Investigating the Linguistic Performance of Large Language Models in Machine Translation

Shushen Manakhimova[1], Vivien Macketanz[1], Eleftherios Avramidis[1], Ekaterina Lapshinova-Koltunski[2], Sergei Bagdasarov[3] and Sebastian Möller[1]

[1]German Research Center for Artificial Intelligence (DFKI)
firstname.lastname@dfki.de
[2]University of Hildesheim, lapshinovakoltun@uni-hildesheim.de
[3]Saarland University, sergeiba@lst.uni-saarland.de

Abstract

This paper summarizes the results of our test suite evaluation on 39 machine translation systems submitted at the Shared Task of the Ninth Conference of Machine Translation (WMT24). It offers a fine-grained linguistic evaluation of machine translation outputs for English–German and English–Russian, resulting from a significant manual linguistic effort. Based on our results, LLMs are inferior to NMT in English–German, both in overall scores and when translating specific linguistic phenomena, such as punctuation, complex future verb tenses, and stripping. LLMs show quite a competitive performance in English-Russian, although top-performing systems might struggle with some cases of named entities and terminology, function words, mediopassive voice, and semantic roles. Additionally, some LLMs generate very verbose or empty outputs, posing challenges to the evaluation process.

1 Introduction

The evolution of large language models (LLMs) has revived interest in machine translation (MT) evaluation, raising the discussion about whether general-purpose LLMs can outperform specialized MT systems. LLMs have demonstrated remarkable performance across various tasks, prompting an urgent need to assess their linguistic capabilities and potential risks (Wang et al., 2024; Guerreiro et al., 2023). Last year's Eighth Conference on Machine Translation findings (WMT23; Kocmi et al., 2023) showed that one LLM performed well across most language pairs. Although GPT-4 excelled in some areas (e.g., translation of user-generated content), it struggled with other aspects, such as speaker gender translation and specific domains (e.g., legal); it ranked lower than encoder-decoder systems when translating from English into less-represented languages, e.g., Czech and Russian.. However, last year's General MT Task included only two LLM-based system submissions (Kocmi et al., 2023). This year marks a noteworthy increase in LLMs participating in the task. As a result, this paper covers a linguistically motivated evaluation of a broad range of LLMs, including Claude-3.5-Sonnet, GPT-4, Llama3-70B, Mistral-Large, and the recently released Unbabel-Tower70B, as well as CUNI-DS, IKUN and IKUN-C, IOL-Research, CommandR-plus, Yandex, and Occiglot.

In this context, we are presenting the results of our participation in the test suite sub-task of the Ninth Conference on Machine Translation (WMT24). Our test suite[1] consists of carefully crafted sentences that assess the ability of MT systems to handle specific linguistic phenomena. It was applied to the MT systems submitted for evaluation in two language directions: English–German and English–Russian.

2 Related Work

Several researchers have adopted test suites or challenge sets to better identify flaws in MT outputs, further contributing to the advancement in MT evaluation. The WMT test suite sub-task has played a significant role by providing a platform for these evaluations.

Chen et al. (2023), for example, developed a systematic method of selecting difficult sentences from the Wiki Corpus, taking into account factors like word difficulty, sentence length, grammatical complexity, and model learnability. Their findings showed significant differences from the official ranking, suggesting that systems performing well on average test sets might not do as well on more challenging ones. Notably, GPT-4 ranked among the top two for Chinese–English translations and between fourth and ninth in the other direction. Other research has focused on difficulties posed by special domains and writing styles. Mukherjee and Shrivastava (2023) designed a test suite for

[1] https://github.com/DFKI-NLP/mt-testsuite

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 355–371
November 15–16, 2024. ©2024 Association for Computational Linguistics

English–German translation across five domains and writing styles. They found that while GPT-4 performed competitively overall, it struggled in the legal domain and with the judgment writing style. The work of Savoldi et al. (2023) looked into gender translation of the English–German and German–English language directions. They found that while systems generally handled gender form translation well, producing gender-inclusive translations still remains a significant challenge. Specifically, GPT-4 exhibited relatively lower accuracy in accurately translating feminine gender in first-person singular references reflecting the speaker's linguistic expression of gender.

Bawden and Sagot (2023) tested the ability of MT systems, including GPT-4, to handle user-generated text from in-domain sources characterized by informal language and various grammatical deviations. Their findings show that although data at such a large scale can provide extensive training data, GPT-4 still does not perform well on consistency and faithfulness to source sentences, implying a hurdle for generalization to out-of-domain text.

The fact that these works indicated weaknesses not apparent on the General MT Shared Task illustrates the critical importance of developing focused test suites beyond general evaluation metrics to measure the capabilities and limitations of MT systems.

3 Method

3.1 Test suite description

We have developed a fine-grained test suite to evaluate the performance of MT systems for the language pairs English–German and English–Russian[2]. While we are only touching on the description of our test suite in the paper, the interested reader can find a detailed description in Macketanz et al. (2022a). Previous submissions of the test suite in WMT can be found in (Macketanz et al., 2018, 2021, 2022b; Avramidis et al., 2019, 2020; Manakhimova et al., 2023).

Our test suite focuses on various linguistic phenomena that are of interest to the respective language pairs. The phenomena are based on extensive research in linguistics, contrastive grammars, and translation studies, covering a wide range of potential translation challenges. The phenomena and their categorization are specific to a language pair and a language direction; however, there is a big overlap of the phenomena between the language pairs for the languages covered so far.

Language Pair	Test Items	Categories	Phenomena
en–de	4,846	13	110
en–ru	1,234	12	51

Table 1: Metadata of the language pairs in the test suite.

The phenomena in the test suite are classified into several categories, grouped by the underlying syntactical/morphological/lexical mechanisms. Each phenomenon is represented by at least 20 (in many cases more) test items. Every test item consists of one or more sentence(s) in the source language and a set of rules to evaluate them. The test items are either handwritten by linguistic experts or taken from existing corpora. The number of test items, phenomena, and categories per language pair can be seen in Table 1. While the English–German test suite has been around and growing since 2017, the English–Russian test suite is newer (from 2022) and, therefore, has fewer test items.

With the change of MT system types over the years (from phrase-based and statistical MT to neural MT, and finally to LLMs), typical MT challenges and errors have also changed. Thus, we have also adapted our test suite over the past few years to accommodate those changes. These adaptations included adding new phenomena, longer/more complex test sentences, and more test items per phenomenon.

MT outputs evaluated by the test suite have been used to produce challenge sets for WMT metrics (Avramidis and Macketanz, 2022; Avramidis et al., 2023).

3.2 Application of the test suite

The test suite can be characterized as semi-automatic, as the evaluation process is based on automatic rules and additional manual evaluation. While this kind of evaluation can be more time-consuming than a fully automatic evaluation process, we assume it to be more accurate as the regular expressions are handwritten by human experts.

For each test item in the test suite, one or more linguist(s) have written regular expressions to cover as many as possible expected correct and incorrect translations. The linguists rely on their years of

[2] Our test suite additionally covers the language pairs German–English and Portuguese–English, but these pairs were not part of the WMT General MT Shared Task.

experience in evaluating MT systems when writing regular expressions. However, of course, not all MT outputs can be covered by the regular expressions as languages, and the MT systems are very diverse. In these cases, the human comes into play again. All outputs that cannot be automatically evaluated by the regular expressions are inspected and hand-evaluated by a linguist.The more unexpected (meaning, in most cases, incorrect) outputs a system creates, the more manual work is involved in the evaluation process. After the evaluation process, the translation accuracy of an MT system specific per phenomenon or category is calculated by dividing the number of correctly translated test items by the total number of test items.

To ensure a fair comparison, only evaluated test items are considered for accuracy calculations. If a test item is not evaluated for one system, it is excluded for all systems, reducing the number of the effective test items.

For the system comparison (per language direction), we first identify the highest-scoring system and then compare it to the other systems. The significance of the comparison is confirmed by a one-tailed Z-test with $\alpha = 0.95$. Systems that do not perform significantly inferior to the best-performing system are grouped into the first performance cluster. The best-performing systems are indicated in boldface in the respective rows of the tables.

To account for variations in the number of test items within each category or phenomenon, average scores are computed in three different methods: The *micro-average* method combines the contributions of all test items to calculate the average percentages. In the *category macro-average*, the percentages are first computed independently per category and subsequently averaged, treating all categories equally. Analogously, for the *phenomenon macro-average* the percentages are computed independently per phenomenon and averaged afterwards, treating all phenomena equally.

4 Experiment Setup

This year, we evaluated a total of 39 systems with our test suite. The systems had been submitted to the General MT Shared Tasl of the Ninth Conference on Machine Translation. 21 systems were evaluated for English–German and 18 systems for English–Russian[3].

It is the fourth time we evaluated the English-German systems and the third time for the English-Russian systems. As described above, the evaluation of the system outputs is only semi-automatic, and therefore, manual work is needed to complement the automatic evaluation by resolving cases in which none of the rules in our rule database can be applied, the so-called *warnings*. Upon receiving the system outputs, there were on average around 25 % of warnings for English–German, varying across systems from 4.7 % to 77.5 %. For English–Russian, there were on average 46.9 %, ranging from 24.5 % to 82.7 %. As we had added several new phenomena and test items to existing phenomena before this year's WMT, we expected more warnings this year. Additionally, several systems this year, particularly LLMs, were more verbose or "creative" with their translations than we are used to from previous years. For example, Mistral sometimes offered several translation options, including explanations. This creativity, however, led to more manual work as the existing evaluation rules could not cover these unexpected outputs.

This year, the manual evaluation was conducted by three linguists who were experts in one or both language pairs. Combined, the linguists spent around 160 person-hours on the manual evaluation within about three weeks. After the manual input, an average of 0.9 % of warnings remained for English–German and 5.7 % for English–Russian.

As mentioned above, test items with one unresolved warning for at least one system were excluded from the comparison. This reduced the number of effective test items to 4219 ($\sim$87 %) test items for English–German, and 994 ($\sim$80 %) for English–Russian.

5 Results

All result tables can be found in the Appendix.

5.1 System comparison

For **English–German**, Online-B, TranssionMT, and Claude-3.5 had the highest micro-average with a score of around 97 %. Furthermore, Online-B and TranssionMT also had the highest macro-average, with a scrore of around 95 %. Whereas little is

[3] There had originally been 25 systems submitted for En-De, and 22 for En-Ru. However, the systems Dubformer and CycleL/CycleL2 had to be excluded from our evaluation for both language pairs due to invalid output

known about Online-B, TranssionMT's good performance may be explained by its optimization for complex grammatical structures and rich morphology through the use of a hyperbolic embedding. The lowest micro-average was reached by TSU-HITs (Mynka and Mikhaylovskiy, 2024) with a score of 38.6 %, and the lowest macro-average was reached by MSLC (Larkin et al., 2024) with a score of 45.8 %. On average, systems reached a micro-average of 81.4 % and a macro-average of 79.7 %.

At this point, it is important to note that two systems, Mistral (Jiang et al., 2023) and Occiglot (Avramidis et al., 2024), produced a (high) number of empty outputs for German–English. While Occiglot only generated 335 empty sentences, Mistral generated as many as 3,624. For the system comparison, we had to mark these sentences as incorrect. Therefore, Mistral appears to have the worst accuracy on micro- and macro-average. However, we conducted an extra analysis for these two systems, only considering the correct and incorrect outputs and ignoring the empty outputs. This resulted in an accuracy of 73.0 % for Occiglot and 85.9 % for Mistral macro-averaged over the non-empty outputs (see Tables in Section A). Since the accuracies are calculated over different test items, they are not comparable with each other and with other systems.

Interestingly, and contrary to previous years, our ranking of the systems according to their linguistically-related performance differs from the preliminary results of the automatic ranking of the General MT Shared Task (Kocmi et al., 2024): While the top 3 systems in the General task were Unbabel-Tower70B (Rei et al., 2024), Dubformer, and TranssionMT, according to our analysis, Online-B and TranssionMT made it to the first significance cluster, with GPT-4 falling in the second position and Unbabel-Tower70B scoring even lower. Furthermore, we had to exclude Dubformer from our analysis due to invalid output. Nonetheless, both analyses have MSLC and TSU-HITs at the bottom of the ranking.

When comparing the human rankings of the General MT Task with our rankings, one can note that in the former, many systems share the cluster of the first position. The fact that our test suite can produce a smaller significance cluster for the first position can be considered a success.

While Unbabel-Tower70B showed exceptional performance across all language directions in the automatic preliminary rankings, our evaluation revealed some potential blind spots. Compared to the top-performing systems, it struggles with less commonly used future tenses (ditransitive—future II progressive, ditransitive—future II simple, reflexive—future II progressive), with the elliptic process of *stripping*, and with *semantic roles*. Future II progressive tense can pose difficulties, likely due to its infrequent occurrence in training data and nuanced nature. An example sentence would be "I will have been baking Tim a cake." Stripping will be explained in further detail below, cf. Sec 5.3 As for semantic roles, English is relatively flexible in assigning semantic roles to subjects. In contrast, German tends to have stricter rules for subject roles regarding agentivity. This difference can cause translation issues when models directly map English constructions onto German without considering these syntactic and semantic differences.

For **English–Russian**, Yandex (Elshin et al., 2024) and Claude-3.5-Sonnet achieved the highest micro-average scores with 91.8 % and 90.4 %, respectively as well as macro-averages with 92.4 % and 90.5 %. This year's poorest-performing system was TSU-HITs, with both micro- and macro-averages of 50 and 49 %. On average, the systems reached a micro-average of 80.1 % and a macro-average of 78.7 %. According to the automatic preliminary results, the top four best-performing systems for English–Russian are Unbabel-Tower70B, Dubformer, Yandex, and Claude-3.5-Sonnet, in that order. As mentioned earlier, Dubformer was excluded from our analysis. Unbabel-Tower70B scores slightly lower than Yandex and Claude-3.5-Sonnet, achieving 89.4 % micro-average and a 90 % macro-average. On the phenomenon level, our evaluation shows that Yandex and Claude-3.5-Sonnet outperform Unbabel-Tower70B, when it comes to *collocations*, *onomatopeia*, *verb valency*, and *passive voice*. If we exclude Cycle and CycleL (Dreano et al., 2024), the worst four performing systems, according to the automatic preliminary ranking, are the same four systems in our ranking, listed here from best to worst: IKUN-C (Liao et al., 2024), CUNI-DS (Semin and Bojar, 2024), NVIDIA-NeMo, and TSU-HITs. GPT-4, one of the best-performing systems last year, falls into the second cluster this year.

Stripping	
John can play the guitar, and Mary too.	
John kann Gitarre spielen und Mary auch.	pass
John kann Gitarre spielen, und Mary auch.	fail
John kann das instrument spielen, und Lucia noch nicht.	fail
Verb Semantics	
"I've missed you so much!" he bawled.	
"Ich habe dich so sehr vermisst!" schluchzte er.	pass
"Ich habe dich so vermisst!" schrie er.	pass
»Ich habe dich so sehr verpasst!«, bawte er.	fail

Table 2: Examples of English–German linguistic phenomena with passing and failing MT outputs.

5.2 Category-level analysis

For **English–German**, two systems are in the cluster of best-performing systems per category in all categories: Online-B and TranssionMT. Furthermore, two systems have the highest accuracies on all but two/three categories, namely GPT-4 and Claude-3.5-Sonnet. The categories with the highest accuracies are *negation*, with 14 systems reaching 100 % accuracy, and *subordination*.

Some of the easiest categories for **English–Russian** include *subordination* (89.7 %), *function words* (89.1 %), where both LLM-based and other MT system score over 95 %. In contrast, *ambiguity* stood out as the most challenging, with an accuracy average of 69.2 % along such categories as *false friends* and *multi-word expressions*, with average accuracies of 70.7 % and 69.5 %, respectively. These indicate more challenges on the lexical rather than the syntactical level.

5.3 Phenomenon-level analysis

For **English–German**, the phenomenon-level macro average is 80 %, which is similar to the category-level macro average and the general micro-average. The phenomena with the highest accuracies (> 90 %) are *negative inversion*, *prepositional MWE*, *date*, *substitution*, *adverbial clause*, *infinitive clause*, and *intransitive future I progressive/simple*; for a detailed overview cf. Table 10.

On the other hand, the phenomena with the lowest accuracies (< 65 %) that a lot of LLM-based model struggled with are *stripping*, *idiom*, *onomatopoeia*, *ditransitive future II progressive/simple*, *reflexive future II progressive/simple*, *transitive future II progressive*, and *semantic roles*. It seems that the future II progressive/simple tense is particularly difficult for systems to translate, no matter the verb type. As mentioned above, this is likely due to this verb tense's uncommonness.

Compound	
The police officer was pregnant.	
Сотрудница полиции была беременна.	pass
Полицейский был беременна.	fail
У полицейской была беременность.	fail
Verb Semantics	
She described the book as a page-turner.	
Она описала книгу как -захватывающую историю.	pass
она описала книгу как страницу-поворотчик.	fail
Она описала книгу как перелистывание страниц.	fail

Table 3: Examples of English–Russian linguistic phenomena with passing and failing MT outputs.

Table 2 contains translation examples from English–German. The first example is a test item for the phenomenon *stripping*. *Stripping* is a type of ellipsis. While stripping exists in both German and English, one aspect that can lead to translation errors is punctuation. In English, there is a comma between the two constituents ("John can play the guitar" and "and Mary too"). In German, however, placing a comma in between the constituents is incorrect; see the first and second translation examples. The third translation contains more errors than the additional comma, as it completely changes the meaning of the second constituent. This translation was produced by the Cycle system and also showcases how these kinds of "creative" translations lead to more manual evaluation work: It is easy to write a regular expression for the incorrect output with the comma before the second constituent, and this regular expression will cover most of the outputs of the incorrect system as this is a very common error. However, it is impossible to predict such an incorrect output as it was produced by Cycle, and therefore, it is impossible to write a regular expression to cover cases like this.

The second example is from the phenomenon of *verb semantics*. This phenomenon refers to semantic components in the verb's semantic structure that do not have formal markers. Some examples of these kinds of verbs are *to stride*, *to rumble*, *to stagger*, or *to bawl*, like in the example at hand. There are usually several correct translations for these verbs, as seen in the Table. However, they might lead to translation errors, with systems sometimes not translating them (because they are not so common) or translating them with an incorrect semantic meaning.

When evaluating the performance across phenomena for **English–Russian**, it was found that the following phenomena posed minimal challenges, with many systems achieving near-perfect accuracy: *catenative verb*, *case government*, *conditional*, *contact clause*, *object clauses*, *personal pronoun coreference*, *prepositional mwe*, and *date*. Notably, *personal pronoun coreference*, a new phenomenon added last year that focuses on the consistency in translating the formal and informal "you" across sentences, as well as ensuring that a past tense "I" retains the correct gender ending. This category attained a remarkable accuracy of 96.8 %, marking a 13 % improvement compared to last year. The phenomena with the lowest accuracies (< 60 %) are *verbal MWE*, *resultative*, *gapping*, *compounds*, *idioms* and *semantic roles*.

Table 3 contains translation examples from English–Russian. The original sentences and their translations have been shortened for the paper. The first example involves the common English compound "police officer". Despite the simplicity of this sentence, a closer examination reveals various issues in the translations. In English, the nominal phrase in question is gender-neutral, with no gender marking on nouns, adjectives, or verbs. However, in contrast to English, Russian has gender marking not only on pronouns but also on other parts of speech, including nouns, adjectives, verbs, determiners, and numbers. The first translation correctly uses the collocation сотрудница полиции (literally, "female police employee") and appropriately pairs it with была беременна ("was pregnant"), both in the feminine form. This nominal phrase construction is necessary to convey the gender within the translation. In the second translation, the word полицейский, typically referring to a male police officer is then followed by the verb был (the masculine form of was), and later by the adjective pregnant in the feminine form. This translation error was produced by an LLM and highlights a gap in the model's understanding of gender agreement rules in Russian and a lack of real-world knowledge. The third translation renders the phrase as "the female police officer had pregnancy," which is not a linguistically acceptable Russian collocation. It also uses the adjective полицейская as a job title, which is not a standard noun for "police officer" in Russian.

The next example comes from the phenomenon of Noun formation with the suffix -er. This process is a part of derivational morphology, where new words are formed by adding affixes to existing words or changing their grammatical category or meaning. This is a highly fruitful suffix in English. In the first example translation, we see it rendered as захватывающую историю or "captivating story." This transformation effectively captures the essence of "page-turner." The second translation has страницу-поворотчик – a literal translation. Перелистывание страниц in the third translation describes the physical action of turning pages. The first translation is accurate as it captures the idiomatic meaning of "page-turner"; the other two translations fail due to overly literal interpretations, a common issue in encoder-decoder models and LLMs.

5.4 Comparison with previous years

We have analyzed some of the best-performing systems' development over the years for systems submitted to the WMT repeatedly in the past years. For **English–German**, we took a closer look at GPT-4, Online-B, Online-W, and Online-A, see Table 8. GPT-4 has seen barely any changes in the accuracy from 2023 to 2024 (although it needs to be noted that the prompting method has changed from 5-shot to 3-shot). Online-B, however, shows an improvement of 2.5 percentage points on the macro-average from 2021 to 2024, while the micro-average stayed almost the same throughout this period. Online-W, similarly to GPT-4, shows almost no changes from 2021 to 2024. And finally, Online-A has slight improvements of 1 and 3 percentage points from 2021 to 2024 on the micro-average and macro-average level, respectively.

While in the past years, Online-B and Online-W were usually in the cluster of best-performing systems together, this year, Online-B has surpassed Online-W as only the former is in the best-performing cluster as of this year, while the latter is not. Furthermore, in 2023, GPT-4, Online-W, and Online-B were together in the cluster of best performing systems, while this year, GPT-4 is also not in that cluster anymore.

As for the scored of the micro-average, the phenomenon macro-average, and the category macro-average, while the first two have almost not changed from 2023 to 2024, the category macro-average has improved about 2.5 percentage points from last year to this year. This suggests that the systems for English—German have undergone a slight improvement compared to last year.

Table 9 compares the performance of Yandex, GPT-4, and Online-G for **English–Russian** from 2022 to 2024. This year's Yandex submission is a trained YandexGPT, an LLM-based model. Their approach includes extensive pre-training, fine-tuning, p-tuning, and structure-preserving techniques, which help ensure contextually accurate translations (Elshin et al., 2024). Over the last two years, Yandex's submission has likely undergone a significant update, as reflected in the 2.59 % accuracy increase. Overall, Yandex shows consistent performance with some improvement. GPT-4, another LLM, demonstrates a generally strong performance compared to last year, with a significant drop in the punctuation category (from 100 % to 60 %). Despite this, GPT-4 has either improved or maintained stable performance across most linguistic categories. Online-G, as we suspect based on encoder-decoder methods, exhibits stable performance without any substantial improvements in any areas.

5.5 LLMs vs. encoder-decoder NMT

NMT systems based on an encoder-decoder (or commercial systems that we assume they use this technology) still exhibit better linguistic performance than LLMs in English–German, whereas in English–Russian the first position is shared indeed by two LLMs. In English-German, LLMs seem to perform worse than the two best-performing NMT systems, regarding *punctuation, future verb tenses* and *stripping*. For English-Russian, Yandex is weaker in *named entities and terminology*, while Claude struggles with *function words*, and Unbabel with *verb valency* that includes error-prone phenomena for all LLMs, such as *semantic roles*, *verb semantics*, *resultative*, and *mediopassive voice*. GPT-4 scores even lower than several commercial NMT-based systems. This suggests that while LLMs are indeed taking over the MT in fine-grained analysis, some still struggle to match the capabilities of specialized NMT systems, which are tailored specifically to the target language and potentially trained on more language-specific data.

6 Conclusions and Outlook

In this paper, we apply a linguistically motivated test suite for the first time to evaluate the translation performance of several LLMs as well as several systems with different architectures. Based on the macro-averaged accuracies, the best systems for English-German are Online-B and TranssionMT, with Claude-3.5-Sonnet also sharing the first position based on micro-averaging. For English–Russian, the best-performing systems are Yandex and Claude-3.5-Sonnet. While LLMs generally perform strongly in MT, systems based on encoder-decoder methods, such as TranssionMT and most probably Online-B may still have an edge in certain areas. What the human evaluations of the main MT task reveal about the systems is still to be determined, pending the official announcement of the rankings. The results underscore the potential of LLMs in MT but also highlight areas for improvement.

Limitations

The current test suite was initially designed to evaluate earlier MT systems, featuring a wide range of linguistic phenomena without challenging the models. However, it is becoming increasingly clear that we need to adapt and potentially eliminate the phenomena that have proven too easy for the systems in recent years. The significance of the averaging is unclear, and adding weights depending on the importance of various phenomena is something to consider. While we have introduced context in some cases and complexity with multi-sentence test items in others, this has not been done for all phenomena and sentences so far. One challenge is that we often encounter correct rendering of the phenomena, but then encounter grave errors in the sentence structure. Internally, it has been concluded that these sentences should be marked as incorrect, as the errors are often too significant for the whole output to be considered correct. Additionally, this year, some models generated responses that resembled those of a classical chatbot, including additional explanations or commentary that mixed correct and incorrect translations, making it challenging to evaluate. Going forward, we plan to further refine the test suite to better capture the nuances of modern translation systems.

Acknowledgements

This research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft; DFG) through the project TextQ. We would like to thank Polina Danilovskaia for her help this year, Hans Uszkoreit, Aljoscha Burchardt, Ursula Strohriegel, Renlong Ai, and He Wang for their

prior contributions to the creation of the test suite.

References

Eleftherios Avramidis, Annika Grützner-Zahn, Manuel Brack, Patrick Schramowski, Pedro Ortiz Suarez, Malte Ostendorff, Fabio Barth, Shushen Manakhimova, Vivien Macketanz, Georg Rehm, and Kristian Kersting. 2024. Occiglot at WMT24: European open-source large language models evaluated on translation. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Eleftherios Avramidis and Vivien Macketanz. 2022. Linguistically motivated evaluation of machine translation metrics based on a challenge set. In *Proceedings of the Seventh Conference on Machine Translation*, pages 514–529, Abu Dhabi. Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, Aljoscha Burchardt, and Sebastian Möller. 2020. Fine-grained linguistic evaluation for state-of-the-art Machine Translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 346–356, Online. Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit. 2019. Linguistic Evaluation of German-English Machine Translation Using a Test Suite. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 445–454, Florence, Italy. Association for Computational Linguistics.

Eleftherios Avramidis, Shushen Manakhimova, Vivien Macketanz, and Sebastian Möller. 2023. Challenging the state-of-the-art machine translation metrics from a linguistic perspective. In *Proceedings of the Eighth Conference on Machine Translation*, pages 713–729, Singapore. Association for Computational Linguistics.

Rachel Bawden and Benoît Sagot. 2023. RoCS-MT: Robustness challenge set for machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 198–216, Singapore. Association for Computational Linguistics.

Xiaoyu Chen, Daimeng Wei, Zhanglin Wu, Ting Zhu, Hengchao Shang, Zongyao Li, Jiaxin Guo, Ning Xie, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023. Multifaceted challenge set for evaluating machine translation performance. In *Proceedings of the Eighth Conference on Machine Translation*, pages 217–223, Singapore. Association for Computational Linguistics.

Sören Dreano, Derek Molloy, and Noel Murphy. 2024. Cyclegn: a cycle consistent approach for neural machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Denis Elshin, Nikolay Karpachev, Boris Gruzdev, Ilya Golovanov, Georgy Ivanov, Alexander Antonov, Nickolay Skachkov, Ekaterina Latypova, Vladimir Layner, Ekaterina Enikeeva, Dmitry Popov, Anton Chekashev, Vladislav Negodin, Vera Frantsuzova, Alexander Chernyshev, and Kirill Denisov. 2024. From general LLM to translation: How we dramatically improve translation quality using human evaluation data for LLM finetuning. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popovic, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. 2024. Preliminary WMT24 Ranking of General MT Systems and LLMs.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Samuel Larkin, Chi-kiu Lo, and Rebecca Knowles. 2024. MSLC24 submissions to the general machine translation task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Baohao Liao, Christian Herold, Shahram Khadivi, and Christof Monz. 2024. IKUN for WMT24 general MT task: Llms are here for multilingual machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, and Hans Uszkoreit. 2018. Fine-grained evaluation of German-English machine translation based on a test suite. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 578–587, Belgium, Brussels. Association for Computational Linguistics.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, and Hans Uszkoreit. 2022a. A linguistically motivated test suite to semi-automatically evaluate German–English machine translation output. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 936–947, Marseille, France. European Language Resources Association.

Vivien Macketanz, Eleftherios Avramidis, Shushen Manakhimova, and Sebastian Möller. 2021. Linguistic Evaluation for the 2021 State-of-the-art Machine Translation Systems for German to English and English to German. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1059–1073, Online. Association for Computational Linguistics.

Vivien Macketanz, Shushen Manakhimova, Eleftherios Avramidis, Ekaterina Lapshinova-koltunski, Sergei Bagdasarov, and Sebastian Möller. 2022b. Linguistically motivated evaluation of the 2022 state-of-the-art machine translation systems for three language directions. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 432–449, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can ChatGPT outperform NMT? In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245, Singapore. Association for Computational Linguistics.

Ananya Mukherjee and Manish Shrivastava. 2023. IIIT HYD's submission for WMT23 test-suite task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 246–251, Singapore. Association for Computational Linguistics.

Vladimir Aleksandrovich Mynka and Nikolay Mikhaylovskiy. 2024. TSU HITS's submissions to the WMT 2024 general machine translation shared task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Ricardo Rei, Jose Maria Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. de Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2023 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2023. Test suites task: Evaluation of gender fairness in MT with MuST-SHE and INES. In *Proceedings of the Eighth Conference on Machine Translation*, pages 252–262, Singapore. Association for Computational Linguistics.

Danil Semin and Ondřej Bojar. 2024. CUNI-DS submission: A naive transfer learning setup for english-to-russian translation utilizing english-to-czech data. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, Xin Zhang, Wei Zhang, Dinggang Shen, Tianming Liu, and Shu Zhang. 2024. A comprehensive review of multimodal large language models: Performance and challenges across different tasks.

A Separate systems

category	items	acc.
Ambiguity	5	100.0
Coordination & ellipsis	35	82.9
False friends	12	100.0
Function word	15	100.0
LDD & interrogatives	54	98.1
Lexical Morphology	17	100.0
MWE	31	90.3
Named entity & terminology	39	94.9
Negation	4	75.0
Non-verbal agreement	31	93.5
Punctuation	8	75.0
Subordination	51	94.1
Verb semantics	4	0.0
Verb tense/aspect/mood	875	96.9
Verb valency	31	87.1
micro-average	1212	95.5
macro-average	1212	85.9

Table 4: Accuracies for the translations of the Mistral-Large system (en-de) considering only the non-empty outputs

category	items	acc.
Ambiguity	22	86.4
Coordination & ellipsis	124	60.5
False friends	40	92.5
Function word	40	75.0
LDD & interrogatives	207	76.3
Lexical Morphology	39	61.5
MWE	123	76.4
Named entity & terminology	112	77.7
Negation	18	66.7
Non-verbal agreement	109	87.2
Punctuation	37	51.4
Subordination	191	85.3
Verb semantics	23	60.9
Verb tense/aspect/mood	3249	71.9
Verb valency	114	65.8
micro-average	4448	72.8
macro-average	4448	73.0

Table 5: Accuracies for the translations of the Occiglot system (en-de) considering only the non-empty outputs

B Analysis based on categories

categ	count	Onl-B	Trans	GPT4	Claud	Unbab	Comma	Onl-W	Llama	Aya23	IOLRe	Onl-A	Onl-G	IKUN	CUNIN	IKUNC	NVIDI	Occig	AISTA	TSUHI	MSLC	Mistr	avg
Ambiguity	24	**100.0**	**100.0**	**95.8**	**100.0**	**95.8**	**100.0**	**95.8**	**91.7**	**91.7**	87.5	**91.7**	87.5	79.2	87.5	83.3	70.8	79.2	50.0	37.5	41.7	20.8	80.4
Coordination & ellipsis	83	**94.0**	**94.0**	**90.4**	71.1	73.5	74.7	74.7	80.7	79.5	65.1	73.5	84.3	62.7	78.3	69.9	66.3	57.8	61.4	42.2	26.5	27.7	69.0
False friends	40	**95.0**	**95.0**	**97.5**	**95.0**	**92.5**	**95.0**	**90.0**	**95.0**	**95.0**	**95.0**	**87.5**	82.5	**97.5**	**95.0**	**90.0**	77.5	**92.5**	70.0	55.0	45.0	30.0	84.2
Function word	42	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**95.2**	**100.0**	**95.2**	92.9	92.9	**97.6**	**97.6**	88.1	88.1	92.9	92.9	66.7	88.1	61.9	47.6	33.3	86.6
LDD & interrogatives	160	**97.5**	**97.5**	**96.3**	**97.5**	**96.9**	**93.8**	**95.6**	92.5	91.9	90.0	**95.6**	93.1	80.6	85.6	80.6	85.6	71.9	75.0	57.5	48.8	25.0	83.3
Lexical Morphology	27	**92.6**	**92.6**	**100.0**	**96.3**	**92.6**	85.2	85.2	88.9	77.8	88.9	81.5	81.5	74.1	66.7	66.7	63.0	63.0	37.0	33.3	18.5	48.1	73.0
MWE	109	**97.2**	**96.3**	91.7	**97.2**	89.0	**93.6**	**93.6**	83.5	90.8	90.8	86.2	86.2	83.5	78.0	74.3	66.1	73.4	62.4	40.4	33.0	22.0	77.6
Named entity & terminology	92	**92.4**	**92.4**	**95.7**	**94.6**	89.1	**96.7**	89.1	**93.5**	89.1	90.2	90.2	85.9	81.5	76.1	81.5	80.4	77.2	77.2	65.2	59.8	33.7	82.5
Negation	19	**100.0**	**100.0**	**100.0**	**100.0**	**94.7**	**89.5**	**94.7**	**100.0**	84.2	**100.0**	**100.0**	**100.0**	**94.7**	**89.5**	**100.0**	**100.0**	63.2	**89.5**	**89.5**	**89.5**	15.8	90.2
Non-verbal agreement	97	**97.9**	**97.9**	**99.0**	**99.0**	**99.0**	93.8	92.8	93.8	92.8	93.8	91.8	91.8	90.7	92.8	85.6	79.4	84.5	78.4	78.4	53.6	23.7	86.2
Punctuation	34	**100.0**	**100.0**	88.2	85.3	**94.1**	91.2	**97.1**	88.2	**100.0**	**94.1**	**94.1**	88.2	88.2	73.5	88.2	88.2	50.0	82.4	61.8	67.6	14.7	82.6
Subordination	148	**98.0**	**98.0**	**98.0**	**99.3**	**96.6**	94.6	**96.6**	95.9	93.2	**97.3**	**97.3**	**96.6**	89.2	95.9	85.8	91.9	81.1	87.8	67.6	66.9	23.0	88.1
Verb semantics	18	**83.3**	**83.3**	**72.2**	**72.2**	**88.9**	**77.8**	**66.7**	**77.8**	**83.3**	**72.2**	50.0	**61.1**	**72.2**	**66.7**	55.6	44.4	55.6	22.2	16.7	11.1	0.0	58.7
Verb tense/aspect/mood	3225	**98.2**	**98.2**	97.6	**98.4**	91.4	96.6	96.6	93.6	96.2	94.9	97.1	**98.7**	77.0	80.7	77.4	82.3	67.1	72.5	33.5	42.2	23.8	81.6
Verb valency	101	**91.1**	**91.1**	**86.1**	**88.1**	**88.1**	**84.2**	**86.1**	78.2	**86.1**	**84.2**	**83.2**	76.2	71.3	72.3	75.2	64.4	63.4	54.5	34.7	34.7	16.8	71.9
micro-average	4219	**97.7**	**97.7**	96.8	**97.3**	91.5	95.3	95.3	92.6	94.7	93.5	95.3	96.3	78.1	81.4	78.1	81.3	68.4	72.1	38.6	43.4	24.0	81.4
macro-average	4219	**95.7**	**95.6**	93.7	92.8	92.0	90.8	90.3	89.9	89.6	89.1	87.8	87.4	82.0	81.8	80.5	76.9	69.8	67.2	51.7	45.8	23.9	79.7

Table 6: Accuracies (%) of successful translations on the categorylevel for English–German. The boldface indicates the significantly best-performing systems per row.

categ	count	Yande	Claud	Unbab	Comma	Onl-G	Onl-W	GPT4	IOLRe	Trans	Onl-B	Onl-A	Aya23	IKUN	Llama	IKUNC	CUNID	NVIDI	TSUHI	avg
Ambiguity	20	**90.0**	**95.0**	**90.0**	**90.0**	70.0	70.0	**90.0**	**90.0**	50.0	50.0	55.0	**85.0**	70.0	**75.0**	60.0	65.0	35.0	15.0	69.2
Coordination & ellipsis	86	**87.2**	**80.2**	**84.9**	**82.6**	**84.9**	**83.7**	**76.7**	75.6	72.1	72.1	72.1	**77.9**	74.4	69.8	65.1	72.1	54.7	47.7	74.1
False friends	15	**86.7**	**86.7**	**86.7**	**86.7**	**86.7**	**73.3**	**66.7**	**60.0**	**66.7**	**66.7**	**80.0**	**66.7**	**66.7**	**66.7**	**60.0**	46.7	**66.7**	**53.3**	70.7
Function word	34	**97.1**	88.2	**94.1**	**100.0**	**94.1**	**100.0**	**97.1**	91.2	**94.1**	**94.1**	88.2	**94.1**	85.3	85.3	82.4	73.5	73.5	70.6	89.1
LDD & interrogatives	81	**97.5**	**93.8**	**97.5**	91.4	**96.3**	**95.1**	91.4	90.1	91.4	91.4	85.2	86.4	82.7	80.2	76.5	82.7	70.4	59.3	86.6
Lexical Morphology	41	**97.6**	**92.7**	**90.2**	**92.7**	82.9	73.2	80.5	73.2	75.6	75.6	70.7	68.3	75.6	75.6	63.4	53.7	34.1	26.8	72.4
MWE	96	**87.5**	**84.4**	**78.1**	**83.3**	**80.2**	71.9	77.1	76.0	72.9	71.9	70.8	67.7	69.8	66.7	66.7	52.1	40.6	33.3	69.5
Named entity & terminology	80	83.8	**95.0**	**87.5**	81.3	80.0	80.0	81.3	73.8	80.0	80.0	77.5	71.3	62.5	77.5	60.0	57.5	56.3	41.3	73.7
Non-verbal agreement	98	**94.9**	**95.9**	**91.8**	**93.9**	**90.8**	**89.8**	**90.8**	**92.9**	80.6	80.6	80.6	**92.9**	83.7	86.7	85.7	81.6	73.5	65.3	86.2
Punctuation	13	**92.3**	**92.3**	**92.3**	**100.0**	**92.3**	76.9	61.5	76.9	**84.6**	**84.6**	**92.3**	**84.6**	**84.6**	61.5	**84.6**	**100.0**	**92.3**	76.9	85.0
Subordination	115	**98.3**	**94.8**	**98.3**	88.7	**95.7**	**96.5**	**94.8**	93.0	**94.8**	**94.8**	93.0	86.1	86.1	88.7	83.5	80.0	80.9	67.0	89.7
Verb semantics	20	**100.0**	**90.0**	**95.0**	70.0	**95.0**	85.0	65.0	80.0	85.0	85.0	85.0	65.0	80.0	75.0	70.0	55.0	30.0	35.0	74.7
Verb tense/aspect/mood	169	**87.0**	**90.5**	**90.5**	**87.0**	**89.3**	**88.8**	**89.9**	**85.2**	**85.2**	**85.2**	**86.4**	**85.2**	81.1	**84.6**	**84.6**	78.1	69.2	45.0	82.9
Verb valency	126	**93.7**	**88.1**	83.3	81.0	84.9	86.5	81.7	81.0	83.3	83.3	77.8	75.4	78.6	76.2	73.8	68.3	58.7	50.0	78.1
micro-average	994	**91.8**	**90.4**	89.4	86.8	87.8	86.1	85.2	83.3	82.3	82.2	80.7	80.4	78.1	79.0	75.1	71.0	62.2	50.0	80.1
macro-average	994	**92.4**	**90.5**	90.0	87.7	87.4	83.6	81.7	81.3	79.7	79.7	79.6	79.0	77.2	76.4	72.6	69.0	59.7	49.0	78.7

Table 7: Accuracies (%) of successful translations on the category-level for English–Russian. The boldface indicates the significantly best-performing systems per row.

C Yearly comparison

Category	items	GPT4		onlineB				OnlineW				onlineA			
		2023	2024	2021	2022	2023	2024	2021	2022	2023	2024	2021	2022	2023	2024
Ambiguity	24	95.8	95.8	91.7	91.7	91.7	100	95.8	95.8	95.8	95.8	91.7	87.5	87.5	91.7
Coordination & ellipsis	88	88.6	87.5	80.7	87.5	8.8	89.8	70.5	70.5	73.9	73.9	71.6	80.7	79.5	72.7
False friends	38	97.4	97.4	84.2	89.5	89.5	94.7	89.5	92.1	89.5	89.5	86.8	86.8	86.8	89.5
Function word	41	100	97.6	100	97.6	97.6	95.1	100	100	97.6	97.6	97.6	97.6	97.6	97.6
MWE	104	97.1	95.2	94.2	95.2	95.2	98.1	96.2	97.1	97.1	97.1	86.5	90.4	92.3	94.2
Named entity & termin.	85	95.3	97.6	92.9	97.6	94.1	98.8	95.3	92.9	94.1	94.1	94.1	94.1	92.9	94.1
Negation	18	100	94.4	94.4	100	94.4	100	100	100	94.4	94.4	94.4	100	100	100
Non-verbal agreement	67	100	100	95.5	95.5	95.5	100	95.5	97	95.5	95.5	95.5	95.5	95.5	97
Punctuation	36	86.1	83.3	83.3	83.3	83.3	100	97.2	94.4	97.2	97.2	97.2	97.2	88.9	91.7
Subordination	163	98.8	97.5	97.5	98.8	98.2	95.7	96.9	96.3	96.9	96.9	98.8	98.2	98.8	98.8
Verb tense & aspect/mood	3076	97.9	97.9	99	98.7	97.9	98.1	96.5	96.3	96.6	96.6	96.1	98.5	98.3	97.2
Verb valency	89	87.6	92.1	86.5	87.6	91	95.5	86.5	86.5	88.8	88.8	84.3	87.6	91	92.1
micro-avg	3829	97.3	97.3	97.5	97.7	97.1	97.8	95.6	95.4	95.8	95.8	95	97.2	97.2	96.3
macro-avg	3829	95.4	94.7	91.7	93.6	93.2	97.2	93.3	93.3	93.1	93.1	91.2	92.8	92.4	93.1

Table 8: Yearly comparison of the systems of WMT24 for English-German, based on the category-level analysis

category	items	Yandex		GPT4		OnlineG		
year		2022	2024	2023	2024	2022	2023	2024
Ambiguity	7	85.7	85.7	100	100	85.7	85.7	85.7
Coordination & ellipsis	30	80	70	80	76.7	80	80	80
False friends	5	80	100	100	100	80	80	80
Function word	10	90	90	80	90	90	90	90
MWE	32	71.9	96.9	75	84.4	71.9	71.9	78.1
Named entity & terminology	21	90.5	90.5	76.2	76.2	90.5	95.2	85.7
Non-verbal agreement	11	81.8	90.9	63.6	72.7	81.8	81.8	90.9
Punctuation	5	100	80	100	60	100	100	100
Subordination	28	89.3	92.9	82.1	85.7	89.3	89.3	89.3
Verb tense/aspect/mood	67	74.6	67.2	77.6	88.1	74.6	82.1	79.1
Verb valency	26	84.6	96.2	84.6	84.6	84.6	80.8	80.8
micro-avg	242	81	83.1	79.8	83.9	81	83.1	82.6
macro-avg	242	84.4	87.3	83.6	83.5	84.4	85.2	85.4

Table 9: Yearly comparison of the systems of WMT24 for English-Russian, based on the category-level analysis

D Detailed analysis on a phenomenon-level

categ	count	Onl-B	Trans	GPT4	Claud	Unbab	Comma	Onl-W	Llama	Aya23	IOLRe	Onl-A	Onl-G	IKUN	CUNIN	IKUNC	NVIDI	Occig	AISTA	TSUHI	MSLC	Mistr	avg
Ambiguity	24	**100.0**	**100.0**	**95.8**	**100.0**	**95.8**	**100.0**	**95.8**	**91.7**	**91.7**	87.5	**91.7**	87.5	79.2	87.5	83.3	70.8	79.2	50.0	37.5	41.7	20.8	80.4
Lexical ambiguity	24	**100.0**	**100.0**	**95.8**	**100.0**	**95.8**	**100.0**	**95.8**	**91.7**	**91.7**	87.5	**91.7**	87.5	79.2	87.5	83.3	70.8	79.2	50.0	37.5	41.7	20.8	80.4
Coordination & ellipsis	83	**94.0**	**94.0**	**90.4**	71.1	73.5	74.7	74.7	80.7	79.5	65.1	73.5	84.3	62.7	78.3	69.9	66.3	57.8	61.4	42.2	26.5	27.7	69.0
Gapping	15	**86.7**	**86.7**	**93.3**	66.7	60.0	80.0	73.3	73.3	**86.7**	46.7	**86.7**	**100.0**	53.3	**93.3**	**86.7**	**93.3**	33.3	73.3	33.3	26.7	40.0	70.2
Pseudogapping	7	**100.0**	**100.0**	**100.0**	**71.4**	**100.0**	**71.4**	**71.4**	**85.7**	**100.0**	**71.4**	**85.7**	57.1	**71.4**	**100.0**	**100.0**	42.9	**100.0**	42.9	42.9	0.0	0.0	72.1
Right node raising	11	**90.9**	**90.9**	**90.9**	**90.9**	**81.8**	**100.0**	**90.9**	**100.0**	**100.0**	**100.0**	**90.9**	**81.8**	72.7	**90.9**	72.7	**81.8**	**81.8**	**81.8**	27.3	45.5	27.3	80.5
Sluicing	18	**100.0**	**100.0**	**100.0**	**94.4**	**94.4**	77.8	**100.0**	**94.4**	**100.0**	**88.9**	77.8	77.8	77.8	66.7	77.8	72.2	72.2	77.8	55.6	33.3	27.8	79.4
Stripping	21	**90.5**	**90.5**	**81.0**	38.1	52.4	47.6	42.9	57.1	47.6	38.1	42.9	**85.7**	42.9	61.9	47.6	42.9	33.3	47.6	42.9	23.8	23.8	51.5
VP-ellipsis	11	**100.0**	**100.0**	**81.8**	**81.8**	72.7	**90.9**	**81.8**	**90.9**	63.6	63.6	**81.8**	**90.9**	72.7	**81.8**	54.5	63.6	63.6	36.4	45.5	18.2	36.4	70.1
False friends	40	**95.0**	**95.0**	**97.5**	**95.0**	**92.5**	**95.0**	**90.0**	**95.0**	**95.0**	**95.0**	**87.5**	82.5	**97.5**	**95.0**	**90.0**	77.5	**92.5**	70.0	55.0	45.0	30.0	84.2
Function word	42	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**95.2**	**100.0**	**95.2**	92.9	92.9	**97.6**	**97.6**	88.1	88.1	92.9	92.9	66.7	88.1	61.9	47.6	33.3	86.6
Focus particle	23	**95.7**	**95.7**	**95.7**	**95.7**	**95.7**	**95.7**	**100.0**	**95.7**	87.0	**95.7**	**95.7**	**95.7**	**95.7**	**91.3**	87.0	**91.3**	73.9	**95.7**	82.6	87.0	39.1	89.9
Question tag	19	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.7**	**100.0**	**94.7**	**100.0**	**89.5**	**100.0**	**100.0**	78.9	84.2	**100.0**	**94.7**	57.9	78.9	36.8	0.0	26.3	82.7
LDD & interrogatives	160	**97.5**	**97.5**	**96.3**	**97.5**	**96.9**	**93.8**	**95.6**	92.5	91.9	90.0	**95.6**	93.1	80.6	85.6	80.6	85.6	71.9	75.0	57.5	48.8	25.0	83.3
Extraposition	16	**93.8**	**93.8**	**87.5**	**81.3**	**81.3**	**93.8**	**87.5**	**81.3**	**75.0**	56.3	**75.0**	62.5	**81.3**	62.5	**75.0**	62.5	50.0	43.8	37.5	12.5	12.5	67.0
Inversion	14	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**92.9**	**85.7**	**92.9**	**85.7**	**85.7**	71.4	28.6	42.9	0.0	85.0
Multiple connectors	16	**93.8**	**93.8**	**100.0**	**100.0**	**100.0**	**93.8**	**100.0**	**93.8**	**100.0**	**100.0**	**100.0**	**100.0**	75.0	**87.5**	81.3	**93.8**	81.3	**93.8**	81.3	**87.5**	31.3	89.9
Negative inversion	14	**92.9**	**92.9**	**92.9**	**92.9**	**100.0**	**92.9**	**100.0**	**92.9**	**92.9**	**100.0**	**100.0**	**100.0**	**92.9**	**92.9**	**100.0**	**92.9**	78.6	**100.0**	**92.9**	**85.7**	21.4	90.8
Pied-piping	14	**100.0**	**100.0**	**92.9**	**100.0**	**100.0**	**85.7**	**92.9**	**92.9**	**92.9**	**92.9**	**100.0**	**100.0**	78.6	**92.9**	71.4	**100.0**	71.4	78.6	57.1	57.1	21.4	84.7
Polar question	18	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.4**	**100.0**	**100.0**	72.2	**100.0**	77.8	**94.4**	72.2	72.2	66.7	38.9	33.3	86.8
Preposition stranding	16	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	81.3	**100.0**	81.3	75.0	75.0	**100.0**	**100.0**	62.5	56.3	75.0	**100.0**	50.0	75.0	43.8	0.0	56.3	77.7
Split infinitive	10	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**80.0**	**90.0**	**100.0**	**80.0**	**90.0**	**80.0**	40.0	**80.0**	30.0	89.0
Topicalization	12	**100.0**	**100.0**	**91.7**	**100.0**	**100.0**	**91.7**	**83.3**	**91.7**	**91.7**	**91.7**	**83.3**	66.7	75.0	**83.3**	50.0	50.0	**83.3**	41.7	50.0	25.0	8.3	74.2
Wh-movement	30	**96.7**	**96.7**	**96.7**	**100.0**	**93.3**	**96.7**	**93.3**	**93.3**	**93.3**	**93.3**	**96.7**	**96.7**	90.0	**96.7**	83.3	86.7	70.0	83.3	63.3	60.0	26.7	86.0
Lexical Morphology	27	**92.6**	**92.6**	**100.0**	**96.3**	**92.6**	85.2	85.2	88.9	77.8	88.9	81.5	81.5	74.1	66.7	66.7	63.0	63.0	37.0	33.3	18.5	48.1	73.0
Functional shift	14	**92.9**	**92.9**	**100.0**	**100.0**	**100.0**	**85.7**	**92.9**	**85.7**	**85.7**	**85.7**	78.6	78.6	**92.9**	57.1	**85.7**	71.4	57.1	35.7	35.7	21.4	64.3	76.2
Noun formation (er)	13	**92.3**	**92.3**	**100.0**	**92.3**	**84.6**	**84.6**	76.9	**92.3**	69.2	**92.3**	**84.6**	**84.6**	53.8	76.9	46.2	53.8	69.2	38.5	30.8	15.4	30.8	69.6
MWE	109	**97.2**	**96.3**	91.7	**97.2**	89.0	**93.6**	**93.6**	83.5	90.8	90.8	86.2	86.2	83.5	78.0	74.3	66.1	73.4	62.4	40.4	33.0	22.0	77.6
Collocation	17	**100.0**	**94.1**	**100.0**	**100.0**	**88.2**	**100.0**	**100.0**	**88.2**	**94.1**	**88.2**	**100.0**	**94.1**	82.4	76.5	58.8	76.5	**94.1**	70.6	41.2	35.3	35.3	81.8
Compound	12	**100.0**	**100.0**	**100.0**	**100.0**	**91.7**	**100.0**	**100.0**	**83.3**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	66.7	**91.7**	66.7	58.3	8.3	88.9
Idiom	19	**89.5**	**89.5**	**63.2**	**84.2**	**68.4**	**73.7**	**73.7**	**63.2**	**68.4**	**68.4**	31.6	31.6	57.9	47.4	47.4	0.0	31.6	**5.3**	0.0	0.0	**5.3**	47.6
Nominal MWE	18	**100.0**	**100.0**	**94.4**	**100.0**	**88.9**	**100.0**	**94.4**	77.8	83.3	**88.9**	**100.0**	**94.4**	83.3	77.8	61.1	77.8	77.8	66.7	33.3	27.8	16.7	78.3
Prepositional MWE	16	**100.0**	**100.0**	**93.8**	**100.0**	**93.8**	**93.8**	**100.0**	**93.8**	**100.0**	**100.0**	**100.0**	**100.0**	**87.5**	**100.0**	**93.8**	**93.8**	81.3	**93.8**	**87.5**	56.3	37.5	90.8
Verbal MWE	27	**96.3**	**96.3**	**100.0**	**100.0**	**100.0**	**96.3**	**96.3**	**92.6**	**100.0**	**100.0**	**92.6**	**100.0**	**92.6**	77.8	88.9	66.7	85.2	63.0	33.3	33.3	25.9	82.7
Named entity & terminology	92	**92.4**	**92.4**	**95.7**	**94.6**	89.1	**96.7**	89.1	**93.5**	89.1	90.2	90.2	85.9	81.5	76.1	81.5	80.4	77.2	77.2	65.2	59.8	33.7	82.5
Date	13	**100.0**	**100.0**	**100.0**	**100.0**	**92.3**	**100.0**	**100.0**	**100.0**	**92.3**	**92.3**	**92.3**	**100.0**	**92.3**	**92.3**	**92.3**	**92.3**	61.5	76.9	**92.3**	**92.3**	30.8	90.1
Domainspecific Term	16	**87.5**	**87.5**	**87.5**	**87.5**	**87.5**	**93.8**	**87.5**	**87.5**	**87.5**	**81.3**	**93.8**	**68.8**	**75.0**	**68.8**	**81.3**	**81.3**	**87.5**	56.3	31.3	37.5	43.8	76.2
Location	18	**100.0**	**100.0**	**100.0**	**100.0**	**94.4**	**100.0**	**94.4**	**100.0**	**94.4**	**100.0**	**94.4**	**100.0**	83.3	77.8	**88.9**	**88.9**	83.3	**94.4**	72.2	83.3	38.9	89.9
Measuring unit	19	**94.7**	**94.7**	**94.7**	**94.7**	**100.0**	**100.0**	**89.5**	**100.0**	**94.7**	**94.7**	**94.7**	**94.7**	84.2	84.2	**89.5**	84.2	73.7	**94.7**	73.7	42.1	21.1	85.5
Onomatopeia	7	**57.1**	**57.1**	**85.7**	**85.7**	**42.9**	**85.7**	**42.9**	**85.7**	**57.1**	**42.9**	**42.9**	**42.9**	**28.6**	**14.3**	**14.3**	0.0	**42.9**	0.0	0.0	0.0	0.0	39.5
Proper name	19	**94.7**	**94.7**	**100.0**	**94.7**	**89.5**	**94.7**	**94.7**	84.2	**89.5**	**100.0**	**94.7**	84.2	**94.7**	84.2	84.2	**89.5**	**89.5**	**89.5**	84.2	73.7	47.4	88.2
Negation	19	**100.0**	**100.0**	**100.0**	**100.0**	**94.7**	**89.5**	**94.7**	**100.0**	84.2	**100.0**	**100.0**	**100.0**	**94.7**	**89.5**	**100.0**	**100.0**	63.2	**89.5**	**89.5**	**89.5**	15.8	90.2
Non-verbal agreement	97	**97.9**	**97.9**	**99.0**	**99.0**	**99.0**	93.8	92.8	93.8	92.8	93.8	91.8	91.8	90.7	92.8	85.6	79.4	84.5	78.4	78.4	53.6	23.7	86.2
Coreference	30	**96.7**	**96.7**	**100.0**	**100.0**	**100.0**	**100.0**	**93.3**	**96.7**	**96.7**	**100.0**	**93.3**	90.0	**93.3**	**93.3**	**93.3**	80.0	90.0	70.0	90.0	43.3	16.7	87.3
Genitive	19	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**89.5**	**94.7**	84.2	**89.5**	84.2	**94.7**	**100.0**	78.9	84.2	68.4	52.6	78.9	73.7	57.9	31.6	31.6	80.7
Personal Pronoun Coreference	12	**91.7**	**91.7**	**100.0**	**100.0**	**100.0**	**91.7**	**83.3**	**91.7**	**91.7**	**91.7**	66.7	66.7	**91.7**	**91.7**	**100.0**	75.0	**83.3**	**83.3**	75.0	58.3	33.3	83.7
Possession	26	**100.0**	**100.0**	**96.2**	**96.2**	**96.2**	**92.3**	**92.3**	**96.2**	**92.3**	**92.3**	**100.0**	**100.0**	**92.3**	**96.2**	76.9	**96.2**	76.9	88.5	76.9	80.8	23.1	88.6
Substitution	10	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**90.0**	**100.0**	**100.0**	**90.0**	**100.0**	**90.0**	**90.0**	**100.0**	**100.0**	**100.0**	**90.0**	**100.0**	**80.0**	**90.0**	50.0	20.0	90.0
Punctuation	34	**100.0**	**100.0**	88.2	85.3	**94.1**	91.2	**97.1**	88.2	**100.0**	**94.1**	**94.1**	88.2	88.2	73.5	88.2	88.2	50.0	82.4	61.8	67.6	14.7	82.6
Quotation marks	34	**100.0**	**100.0**	88.2	85.3	**94.1**	91.2	**97.1**	88.2	**100.0**	**94.1**	**94.1**	88.2	88.2	73.5	88.2	88.2	50.0	82.4	61.8	67.6	14.7	82.6
Subordination	148	**98.0**	**98.0**	**98.0**	**99.3**	**96.6**	94.6	**96.6**	95.9	93.2	**97.3**	**97.3**	**96.6**	89.2	95.9	85.8	91.9	81.1	87.8	67.6	66.9	23.0	88.1

categ	count	Onl-B	Trans	GPT4	Claud	Unbab	Comma	Onl-W	Llama	Aya23	IOLRe	Onl-A	Onl-G	IKUN	CUNIN	IKUNC	NVIDI	Occig	AISTA	TSUHI	MSLC	Mistr	avg
Adverbial clause	6	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**66.7**	**100.0**	**100.0**	**100.0**	**83.3**	**83.3**	33.3	93.7
Cleft sentence	12	**91.7**	**91.7**	**100.0**	**100.0**	**91.7**	**83.3**	**100.0**	**91.7**	**91.7**	**100.0**	**100.0**	**91.7**	**91.7**	**91.7**	**83.3**	75.0	66.7	**83.3**	75.0	75.0	25.0	85.7
Contact clause	21	**95.2**	**95.2**	**95.2**	**100.0**	**100.0**	**95.2**	**95.2**	**100.0**	**100.0**	**100.0**	**100.0**	**95.2**	81.0	**100.0**	81.0	85.7	61.9	71.4	61.9	42.9	14.3	84.4
Indirect speech	16	**100.0**	**100.0**	**93.8**	**100.0**	**93.8**	**87.5**	**87.5**	**93.8**	**93.8**	**100.0**	**93.8**	**100.0**	**93.8**	**93.8**	81.3	**100.0**	**87.5**	**100.0**	68.8	**87.5**	18.8	89.3
Infinitive clause	16	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**93.8**	**100.0**	**100.0**	**93.8**	**100.0**	**100.0**	**100.0**	**93.8**	**93.8**	**87.5**	**93.8**	81.3	**87.5**	75.0	**87.5**	25.0	91.1
Object clause	16	**100.0**	**100.0**	**100.0**	**93.8**	**93.8**	**100.0**	**100.0**	**87.5**	**87.5**	**93.8**	**93.8**	**87.5**	75.0	**100.0**	**93.8**	**87.5**	**93.8**	**87.5**	62.5	68.8	18.8	86.9
Pseudo-cleft sentence	14	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**92.9**	**100.0**	**100.0**	**100.0**	**100.0**	**85.7**	**92.9**	78.6	**100.0**	**85.7**	**92.9**	50.0	71.4	14.3	88.8
Relative clause	34	**97.1**	**97.1**	**97.1**	**100.0**	**97.1**	**94.1**	**97.1**	**97.1**	88.2	**94.1**	**97.1**	**97.1**	**97.1**	**94.1**	**94.1**	**94.1**	85.3	88.2	76.5	61.8	26.5	89.1
Subject clause	13	**100.0**	**100.0**	**100.0**	**100.0**	**92.3**	**100.0**	**92.3**	**100.0**	**92.3**	**92.3**	**92.3**	**100.0**	**84.6**	**100.0**	**84.6**	**92.3**	76.9	**92.3**	53.8	46.2	38.5	87.2
Verb semantics	18	**83.3**	**83.3**	**72.2**	**72.2**	**88.9**	**77.8**	**66.7**	**77.8**	**83.3**	**72.2**	50.0	**61.1**	**72.2**	**66.7**	55.6	44.4	55.6	22.2	16.7	11.1	0.0	58.7
Verb tense/aspect/mood	3225	**98.2**	**98.2**	97.6	**98.4**	91.4	96.6	96.6	93.6	96.2	94.9	97.1	**98.7**	77.0	80.7	77.4	82.3	67.1	72.5	33.5	42.2	23.8	81.6
Conditional	19	**94.7**	**94.7**	**94.7**	**94.7**	**94.7**	**94.7**	**89.5**	**89.5**	**94.7**	**94.7**	**94.7**	**94.7**	**89.5**	**94.7**	**84.2**	**84.2**	68.4	**78.9**	42.1	63.2	26.3	83.7
Ditransitive - conditional I progressive	60	**100.0**	**100.0**	**98.3**	95.0	93.3	95.0	**100.0**	**100.0**	**98.3**	**96.7**	95.0	**100.0**	73.3	86.7	85.0	75.0	63.3	58.3	30.0	6.7	15.0	79.3
Ditransitive - conditional I simple	52	**100.0**	**100.0**	**100.0**	**98.1**	86.5	**98.1**	**100.0**	**100.0**	**100.0**	**100.0**	57.7	**100.0**	40.4	38.5	71.2	67.3	71.2	57.7	17.3	5.8	21.2	72.9
Ditransitive - conditional II progressive	59	**100.0**	**100.0**	94.9	**100.0**	76.3	**100.0**	**98.3**	91.5	**100.0**	**100.0**	**98.3**	**100.0**	67.8	83.1	57.6	83.1	45.8	76.3	42.4	10.2	30.5	78.9
Ditransitive - conditional II simple	59	**100.0**	**100.0**	**100.0**	**100.0**	86.4	**100.0**	**100.0**	**98.3**	**100.0**	**100.0**	**100.0**	**100.0**	78.0	**98.3**	79.7	88.1	71.2	71.2	35.6	27.1	35.6	84.3
Ditransitive - future I progressive	57	**100.0**	**100.0**	94.7	**96.5**	89.5	**96.5**	**100.0**	**96.5**	**98.2**	**98.2**	**96.5**	**100.0**	70.2	86.0	86.0	86.0	77.2	80.7	54.4	21.1	19.3	83.2
Ditransitive - future I simple	112	**100.0**	**100.0**	**99.1**	95.5	87.5	95.5	**100.0**	**99.1**	**100.0**	**99.1**	94.6	**100.0**	87.5	75.9	89.3	76.8	67.9	80.4	33.0	24.1	28.6	82.6
Ditransitive - future II progressive	52	94.2	94.2	94.2	**100.0**	53.8	**96.2**	88.5	86.5	90.4	80.8	**100.0**	**98.1**	0.0	34.6	5.8	26.9	28.8	3.8	0.0	1.9	26.9	57.4
Ditransitive - future II simple	57	**98.2**	**98.2**	**96.5**	**100.0**	59.6	**100.0**	**100.0**	94.7	87.7	91.2	**98.2**	**100.0**	0.0	43.9	7.0	52.6	26.3	7.0	0.0	1.8	28.1	61.5
Ditransitive - past perfect progressive	53	92.5	92.5	**98.1**	**100.0**	86.8	92.5	92.5	71.7	83.0	83.0	**100.0**	**100.0**	26.4	67.9	43.4	88.7	54.7	73.6	47.2	22.6	24.5	73.4
Ditransitive - past perfect simple	56	**98.2**	**98.2**	**96.4**	**100.0**	89.3	92.9	**96.4**	80.4	87.5	82.1	**100.0**	**100.0**	51.8	69.6	73.2	76.8	58.9	76.8	39.3	46.4	21.4	77.9
Ditransitive - past progressive	54	**100.0**	**100.0**	92.6	**100.0**	92.6	**96.3**	85.2	88.9	**96.3**	**96.3**	**100.0**	**100.0**	72.2	83.3	**98.1**	77.8	72.2	72.2	33.3	24.1	20.4	81.0
Ditransitive - present perfect progressive	56	**100.0**	**100.0**	92.9	**98.2**	89.3	**100.0**	**100.0**	87.5	**96.4**	**100.0**	**96.4**	**100.0**	60.7	**96.4**	66.1	89.3	71.4	48.2	30.4	28.6	12.5	79.3
Ditransitive - present perfect simple	57	**100.0**	**100.0**	91.2	**100.0**	93.0	**100.0**	**96.5**	**96.5**	**100.0**	**100.0**	**100.0**	**100.0**	78.9	94.7	93.0	94.7	71.9	77.2	42.1	47.4	22.8	85.7
Ditransitive - present progressive	59	**96.6**	**96.6**	86.4	91.5	94.9	93.2	**96.6**	91.5	**100.0**	89.8	**98.3**	**100.0**	74.6	86.4	93.2	71.2	78.0	64.4	50.8	16.9	33.9	81.2
Ditransitive - simple past	79	**98.7**	**98.7**	96.2	**97.5**	94.9	**97.5**	**97.5**	94.9	**100.0**	94.9	**98.7**	**100.0**	75.9	84.8	93.7	92.4	74.7	86.1	41.8	22.8	22.8	84.0
Ditransitive - simple present	55	**98.2**	**98.2**	90.9	**96.4**	**98.2**	**96.4**	**98.2**	87.3	**100.0**	92.7	**96.4**	**100.0**	72.7	81.8	92.7	83.6	70.9	78.2	60.0	10.9	12.7	81.7
Gerund	24	**95.8**	**95.8**	**95.8**	**95.8**	**100.0**	**100.0**	**100.0**	83.3	**100.0**	**95.8**	**100.0**	**100.0**	**91.7**	79.2	83.3	87.5	70.8	66.7	45.8	41.7	37.5	84.1
Imperative	14	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	78.6	**100.0**	**100.0**	**92.9**	**85.7**	**92.9**	**92.9**	57.1	57.1	71.4	71.4	28.6	71.4	28.6	14.3	35.7	75.2
Intransitive - conditional I progressive	29	**100.0**	**100.0**	**100.0**	**100.0**	**93.1**	**100.0**	89.7	**96.6**	**100.0**	**100.0**	**96.6**	**100.0**	79.3	**100.0**	82.8	**100.0**	82.8	**96.6**	41.4	86.2	37.9	89.7
Intransitive - conditional I simple	27	**92.6**	**92.6**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**96.3**	**100.0**	**100.0**	85.2	**100.0**	81.5	70.4	81.5	**100.0**	88.9	**92.6**	74.1	77.8	22.2	88.4
Intransitive - conditional II progressive	29	**100.0**	**100.0**	**100.0**	**100.0**	**96.6**	**100.0**	**96.6**	82.8	**100.0**	**96.6**	**100.0**	**100.0**	89.7	51.7	69.0	58.6	62.1	58.6	6.9	13.8	24.1	76.5
Intransitive - conditional II simple	29	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	82.8	**100.0**	89.7	**100.0**	**100.0**	89.7	62.1	69.0	62.1	69.0	82.8	41.4	24.1	20.7	80.6
Intransitive - future I progressive	30	**96.7**	**96.7**	**100.0**	**96.7**	**100.0**	**100.0**	86.7	90.0	**96.7**	**100.0**	90.0	**100.0**	**93.3**	**100.0**	**100.0**	**100.0**	83.3	**93.3**	50.0	**96.7**	20.0	90.0
Intransitive - future I simple	69	**98.6**	**98.6**	**97.1**	**97.1**	**98.6**	**100.0**	89.9	95.7	**97.1**	**100.0**	95.7	**98.6**	**98.6**	**98.6**	92.8	**100.0**	88.4	**100.0**	60.9	84.1	20.3	91.0
Intransitive - future II progressive	27	**100.0**	**100.0**	**100.0**	**100.0**	**96.3**	**96.3**	88.9	**96.3**	**100.0**	**100.0**	**100.0**	**100.0**	14.8	74.1	29.6	77.8	48.1	18.5	3.7	0.0	22.2	69.8
Intransitive - future II simple	31	**100.0**	**100.0**	**100.0**	**100.0**	**93.5**	80.6	**96.8**	**100.0**	90.3	90.3	**100.0**	**100.0**	19.4	61.3	16.1	80.6	48.4	38.7	35.5	16.1	16.1	70.7
Intransitive - past perfect progressive	24	87.5	87.5	**100.0**	**100.0**	**100.0**	87.5	79.2	79.2	70.8	87.5	**100.0**	**95.8**	70.8	75.0	75.0	87.5	50.0	83.3	12.5	8.3	16.7	74.0
Intransitive - past perfect simple	34	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	91.2	**94.1**	82.4	79.4	76.5	**100.0**	**100.0**	85.3	**94.1**	85.3	88.2	73.5	88.2	38.2	58.8	17.6	83.5
Intransitive - past progressive	32	90.6	90.6	**100.0**	**100.0**	**96.9**	**96.9**	**100.0**	**100.0**	**96.9**	**100.0**	**96.9**	**93.8**	**96.9**	87.5	90.6	37.5	75.0	84.4	21.9	12.5	37.5	81.3
Intransitive - present perfect progressive	25	**100.0**	**100.0**	**100.0**	**96.0**	80.0	**92.0**	**100.0**	88.0	**100.0**	**92.0**	**100.0**	**100.0**	**100.0**	80.0	**92.0**	68.0	80.0	72.0	32.0	24.0	0.0	80.8
Intransitive - present perfect simple	30	**100.0**	**100.0**	**100.0**	**100.0**	**93.3**	**100.0**	**100.0**	86.7	**100.0**	90.0	**100.0**	**100.0**	**100.0**	**100.0**	**93.3**	83.3	70.0	86.7	66.7	33.3	13.3	86.5
Intransitive - present progressive	64	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	93.8	**100.0**	**100.0**	**98.4**	**100.0**	**100.0**	95.3	**98.4**	95.3	93.8	79.7	95.3	31.3	51.6	29.7	88.7
Intransitive - simple past	39	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**97.4**	**100.0**	**100.0**	**97.4**	**100.0**	**100.0**	**97.4**	**100.0**	84.6	69.2	87.2	53.8	17.9	20.5	86.9
Intransitive - simple present	38	**100.0**	**100.0**	**100.0**	**100.0**	**97.4**	**97.4**	**100.0**	**97.4**	**100.0**	**100.0**	**100.0**	**100.0**	**97.4**	**97.4**	**94.7**	**94.7**	65.8	81.6	55.3	71.1	18.4	89.0
Modal	294	98.6	98.6	98.6	99.0	98.3	99.0	98.6	98.0	**99.7**	**100.0**	**99.3**	**99.7**	96.9	92.5	94.2	95.9	65.0	94.2	36.4	87.4	23.5	89.2
Modal negated	288	**98.3**	**98.3**	95.8	**98.3**	96.5	92.4	**99.7**	97.2	96.2	93.1	97.9	96.5	91.3	88.2	88.9	89.9	71.2	95.1	56.3	83.0	22.9	87.9
Reflexive - conditional I progressive	34	**100.0**	**100.0**	**100.0**	**100.0**	**97.1**	**100.0**	**100.0**	**100.0**	**100.0**	91.2	**100.0**	**97.1**	**97.1**	**94.1**	73.5	88.2	70.6	58.8	5.9	35.3	8.8	81.8
Reflexive - conditional I simple	28	**100.0**	**100.0**	**100.0**	**100.0**	**92.9**	**100.0**	**96.4**	**100.0**	**100.0**	**96.4**	78.6	**100.0**	**96.4**	60.7	**92.9**	78.6	60.7	75.0	28.6	35.7	14.3	81.3
Reflexive - conditional II progressive	31	**100.0**	**100.0**	**100.0**	**100.0**	71.0	**93.5**	**93.5**	**96.8**	87.1	**100.0**	**100.0**	90.3	74.2	45.2	16.1	80.6	54.8	64.5	3.2	35.5	32.3	73.3
Reflexive - conditional II simple	35	**100.0**	**100.0**	**100.0**	**100.0**	**94.3**	**100.0**	**100.0**	**97.1**	**100.0**	**100.0**	**100.0**	**94.3**	80.0	91.4	62.9	85.7	71.4	57.1	5.7	37.1	34.3	81.5
Reflexive - future I progressive	32	**93.8**	**93.8**	**100.0**	**96.9**	90.6	**96.9**	**96.9**	**96.9**	**93.8**	**96.9**	**96.9**	**100.0**	**93.8**	75.0	84.4	84.4	50.0	59.4	25.0	34.4	31.3	80.5
Reflexive - future I simple	68	**98.5**	**98.5**	**100.0**	**98.5**	88.2	**100.0**	**97.1**	95.6	**100.0**	**98.5**	**100.0**	**97.1**	82.4	82.4	82.4	94.1	50.0	83.8	33.8	27.9	26.5	82.6

categ	count	Onl-B	Trans	GPT4	Claud	Unbab	Comma	Onl-W	Llama	Aya23	IOLRe	Onl-A	Onl-G	IKUN	CUNIN	IKUNC	NVIDI	Occig	AISTA	TSUHI	MSLC	Mistr	avg
Reflexive - future II progressive	29	**93.1**	**93.1**	**100.0**	**100.0**	48.3	**96.6**	**96.6**	**93.1**	**93.1**	**96.6**	**96.6**	**100.0**	0.0	34.5	6.9	27.6	34.5	6.9	0.0	0.0	31.0	59.4
Reflexive - future II simple	33	81.8	81.8	**100.0**	**100.0**	81.8	**97.0**	**100.0**	90.9	**97.0**	**93.9**	**93.9**	**100.0**	3.0	63.6	9.1	54.5	39.4	36.4	0.0	3.0	15.2	63.9
Reflexive - past perfect progressive	29	**93.1**	**93.1**	**100.0**	**100.0**	79.3	**100.0**	89.7	62.1	65.5	69.0	**100.0**	**100.0**	79.3	72.4	55.2	69.0	44.8	58.6	17.2	34.5	27.6	71.9
Reflexive - past perfect simple	27	**100.0**	**100.0**	**96.3**	**100.0**	**96.3**	**96.3**	**92.6**	74.1	70.4	77.8	**100.0**	**100.0**	85.2	74.1	63.0	74.1	63.0	70.4	3.7	48.1	14.8	76.2
Reflexive - past progressive	32	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**96.9**	**96.9**	**93.8**	**100.0**	90.6	84.4	87.5	87.5	68.8	81.3	71.9	50.0	37.5	21.9	84.2
Reflexive - present perfect progressive	26	**100.0**	**100.0**	**100.0**	**100.0**	**92.3**	**96.2**	**92.3**	**92.3**	**100.0**	**100.0**	**100.0**	**100.0**	**92.3**	88.5	80.8	**100.0**	69.2	50.0	23.1	34.6	30.8	83.0
Reflexive - present perfect simple	32	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	75.0	90.6	**96.9**	75.0	71.9	31.3	37.5	9.4	85.1
Reflexive - present progressive	32	**96.9**	**96.9**	**100.0**	**100.0**	87.5	**96.9**	**100.0**	90.6	**96.9**	90.6	**96.9**	**93.8**	87.5	78.1	90.6	84.4	75.0	84.4	15.6	53.1	43.8	83.8
Reflexive - simple past	32	**96.9**	**96.9**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**93.8**	84.4	87.5	**96.9**	90.6	81.3	78.1	21.9	34.4	25.0	85.1
Reflexive - simple present	32	**96.9**	**96.9**	**100.0**	**100.0**	81.3	**100.0**	**100.0**	75.0	**93.8**	81.3	**96.9**	90.6	84.4	78.1	**93.8**	81.3	59.4	78.1	31.3	37.5	18.8	79.8
Transitive - future II progressive	30	**100.0**	**100.0**	**100.0**	**100.0**	83.3	**100.0**	86.7	**100.0**	**100.0**	86.7	**100.0**	**100.0**	3.3	53.3	3.3	46.7	36.7	3.3	0.0	0.0	13.3	62.7
Transitive - conditional I progressive	28	**96.4**	**96.4**	**100.0**	**100.0**	**96.4**	**92.9**	82.1	**96.4**	**100.0**	**100.0**	**100.0**	**100.0**	89.3	82.1	67.9	78.6	89.3	57.1	7.1	35.7	28.6	80.8
Transitive - conditional I simple	18	**100.0**	**100.0**	**100.0**	**100.0**	**94.4**	**94.4**	**88.9**	**100.0**	**100.0**	**100.0**	55.6	**100.0**	83.3	50.0	61.1	**88.9**	72.2	50.0	27.8	11.1	22.2	76.2
Transitive - conditional II progressive	27	**100.0**	**100.0**	**96.3**	**100.0**	**92.6**	**96.3**	88.9	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	70.4	59.3	85.2	88.9	74.1	7.4	44.4	33.3	82.7
Transitive - conditional II simple	30	**100.0**	**100.0**	**100.0**	**100.0**	**96.7**	**96.7**	**96.7**	**93.3**	**100.0**	**93.3**	**100.0**	**100.0**	**93.3**	76.7	**93.3**	**96.7**	66.7	73.3	16.7	66.7	23.3	84.9
Transitive - future I progressive	30	**100.0**	**100.0**	**96.7**	**96.7**	**93.3**	**93.3**	83.3	**93.3**	**100.0**	**100.0**	**100.0**	**100.0**	86.7	80.0	90.0	86.7	73.3	70.0	23.3	36.7	23.3	82.2
Transitive - future I simple	57	**100.0**	**100.0**	94.7	**96.5**	94.7	93.0	**96.5**	**96.5**	**100.0**	**98.2**	**100.0**	**100.0**	93.0	94.7	94.7	89.5	80.7	71.9	24.6	52.6	22.8	85.5
Transitive - future II simple	35	**97.1**	**97.1**	**100.0**	**100.0**	**94.3**	**97.1**	**100.0**	**97.1**	**97.1**	**100.0**	**100.0**	**100.0**	14.3	57.1	20.0	77.1	60.0	5.7	14.3	0.0	40.0	69.9
Transitive - past perfect progressive	24	**91.7**	**91.7**	**95.8**	**100.0**	**95.8**	**100.0**	79.2	75.0	70.8	79.2	**95.8**	**100.0**	50.0	79.2	41.7	75.0	87.5	58.3	8.3	62.5	37.5	75.0
Transitive - past perfect simple	25	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**96.0**	**96.0**	**92.0**	88.0	88.0	**100.0**	**100.0**	88.0	84.0	76.0	88.0	76.0	68.0	8.0	68.0	8.0	82.1
Transitive - past progressive	38	**97.4**	**97.4**	84.2	78.9	71.1	84.2	**97.4**	81.6	76.3	76.3	78.9	**86.8**	71.1	63.2	78.9	71.1	68.4	68.4	21.1	42.1	23.7	72.3
Transitive - present perfect progressive	30	**100.0**	**100.0**	**100.0**	**100.0**	86.7	**100.0**	**93.3**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**96.7**	70.0	83.3	73.3	80.0	46.7	10.0	26.7	30.0	80.8
Transitive - present perfect simple	31	**100.0**	**100.0**	**100.0**	**93.5**	90.3	**96.8**	**96.8**	**93.5**	**96.8**	**100.0**	**100.0**	**100.0**	**100.0**	83.9	80.6	87.1	64.5	71.0	25.8	58.1	22.6	83.9
Transitive - present progressive	40	**97.5**	**97.5**	**100.0**	**100.0**	87.5	**100.0**	**100.0**	**95.0**	**100.0**	**100.0**	**100.0**	**100.0**	92.5	80.0	92.5	82.5	82.5	77.5	22.5	50.0	25.0	84.9
Transitive - simple past	38	**100.0**	**100.0**	**97.4**	**94.7**	**94.7**	**94.7**	**100.0**	**94.7**	**94.7**	**100.0**	**100.0**	**97.4**	**97.4**	89.5	92.1	78.9	81.6	68.4	52.6	31.6	26.3	85.1
Transitive - simple present	39	**97.4**	**97.4**	**100.0**	**100.0**	92.3	**94.9**	**100.0**	**97.4**	**100.0**	87.2	**100.0**	**100.0**	**94.9**	92.3	**97.4**	69.2	76.9	76.9	51.3	51.3	25.6	85.8
Verb valency	101	**91.1**	**91.1**	**86.1**	**88.1**	**88.1**	**84.2**	**86.1**	78.2	**86.1**	**84.2**	**83.2**	76.2	71.3	72.3	75.2	64.4	63.4	54.5	34.7	34.7	16.8	71.9
Case government	20	**95.0**	**95.0**	**95.0**	**95.0**	**100.0**	**95.0**	**90.0**	**95.0**	**95.0**	**95.0**	**95.0**	**90.0**	85.0	**95.0**	**95.0**	85.0	60.0	75.0	25.0	55.0	15.0	82.1
Catenative verb	16	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**93.8**	**93.8**	**87.5**	**100.0**	**100.0**	**100.0**	**100.0**	81.3	**87.5**	81.3	**87.5**	81.3	62.5	37.5	43.8	12.5	83.3
Mediopassive voice	18	**88.9**	**88.9**	**77.8**	**83.3**	**77.8**	**83.3**	**77.8**	55.6	**83.3**	**83.3**	**66.7**	50.0	**66.7**	38.9	55.6	27.8	55.6	22.2	27.8	**5.6**	22.2	59.0
Passive voice	16	**93.8**	**93.8**	**93.8**	**100.0**	**87.5**	**87.5**	**93.8**	**100.0**	**100.0**	**93.8**	**93.8**	**93.8**	**87.5**	**87.5**	**87.5**	75.0	75.0	81.3	43.8	75.0	25.0	84.2
Resultative	19	**89.5**	**89.5**	**89.5**	**94.7**	**94.7**	**84.2**	**84.2**	**84.2**	**89.5**	**78.9**	**84.2**	**73.7**	52.6	**73.7**	68.4	**73.7**	63.2	47.4	36.8	10.5	5.3	69.9
Semantic roles	12	75.0	75.0	50.0	41.7	58.3	50.0	75.0	33.3	33.3	41.7	50.0	41.7	50.0	41.7	58.3	25.0	41.7	33.3	41.7	16.7	25.0	45.6
micro-average	4219	**97.7**	**97.7**	96.8	**97.3**	91.5	95.3	95.3	92.6	94.7	93.5	95.3	96.3	78.1	81.4	78.1	81.3	68.4	72.1	38.6	43.4	24.0	81.4
phen. macro-average	4219	**96.8**	**96.8**	**96.3**	96.1	91.1	93.9	93.2	91.0	92.5	91.4	93.0	93.5	77.3	80.1	75.9	78.5	69.4	68.3	40.4	40.5	24.2	80.0
categ. macro-average	4219	**95.7**	**95.6**	93.7	92.8	92.0	90.8	90.3	89.9	89.6	89.1	87.8	87.4	82.0	81.8	80.5	76.9	69.8	67.2	51.7	45.8	23.9	79.7

Table 10: Accuracies (%) of successful translations on the phenomenon-level for English–German. The boldface indicates the significantly best-performing systems per row.

categ	count	Yande	Claud	Unbab	Comma	Onl-G	Onl-W	GPT4	IOLRe	Trans	Onl-B	Onl-A	Aya23	IKUN	Llama	IKUNC	CUNID	NVIDI	TSUHI	avg
Ambiguity	20	**90.0**	**95.0**	**90.0**	**90.0**	70.0	70.0	**90.0**	**90.0**	50.0	50.0	55.0	**85.0**	70.0	**75.0**	60.0	65.0	35.0	15.0	69.2
Lexical ambiguity	20	**90.0**	**95.0**	**90.0**	**90.0**	70.0	70.0	**90.0**	**90.0**	50.0	50.0	55.0	**85.0**	70.0	**75.0**	60.0	65.0	35.0	15.0	69.2
Coordination & ellipsis	86	**87.2**	**80.2**	**84.9**	**82.6**	**84.9**	**83.7**	**76.7**	75.6	72.1	72.1	72.1	**77.9**	74.4	69.8	65.1	72.1	54.7	47.7	74.1
Gapping	18	**77.8**	**66.7**	**77.8**	**72.2**	**88.9**	**77.8**	**66.7**	55.6	55.6	55.6	55.6	**72.2**	**66.7**	44.4	44.4	**72.2**	38.9	**5.6**	60.8
Pseudogapping	13	**76.9**	**84.6**	**84.6**	**76.9**	**84.6**	**76.9**	**53.8**	**76.9**	**61.5**	**61.5**	**76.9**	**61.5**	**61.5**	**53.8**	**53.8**	**69.2**	23.1	38.5	65.4
Right node raising	14	**92.9**	**78.6**	**92.9**	**78.6**	**78.6**	**85.7**	**78.6**	**78.6**	**92.9**	**92.9**	**71.4**	**92.9**	**78.6**	**71.4**	**78.6**	**85.7**	57.1	57.1	80.2
Sluicing	8	**100.0**	**100.0**	**75.0**	**87.5**	**75.0**	**87.5**	**87.5**	**75.0**	**87.5**	**87.5**	62.5	62.5	**100.0**	**75.0**	**75.0**	37.5	37.5	**75.0**	77.1
Stripping	19	**89.5**	84.2	**89.5**	**100.0**	**100.0**	**94.7**	**89.5**	84.2	84.2	84.2	**89.5**	84.2	**89.5**	84.2	78.9	84.2	**89.5**	63.2	86.8
VP-ellipsis	14	**92.9**	**78.6**	**85.7**	**78.6**	**71.4**	**78.6**	**85.7**	**85.7**	57.1	57.1	**71.4**	**85.7**	57.1	**92.9**	**64.3**	**64.3**	**64.3**	**64.3**	74.2
False friends	15	**86.7**	**86.7**	**86.7**	**86.7**	**86.7**	**73.3**	**66.7**	**60.0**	**66.7**	**66.7**	**80.0**	**66.7**	**66.7**	**66.7**	**60.0**	46.7	**66.7**	**53.3**	70.7

categ	count	Yande	Claud	Unbab	Comma	Onl-G	Onl-W	GPT4	IOLRe	Trans	Onl-B	Onl-A	Aya23	IKUN	Llama	IKUNC	CUNID	NVIDI	TSUHI	avg
Function word	34	**97.1**	88.2	**94.1**	**100.0**	**94.1**	**100.0**	**97.1**	91.2	**94.1**	**94.1**	88.2	**94.1**	85.3	85.3	82.4	73.5	73.5	70.6	89.1
Focus particle	15	**93.3**	73.3	**86.7**	**100.0**	**86.7**	**100.0**	**93.3**	80.0	**86.7**	**86.7**	**86.7**	**86.7**	73.3	73.3	73.3	66.7	80.0	73.3	83.3
Question tag	19	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**89.5**	**100.0**	**94.7**	**94.7**	**89.5**	78.9	68.4	68.4	93.6
LDD & interrogatives	81	**97.5**	**93.8**	**97.5**	91.4	**96.3**	**95.1**	91.4	90.1	91.4	91.4	85.2	86.4	82.7	80.2	76.5	82.7	70.4	59.3	86.6
Inversion	22	**95.5**	**95.5**	**90.9**	**90.9**	**100.0**	**95.5**	**90.9**	**90.9**	**95.5**	**95.5**	86.4	86.4	**90.9**	81.8	72.7	**90.9**	68.2	54.5	87.4
Modifying Comparison	4	100.0	100.0	100.0	100.0	75.0	100.0	100.0	100.0	75.0	75.0	75.0	75.0	75.0	75.0	100.0	75.0	75.0	50.0	84.7
Multiple connectors	13	**100.0**	**100.0**	**100.0**	**100.0**	**92.3**	**100.0**	**100.0**	**92.3**	**84.6**	**84.6**	**84.6**	**84.6**	**100.0**	**84.6**	76.9	**92.3**	76.9	76.9	90.6
Pied-piping	14	**92.9**	**92.9**	**100.0**	**100.0**	**100.0**	**92.9**	**92.9**	**100.0**	**100.0**	**100.0**	**92.9**	**92.9**	**85.7**	**85.7**	**85.7**	78.6	**100.0**	**92.9**	93.7
Preposition stranding	17	**100.0**	**100.0**	**100.0**	**88.2**	**94.1**	**100.0**	**88.2**	**88.2**	**100.0**	**100.0**	82.4	**88.2**	76.5	76.5	82.4	64.7	52.9	47.1	85.0
Topicalization	11	**100.0**	72.7	**100.0**	72.7	**100.0**	**81.8**	**81.8**	72.7	72.7	72.7	**81.8**	**81.8**	54.5	72.7	54.5	**90.9**	54.5	27.3	74.7
Lexical Morphology	41	**97.6**	**92.7**	**90.2**	**92.7**	82.9	73.2	80.5	73.2	75.6	75.6	70.7	68.3	75.6	75.6	63.4	53.7	34.1	26.8	72.4
Functional shift	17	**100.0**	**100.0**	**100.0**	**94.1**	**88.2**	**88.2**	**100.0**	**94.1**	**88.2**	**88.2**	76.5	82.4	**88.2**	**88.2**	82.4	64.7	41.2	41.2	83.7
Noun formation (er)	24	**95.8**	**87.5**	**83.3**	**91.7**	**79.2**	62.5	66.7	58.3	66.7	66.7	66.7	58.3	66.7	66.7	50.0	45.8	29.2	16.7	64.4
MWE	96	**87.5**	**84.4**	**78.1**	**83.3**	**80.2**	71.9	77.1	76.0	72.9	71.9	70.8	67.7	69.8	66.7	66.7	52.1	40.6	33.3	69.5
Collocation	13	**100.0**	**84.6**	76.9	69.2	**92.3**	**84.6**	76.9	76.9	69.2	69.2	**84.6**	69.2	76.9	61.5	69.2	53.8	38.5	15.4	70.5
Compound	14	**71.4**	**78.6**	**57.1**	**71.4**	**64.3**	**42.9**	**50.0**	**78.6**	**57.1**	**50.0**	**42.9**	**64.3**	**50.0**	**50.0**	**42.9**	**50.0**	28.6	14.3	53.6
Idiom	17	**94.1**	**88.2**	**70.6**	**88.2**	52.9	47.1	**76.5**	**70.6**	52.9	52.9	47.1	41.2	52.9	41.2	47.1	41.2	5.9	11.8	54.6
Nominal MWE	17	76.5	**88.2**	**88.2**	**88.2**	**100.0**	**94.1**	82.4	70.6	**100.0**	**100.0**	**94.1**	**88.2**	76.5	70.6	70.6	58.8	58.8	52.9	81.0
Prepositional MWE	18	**94.4**	**88.9**	**94.4**	**94.4**	**94.4**	**94.4**	**88.9**	77.8	83.3	83.3	**88.9**	**100.0**	**88.9**	**94.4**	**88.9**	61.1	83.3	72.2	87.3
Verbal MWE	17	**88.2**	**76.5**	**76.5**	**82.4**	**76.5**	**64.7**	**82.4**	**82.4**	**70.6**	**70.6**	**64.7**	41.2	**70.6**	**76.5**	**76.5**	47.1	23.5	23.5	66.3
Named entity & terminology	80	83.8	**95.0**	**87.5**	81.3	80.0	80.0	81.3	73.8	80.0	80.0	77.5	71.3	62.5	77.5	60.0	57.5	56.3	41.3	73.7
Date	20	**100.0**	**100.0**	**95.0**	**95.0**	**95.0**	85.0	**95.0**	85.0	**95.0**	**95.0**	**95.0**	85.0	80.0	**95.0**	80.0	85.0	70.0	75.0	89.2
Domainspecific Term	5	**40.0**	**80.0**	**60.0**	**60.0**	**60.0**	**80.0**	**40.0**	**60.0**	**60.0**	**60.0**	**60.0**	**60.0**	0.0	**40.0**	**20.0**	**60.0**	**60.0**	0.0	50.0
Measuring Unit	18	72.2	**100.0**	**100.0**	**94.4**	**94.4**	**94.4**	**94.4**	**94.4**	**94.4**	**94.4**	**88.9**	**88.9**	83.3	**100.0**	77.8	77.8	72.2	44.4	87.0
Onomatopeia	11	**100.0**	**100.0**	72.7	63.6	54.5	72.7	**81.8**	45.5	54.5	54.5	54.5	63.6	45.5	54.5	45.5	45.5	18.2	0.0	57.1
Proper Name & Location	26	**80.8**	**88.5**	**84.6**	**73.1**	**73.1**	**69.2**	**69.2**	65.4	**73.1**	**73.1**	**69.2**	53.8	53.8	65.4	46.2	26.9	50.0	38.5	64.1
Non-verbal agreement	98	**94.9**	**95.9**	**91.8**	**93.9**	**90.8**	**89.8**	**90.8**	**92.9**	80.6	80.6	80.6	**92.9**	83.7	86.7	85.7	81.6	73.5	65.3	86.2
Coreference	24	**87.5**	**91.7**	**83.3**	**87.5**	**75.0**	**83.3**	**83.3**	**83.3**	54.2	54.2	66.7	**83.3**	**79.2**	**75.0**	**83.3**	70.8	54.2	58.3	75.2
Genitive	16	93.8	93.8	81.3	87.5	93.8	81.3	81.3	87.5	87.5	87.5	93.8	93.8	75.0	87.5	81.3	81.3	68.8	68.8	84.7
Personal Pronoun Coreference	19	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.7**	**100.0**	**100.0**	**89.5**	**89.5**	78.9	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**89.5**	96.8
Possessive Pronouns	22	**95.5**	**95.5**	**95.5**	**100.0**	**90.9**	**95.5**	**95.5**	**100.0**	**95.5**	**95.5**	**90.9**	**90.9**	86.4	**90.9**	86.4	77.3	86.4	40.9	89.4
Substitution	17	**100.0**	**100.0**	**100.0**	**94.1**	**100.0**	**94.1**	**94.1**	**94.1**	82.4	82.4	76.5	**100.0**	76.5	82.4	76.5	82.4	58.8	76.5	87.3
Punctuation	13	**92.3**	**92.3**	**92.3**	**100.0**	**92.3**	76.9	61.5	76.9	**84.6**	**84.6**	**92.3**	**84.6**	**84.6**	61.5	**84.6**	**100.0**	**92.3**	76.9	85.0
Quotation marks	13	**92.3**	**92.3**	**92.3**	**100.0**	**92.3**	76.9	61.5	76.9	**84.6**	**84.6**	**92.3**	**84.6**	**84.6**	61.5	**84.6**	**100.0**	**92.3**	76.9	85.0
Subordination	115	**98.3**	**94.8**	**98.3**	88.7	**95.7**	**96.5**	**94.8**	93.0	**94.8**	**94.8**	93.0	86.1	86.1	88.7	83.5	80.0	80.9	67.0	89.7
Adverbial clause	9	**88.9**	**100.0**	**100.0**	**88.9**	**100.0**	**88.9**	**100.0**	**88.9**	**100.0**	**100.0**	**88.9**	66.7	66.7	**88.9**	**88.9**	**88.9**	66.7	55.6	87.0
Cleft sentence	17	**100.0**	**94.1**	**100.0**	82.4	**100.0**	**100.0**	**88.2**	**88.2**	**94.1**	**94.1**	**88.2**	76.5	82.4	58.8	64.7	70.6	76.5	64.7	84.6
Complex object	18	**100.0**	**94.4**	**100.0**	**94.4**	**94.4**	**100.0**	**100.0**	**100.0**	**94.4**	**94.4**	**88.9**	**100.0**	**94.4**	**100.0**	**94.4**	83.3	83.3	61.1	93.2
Contact clause	12	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**91.7**	**100.0**	**91.7**	**100.0**	**100.0**	**100.0**	**91.7**	**91.7**	**91.7**	75.0	**83.3**	**91.7**	**83.3**	94.0
Infinitive clause	25	**96.0**	88.0	**96.0**	88.0	**92.0**	**100.0**	**92.0**	**100.0**	**92.0**	**92.0**	**100.0**	**96.0**	**92.0**	**100.0**	**92.0**	72.0	84.0	68.0	91.1
Participle clause	22	**100.0**	**95.5**	**95.5**	86.4	**90.9**	**90.9**	**95.5**	**90.9**	**90.9**	**90.9**	86.4	77.3	81.8	81.8	72.7	86.4	81.8	59.1	86.4
Subject clause	12	**100.0**	**100.0**	**100.0**	**83.3**	**100.0**	**100.0**	**91.7**	**83.3**	**100.0**	**100.0**	**100.0**	**83.3**	**83.3**	**100.0**	**100.0**	**83.3**	75.0	**83.3**	92.6
Verb semantics	20	**100.0**	**90.0**	**95.0**	70.0	**95.0**	85.0	65.0	80.0	85.0	85.0	85.0	65.0	80.0	75.0	70.0	55.0	30.0	35.0	74.7
Verb tense/aspect/mood	169	**87.0**	**90.5**	**90.5**	**87.0**	**89.3**	**88.8**	**89.9**	**85.2**	**85.2**	**85.2**	**86.4**	**85.2**	81.1	**84.6**	**84.6**	78.1	69.2	45.0	82.9
Conditional	25	**96.0**	**100.0**	**100.0**	**96.0**	**96.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**92.0**	88.0	88.0	76.0	80.0	72.0	60.0	91.3
Ditransitive	34	82.4	**94.1**	**97.1**	**91.2**	**91.2**	**88.2**	**94.1**	**91.2**	**91.2**	**91.2**	**97.1**	**94.1**	**88.2**	**91.2**	**94.1**	**94.1**	82.4	50.0	89.1
Gerund	19	**94.7**	**84.2**	**94.7**	**78.9**	**94.7**	**73.7**	**89.5**	**73.7**	**94.7**	**94.7**	**89.5**	**78.9**	**73.7**	**89.5**	**94.7**	**73.7**	47.4	42.1	81.3
Imperative	24	**91.7**	**83.3**	**87.5**	**83.3**	**91.7**	**87.5**	**83.3**	**83.3**	70.8	70.8	70.8	**83.3**	**75.0**	**75.0**	**79.2**	66.7	62.5	20.8	75.9
Intransitive	29	**82.8**	**86.2**	**82.8**	**89.7**	**79.3**	**89.7**	**86.2**	**82.8**	69.0	69.0	**75.9**	**75.9**	**72.4**	**86.2**	**79.3**	**86.2**	**79.3**	55.2	79.3
Reflexive	19	**89.5**	**89.5**	**84.2**	**84.2**	**89.5**	**89.5**	**78.9**	**78.9**	**84.2**	**84.2**	**84.2**	**73.7**	**84.2**	**68.4**	**78.9**	52.6	**68.4**	31.6	77.5
Transitive	19	**73.7**	**94.7**	**84.2**	**78.9**	**84.2**	**89.5**	**94.7**	**78.9**	**89.5**	**89.5**	**84.2**	**94.7**	**84.2**	**89.5**	**89.5**	**78.9**	57.9	47.4	82.5
Verb valency	126	**93.7**	**88.1**	83.3	81.0	84.9	86.5	81.7	81.0	83.3	83.3	77.8	75.4	78.6	76.2	73.8	68.3	58.7	50.0	78.1
Case government	24	**100.0**	**100.0**	**95.8**	**95.8**	**100.0**	**100.0**	**100.0**	**100.0**	**95.8**	**95.8**	**95.8**	**91.7**	**100.0**	**95.8**	87.5	87.5	87.5	79.2	94.9
Catenative verb	25	**96.0**	**96.0**	**92.0**	**84.0**	**92.0**	**96.0**	**84.0**	**88.0**	**96.0**	**96.0**	**92.0**	**88.0**	**84.0**	**92.0**	**88.0**	**84.0**	**84.0**	72.0	89.1

categ	count	Yande	Claud	Unbab	Comma	Onl-G	Onl-W	GPT4	IOLRe	Trans	Onl-B	Onl-A	Aya23	IKUN	Llama	IKUNC	CUNID	NVIDI	TSUHI	avg
Mediopassive voice	18	**94.4**	**83.3**	**83.3**	**83.3**	**77.8**	**72.2**	**77.8**	**77.8**	**83.3**	**83.3**	**72.2**	61.1	**83.3**	**72.2**	**77.8**	**77.8**	33.3	38.9	74.1
Passive voice	25	**100.0**	**96.0**	88.0	**96.0**	**92.0**	**92.0**	88.0	88.0	**96.0**	**96.0**	**92.0**	**92.0**	**96.0**	76.0	80.0	76.0	80.0	52.0	87.6
Resultative	18	**88.9**	**83.3**	**83.3**	**61.1**	**72.2**	**83.3**	**83.3**	**66.7**	**61.1**	**61.1**	44.4	44.4	50.0	**72.2**	55.6	33.3	16.7	22.2	60.2
Semantic roles	16	**75.0**	**56.3**	**43.8**	**50.0**	**62.5**	**62.5**	**43.8**	**50.0**	**50.0**	**50.0**	**50.0**	**56.3**	**37.5**	31.3	**37.5**	31.3	18.8	12.5	45.5
micro-average	994	**91.8**	**90.4**	89.4	86.8	87.8	86.1	85.2	83.3	82.3	82.2	80.7	80.4	78.1	79.0	75.1	71.0	62.2	50.0	80.1
phen. macro-average	994	**91.4**	**90.2**	88.8	86.1	87.1	85.7	84.3	82.4	81.6	81.4	79.6	79.2	76.2	77.4	73.8	70.3	60.8	49.5	79.2
categ. macro-average	994	**92.4**	**90.5**	90.0	87.7	87.4	83.6	81.7	81.3	79.7	79.7	79.6	79.0	77.2	76.4	72.6	69.0	59.7	49.0	78.7

Table 11: Accuracies (%) of successful translations on the phenomenon-level for English–Russian. The boldface indicates the significantly best-performing systems per row.