# Fine-Grained Evaluation of English-Russian MT in 2025: Linguistic Challenges Mirroring Human Translator Training

**Shushen Manakhimova[1], Maria Kulinovskaya[2], Ekaterina Lapshinova-Koltunski[3], Eleftherios Avramidis[1],**
[1]German Research Center for Artificial Intelligence (DFKI)
firstname.lastname@dfki.de
[2]Saarland University, maria.kunilovskaya@uni-saarland.de
[3]University of Hildesheim, lapshinovakoltun@uni-hildesheim.de

## Abstract

We analyze how English–Russian machine translation (MT) systems submitted to WMT25 perform on linguistically challenging translation tasks, similar to problems used in university professional translator training. We assessed the ten top-performing systems using a fine-grained test suite containing 465 manually devised test items, which cover 55 lexical, grammatical, and discourse phenomena, in 13 categories. By applying pass/fail rules with human adjudication and micro/macro aggregates, we observe three performance tiers. Compared with the official WMT25 ranking, our ranking broadly aligns but reveals notable shifts.

Our findings show that in 2025, even top-performing MT systems still struggle with translation problems that require deep understanding and rephrasing, much like human novices do. The best systems exhibit creativity and can be very good at handling such challenges, often producing more natural translations rather than producing word-for-word renditions. However, persistent structural and lexical problems remain: literal word order carryovers, misused verb forms, and rigid phrase translations were common, mirroring errors typically seen in beginner translator assignments.

## 1 Introduction

Unlike standard test sets, which consist of randomly selected source material, a test suite contains 'extra-credit' problems: the source items, deliberately designed to be challenging in translation, similar to the 'rich points' method (Nord, 1997) or preselected items method (Egdom et al., 2019) in translator training. These focused items are often lexical units or grammatical structures requiring non-literal approaches based on deep situational understanding and coherent target-language encoding. Successful translations are expected to abstract away from source form, displaying both familiarity with standard translation techniques and creativity in adapting them to individual contexts. While isomorphic translations may be formally grammatical, they are typically judged suboptimal. In translation didactics, such tasks test translators' ability to identify and resolve translation problems.

The fine-grained linguistic test suite has been partially in development over the last few years (Macketanz et al., 2022; Manakhimova et al., 2023, 2024).[1] Its original purpose was to track MT systems' ability to handle specific source language phenomena, with evaluation strategies focused on translating those targeted items. This approach provides valuable, fine-grained insights into MT systems, highlighting system strengths across a wide range of phenomena and establishing objective grounds for comparison. However, the recent dominance of large language models (LLMs) in MT has prompted us to reconceptualize the test suite's role. Beyond its original function of testing specific linguistic categories, we also view it as a **diagnostic tool for identifying overarching translation challenges** that persist across current English–Russian MT systems. This evolution reflects the need to understand systemic issues in MT. Accordingly, we have tightened **our quality requirements to match those of a university professional translation training**. Our evaluation focus has shifted from assessing the handling of specific linguistic categories to evaluating overall translation quality when processing these 'extra-credit' problems. We no longer make concessions for 'gist translation quality,' instead expecting the publication-standard quality of human translation as defined by Ahrenberg (2017, p. 21). This refined approach maintains our ability to compare systems based on the ratio of successfully handled

[1]The wider project also maintains suites for other directions (e.g., German-English, English–German, English–Portuguese in Avramidis et al., 2020; Macketanz et al., 2021; Avelino et al., 2022), which we do not analyze here.



*Proceedings of the Tenth Conference on Machine Translation (WMT)*, Volume 2: Shared Task Papers, pages 866–877
November 8–9, 2025. 

items across the entire test suite, while simultaneously revealing persistent issues and blind spots that transcend formal categories. We supplement our qualitative analysis with statistical results and system rankings, comparing these findings with WMT's official human evaluation.

## 2 Method

### 2.1 Test suite overview

For the English–Russian part of the test suite, the 13 evaluation categories cover a broad range of translation challenges. They include ambiguity, collocations, compounds, false friends, and multi-word expressions, which test lexical precision. Morphosyntactic control is addressed through case government, function words, passive voice, subordination, and verb valency. The suite also targets discourse as well as stylistic aspects via personal pronoun coreference and onomatopoeia. Together, these categories span the main lexical, grammatical, and pragmatic difficulties that can arise when translating from English to Russian.

Test items consist of one (or occasionally more) source sentence(s) plus an associated set of evaluation rules. These rules comprise hand-crafted regular expressions and fixed strings of translation outputs. Test items are either created manually by linguists or sourced from existing corpora and curated for the target phenomenon. An ideal example for a test suite should require the interpretation of the message and deep restructuring in the target language without being vague or context-dependent. Such examples also give rise to greater variability in translation, and can be identified as having a higher entropy of translation solutions, a known measure of the source item difficulty in translation (Carl and Schaeffer, 2017; Wei, 2022; Kunilovskaya et al., 2025). Examples from the *Resultative* subcategory (category: Verb valency) seem to represent true challenges in the English–Russian translation. For example, *The skiers skied the trail clean of snow.*

As MT technology has shifted from phrase-based systems to NMT and now LLMs, error profiles have also changed. The suite has therefore been revised over time: we have added phenomena, increased item counts, and introduced longer or structurally richer sentences to stress contemporary systems. MT outputs, after being evaluated by the test suite, have also been utilized to create challenge sets for WMT metrics (Avramidis et al., 2023, 2024). This year's revision of the test suite excluded several items that either misrepresented their category or lacked context independence to be fairly evaluated.

Given the increased variation in translation solutions offered by the submitted systems, the automatic pass-fail rules and annotation guidelines were tightened this year to reflect the requirements that would be applied to translations considered as part of a university professional translator training.

### 2.2 Scoring

The evaluation results for the categorized items are produced semi-automatically: hand-written regular expressions as well as fixed strings (correct and incorrect translations from earlier MT system outputs) capture expected *correct* and *incorrect* renderings, and any remaining cases are adjudicated by a linguist. Regular-expression design leverages prior experience with MT outputs and aims to maximize coverage; however, novel outputs routinely require human judgment.

Since this evaluation aims to compare the systems fairly, only the test items that have a valid judgment for all systems are included in the calculation. If a test item has a judgment neither by regular expressions nor by the annotators for any MT systems, we exclude it from the calculation. As a result, not all test items that had originally been designed can be used in our calculations.

For system comparison, we first identify the highest-scoring system and then test all others against it using a one-tailed Z-test with $\alpha = 0.95$. Systems not significantly worse than the top system form the first performance cluster; we mark the best systems in bold in the result tables. Because categories and phenomena (subcategories) differ in size, we report three complementary aggregates: micro-average (accuracy over all items, item-weighted), category macro-average (mean of category-level accuracies), and phenomenon macro-average (mean of phenomenon-level accuracies).

### 2.3 Manual annotation procedure

This year marks the third evaluation of English–Russian systems using our test suite. As in previous years, manual intervention was necessary to process system outputs that could not be automatically evaluated. At the beginning of this year's evaluation, this referred to $44.83\%$ of outputs on average. Three annotators divided the workload,

with each responsible for a disjoint subset of the data.

We did not compute inter-annotator agreement (IAA), as the evaluation workflow did not involve multiple annotators independently labeling the same outputs. Instead of computing IAA, we relied on the extensive, linguistically motivated rule set refined over several years. These rules served as a shared reference, reducing subjectivity; borderline cases were resolved in group discussion.

Annotators were instructed to focus on the targeted source-language phenomenon, provided that a translation candidate meets basic standards of accuracy and fluency. Above all, the translation should faithfully convey the original message while adhering to the norms and conventions of the target language.

The evaluation is relative: in disputable cases, comparing translation candidates helps determine what is achievable for a given item. In professional translator training, it is common to assess solutions against available or hypothesized alternatives. For example, Bittner (2020, p. 172) observes: *"Good translation quality can only be better translation quality, just as bad translation quality can only be worse translation quality. There is no use dismissing a translation solution as unacceptable unless a better alternative can be produced."* At the same time, it is possible that none of the proposed solutions is acceptable, or that two solutions using different techniques are equally valid.

When a source item is aimed at evaluating more than one phenomenon, a translation is considered correct even if some parts are suboptimal (but not unacceptable). This ensures that effective strategies and creative handling of the target phenomenon are recognized. To illustrate the kinds of defects that were tolerated, consider Example (1), where the source is categorized as *Modifying Comparison* and both translations are accepted as correct. The second version, however, is preferable: it omits the possessive pronoun её ("her") in the subordinate clause, avoids semantic tautology in rendering *expertise*, and dispenses with the redundant demonstrative pronoun того. Importantly, both variants successfully address the lexical challenges that weaker systems often mishandle. For instance, many systems reproduced the English collocation *extensive level of expertise* as обширный уровень экспертизы.

(1) Her level of expertise was not as extensive as her employer had hoped.

a. Уровень её профессиональной квалификации оказался не таким высоким, как того ожидал её работодатель.

b. Уровень её квалификации оказался не таким высоким, как надеялся работодатель.

In cases where the source was ambiguous or difficult to interpret (e.g., *he ran under the porch*), a translation was judged correct if it provided a plausible and contextually logical reading consistent with real-world knowledge. For example, он (по)бежал под крыльцо for the source above remains questionable.

Although the items are designed to be context-independent, some may admit multiple interpretations in a broader context. In such cases, evaluation favors the most prototypical or expected reading, while unusual or exotic contexts are disregarded. Finally, translations that ignore potential of the target language for optimal information packaging – often requiring creative reconceptualization of the message – and instead follow the source language structures in a routine, linear manner, were not accepted. They might be grammatically correct, but in dissonance with the conventional usage of the target language.

### 2.4 Experiment Setup

Although the full test suite was applied to 42 systems submitted to the WMT25 Shared Task, this paper reports statistical comparisons for a representative subset of 10 systems. The selection is based on the official Error Span Annotation

## 3 Results

### 3.1 System Performance Overview

This section reports system-level performance (overview and hierarchy) and category-level difficulty, following the scoring protocol in Section 2.2 Aggregate accuracies per system and per category/phenomenon are provided in the Appendix tables, along with the test suite system ranking and the official WMT25 ranking (Kocmi et al., 2025b).

The performance distribution reveals three distinct tiers: high performers Wenyiil (Wang, 2025), Algharb (Xu, 2025), Yandex (Karpachev et al., 2025), and Gemini (Finkelstein et al., 2025);

mid-tier systems Claude-4, DeepSeek-V3, GPT-4.1, and Shy-hunyuan-MT (Zheng et al., 2025, 89.4–91.9%); and lower-tier systems CommandA (Kocmi et al., 2025a) and UvA-MT (Wu et al., 2025, 83.2–88.4%).

Compared to the WMT25 official ESA ranking, Yandex moves from the 8–10 cluster into the top cluster. Conversely, Shy-hunyuan-MT-hunyuan-MT, ranked 2 under ESA, falls into the middle cluster in our suite, Claude-4 and GPT-4.1 also belong to the middle cluster; both exhibit the same weakness on verb semantics (71.4%), and GPT-4.1 additionally scores low on long-distance dependencies and interrogatives (81.5%).

**Peformance of Constrained vs. unconstrained models.** In our test suite, constrained MT systems frequently occupy the top cluster, consistent with scoring that rewards precise handling of hard, localized phenomena and conservative choices under ambiguity. In the WMT25 official ranking, however, unconstrained systems rise, reflecting strengths in fluency, stylistic naturalness, and document-level coherence. Systems that bridge both profiles narrow the gap between the two.

### 3.2 Category Difficulty Analysis

*MWE* represents the most challenging category with only 80.4% average accuracy across all systems, followed by *Verb semantics* (81.4%) and *Verb valency* (87.4%). These categories demonstrate the complexity of handling idiomatic expressions and verb–argument structures in English–Russian translation. Conversely, *Lexical Ambiguity* proves easiest (97.0% average), with eight systems achieving perfect scores, indicating strong disambiguation capabilities across translation systems. *Function words* and *Subordination* also show high accuracy (93.6% and 94.3% respectively), suggesting robust handling of grammatical structures.

*Verb semantics* exhibits the largest performance gap (57.1%) between systems, with Wenyiil achieving 100% while UvAMT manages only 42.9%.

### 3.3 Linguistic Analysis

In this section, we summarize the overall patterns and translation strategies revealed in the manual analysis, along with some notable peculiarities of individual systems. We then turn to the most persistent challenges: (a) difficulties rooted in English source structures and (b) recurring problems with Russian target-language conventions that MT systems struggle to master. None of the highlighted issues is ubiquitous; for each example, we provide a more acceptable version drawn from the available translations. Translations marked with an asterisk are considered suboptimal. Generally, a comparison with previous years' submissions indicates noticeable improvements across most problem areas.

**Overall translation patterns.** This paragraph offers some high-level observations from annotating the 2025 submissions in comparison with previous years. We note an improvement in the variation of generated output, suggesting greater creativity and a stronger ability to recast the original message in new forms, rather than reproducing the formal and semantic structures of the source language.

Translations from the strongest 2025 systems demonstrate an increased capacity to do what the test suite examples force them to do: they **move beyond literal translation** and re-package the original message into a form that is natural in the target language. For a human translator, accomplishing this requires careful extraction of the intended meaning, imagining the described real-world situation, and expressing it in a way that aligns with conventional norms and expectations. The distinction between a literal strategy and a more interpretative approach is illustrated in Example (2). The example highlights the ability of the system to infer the contextual meaning of the descriptive verb "shuddered" and generate a plausible rendering of the situation.

(2) They shuddered home under the hailstorm.
   a. Они брели домой под градом, ежась от холода.
   b. *Они дрожащими вернулись домой под градом.

**Re-creating a situation in another language** may require modifying the set of properties by which it is recognized in the target language. In 2025, automatic translation systems are better at adding necessary elements and omitting redundant ones. For instance, *he's a fabulous inspiration* is rendered as он потрясающий <источник> вдохновения, while *Many people are concerned about High Street* becomes Многих беспокоит <состояние> главной улицы.

The ability of a system to take context into account and coordinate elements into a coherent

whole can also be observed at the sentence level. The source in Example (3) evokes a snapshot-like scene. Re-creating this scene in Russian requires abandoning the English mode of depiction and adopting a different strategy in the second clause. The asterisked translation illustrates a common problem: mismatched aspect forms in coordinated verbs, which disrupts the natural flow of information.

(3) Paula entered the small souvenir shop and took her time browsing through the magazines.
   a. Паула зашла в небольшой сувенирный магазин и принялась рассматривать журналы.
   b. *Паула вошла в небольшой сувенирный магазин и не спеша просматривала журналы.

This enhanced interpretative capacity reflects greater sensitivity to the functional potential of expressive means and the ability to deploy alternative but appropriate forms in the target language. As illustrated in Example (4), Algharb recognizes that paired synonyms are typical in English but generally avoided in Russian, while in Example (5), Algharb creatively re-packages the information to arrive at a conventional Russian rendition.

(4) Despite the neat and tidy ending to Season 3 . . .
   a. Несмотря на вполне законченную концовку третьего сезона, . . .

(5) . . . there was a delivery charge on top.
   a. . . . к сумме добавилась плата за доставку.

These examples are presented to highlight some of the advances in the technology. At the same time, they do not provide a complete picture, as persistent problems remain.

Where possible, the same systems resort to suboptimal **crude unpacking of English secondary predicates into full clauses**, producing wordy, redundant, and clumsy (but not ungrammatical) sentences. This strategy disrupts the information flow: in discourse, each full predicate conventionally signals a step forward in the narrative. By upgrading secondary predicates to main clauses, the translation introduces artificial shifts in topic–comment structure, resulting in unmotivated changes of focus and sentences that can appear contradictory or unclear. It is not uncommon that this tendency is coupled with a known redundancy of functional words such as auxiliaries, pronouns, and connectives (esp. consecutive connectives перед тем как, после того как).

Newer systems increasingly prefer **gender-neutral realization in the cases where gender is not explicit**, e.g. in generic contexts or in first person in Russian outputs. In contrast to English, Russian has explicit grammatical gender marking not only on pronouns, but also on past verb forms, participles, and adjectives. They are congruent with subjects. In contrast to previous years, systems for the first time consistently produced *explicit inclusive* forms, i.e., forms that include both masculine and feminine forms: купил(а) (“bought (m/f)”), мог(ла) (“could (m/f)”), готов(а) (“ready (m/f)”). The pattern is consistent across 33 systems (out of 42 evaluated systems) and various categories, indicating that newer models increasingly prefer gender-neutral realizations. Additionally, some systems (GPT-4.1, Algharb, Gemini) demonstrate the use of gendered profession names where the associated person is female (учительница, предпринимательница).

When faced with a **faulty input**, some models return a translation of the more plausible corrected source version, like a human translator would do. This is counteracting the automated tendency known as ‘garbage in, garbage out’.

**Source challenges and target issues.** In this part of the analysis, we first describe recurrent and problematic translation patterns in Russian that are triggered by specific source-language items. We then turn to a second group: target-language categories that consistently prove difficult for MT systems regardless the source.

**(a) Source phenomena as error triggers.** The most prominent defects in automatic translation stem from the literal transfer of source-language lexical and grammatical features. In particular, we highlight issues arising from (i) reproducing the word order of the source sentence, (ii) neglecting the contrastive use of pronouns and connectives, and (iii) calquing lexical frequency patterns and collocations. These problems complement those discussed in Section 3.2.

Unlike English, Russian relies heavily on **word order** to structure information. The most important, focused information typically occurs at the end of a

sentence, whereas English allows more flexibility; the sentence-final position in English can be filled with adverbials of time and place, prepositional objects, and other elements. Failure to identify the focused element in the English source and promote it to the sentence-final position in Russian, therefore, disrupts the natural flow of information, even in isolated sentences. A typical case is presented in Example (6). The topical sentence member им ("them") is awkward at the end of the Russian sentence in (6-b).

(6) When students walk into our classrooms, the course objectives are given to them right up front.
   a. Когда студенты заходят в наши аудитории, им сразу же сообщают цели курса.
   b. *Когда студенты заходят в наши аудитории, цели курса сразу же сообщаются <им>.

Misalignment in information structure is particularly noticeable in cleft sentences (e.g., *It wasn't until ...*) and elliptical constructions (e.g., *She asked the kids to stay, and the adults too; Laura drank the milk last night, or perhaps the juice; I met Aisha yesterday, but not her daughter*). As illustrated in Example (7), failing to place emphasized information at the end produces sentences like in (7-b) that are immediately recognizable as translations.

(7) After all it was not war that completely ravaged East Asian states in 1997.
   a. В конце концов, в 1997 году государства Восточной Азии разорила вовсе не война.
   b. *В конце концов, это была не война, которая полностью опустошила восточноазиатские государства в 1997 году.

Automatic translations into Russian often **overuse possessive pronouns**, mirroring their higher frequency in English. Example (8) shows sentences where *their* is rendered as <их> in Russian. While grammatically correct, these translations add possessive markers that a human translator would likely omit, resulting in a style that is formally acceptable but less natural (see also Example (1)).

(8) Despite <their> intense feelings for one another, it seems as though the two heroes might never remain together.
   a. Несмотря на <их> сильные чувства друг к другу, кажется, что этим двум героям никогда не быть вместе.

**Indefinite pronouns** (someone, anywhere, every, all) also contribute to a significant level of disfluency in machine translation. The apparent one-to-one correspondences have different usage patterns and frequencies (every $\neq$ каждый, all $\neq$ все).

Finally, **lexical problems** – the choice of words, collocation and idioms – are as pervasive as structural difficulties. Occasionally, many systems would find a particular word in the source language difficult and fall victim to literal translation, false friends, undetected idioms or terms. Example (9) shows a typical MT output, where sponsor is translated as спонсор. At the same time, in Russian, this word and derivatives from the same root rarely carry the "legislative initiator" meaning.

(9) They persuaded Kennedy and some other Senator to jointly sponsor the legislation, but I can't remember which one.
   a. *Они убедили Кеннеди и ещё одного сенатора совместно выступить спонсорами законопроекта, но я не помню, кого именно.

The hallmark of low-quality machine translation is translating every occurrence of *enjoy* with наслаждаться, and *people* with люди, to give examples of typical frequency calques seen in the analysis of this year.

**(b) Target phenomena as persistent difficulties.** A number of Russian categories can be problematic because they are not directly marked in the source but are obligatory in Russian. These categories are known to be difficult for Russian language learners, too. For example, English often encodes *verbal aspect* (a grammatical category that characterizes an action with regard to its internal temporal structure, such as whether it is ongoing, completed, repeated, or habitual) through contextual or grammatical means, while Russian uses a *lexico-grammatical system* (perfective, imperfective verbs). The translator is compelled to make a lexical choice that cannot be carried over from the source, since the category is not explicitly expressed there. Instead, the decision must be guided

by contextual cues and world knowledge, with the challenge lying in correctly reading those signals.

**Verbal aspect:** Automatic translation into Russian often struggles with maintaining aspectual coherence when rendering coordinated English verbs in the past tense. In Example (10), the first translation maintains consistent temporal and aspectual framing by using two perfective verbs (раздал and получил). By contrast, the incorrect version (10-b) uses an imperfective verb (раздавал) in the first clause. This creates a mismatch with the following perfective verb, producing an incoherent sequence: the first action is presented as ongoing or habitual, while the second is punctual and bounded. The result is an aspectual clash that disrupts the event structure of the sentence.

(10) The teacher handed out worksheets, but I didn't get one.
   a. Учитель раздал рабочие листы, но мне не досталось.
   b. *Учитель раздавал рабочие листы, но я их не получил.

In Example (11), the source communicates a polite encouragement and requires an imperfective verb (снимать) used in (11-a). Translations with possessive pronouns and perfective verb (снять) are suboptimal, because they sound like a command or instruction like in (11-b).

(11) Do take your coat off.
   a. Снимайте пальто (прошу вас).
   b. *Снимите своё пальто.

Another interesting example related to aspect is given in (12). The verb разъедайтесь used in (12-b) is an imperfective form of the verb разъесться, one of the meanings of which is 'to enjoy'. However, this verb is never used in the imperfective form with this meaning. Moreover, the given example contains an idiomatic expression which should not be translated literally word by word. Instead, a corresponding idiomatic expression should be used in Russian, as in (12-a).

(12) Enjoy your meal.
   a. Приятного аппетита.
   b. *Разъедайтесь обедом.

Beyond that, this output also contained the explanation by the system: *Note "Enjoy your meal" can be translated more literally as "Eat your meal with pleasure", but* Разъедайтесь обедом *is a more common colloquial way to say it in Russian*. This explanation is wrong.

**Nominalisations:** One important component of human translator training is drawing attention to linguistic categories that tend to be underrepresented in translation. Based on our analysis, in MT, these include *nominalizations* and *ellipsis*.

A variety of English subordinate clauses can be rendered as nominal phrases in Russian. This strategy helps avoid unnecessary nesting and reduces sequences of functional words, such as в том, что; это что-то, над чем; до тех пор, пока, resulting in a text that reads more naturally in Russian. In Example (13), the subordinate clause *he's retired* is rendered as the phrase после завершения карьеры in the accepted translation, whereas a weaker system (13-b) fails to apply this transformation.

(13) As previously documented, he discussed what his next move will be now that he's retired from in-ring competition.
   a. Как уже сообщалось ранее, он рассказал о своих дальнейших планах после завершения карьеры рестлера.
   b. *Как было задокументировано ранее, он обсудил, каким будет его следующий шаг теперь, когда он завершил карьеру активного борца.

Finally, a translated text can often be recognized by its unnaturally complete structure, with elements such as pronouns, copula verbs, and connectives explicitly spelled out where they could easily be inferred from the context. In other words, translated language underuses ellipsis.

There is a clear distinction between weaker and stronger MT systems in this regard. Stronger systems are less likely to produce a subject in subsequent clauses when it is identical to the subject of the main clause, which aligns with natural Russian usage and enhances fluency. For example, in Мы сделаем А, как только мы получим Б ("We shall do A, as soon as we get B"), the second мы would normally be omitted. Similarly, the copula verb in sentences such as Он заметил, что она <была> печальна ("He noticed that she was sad") should be omitted; however, failure to follow this pattern is pervasive even among strong

systems. These issues rarely impede understanding and are generally tolerated in evaluation, but they signal a lack of expected quality and function as an indicator of professional translation proficiency.

Due to space constraints, this description includes only the more pervasive defects of MT. Other flaws, which we want to flag, include contrastive connectives (esp. but translated as но where а is required), confusion caused by epistemic *would* and *could*, the limited use of short adjectives in predicative function, failures to build adverbial participial clauses as required by Russian school grammar, etc.

## 4 Discussion: Lessons Learned

The test suite was revised to exclude items with questionable categorization or insufficient context-independence for translation in isolation. The rules have become less permissive in terms of fluency and accuracy.

There are some similarities between the translation patterns produced by NMT and by learner translators (see the detailed analysis in Kunilovskaya et al., 2023). These similarities are most visible in source-language–triggered issues, such as the placement of adverbials and prepositional objects, or the preference for analytical (instead of synthetic) forms of future and passive verbs. This points to the phenomenon of shining through (as defined by Teich, 2003), which belongs to the phenomena of translationese (Gellerstam, 1986), i.e., specific features of translated language that make it different from non-translated original language production. However, compared to human translators, NMT systems (especially those outside the top tier) more often generate sentences that obscure the message, overcomplicate structure, or introduce redundancy. Such output departs from target-language conventions in ways that sound recognizably non-human, often failing to produce a coherent text that conveys a clear, plausible situation.

We observed that the same systems might pursue different strategies depending on the conditions. Faced with the absence of isomorphic structures in the target language, they are capable of impressively creative solutions. In less demanding contexts, however, they tend to revert to routine, near-literal strategies that overlook the target language's potential for more optimal expression.

## 5 Conclusion

In 2025, state-of-the-art English–Russian MT presented by the latest LLM-based systems shows substantial progress in performance, yet it continues to display important weaknesses. Our fine-grained evaluation revealed that even top-performing models still falter on translations requiring deep comprehension and nuanced rephrasing, much like human novice translators. At the same time, the best systems exhibit marked improvements on such 'extra-credit' items by re-structuring and wording translations in a more natural Russian style instead of relying on word-for-word renditions. This shift toward non-literal, context-aware translation indicates that current MT can approximate some of the flexible strategies employed by skilled human translators. Notably, general-purpose LLMs (e.g., GPT-4.1, Claude-4) only attained mid-tier accuracy on our test suite, underscoring that even massive generalist models have not fully solved certain linguistic subtleties Specialized MT engines thus continue to hold an edge on fine-grained challenges, though the gap is beginning to narrow as new models adopt more human-like problem-solving approaches.

## Limitations

A limitation of our current evaluation design is its reliance on a binary correctness—each test item is marked as either correct or incorrect based on regular-expression matching or manual adjudication. While this design facilitates scoring and result aggregation, it inevitably lacks the granularity needed for a more nuanced evaluation of translation quality, especially when annotators are faced with a human-like variation of MT outputs for less straightforward examples.

The second most notable limitation is the sentence-level nature of examples, which provides a reduced opportunity to track translation problems that might arise from the discourse level. It is not clear whether MT models would employ sentence splitting and merging as well as redistribution of semantics across several sentences if they were faced with larger spans of text to operate on.

Next, the test suite requires further revision to strengthen its construct validity. In particular, source items should foreground the targeted source-language phenomenon as the primary translation challenge, without being obscured by additional difficulties in other parts of the sentence, insofar as this is possible.

The current scoring approach does not differentiate sources by their translation difficulty. In future work, we plan to introduce a weighting scheme informed by the entropy of submitted translation solutions.

## Acknowledgments

We would like to thank Vladimir Kropivnitskiy for his contribution to the manual annotation of the test suite this year. We would also like to thank Vivien Macketanz, Sergei Bagdasarov, Hans Uszkoreit, Aljoscha Burchardt, Ursula Strohriegel, Renlong Ai, and He Wang for their prior contributions to the creation of the test suite.

## References

Lars Ahrenberg. 2017. Comparing machine translation and human translation: A case study. In *Proceedings of the First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT)*, pages 21–28, Varna, Bulgaria.

Mariana Avelino, Vivien Macketanz, Eleftherios Avramidis, and Sebastian Möller. 2022. A Test Suite for the Evaluation of Portuguese-English Machine Translation. In *Computational Processing of the Portuguese Language*, pages 15–25, Cham. Springer International Publishing.

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*. 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.

Eleftherios Avramidis, Shushen Manakhimova, Vivien Macketanz, and Sebastian Möller. 2024. Machine translation metrics are better in evaluating linguistic errors on LLMs than on encoder–decoder systems. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, Florida, USA. Association for Computational Linguistics.

Hanna Bittner. 2020. *Evaluating the Evaluator: A Novel Perspective on Translation Quality Assessment*. Routledge, New York and London.

Michael Carl and Moritz Jonas Schaeffer. 2017. Why translation is difficult: A corpus-based study of non-literality in post-editing and from-scratch translation. *Hermes (Denmark)*, 56:43–57.

Gys-Walt Van Egdom, Heidi Verplaetse, Iris Schrijver, Hendrik J. Kockaert, Winibert Segers, Jasper Pauwels, Bert Wylin, and Henri Bloemen. 2019. How to Put the Translation Test to the Test? On Preselected Items Evaluation and Perturbation. In Elsa Huertas-Barros, Sonia Vandepitte, and Emilia Iglesias-Fernández, editors, *Quality Assurance and Assessment Practices in Translation and Interpreting*, pages 26–56. IGI Global, Hershey, PA, USA.

Mara Finkelstein, Geza Kovacs, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Markus Freitag, and David Vilar. 2025. Google Translate's Research Submission to WMT2025. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In L. Wollin and H. Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund.

Nikolay Karpachev, Ekaterina Enikeeva, Dmitry Popov, Arsenii Bulgakov, Daniil Panteleev, Dmitrii Ulianov, Artem Kryukov, and Artem Mekhraliev. 2025. Yandex Submission to the WMT25 General Machine Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Arkady Arkhangorodsky, Alexandre Berard, Phil Blunsom, Samuel Cahyawijaya, Théo Dehaze, Marzieh Fadaee, Nicholas Frosst, Matthias Galle, Aidan Gomez, Nithya Govindarajan, Wei-Yin Ko, Julia Kreutzer, Kelly Marchisio, Ahmet Üstün, Sebastian Vincent, and Ivan Zhang. 2025a. Command-A-Translate: Raising the Bar of Machine Translation with Difficulty Filtering. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, and 9 others. 2025b. Preliminary Ranking of WMT25 General Machine Translation Systems.

Maria Kunilovskaya, Tatyana Ilyushchenya, Natalia Morgoun, and Ruslan Mitkov. 2023. Source language difficulties in learner translation: Evidence from an error-annotated corpus. *Target*, 35(1):34–62.

Maria Kunilovskaya, Iuliia Zaitova, Wei Xue, Irina Stenger, and Tania Avgustinova. 2025. Predictability of microsyntactic units across slavic languages: A translation-based study. In *The Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies*, pages 313–322.

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. (WMT21)*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vivien Macketanz, Shushen Manakhimova, Eleftherios Avramidis, Ekaterina Lapshinova-koltunski, Sergei Bagdasarov, and Sebastian Möller. 2022. 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.

Shushen Manakhimova, Vivien Macketanz, Eleftherios Avramidis, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2024. Investigating the linguistic performance of large language models in machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 355–371, Miami, Florida, USA. Association for Computational Linguistics.

Christiane Nord. 1997. *Translating as a Purposeful Activity: Functionalist Approaches Explained*. St. Jerome, Manchester, UK.

Elke Teich. 2003. *Cross-Linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, Berlin.

Hao Wang. 2025. Wenyiil's Submissions to the WMT 2025 General Machine Translation Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Yuxiang Wei. 2022. Entropy as a measurement of cognitive load in translation. In *AMTA 2022 - 15th Conference of the Association for Machine Translation in the Americas, Proceedings - Workshop on Empirical Translation Process Research*, volume 1, pages 75–86.

Di Wu, Yan Meng, Maya Konstantinovna Nachesa, Seth Aycock, and Christof Monz. 2025. UvA-MT's Participation in the WMT25 General Translation Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Linlong Xu. 2025. Algharb at WMT25 Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Mao Zheng, Zheng Li, Yang Du, Bingxin Qu, and Mingyang Song. 2025. Shy-hunyuan-MT at WMT25 General Machine Translation Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

| categ | count | Wenyi | Algha | Yande | Gemin | Claud | DeepS | GPT41 | Shy | Comma | UvAMT | avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ambiguity | 10 | 100.0 | 100.0 | 90.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 80.0 | 97.0 |
| Coordination & ellipsis | 27 | 85.2 | 85.2 | 92.6 | 88.9 | 85.2 | 81.5 | 85.2 | 88.9 | 96.3 | 88.9 | 87.8 |
| False friends | 8 | 100.0 | 100.0 | 100.0 | 87.5 | 100.0 | 87.5 | 100.0 | 75.0 | 75.0 | 100.0 | 92.5 |
| Function word | 16 | 93.8 | 93.8 | 87.5 | 93.8 | 100.0 | 100.0 | 93.8 | 93.8 | 87.5 | 93.8 | 93.8 |
| LDD & interrogatives | 27 | 96.3 | 96.3 | 88.9 | 92.6 | 88.9 | 85.2 | 81.5 | 92.6 | 85.2 | 85.2 | 89.3 |
| Lexical Morphology | 16 | 87.5 | 87.5 | 93.8 | 93.8 | 87.5 | 87.5 | 87.5 | 87.5 | 75.0 | 93.8 | 88.1 |
| MWE | 55 | **85.5** | **85.5** | **92.7** | 80.0 | 78.2 | 78.2 | **83.6** | **83.6** | 69.1 | 67.3 | 80.4 |
| Named entity & terminology | 47 | **91.5** | **91.5** | **97.9** | **93.6** | **93.6** | **91.5** | **95.7** | **89.4** | **89.4** | 85.1 | 91.9 |
| Non-verbal agreement | 39 | **100.0** | **100.0** | 89.7 | **97.4** | 92.3 | **94.9** | 92.3 | 92.3 | **94.9** | 89.7 | 94.4 |
| Subordination | 49 | **98.0** | **98.0** | **95.9** | **100.0** | 93.9 | 93.9 | 93.9 | **95.9** | 87.8 | 85.7 | 94.3 |
| Verb semantics | 7 | **100.0** | **100.0** | **100.0** | **85.7** | **71.4** | **85.7** | **71.4** | **71.4** | **85.7** | 42.9 | 81.4 |
| Verb tense/aspect/mood | 94 | **94.7** | **94.7** | **94.7** | **94.7** | **89.4** | **92.6** | **90.4** | **91.5** | **93.6** | 87.2 | 92.3 |
| Verb valency | 70 | 91.4 | 91.4 | 90.0 | 87.1 | 90.0 | 87.1 | 87.1 | 87.1 | 81.4 | 81.4 | 87.4 |
| micro-average | 465 | **93.1** | **93.1** | **93.1** | **91.8** | 89.5 | 89.2 | 89.5 | 89.7 | 86.5 | 83.7 | 89.9 |
| macro-average | 465 | **94.1** | **94.1** | **93.4** | **91.9** | 90.0 | 89.7 | 89.4 | 88.4 | 86.2 | 83.1 | 90.0 |
| our rank | | 1 | 1 | 1 | 1 | 5 | 5 | 5 | 5 | 9 | 10 | |
| WMT25 human rank | | 3 | 5 | 8 | 1 | 3 | 6 | 3 | 2 | 6 | 10 | |

| phenomenon | count | Wenyi | Algha | Yande | Gemin | Claud | DeepS | GPT41 | Shy | Comma | UvAMT | avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ambiguity | 10 | 100.0 | 100.0 | 90.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 80.0 | 97.0 |
| Lexical ambiguity | 10 | 100.0 | 100.0 | 90.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 80.0 | 97.0 |
| Coordination & ellipsis | 27 | 85.2 | 85.2 | 92.6 | 88.9 | 85.2 | 81.5 | 85.2 | 88.9 | 96.3 | 88.9 | 87.8 |
| Gapping | 5 | 80.0 | 80.0 | 100.0 | 100.0 | 80.0 | 80.0 | 60.0 | 60.0 | 80.0 | 60.0 | 78.0 |
| Pseudogapping | 7 | 71.4 | 71.4 | 85.7 | 100.0 | 85.7 | 71.4 | 100.0 | 100.0 | 100.0 | 100.0 | 88.6 |
| Right node raising | 5 | 80.0 | 80.0 | 100.0 | 80.0 | 80.0 | 80.0 | 80.0 | 80.0 | 100.0 | 80.0 | 84.0 |
| Sluicing | 2 | 100.0 | 100.0 | 100.0 | 50.0 | 100.0 | 50.0 | 50.0 | 100.0 | 100.0 | 100.0 | 85.0 |
| Stripping | 6 | 100.0 | 100.0 | 83.3 | 83.3 | 83.3 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 95.0 |
| VP-ellipsis | 2 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| False friends | 8 | 100.0 | 100.0 | 100.0 | 87.5 | 100.0 | 87.5 | 100.0 | 75.0 | 75.0 | 100.0 | 92.5 |
| Function word | 16 | 93.8 | 93.8 | 87.5 | 93.8 | 100.0 | 100.0 | 93.8 | 93.8 | 87.5 | 93.8 | 93.8 |
| Focus particle | 4 | 75.0 | 75.0 | 50.0 | 75.0 | 100.0 | 100.0 | 75.0 | 75.0 | 50.0 | 100.0 | 77.5 |
| Question tag | 12 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 91.7 | 99.2 |
| LDD & interrogatives | 27 | 96.3 | 96.3 | 88.9 | 92.6 | 88.9 | 85.2 | 81.5 | 92.6 | 85.2 | 85.2 | 89.3 |
| Inversion | 8 | 87.5 | 87.5 | 87.5 | 87.5 | 87.5 | 87.5 | 75.0 | 100.0 | 75.0 | 87.5 | 86.3 |
| Multiple connectors | 3 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 66.7 | 100.0 | 100.0 | 96.7 |
| Pied-piping | 7 | 100.0 | 100.0 | 85.7 | 100.0 | 85.7 | 85.7 | 85.7 | 100.0 | 85.7 | 100.0 | 92.9 |
| Preposition stranding | 3 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| Topicalization | 3 | 100.0 | 100.0 | 66.7 | 66.7 | 100.0 | 66.7 | 66.7 | 100.0 | 100.0 | 66.7 | 83.3 |
| Wh-movement | 3 | 100.0 | 100.0 | 100.0 | 100.0 | 66.7 | 66.7 | 66.7 | 66.7 | 66.7 | 33.3 | 76.7 |
| Lexical Morphology | 16 | 87.5 | 87.5 | 93.8 | 93.8 | 87.5 | 87.5 | 87.5 | 87.5 | 75.0 | 93.8 | 88.1 |
| Functional shift | 8 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| Noun formation (er) | 8 | 75.0 | 75.0 | 87.5 | 87.5 | 75.0 | 75.0 | 75.0 | 75.0 | 50.0 | 87.5 | 76.3 |
| MWE | 55 | **85.5** | **85.5** | **92.7** | 80.0 | 78.2 | 78.2 | **83.6** | **83.6** | 69.1 | 67.3 | 80.4 |
| Collocation | 9 | 88.9 | 88.9 | 88.9 | 88.9 | 77.8 | 77.8 | 77.8 | 88.9 | 66.7 | 66.7 | 81.1 |
| Compound | 6 | 66.7 | 66.7 | 66.7 | 33.3 | 50.0 | 66.7 | 66.7 | 66.7 | 66.7 | 50.0 | 60.0 |
| Idiom | 13 | **92.3** | **92.3** | **100.0** | **84.6** | **84.6** | 76.9 | **100.0** | 69.2 | 61.5 | 76.9 | 83.8 |
| Nominal MWE | 8 | **100.0** | **100.0** | **100.0** | **100.0** | **100.0** | **100.0** | **100.0** | **100.0** | **87.5** | 50.0 | 93.8 |
| Prepositional MWE | 7 | 100.0 | 100.0 | 100.0 | 100.0 | 71.4 | 100.0 | 85.7 | 100.0 | 85.7 | 71.4 | 91.4 |
| Verbal MWE | 12 | 66.7 | 66.7 | 91.7 | 66.7 | 75.0 | 58.3 | 66.7 | 83.3 | 58.3 | 75.0 | 70.8 |
| Named entity & terminology | 47 | **91.5** | **91.5** | **97.9** | **93.6** | **93.6** | **91.5** | **95.7** | **89.4** | **89.4** | 85.1 | 91.9 |
| Date | 17 | 88.2 | 88.2 | 100.0 | 94.1 | 94.1 | 88.2 | 100.0 | 94.1 | 100.0 | 94.1 | 94.1 |
| Domainspecific Term | 2 | 50.0 | 50.0 | 100.0 | 50.0 | 50.0 | 50.0 | 50.0 | 50.0 | 100.0 | 50.0 | 60.0 |
| Measuring Unit | 9 | 88.9 | 88.9 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 88.9 | 88.9 | 100.0 | 95.6 |
| Onomatopeia | 4 | 100.0 | 100.0 | 75.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 75.0 | 95.0 |
| Proper Name & Location | 15 | **100.0** | **100.0** | **100.0** | **93.3** | **93.3** | **93.3** | **93.3** | **86.7** | 73.3 | 73.3 | 90.7 |
| Non-verbal agreement | 39 | **100.0** | **100.0** | 89.7 | **97.4** | 92.3 | **94.9** | 92.3 | 92.3 | **94.9** | 89.7 | 94.4 |
| Coreference | 14 | **100.0** | **100.0** | **85.7** | **92.9** | **85.7** | **92.9** | **85.7** | **92.9** | **92.9** | 78.6 | 90.7 |
| Genitive | 10 | 100.0 | 100.0 | 100.0 | 100.0 | 90.0 | 90.0 | 90.0 | 90.0 | 90.0 | 90.0 | 94.0 |
| Personal Pronoun Coreference | 4 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| Substitution | 11 | 100.0 | 100.0 | 81.8 | 100.0 | 100.0 | 100.0 | 100.0 | 90.9 | 100.0 | 100.0 | 97.3 |
| Subordination | 49 | **98.0** | **98.0** | **95.9** | **100.0** | 93.9 | 93.9 | 93.9 | **95.9** | 87.8 | 85.7 | 94.3 |
| Adverbial clause | 1 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| Cleft sentence | 3 | 100.0 | 100.0 | 66.7 | 100.0 | 66.7 | 100.0 | 66.7 | 100.0 | 66.7 | 33.3 | 80.0 |
| Complex object | 9 | 88.9 | 88.9 | 100.0 | 100.0 | 88.9 | 77.8 | 88.9 | 88.9 | 88.9 | 88.9 | 90.0 |
| Contact clause | 3 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| Infinitive clause | 12 | **100.0** | **100.0** | **91.7** | **100.0** | **100.0** | **100.0** | **100.0** | **100.0** | 75.0 | **91.7** | 95.8 |
| Object clause | 3 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| Participle clause | 12 | 100.0 | 100.0 | 100.0 | 100.0 | 91.7 | 100.0 | 91.7 | 91.7 | 91.7 | 83.3 | 95.0 |
| Subject clause | 6 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 83.3 | 100.0 | 100.0 | 100.0 | 83.3 | 96.7 |
| Verb semantics | 7 | **100.0** | **100.0** | **100.0** | **85.7** | **71.4** | **85.7** | **71.4** | **71.4** | **85.7** | 42.9 | 81.4 |
| Verb tense/aspect/mood | 94 | **94.7** | **94.7** | **94.7** | **94.7** | **89.4** | **92.6** | **90.4** | **91.5** | **93.6** | 87.2 | 92.3 |
| Conditional | 12 | 100.0 | 100.0 | 100.0 | 100.0 | 91.7 | 100.0 | 100.0 | 100.0 | 100.0 | 91.7 | 98.3 |
| Ditransitive | 22 | **100.0** | **100.0** | 86.4 | **100.0** | **95.5** | **95.5** | **100.0** | 86.4 | **100.0** | 86.4 | 95.0 |
| Gerund | 5 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| Imperative | 16 | **87.5** | **87.5** | **100.0** | **87.5** | 81.3 | **87.5** | 81.3 | **87.5** | **93.8** | **87.5** | 88.1 |
| Intransitive | 22 | 90.9 | 90.9 | 90.9 | 90.9 | 86.4 | 90.9 | 86.4 | 90.9 | 95.5 | 86.4 | 90.0 |
| Reflexive | 11 | 90.9 | 90.9 | 100.0 | 90.9 | 90.9 | 90.9 | 90.9 | 90.9 | 81.8 | 90.9 | 90.9 |
| Transitive | 6 | 100.0 | 100.0 | 100.0 | 100.0 | 83.3 | 83.3 | 66.7 | 100.0 | 66.7 | 66.7 | 86.7 |
| Verb valency | 70 | 91.4 | 91.4 | 90.0 | 87.1 | 90.0 | 87.1 | 87.1 | 87.1 | 81.4 | 81.4 | 87.4 |
| Case government | 20 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| Catenative verb | 13 | **92.3** | **92.3** | **92.3** | **92.3** | **100.0** | **92.3** | **100.0** | **92.3** | 76.9 | **84.6** | 91.5 |

| phenomenon | count | Wenyi | Algha | Yande | Gemin | Claud | DeepS | GPT41 | Shy | Comma | UvAMT | avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mediopassive voice | 6 | 100.0 | 100.0 | 83.3 | 83.3 | 100.0 | 100.0 | 100.0 | 83.3 | 66.7 | 66.7 | 88.3 |
| Passive voice | 12 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 91.7 | 91.7 | 100.0 | 100.0 | 100.0 | 98.3 |
| Resultative | 8 | **100.0** | **100.0** | **87.5** | 62.5 | **87.5** | **87.5** | **75.0** | **87.5** | **75.0** | **75.0** | 83.8 |
| Semantic roles | 11 | 54.5 | 54.5 | 63.6 | 63.6 | 45.5 | 45.5 | 45.5 | 45.5 | 45.5 | 36.4 | 50.0 |
| micro-average | 465 | **93.1** | **93.1** | **93.1** | **91.8** | 89.5 | 89.2 | 89.5 | 89.7 | 86.5 | 83.7 | 89.9 |
| phen. macro-average | 465 | **93.0** | **93.0** | **92.5** | **90.5** | 89.0 | 88.3 | 87.4 | 89.4 | 86.6 | 82.4 | 89.2 |
| categ. macro-average | 465 | **94.1** | **94.1** | **93.4** | **91.9** | 90.0 | 89.7 | 89.4 | 88.4 | 86.2 | 83.1 | 90.0 |