855, Suzhou, China. Association for Computational Linguistics.
- Norbert Hornstein and Amy Weinberg. 1981. Case Theory and Preposition Stranding. *Linguistic Inquiry*, 12(1):55–91.
- Rodney Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge.
- Hasan M.S. Jaashan and Wagdi Rashad Ali Bin-Hady. 2025. *Stylometric analysis of AI-generated texts: A comparative study of ChatGPT and DeepSeek*. *Cogent Arts & Humanities*, 12(1):2553162.
- E Jamieson, Jennifer Smith, David Adger, Caroline Heycock, and Gary Thoms. 2025. ‘When intuitions (don’t) fail’: combining syntax and sociolinguistics in the analysis of Scots. *English Language and Linguistics*, 29(1):1–34.
- Stefan Keine and Ethan Poole. 2017. Intervention in tough-constructions revisited. *The Linguistic Review*, 34(2):295–329.
- Eva Klingvall. 2012. Topics in pseudo-passives. *Working Papers in Scandinavian Syntax*, 90:53–80.
- 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 Lakounga, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 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*, pages 355–413, Suzhou, China. Association for Computational Linguistics.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Sergey Dukanov, Anton Dvorkovich, Mark Fishel, Markus Freitag, Samuel Frontull, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Stan Kharevich, Philipp Koehn, Zheng Li, Jean Maillard, Christof Monz, Alexander Murauski, and 12 others. 2026. Findings of the WMT26 general machine translation shared task: Contrastive dynamic human evaluation at scale. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Angelika Kratzer. 1977. What ‘must’ and ‘can’ must and can mean. *Linguistics and Philosophy*, 1(3):337–355.
- William Labov. 1996. When intuitions fail. In *CLS 32: Papers from the Parasession on Theory and Data in Linguistics*, pages 77–106. Chicago Linguistic Society.
- J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

- Beth Levin and Malka Rappaport Hovav. 1995. *Unaccusativity: At the Syntax–Lexical Semantics Interface*. MIT Press, Cambridge, MA.
- Anne Lobeck. 1995. *Ellipsis: Functional Heads, Licensing, and Identification*. Oxford University Press, New York.
- Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, and Hans Uszkoreit. 2022. [A Linguistically Motivated Test Suite to Semi-Automatically Evaluate German–English Machine Translation Output](#). In *Proceedings of the Language Resources and Evaluation Conference*, pages 936–947, Marseille, France. European Language Resources Association.
- Joan Maling. 2001. [Dative: The heterogeneity of the mapping among morphological case, grammatical functions, and thematic roles](#). *Lingua*, 111(4):419–464.
- Joan Maling and Sigríður Sigurjónsdóttir. 2002. [The ‘new impersonal’ construction in Icelandic](#). *Journal of Comparative Germanic Linguistics*, 5:97–142.
- Shushen Manakhimova, Maria Kunilovskaya, Ekaterina Lapshinova-Koltunski, and Eleftherios Avramidis. 2025. [Fine-Grained Evaluation of English-Russian MT in 2025: Linguistic Challenges Mirroring Human Translator Training](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 866–877, Suzhou, China. Association for Computational Linguistics.
- Iris Nowenstein and Sigríður Sigurjónsdóttir. 2024. [Perceived threats to the future of Icelandic and the importance of language acquisition research](#). *Sociolinguistica*, 38(2):143–165.
- Liina Pylkkänen. 2008. *Introducing arguments*. The MIT Press, Cambridge, MA.
- Daniel Ross. 2021. *Pseudocoordination, Serial Verb Constructions and Multi-verb Predicates: The Relationship Between Form and Structure*. Ph.D. thesis, University of Illinois Urbana-Champaign.
- Eiríkur Rögnvaldsson. 1993. Coordination, ATB-extractions, and the Identification of *pro*. *Harvard Working Papers in Linguistics*, 3:153–180.
- Ivan A. Sag. 1997. [English relative clause constructions](#). *Journal of Linguistics*, 33:431–483.
- Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. [Do Multilingual LLMs Think In English?](#) *Preprint*, arXiv:2502.15603.
- Einar Freyr Sigurðsson, Magnús 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*, pages 856–865, Suzhou, China. Association for Computational Linguistics.
- Einar Freyr Sigurðsson and Eiríkur Rögnvaldsson. 2018. Flókin færsla. A talk given at the 32nd Rask conference on Icelandic and general linguistics. University of Iceland, 27 January.
- Agnes Sólmundsdóttir, Dagbjört Guðmundsdóttir, Lilja Björk Stefánsdóttir, and Anton Karl Ingason. 2022. [Mean Machine Translations: On Gender Bias in Icelandic Machine Translations](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3113–3121, Marseille, France. European Language Resources Association.
- Jon Sprouse and Diogo Almeida. 2012. [Assessing the reliability of textbook data in syntax: Adger’s Core Syntax](#). *Journal of Linguistics*, 48(3):609–652.
- Jon Sprouse and Diogo Almeida. 2013. [The empirical status of data in syntax: A reply to Gibson and Fedorenko](#). *Language and Cognitive Processes*, 28(3):222–228.
- Jon Sprouse and Diogo Almeida. 2017. [Setting the empirical record straight: Acceptability judgments appear to be reliable, robust, and replicable](#). *Behavioral and Brain Sciences*, 40:e311.
- Steinþór Steingrímsson and Einar Freyr Sigurðsson. 2026. [Scoring the Translation: On Target Automatic Keyword-Based Evaluation of Machine Translation in the Sports Domain](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 8862–8871, Palma, Mallorca, Spain. European Language Resources Association.
- Matthew Tyler. 2018. *Try and*. In *Yale Grammatical Diversity Project: English in North America*. Available online at <http://ygdg.yale.edu/phenomena/try-and>. Updated by Katie Martin (2018).
- Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. [Do Llamas work in English? On the Latent Language of Multilingual Transformers](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.
- Jim Wood. 2022. [Reflexive datives and argument structure](#). Manuscript, Yale University.
- Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. [A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models](#). In *The Twelfth International Conference on Learning Representations*.
- Vilém Zouhar, Roman Grundkiewicz, Sara Rajaei, Parker Riley, Martin Popel, Rachel Bawden, Philipp Koehn, Marine Carpuat, and Tom Kocmi. 2026. [Contrastive ESA: Human Evaluation of Multiple Translations at Once](#). *Preprint*, arXiv:2607.26640.

Höskuldur Þráinsson, Matthew J. Whelpton, and Jóhannes Gísli Jónsson. 2013. Ýmislegt. In Höskuldur Þráinsson, Ásgrímur Angantýsson, and Einar Freyr Sigurðsson, editors, *Tilbrigði í íslenskri setningagerð II. Helstu niðurstöður. Tölfræðilegt yfirlit*, pages 331–364. Málvísindastofnun Háskóla Íslands, Reykjavík.

A Appendix: Categories

Below we list the example categories and show sentences for each of them. The description of the categories was similar in the evaluation process, although we did not cite references there and we did not gloss the Icelandic examples. The description for each category is rather short which may simplify greatly their syntactic nature. The Icelandic versions give an idea how the constructions can be translated; they are not intended to represent a single definitive way of translating them.

A.1 Applicatives

Icelandic has a wide range of dative-case use (e.g., Barðdal 2001; Maling 2001). Indirect objects are frequently found in the dative case, as in *Jón gaf mér þessa bók* ‘John gave me this book’. In that example, *mér* ‘me’ is what is sometimes classified as a low applicative (e.g., Pylkkänen 2008). Nonetheless, Icelandic is more restricted with respect to various applicative arguments, such that for various verbs, indirect objects are not possible unless they are reflexive (Wood, 2022). Let us look at two English examples with an applied indirect object, one of which contains a reflexive indirect object, see (1-a), and one of which contains a non-reflexive indirect object, see (1-b).

- (1) a. He boiled himself an egg.
- b. He boiled her an egg.

The translation of (1-a), see (2-a), contains an applied reflexive dative *sér* (an indirect object) whereas the applied non-reflexive indirect object of (1-b) is translated within a prepositional phrase in (2-b) and (2-c).¹⁰

- (2) a. Hann sauð sér egg.
he boiled REFL.DAT egg

¹⁰We use the following abbreviations in glosses of Icelandic examples: ACC = accusative, DAT = dative, DEM = demonstrative, EXPL = expletive, FEM = feminine, GEN = genitive, INF = infinitive, INTR = intransitive, MID = middle voice, NEUT = neuter, NOM = nominative, PL = plural, PRS = present tense, PST = past tense, PTCP = participle, REFL = reflexive, SBJV = subjunctive, SG = singular, TR = transitive.

- b. Hann sauð egg handa henni.
he boiled egg for her.DAT
- c. Hann sauð egg fyrir hana.
he boiled egg for her.ACC

A.2 Pseudo-passives

English pseudo-passives (e.g., Hornstein and Weinberg 1981) involve the passive together with a stranded preposition. What corresponds to the direct object of the preposition is in subject position and triggers verb agreement like regular subjects in English do (for similar facts, e.g., in Swedish, see Klingvall 2012).

- (3) a. The players were screamed at.
- b. Which players were screamed at?

Icelandic allows stranded prepositions where the object of the preposition is topicalized but does not move to subject position. In the translation of the sentence in (3-a), see (4-a), the noun phrase ‘the players’ does not move at all and stays within the prepositional phrase. We show movement in the two translations of (3-b): In (4-b) the whole prepositional phrase, containing ‘which players’, is topicalized whereas only the wh-phrase moves in (4-c). Note that this noun phrase is topicalized and does not move to subject position in (4-c) (we do not go further into that here but it should be noted that the noun phrase is not in the nominative case but preserves accusative case from the preposition).

- (4) a. Það var öskrað á leikmennina.
EXPL was screamed at players.the.ACC
- b. Á hvaða leikmenn var öskrað?
at which players.ACC was screamed
- c. Hvaða leikmenn var öskrað á?
which players.ACC was screamed at

A.3 Topicalization and word order

In the V2-language Icelandic, when, e.g., an adverbial phrase or a prepositional phrase is topicalized, it is followed by a finite verb. This means that English is different in examples like the following, where the topicalized adverb *usually* in (5-a) and the prepositional phrase *of course* in (5-b) are followed by the subject *I* and *you*, respectively.

- (5) a. Usually I wake up early.
- b. Of course you can have more.

In the translation of these, however, the topicalized adverb and prepositional phrase are followed by a finite verb and not the subject.

- (6) a. Venjulega vakna ég snemma.
usually wake.up I early
b. Að sjálfsögðu getur þú fengið meira.
at obvious can you get more

A.4 Anaphoric one

Instead of repeating *book* in (7) (i.e., *the new book*), anaphoric *one* is used.

- (7) The first book is great but I still find the new one better.

Anaphoric *one* (e.g., [Goldberg and Michaelis 2017](#)) is different from cardinal number *one*. When a construction containing anaphoric *one* is translated to Icelandic, *one* should at least not be translated as the cardinal number *einn* ‘one’.

The translation of *the new one* in (7), see (8) below, uses a demonstrative in *sú nýja*, *sú* being a feminine form of a demonstrative in the singular, nominative case.

- (8) Fyrsta bókin er frábær en mér
first book.the.FEM is great but me.DAT
finnst samt sú
finds still DEM.FEM.SG.NOM
nýja betri.
new.FEM.SG.NOM better

There can be various ways of translating anaphoric *one* in different examples but we will not go into that here.

A.5 Exceptional case marking

English and Icelandic have so-called exceptional case marking (ECM) verbs (e.g., [Chomsky 1981](#)) that select an infinitival clause, whose subject in turn is in an object/accusative case (that is, when the ECM verb is used in the active voice), cf. Lat. *accusativus cum infinitivo*. The English sentence in (9) and its translation to Icelandic in (10) demonstrate the construction; in (9), *consider* is the ECM verb and *her* is the subject of the infinitival clause *her to be clever*.

- (9) I consider her to be clever.
(10) Ég tel hana vera snjalla.
I consider her.ACC be.INF clever.ACC

However, not all ECM verbs in English have an equivalent ECM verb in Icelandic.

- (11) We want our parents to be happy.

While *want* is an ECM verb, Icelandic *vilja*, which we can use to translate English *want*, is not. There-

fore, the translation in (12) has a different structure than the English sentence; here, *vilja* takes a finite embedded clause and not an infinitival clause.

- (12) Við viljum að foreldrar okkar séu
we want that parents.NOM our are.SBJV
hamingjusamir.
happy.NOM

We also look at the translation of ECM examples that contain wh-movement from the subject position of the infinitival clause; see (13) and its translation in (14). As before, the focus is on English ECM verbs whose equivalent verbs in Icelandic are not ECM verbs.

- (13) Who do you want to be happy?
(14) Hver vilt þú að sé
who.NOM want you that is.SBJV
hamingjusamur?
happy.NOM

A.6 Complex wh-noun phrases

These come in three types.

A.6.1 Wh-movement with *which*+noun

These are wh-phrases, moved from the object position of either a verb or a preposition. When translated to Icelandic, the phrases are assigned case in object position which is preserved under movement.

- (15) a. Which chapter did you read?
b. Which employee did you talk to?
(16) a. Hvaða kafla lastu?
which chapter.ACC read.you
b. Hvaða starfsmann talaðirðu við?
which employee.ACC talked.you to

A.6.2 Wh-movement with *whose*+noun

These examples contain passive structures where a noun phrase that originates in the object position and consists of *whose* and a noun, moves to subject position.

- (17) Whose bike was stolen?

In the translation of (17), see (18), the word order within the noun phrase is different from the English noun phrase. Instead of *whose* + noun, the word order is noun + *hvers* ‘whose’.

- (18) Hjóli hvers var stolið?
bike.DAT whose was stolen

A.6.3 Wh-movement with *whose+noun's+noun*

These examples contain passive structures where a noun phrase that originates in the object position and consists of *whose* + possessive noun + noun (the head of the whole noun phrase), moves to subject position.

- (19) Whose brother's bike was stolen?

In the translation of (19), see (20), the word order within the noun phrase is different from the English noun phrase. Instead of *whose* + possessive noun + noun, the word order is noun + possessive noun + *hvers* 'whose'.

- (20) Hjóli bróður hvers var stolið?
bike.DAT brother.GEN whose was stolen

A.7 Tough movement

In *tough* movement as in (21), "the adjective combines with two elements: an infinitival clause containing a gap and a subject DP that is semantically interpreted in the gap position" (Keine and Poole 2017).

- (21) a. My pets are not easy to forget ____.
b. Which pets are not easy to forget ____?

In a corresponding expletive construction, "the adjective takes an infinitival clause as its argument and the matrix subject position is filled with an expletive" (Keine and Poole 2017).

- (22) It is not easy to forget my pets.

Some speakers of Icelandic find the *tough* construction as in (23) unacceptable and prefer, e.g., the expletive construction in (24-a) (e.g., Þráinsson et al. 2013; Sigurðsson and Rögnvaldsson 2018).

- (23) a. Gæludýrin mín eru ekki
pets.the.NEUT.NOM.PL my are not
auðvelt að gleyma.
easy.NEUT.NOM.PL to forget.INF
b. Hvaða gæludýr eru ekki
which pets.NEUT.NOM.PL are not
auðvelt að gleyma?
easy.NEUT.NOM.PL to forget.INF
(24) a. Það er ekki auðvelt að
EXPL is not easy.NEUT.NOM.SG to
gleyma gæludýrunum mínum.
forget.INF pets.the.DAT my
b. Hvaða gæludýrum er ekki
which pets.DAT is not

auðvelt að gleyma?
easy.NEUT.NOM.SG to forget.INF

Note that *gleyma* 'forget' takes a dative-case object in Icelandic. In the *tough* construction in (23), dative case is not preserved and instead 'my pets'/'which pets' is in the nominative case in subject position, triggering plural on the finite verb and agreement in case, number and gender on the adjective. In the expletive construction in (24-a), however, as well as in (24-b), dative case on the object is preserved and the finite verb and the adjective do not agree with 'my pets'/'which pets'.

A.8 Pseudo-coordination

Pseudo-coordination "refers to any construction with the appearance of coordination in form despite subordinative (or otherwise non-coordinative) properties" (Ross 2021). (25) exhibits verbal pseudo-coordination with *try and* (followed by a verb phrase), with a similar or the same meaning as *try to* (see, e.g., Tyler 2018).

- (25) I'm gonna try and help them.

The Icelandic translation in (26) does not contain pseudo-coordination; rather, *reyna* 'try', which is not compatible with pseudo-coordination in Icelandic, selects an infinitival clause with the infinitival marker *að* 'to'.

- (26) Ég ætla að reyna að hjálpa þeim.
I intend to try.INF to help.INF them

A.9 Inalienable body-part possession

Inalienable body-part possession in English sometimes differs from Icelandic. For example, English sometimes denotes possessors in inalienable body-part possession where Icelandic may not use them at all or the possessor may be located within a prepositional phrase.

- (27) María shook her head.

In the translation of (27), see (28), a pronominal possessor is not used (which would be an equivalent of *her*), so the literal meaning of the Icelandic translation is 'María shook the head'.

- (28) María hristi höfuðið.
María shook head.the

See, e.g., Guéron 2006 for a comparison of English vs. Romance in this respect.

A.10 Causative alternation verbs

Various verbs in English that participate in a causative alternation (Levin and Rappaport Hovav, 1995), such as *open*, do not show different morphology depending on whether they occur in a transitive (29-a) or an intransitive (29-b) structure.

- (29) a. I opened the door.
b. The door opened.

They often do in Icelandic, however, as the translation of (29) in (30) shows.

- (30) a. Ég opnaði dyrnar.
I opened.TR door.the
b. Dyrnar opnuðust.
door.the opened.INTR

A.11 Modal verbs

There are various modal verbs in both English and Icelandic. Some English modal verbs might have straightforward Icelandic translations whereas some other modal verbs, like English *must*, can have different Icelandic translations, depending on the context – *must* can be deontic, epistemic, etc. (e.g., Kratzer 1977). *Must* in (31-a) has a deontic reading, whereas it has an epistemic reading in (31-b). Even though the two uses have the same form in English, they have two different translations in Icelandic, see (32-a) and (32-b).

- (31) a. You must leave.
b. You must have left early.
(32) a. Þú verður að fara.
you must to leave
b. Þú hlýtur að hafa farið snemma.
you must to have left early

A.12 Verb phrase ellipsis

Verb phrases (VPs) in English can be elided after at least modal verbs, such as *can*, and auxiliaries, such as *have* (e.g., Lobeck 1995). (33) is an example of such a VP ellipsis in English.

- (33) a. I want to help you, but I don't think I can.
b. I have never eaten here before, but you have.

Icelandic usually does not allow these structures in the same way. (34-a) shows a translation of (33-a) where *það* 'it, that' is added at the end, and in the translation of (33-b), *gert það* 'done that' is used, see (34-b). Other variations are possible

when translating the test suite examples but we do not discuss them here.

- (34) a. Ég vil hjálpa þér en ég held ég geti
I want help you but I think I can
það ekki.
that not
b. Ég hef aldrei borðað hér en þú
I have never eaten here but you
hefur gert það.
have done that

A.13 Subjunctive mood

English has a rather limited use of a subjunctive mood compared to Icelandic. The test suite contains sentences in which English finite verbs in the present or past tense should be translated in the subjunctive in Icelandic rather than the indicative mood.

- (35) a. It is likely that she can do this.
b. If I had to choose between them, I would probably choose her.
(36) a. Það er líklegt að hún geti
it is likely that she can.PRS.SBJV
gert þetta.
do.PTCP this
b. Ef ég þyrfti að velja
if I had.to.PST.SBJV to choose
milli þeirra myndi ég sennilega
between them would I probably
kjósa hana.
choose her

A.14 Bare and reduced relative clauses and garden path sentences

English has both bare relative clauses, as in (37-a), and reduced relative clauses, as in (37-b), both of which lack a relative complementizer or a relative pronoun (e.g., Sag 1997).

- (37) a. The movie she saw changed her life.
b. The book taught in the class was not popular.

Icelandic does not have bare relative clauses like the one in (37-a) and even though reduced relative clauses are found in Icelandic, examples like the one in (37-b) can generally not be translated as reduced relative clauses. A translation of both examples therefore contains a relative complementizer *sem* 'who, that', see (38).

- (38) a. Kvikmyndin sem hún sá breytti lífi
movie.the that she saw changed life

- hennar.
her
- b. Bókin sem var kennd í
book.the that was taught in
námskeiðinu var ekki vinsæl.
class.the was not popular

Some of the reduced relative clauses in our test suite are garden path sentences. The famous garden path sentence in (39) is in fact a reduced relative clause.

- (39) The horse raced past the barn fell.

When translating (39) – and to parse the sentence correctly – *raced* must be taken to be a participle in a reduced relative clause (*raced past the barn*) and not a finite verb.

A.15 The understood initiator of passives

The passive voice in Icelandic has an understood agent whereas the demoted participant (initiator, external argument) in the English passive can also be, e.g., a natural force (see discussion in [Maling and Sigurjónsdóttir 2002](#), 132).

- (40) The house was destroyed by the avalanche.

By-phrases (*af*-phrases in Icelandic) are a hallmark of passives. Passives in English are possible where a natural force is denoted within a *by*-phrase but the translation of (40) in (41) uses the middle voice and a prepositional phrase headed by a different preposition, *í* ‘in’.

- (41) Húsið eyðilagðist í snjóflóðinu.
house.the destroyed.MID in avalanche.the

A.16 Parasitic gaps

Using [Engdahl’s \(1983, 5\)](#) tentative definition, a parasitic gap is “a gap that is dependent on the existence of another gap”. The first gap in the following is the “real” or “true gap” (rg), linked with the moved *wh*-phrase. The second gap is the parasitic one (pg).

- (42) Which books has John burnt __ (rg) without reading __ (pg)?

If the *wh*-phrase is in situ (does not move), a parasitic gap is not licensed (cf. [Engdahl 1983](#), 14), as shown in the echo question in (43) (the asterisk denotes ungrammaticality).

- (43) *John has burnt which books, without reading __ (pg)?

[Rögnvaldsson \(1993\)](#) notes that “[a]dmittedly, parasitic gaps are never fully acceptable in Icelandic”. In the translation of (42), see (44), we have added a pronoun *þær* (referring to *bækur* ‘books’) so there is no parasitic gap in the example.

- (44) Hvaða bækur hefur Jón brennt __ (rg)
which books.FEM has Jón burnt
án þess að lesa þær?
without it to read them.FEM

B Appendix: LLM Prompt for Automated Evaluation

To conduct the automated evaluation, the LLM was provided with a system message defining its role and output constraints, followed by a dynamically generated user message containing the specific test item. The user message is the same as the questions asked in our manual evaluations, with the same description of the category. Below is the system message and an example of a user message targeting the ‘Pseudo-coordination’ category.

System Message

You are an expert linguist and translation evaluator specializing in English to Icelandic translation. Your task is to evaluate a machine-translated Icelandic sentence against its English source.

You will be provided with:

1. A specific syntactic category and its description.
2. The original English sentence.
3. The Icelandic translation to evaluate.

You must evaluate the translation strictly on the following three criteria:

1. Construction Acceptability: Is the translation of the specific syntactic construction described acceptable in Icelandic? Focus only on the described construction. (Acceptable/Unacceptable)
2. Meaning Preserved: Does the translation preserve the overall meaning of the source sentence, regardless of grammatical acceptability? (True/False)
3. General Translation Acceptability: Not taking into account the specific construction evaluated in Step 1, are the other parts of the translation

acceptable, such as vocabulary selection, morphology or other syntactic structures? (Acceptable/Unacceptable)

You must return your evaluation in strictly valid JSON format using this structure:

```
{
  "construction_acceptability":
  {"judgment": "...", "reasoning":
  "..."},
  "meaning_preserved": {"judgment":
  "...", "reasoning": "..."},
  "general_translation_acceptability":
  {"judgment": "...", "reasoning": "..."}
}
```

User Message (Example)

Category: Pseudo-coordination

Description: Pseudo-coordination refers to any construction with the appearance of coordination in form despite subordinative (or otherwise non-coordinative) properties. (1) exhibits verbal pseudo-coordination with "try and" (followed by a verb phrase), with a similar or the same meaning as "try to".

(1) I'm gonna try and help them.

The Icelandic translation in (2) does not contain pseudo-coordination; rather, "reyna" 'try', which is not compatible with pseudo-coordination in Icelandic, selects an infinitival clause with the infinitival marker "að" 'to'.

(2) Ég ætla að reyna að hjálpa þeim.

Source Sentence:

\$\$\$What conference are you going to try and get into?\$\$\$

Translation to Evaluate:

\$\$\$Hvaða ráðstefnu ætla þú að reyna að komast inn á?\$\$\$