Linguistically Motivated Evaluation of the 2023 State-of-the-art Machine Translation: Can GPT-4 Outperform NMT?

Shushen Manakhimova[1], Eleftherios Avramidis[1], Vivien Macketanz[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 offers a fine-grained analysis of the machine translation outputs in the context of the Shared Task at the 8th Conference of Machine Translation (WMT23). Building on the foundation of previous test suite efforts, our analysis includes Large Language Models and an updated test set featuring new linguistic phenomena. To our knowledge, this is the first fine-grained linguistic analysis for the GPT-4 (5-shot) translation outputs. Our evaluation spans German–English, English–German, and English–Russian language directions. Some of the phenomena with the lowest accuracies for German–English are *idioms* and *resultative predicates*. For English–German, these include *mediopassive voice*, and *noun formation(er)*. As for English–Russian, these included *idioms* and *semantic roles*. GPT-4 (5-shot) performs equally or comparably to the best systems in German–English and English–German but falls in the second significance cluster for English–Russian.

1 Introduction

Over the past few years, we have witnessed substantial advancements in Machine Translation (MT) alongside the rapid expansion of Large Language Models (LLMs). These developments have brought translation quality up to par with human capabilities. However, these seemingly perfect translations might contain fine-grained linguistic errors that go unnoticed or get overlooked entirely in automated evaluation. A more structured approach to identifying linguistic issues in the outputs involves the use of *test suites* or *challenge sets* to systematically evaluate the system's performance on specific tasks. The current study focuses on providing a fine-grained evaluation of the translation proficiency of the latest generation of Neural Machine Translation (NMT) against the latest generation of LLMs, exemplified by GPT-4 (5-shot). One of the objectives is therefore to assess whether GPT-4, as an LLM, excels NMT in managing specific linguistic phenomena. Although our focus lies on GPT-4, we are aware that there might be other LLMS participating in the sub-task.

In this context, we are presenting the results of the test suites analyzing state-of-the-art systems in terms of numerous linguistically motivated phenomena. These test suites[1] were applied to the MT systems submitted for evaluation at the 8th Conference on Machine Translation (WMT23; Kocmi et al., 2023) across multiple language directions: German–English, English–German, and English–Russian.

This paper is structured as follows: Section 2 goes through related work, whereas Section 3 explains how the test suite was created and applied. Section 4 outlines the setup of this year's experiment, whose results are detailed in Section 5. Section 6 concludes the paper with an outlook to future research.

2 Related Work

The origins of test suites can be traced back to the early days of machine translation in the 1990s (King and Falkedal, 1990; Way, 1991; Heid and Hildenbrand, 1991). Several researchers have adopted the use of test suites to achieve their goals. For instance, Guillou and Hardmeier (2016) employed test suites to evaluate pronoun translation. Other studies (e.g. Isabelle et al., 2017; Burchardt et al., 2017) compared different MT technologies, while Avramidis et al. (2018) explored their applicability in Quality Estimation methods.

The Machine Translation test suite track has played a significant role in this context, leading to the creation of test suites focusing on specific translation-related phenomena. For example, the work by Weller-di Marco and Fraser (2022) addressed the translation of morphologically complex

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

words from German into English. Additionally, Semenov and Bojar's research delved into document-level translation quality assessment. These test suites, however, focus on one or at most a few phenomena per test suite, including the works by Cinkova and Bojar (2018), Bojar et al. (2018), Burlot et al. (2018), Guillou et al. (2018), Rios et al. (2018), Popović (2019), Raganato et al. (2019), Rysová et al. (2019), Vojtěchová et al. (2019), Kocmi et al. (2020), Scherrer et al. (2020), Zouhar et al. (2020). Test suites, in conjunction with human evaluation, are also instrumental in assessing the quality of machine translation metrics (Freitag et al., 2021; Avramidis and Macketanz, 2022). Our approach enables a comprehensive analysis that spans over a hundred linguistic phenomena across three language pairs (Macketanz et al., 2022a). It incorporates semi-automated human evaluation, combining efficiency with in-depth analysis. Due to our participation in past shared tasks since 2018 (Macketanz et al., 2018b), we are able to analyze the development of machine translation systems over the years.

With the growing interest surrounding LLMs, researchers have been increasingly focused on evaluating GPT-'s performance in MT. For instance, the paper by Jiao et al. (2023) concludes that ChatGPT performs competitively with commercial translation products on high-resource European languages. A comprehensive evaluation across 18 languages of GPT models versus best-performing WMT-22 systems including human evaluations by Hendy et al. (2023) supports the previous finding. Other research explores these differences in terms of the literalness of translations produced by standard NMT and GPT-3 (Raunak et al., 2023). Castilho et al. (2023) have tested ChatGPT for handling context-related linguistic phenomena such as coreference, terminology, etc. to show that it performed even better than other MT engines. This current paper also places a specific focus on evaluating GPT's performance compared to other systems in the shared task.

3 Method

3.1 Test suite description

This paper focuses on three language pairs: German–English, English–German, and English–Russian. The test suite is built around specific linguistic categories, further divided into more detailed linguistic phenomena. While these categories and phenomena are specific to each language pair or direction, they may overlap across different directions. Although the logic of the test suite does not follow a particular linguistic theory, the categorization is based on linguistic research, established contrastive grammars, and findings from translation studies. The test suite was designed to cover a wide range of potential translation challenges, and its categories and phenomena were internally reviewed for objectivity by linguists and professional translators.

Test set	Test sentences	Categories	Phenomena
De–En	∼5,500	14	106
En–De	∼4,785	13	110
En–Ru	∼1232	12	51

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

Table 1 provides an overview of the number of test sentences, categories, and phenomena for each language pair. Notably, our English–Russian test set has more than doubled compared to last year, from 350 sentences (Macketanz et al., 2022b) to 1232. The new categories and phenomena have been added to the English–German direction as well.

To allow the evaluation of test sentences to operate semi-automatically, we have written rules that determine translation correctness. These rules include hand-crafted regular expressions and predefined translation outputs, applied using an internal evaluation tool (Macketanz et al., 2018a). Figure 1 illustrates the workflow of the preparation and application of our test suite.

3.2 Application of the test suite

The details regarding the development and application of our test suite are available in prior publications within the test suite track.(Macketanz et al., 2018c, 2021, 2022b; Avramidis et al., 2019, 2020). In this paper, we present an overview of the complete system. As shown in Figure 1, the building of the test suite follows steps a to c. Once test sentences are input to MT systems (step d), the test suite is applied, and automatic evaluation begins. This is done using predefined rules (step e). These rules are made of regular expressions and fixed strings, indicating correct and incorrect translations based on previous MT system outputs. Regular expressions are designed to evaluate translation accuracy for specific phenomena, possibly excluding unrelated errors. Sentences are flagged with warn-

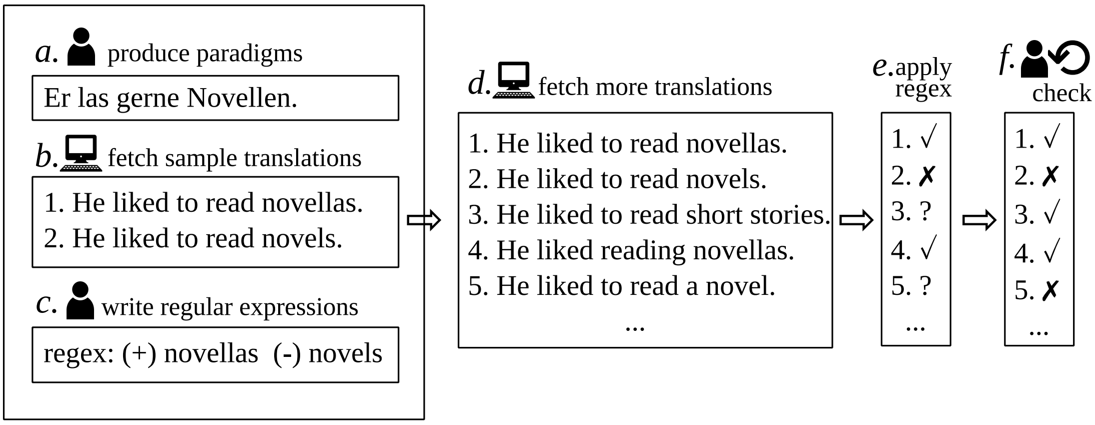
a. produce paradigms
Er las gerne Novellen.
b. fetch sample translations
1. He liked to read novellas.
2. He liked to read novels.
c. write regular expressions
regex: (+) novellas (-) novels
d. fetch more translations
1. He liked to read novellas.
2. He liked to read novels.
3. He liked to read short stories.
4. He liked reading novellas.
5. He liked to read a novel.
...
e. apply regex
1. √
2. ✗
3. ?
4. √
5. ?
...
f. check
1. √
2. ✗
3. √
4. √
5. ✗
...

Figure 1: Example of the preparation and application of the test suite for one test sentence

ings when they cannot be automatically sorted as correct or incorrect. Human linguist annotators review and adjust the rules, while sentences with critical language errors unrelated to the phenomena are deemed incorrect.

Subsequently, the translation accuracy specific to the phenomenon is calculated by dividing the number of correctly translated test sentences for that phenomenon by the total number of test sentences for that same phenomenon:

$$\text{accuracy} = \frac{\text{correct translations}}{\text{sum of test items}}$$

Since the goal is to ensure a fair comparison among systems, only the test items that do not have any warnings are included in the calculation. If a test item has at least one unresolved warning, we exclude it from the calculation. Such an approach reduces the total number of test items, which was crucial this year, as there were many problematic outputs.

We begin by identifying the highest-scoring system in each language direction and then compare it to other systems. To do so, we confirm the significance of the comparison with a one-tailed Z-test with $\alpha = 0.95$. Systems that do not significantly differ from the top-performing system are grouped into the first performance cluster, which is indicated with boldface in the respective rows of the tables.

Average scores are computed using three distinct methods to account for variations in the number of test items within each category or phenomenon. The micro-average method aggregates the contributions of all test items to calculate average percentages. Category macro-average computes the percentages independently for each category and then averages them, treating all categories equally. Similarly, the phenomenon macro-average computes percentages independently for each phenomenon and then averages them, treating all phenomena equally.

3.3 Addition of new phenomena

This year, we added some new phenomena and made an effort to make the new test items more challenging for the systems. For instance:

- Some test items are now spanned across multiple sentences. Previously, the *coreference* category had only one sentence test items e.g., *Susan dropped the plate, and it shattered loudly.* This year, some new test items divided into two sentences had been added e.g., *The cat climbed up a tree. It was afraid.*

- There was an effort to include sentences that vary in their length, ambiguity as well syntax complexity. For example, *He was also seen wearing harem-style trousers as he tapped his feet along with his new track* as well as

- to add phenomena that require inventive approach and cultural knowledge e.g., *onomatopoeia.*

4 Experiment Setup

In this paper, we present the evaluation of 37 systems with our test suite. The systems were submitted to the *news translation task* of the Eighth Conference on Machine Translation (WMT23; Kocmi et al., 2023): 13 systems for German–English, 12 systems for English–German, and 12 systems for English–Russian.

This year is the third time that the English–German systems are being evaluated with our test suite and the second time for the English–Russian systems. Every year, manual work is involved upon receiving the system translations as there are usually a number of translation outputs that are not yet covered by the existing rules in the database (the *warnings*). At the beginning of the evaluation process this year, there were on average 10.7 % of warnings for German–English, 15.6 % for English–German, and 70.6 % for English–Russian. The English–Russian test has grown significantly since last year and in comparison with the other sets had more new items that had not been evaluated before. It was also expected that English–German would have a higher amount of warnings than German–English as there were some new categories added to the English–German test suite.

One annotator with extensive linguistic knowledge of the three languages conducted the manual evaluation of the warnings; problematic cases were discussed with several translation experts to exclude subjectivity. The manual evaluation took around three and a half weeks and involved around 55 person-hours. After the manual evaluation, there were on average 7 % of warnings left for German–English, 6.8 % for English–German, and 6.9 % for English–Russian.

As mentioned above, test sentences with at least one warning by one system were excluded from the analysis to achieve a fair comparison between the systems under inspection. As this year, we saw a lot of problematic outputs that could not be properly evaluated, this report deals with a significantly less number of test items than in the previous years. We suspect that some of these can be explained by possible models' hallucinations: a number of the MT outputs this year had some parts of the sentences repeated twice or parts of the test items were not translated at all or seemed out of place altogether. To illustrate, one unevaluated output was from the phenomenon *intransitive-perfect* “Ich bin gerannt" (“I ran" or“I was running") that in the submission of Lan-Bridge (Wu and Hu, 2023) was rendered “I'm a manager".

As a result, our analysis was conducted on 3234 (58.9 %) test sentences for German–English, 3109 (64.8 %) test sentences for English–German, and 909 (73.8 %) test sentences for English–Russian.

5 Results

All result tables can be found in the Appendix.

5.1 System comparison

For **German–English**, GPT-4 (5-shot) produced micro and macro scores of 92.5 % and 91.6 % respectively, which puts GPT-4 (5-shot) into the cluster of top-performing systems. The highest micro averages ranging from 95.9-93 % were achieved by the systems Online-W, Online-A, and Online-Y. In terms of the macro average, Online-W, Online-A, and Online-B demonstrated the highest scores, ranging from 91.8 % to 92.7 %. The system with the lowest performance on the micro average this year was Lan-Bridge with 81.2 %, while the system with the lowest macro average was AIRC with 74.3 %.

For the **English–German** direction, GPT-4 (5-shot) leads with a micro average of 97.8 %, followed closely by Online-Y at 97.4 % and Online-B at 97.2 %. GPT-4, on the macro average, displays the highest score 92.9 %, followed by Online-W with 92.6 % and Online-B with 92 %. The system AIRC achieved the lowest scores: 87.1 % for micro and 71 % for macro. On average, systems get micro average of 95.4 % and macro average 86.7 %.

For **English–Russian**, only Online-G and Online-W stand out with the highest scores. Online-G achieves a micro average of 86.9 % and a macro average of 86.3 %, while Online-W achieves 86.8 % and 85.5 % respectively. GPT-4 doesn't end up in the top-performing cluster and GPT-4 gets the same micro average as Online-B 81.7 %. Online-B achieves 81.3 % on macro average and outperforms GPT-4 by 3.4 %. LanguageX and Lan-Bridge as the two systems with the lowest scores achieve micro scores of 65-65.7 % and macro of 61.1 %. Several factors, such as limited training data and substantial structural differences between the languages, contribute to the translation challenges for this language pair, compared to the relatively similar English–German pair.

5.2 Category-level analysis

In **German–English**, a few models achieve 100 % in categories such as *composition*, *named entity & terminology*, and *negation*. This might be attributed to the fact that these categories have well-defined rules that the models have mastered. Categories like *ambiguity* and *false friends* still show varied results, indicating their complexity. GPT-4 (5-shot) excels in many categories, scoring 91.0 % in *ambiguity* and 95.5 % in *ldd & interrogatives*. *Punctuation* is the most difficult category for GPT-4 (5-shot) achieving 76 % accuracy. One possible explanation is that GPT translations frequently include punctuation and other content not present in the original text (Hendy et al., 2023).

For **English–German**, the categories with the highest scores are *negation*, *verb tense/aspect/mood*, and *function word*. GPT-4 (5-shot) performs well in *function word* (97.6 %) and *ldd & interrogatives*, although NLLBG still outperforms GPT-4 in *ldd & interrogatives*. GPT-4 and NMT models can improve in categories like *subordination* and *verb valency*, where scores are often below 90 %.

For **English–Russian**, the category with the highest average score (89.4 %) is *punctuation*. Categories like *verb semantics* and *lexical Morphology* pose significant challenges. The categories with the lowest accuracy are *ambiguity* with 51.8 %, followed by *coordination & ellipsis*. However, GPT-4 achieves the lowest results in the category *false friends* with 61.5 % accuracy. GPT-4 performs best in *function word* (93.1 %) and *verb tense/aspect/mood* (85.9 %). The most challenging phenomenon for GPT-4 is *verb semantics* with a score of 47.1 %.

5.3 Phenomenon-level analysis

For **German–English**, the phenomenon macro-average for GPT-4 is 91.5 % with over 40 phenomena reaching a 100 % accuracy. There are no phenomena that reach 100 % accuracy across all models but some of the easier phenomena for most models include *phrasal verb*, *sluicing*, *polar question*, *ditransitive future I*, *passive voice* and other. The phenomena with the lowest accuracies are *idioms*, *modal negated - pluperfect*, and *resultative predicates*. In terms of *idioms*, GPT-4 performs better than most systems with 57.9 % accuracy.

Table 2 contains example outputs from two different phenomena for German–English. The first example comes from the phenomenon *extended adjective construction*, a frequent construction in German grammar, where the adjective is modified prepositional phrases or attributes. This structure tends to complicate the syntactic structure, making MT more challenging. The first translation is incorrect as it doesn't accurately convey the meaning of the original sentence. The second translation accurately conveys the meaning of the original sentence and uses correct English grammar. The third translation is also inaccurate due to the wrong word order and the incorrect use of an adjective instead of an adverb.

The second example contains a *resultative predicate*. The first translation is incorrect because it does not follow the correct word order in English. The word-to-word translation of the German sentence is taken too directly, resulting in an awkward and non-sensical English sentence. The second translation is correct. It accurately conveys the meaning of the original German sentence and uses a natural English construction to do so. The third translation is also incorrect as “having a wet effect" is not typically used to describe things that are “wet" or that “get wet".

Extended Adjective Construction	
Auf der anderen Straßenseite stand ein laut weinendes Kind.	
On the other side of the street was a noisy child.	fail
A child was crying loudly across the street.	pass
Across the street stood a loud crying child.	fail
Resultative Predicate	
Es regnete die Stühle nass.	
It rained wet the chairs.	fail
It rained and the chairs got wet.	pass
It had a wet effect on the chairs.	fail

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

Functional Shift	
You can whatsapp me on this number.	
Sie können mich per Whatsapp unter dieser Nummer erreichen.	pass
Sie können mich auf dieser Nummer wassappieren.	fail
Du kannst mich auf dieser Nummer aufpassen.	fail
Semantic Roles	
The bike accident broke Sarah’s arm.	
Der Fahrradunfall brach Sarah den Arm.	fail
Bei dem Fahrradunfall brach sich Sarah den Arm.	pass

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

For **English–German**, the phenomenon-level macro average is similarly high as for the other language direction with 93 %. The phenomena for which all systems reach near 100 % accuracy include *inversion*, *multiple connectors*, *pied-piping*, *prepositional mwe*, *substitution*, *adverbial clause* and others. Most of the phenomena achieve high accuracies over 85 %, with some exceptions including *stripping*, *topicalization*, *verb semantics*, *mediopassive voice*, and *noun formation(er)*.

Table 3 contains translation examples from English–German. The first example contains a *functional shift*. Functional shift, or conversion, is when a word switches from one word class, or part of speech without changing its form Cannon (1985). In the first output, we can observe a correct structural change with the use of a common German prepositional phrase. In the second output, however, the word "wassappieren" is not a valid German word, resulting in an incomprehensible translation. Similarly, the third translation is also not a valid German sentence, it introduces a different verb, "aufpassen", which means "to look after" and doesn't fit the original meaning of the sentence. The second example deals with the problem of *semantic roles* also known as *thematic relations*. English has a broad range of semantic roles in the subject position and while German also allows for non-agentive semantic roles to be expressed as subjects, it may be more restrictive than English. In the incorrect translation, the accident itself is depicted as the direct agent of the action, which is unusual for German. According to the accurate translation, which follows the typical German sentence form, "Sarah's arm broke as a result of the accident".

For **English–Russian**, the phenomenon level macro-average accuracy lies at 77 %. In this year's submission, the following phenomena reached 97-100 % accuracy: *prepositional mwe*, *contact clause*, *object clause*. The two phenomena reaching the lowest accuracies were *idioms* and *semantic roles* with less than 40 % averages. The low accuracy for *idioms* and *semantic roles* are not surprising as t expressions still cause translation errors across all language pairs. GPT-4 performs as the fourth-best system in all the averages, showing the lowest result for *semantic roles* as well.

Table 4 covers translation examples in English–Russian. For instance, the translation of a problematic English *compound* "skin-deep" into Russian. The first translation "Он отрицает, что расизм — это просто глубинка" means in Russian "He denies that racism is just a small rural town." "Глубинка" does have the same root as the word "deep" in Russian but has a completely different meaning, which makes this translation incorrect. The second structure is correct as it uses the adjective "поверхностен" or "superficial". The third translation is also incorrect as it means "He denies that racism is only about skin color" and states that the issue of racism is related to skin color, which was not present in the test item. The second example comes from the phenomenon *idiom*. This example includes a very common English non-literal expression "silver lining" meaning that there might be a positive aspect to a situation that may initially appear depressing or hopeless. The first translation correctly interprets the English idiom using a popular expression in Russian, "луч надежды" (ray of hope), reflecting the idea that even in bad times, there is always hope for something positive. The second translation renders the idiom literally. The Russian phrase "серебряная подкладка"(silver underlay) is not commonly used and does not accurately express the original meaning. In the third translation, an appropriate Russian proverb "There is no bad without good" is used to convey the meaning, but there's an error in the Russian expression: instead of "худа", there is a non-existent word "худ", making this translation incorrect.

Compound	
He denies that racism is just skin-deep.	
Он отрицает, что расизм — это просто глубинка.	fail
Он отрицает, что расизм поверхностен.	pass
Он отрицает, что расизм сводится только к цвету кожи.	fail
Idiom	
When things look black, there's always a silver lining.	
Когда все выглядит мрачно, всегда есть луч надежды.	pass
Когда все выглядит черным, всегда есть серебряная подкладка.	fail
Когда все выглядит черным, всегда есть худ без добра.	fail

Table 4: Examples of English–Russian linguistic phenomena with passing and failing MT outputs.

5.4 Comparison with previous years

The progress of the systems' accuracy for particular categories through the last years can be seen in Table 8 for German–English (since 2018), Table 9 for English–German (since 2021) and Table 10 for English–Russian (since 2022). The calculation has been done based on the common test items without warnings over the years. Compared to last year, the micro- and macro-average scores for the German-English systems included in the comparison have either shown very small improvement or remained the same. For English–German, 3 systems (Online-G, Y, and W) showed an improvement, which in some categories sums up to several percentage points. In English–Russian, 5 out of the 7 the systems (Online-A, G, W, Y, and PROMT) showed an improvement which averages to 1-5 %. Whereas we have little information about the development behind the online systems, we can assume that English–Russian is still in active development, English–German has undergone minor improvements, whereas there seems to have been no development for German-English.

Interestingly enough, the Lan-Bridge performance has gotten worse both in micro and macro averages compared to last year. The drop in performance is important in light of Lan-Bridge's own system description. Their approach in the WMT23 competition has been shaped by the shift towards large-scale models and lies on prompt-based experiments. To understand the specific reasons for Lan-Bridge's drop in performance, a detailed analysis of their models, data, experiment designs, and evaluation metrics would be necessary.

6 Conclusions and Outlook

This paper presents a fine-grained, linguistically motivated test suite to evaluate machine translation outputs. The test suite was applied to evaluate and compare the outputs of 37 machine translation systems in three different language pairs: German–English, English–German, and English–Russian.

While the evaluation showed high scores for all language pairs, there was a clear drop in accuracy when dealing with structurally different languages, such as English and Russian. For this language pair, GPT-4's performance falls in the second significance cluster. Although we didn't observe a systematic significant difference between GPT-4 (5-shot) and other systems, it is important to highlight that GPT-4 shows competitive results in the context of our evaluation. This indicates that GPT-4, a general model, remains competitive in MT and sometimes performs better than some specialized NMT systems. Nevertheless, many linguistic nuances still pose difficulties for these models, demonstrating the continuous need for study and improvement in the field of MT. In terms of linguistic coverage, the current test suite stands out as one of the most extensive available. The semi-automated approach offers a more effective, while still fine-grained analysis in comparison to a typical human evaluation. When paired with other automated metrics or MQM analysis, this method can be seen as a valuable addition offering deeper insights into translation quality. The test suite approach is also highly versatile, allowing for the analysis of various tasks performed by LLMs in different contexts.

Limitations

The current test suite, evolving since 2016, was originally designed to evaluate weaker MT systems and focused on simpler linguistic phenomena. While we've introduced complexity with multi-sentence test items and more intricate sentences, it could be done only for a handful of phenomena and sentences. There are other limitations to consider. Firstly, this analysis is mostly limited to a sentence-level analysis. Secondly, all phenomena and categories are treated equally, although they may vary in their complexity. As mentioned earlier, the current evaluation rules prioritize accuracy in translating specific linguistic phenomena, sometimes at the expense of overall natural fluency, resulting in technically correct but less fluent outputs. To address some of these limitations, we consider including a linguistic acceptability score and an inter-annotator agreement score in future evaluations.

Acknowledgements

This research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft; DFG) through the project TextQ, by the German Federal Ministry of Education through the project SocialWear (grant num. 01IW20002). We would like to thank 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 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, Arle Lommel, and Hans Uszkoreit. 2018. Fine-grained evaluation of quality estimation for machine translation based on a linguistically motivated test suite. In *Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing*, pages 243–248, Boston, MA. Association for Machine Translation in the Americas.

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.

Ondřej Bojar, Jiří Mírovský, Kateřina Rysová, and Magdaléna Rysová. 2018. EvalD Reference-Less Discourse Evaluation for WMT18. In *Proceedings of the Third Conference on Machine Translation*, pages 545–549, Belgium, Brussels. Association for Computational Linguistics.

Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter, and Philip Williams. 2017. A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. *The Prague Bulletin of Mathematical Linguistics*, 108:159–170.

Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondřej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen, and François Yvon. 2018. The WMT'18 Morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English. In *Proceedings of the Third Conference on Machine Translation*, pages 550–564, Belgium, Brussels. Association for Computational Linguistics.

Garland Cannon. 1985. Functional shift in english.

Sheila Castilho, Clodagh Quinn Mallon, Rahel Meister, and Shengya Yue. 2023. Do online Machine Translation Systems Care for Context? What About a GPT Model? In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 393–417, Tampere, Finland.

Silvie Cinkova and Ondřej Bojar. 2018. Testsuite on Czech–English Grammatical Contrasts. In *Proceedings of the Third Conference on Machine Translation*, pages 565–575, Belgium, Brussels. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondrej Bojar. 2021. Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia. European Language Resources Association (ELRA).

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A Pronoun Test Suite Evaluation of the English–German MT Systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation*, pages 576–583, Belgium, Brussels. Association for Computational Linguistics.

Ulrich Heid and Elke Hildenbrand. 1991. Some practical experience with the use of test suites for the evaluation of SYSTRAN. In *the Proceedings of the Evaluators' Forum, Les Rasses*. Citeseer.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

Margaret King and Kirsten Falkedal. 1990. Using test suites in evaluation of machine translation systems. In *Proceedings of the 13th conference on Computational Linguistics*, volume 2, pages 211–216, Morristown, NJ, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Christof Monz, Makoto Morishita, Murray Kenton, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23). In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at wmt 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 357–364, Online. Association for Computational Linguistics.

Vivien Macketanz, Renlong Ai, Aljoscha Burchardt, and Hans Uszkoreit. 2018a. TQ-AutoTest – an automated test suite for (machine) translation quality. In *Proceedings of the Eleventh International Conference on*

Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, and Hans Uszkoreit. 2018b. 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, and Hans Uszkoreit. 2018c. Fine-grained evaluation of German-English Machine Translation based on a Test Suite. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 584–593, 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 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.

Maja Popović. 2019. Evaluating Conjunction Disambiguation on English-to-German and French-to-German WMT 2019 Translation Hypotheses. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 464–469, Florence, Italy. Association for Computational Linguistics.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The MuCoW Test Suite at WMT 2019: Automatically Harvested Multilingual Contrastive Word Sense Disambiguation Test Sets for Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan Awadalla. 2023. Do gpts produce less literal translations?

Annette Rios, Mathias Müller, and Rico Sennrich. 2018. The Word Sense Disambiguation Test Suite at WMT18. In *Proceedings of the Third Conference on Machine Translation*, pages 594–602, Belgium, Brussels. Association for Computational Linguistics.

Kateřina Rysová, Magdaléna Rysová, Tomáš Musil, Lucie Poláková, and Ondřej Bojar. 2019. A Test Suite and Manual Evaluation of Document-Level NMT at WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 455–463, Florence, Italy. Association for Computational Linguistics.

Yves Scherrer, Alessandro Raganato, and Jörg Tiedemann. 2020. The MUCOW word sense disambiguation test suite at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 365–370, Online. Association for Computational Linguistics.

Kirill Semenov and Ondřej Bojar. 2022. Automated evaluation metric for terminology consistency in MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 450–457, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tereza Vojtěchová, Michal Novák, Miloš Klouček, and Ondřej Bojar. 2019. SAO WMT19 Test Suite: Machine Translation of Audit Reports. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 481–493, Florence, Italy. Association for Computational Linguistics.

Andrew Way. 1991. Developer-Oriented Evaluation of MT Systems. In *Proceedings of the Evaluators' Forum*, pages 237–244, Les Rasses, Vaud, Switzerland. ISSCO.

Marion Weller-di Marco and Alexander Fraser. 2022. Test suite evaluation: Morphological challenges and pronoun translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 458–468, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yangjian Wu and Gang Hu. 2023. Exploring Prompt Engineering with GPT Language Models for Document-Level Machine Translation: Insights and Findings. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Vilém Zouhar, Tereza Vojtěchová, and Ondřej Bojar. 2020. WMT20 Document-Level Markable Error Exploration. In *Proceedings of the Fifth Conference on Machine Translation*, pages 371–380, Online. Association for Computational Linguistics.

A Analysis based on categories

categ	count	Onl-W	Onl-A	Onl-B	ChatG	Onl-M	Onl-Y	NLLBM	NLLBG	Onl-G	LanBr	GTCOM	ZengH	AIRC	avg
Ambiguity	78	**85.9**	**88.5**	**93.6**	**91.0**	84.6	**87.2**	**87.2**	84.6	**87.2**	78.2	75.6	**88.5**	62.8	84.2
Composition	45	**100.0**	**100.0**	**97.8**	**100.0**	**97.8**	**100.0**	93.3	**95.6**	**95.6**	91.1	**95.6**	**95.6**	77.8	95.4
Coordination & ellipsis	49	**93.9**	**93.9**	**91.8**	**89.8**	77.6	**91.8**	**85.7**	**83.7**	**93.9**	77.6	**91.8**	**87.8**	81.6	87.8
False friends	36	**91.7**	**86.1**	**77.8**	**83.3**	**83.3**	69.4	**83.3**	**80.6**	**80.6**	75.0	75.0	72.2	52.8	77.8
Function word	61	**90.2**	**93.4**	**93.4**	**91.8**	**91.8**	**88.5**	**95.1**	**91.8**	**90.2**	78.7	83.6	52.5	65.6	85.1
LDD & interrogatives	154	87.0	90.3	88.3	**95.5**	87.7	87.7	87.0	89.6	90.3	79.2	85.1	72.1	66.2	85.1
MWE	76	**90.8**	**82.9**	**82.9**	**88.2**	77.6	80.3	**81.6**	**82.9**	80.3	71.1	76.3	**84.2**	53.9	79.5
Named entity & terminology	20	**95.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.0**	50.0	0.0	**90.0**	86.9
Negation	19	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	94.7	99.6
Non-verbal agreement	60	**93.3**	90.0	**96.7**	**93.3**	**95.0**	**98.3**	**96.7**	**96.7**	88.3	86.7	81.7	**95.0**	71.7	91.0
Punctuation	50	**100.0**	**100.0**	94.0	76.0	**100.0**	74.0	74.0	74.0	64.0	84.0	70.0	50.0	94.0	81.1
Subordination	158	**91.1**	89.2	**92.4**	**91.8**	**92.4**	**93.7**	**94.9**	**93.0**	**92.4**	75.9	86.7	85.4	76.6	88.9
Verb tense/aspect/mood	2347	**93.7**	**94.0**	91.4	**92.9**	88.0	**94.0**	84.3	84.4	**93.1**	81.8	**93.8**	**93.8**	86.6	90.1
Verb valency	81	**84.0**	**84.0**	**85.2**	**88.9**	**85.2**	**84.0**	**85.2**	**87.7**	77.8	77.8	**82.7**	**81.5**	65.4	82.2
micro-average	3234	**92.9**	**93.0**	91.2	**92.5**	88.3	**92.5**	85.6	85.6	91.5	81.2	90.7	89.4	82.2	89.0
macro-average	3234	**92.6**	**92.3**	**91.8**	**91.6**	90.1	89.2	89.2	88.9	88.1	82.3	82.0	75.6	74.3	86.8

Table 5: Accuracies (%) of successful translations on the category level for German–English. Boldface indicates the significantly best performing systems per row.

categ	count	ChatG	Onl-W	Onl-B	Onl-A	Onl-Y	NLLBG	Onl-G	NLLBM	Onl-M	ZengH	LanBr	AIRC	avg
Ambiguity	24	**95.8**	**95.8**	**91.7**	**87.5**	**83.3**	**83.3**	**87.5**	**83.3**	**87.5**	**91.7**	75.0	50.0	84.4
Coordination & ellipsis	74	**90.5**	78.4	**93.2**	**85.1**	**93.2**	70.3	**90.5**	67.6	82.4	74.3	71.6	63.5	80.1
False friends	33	**93.9**	**93.9**	**97.0**	**93.9**	**93.9**	**97.0**	**90.9**	**97.0**	**93.9**	**93.9**	**93.9**	81.8	93.4
Function word	41	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	75.6	**97.6**	85.4	94.7
LDD & interrogatives	131	**96.9**	**96.2**	**95.4**	**96.9**	**96.9**	**94.7**	**93.9**	**93.9**	**93.9**	88.5	**92.4**	84.7	93.7
Lexical Morphology	28	**85.7**	**85.7**	**82.1**	**75.0**	**67.9**	**67.9**	**64.3**	**64.3**	57.1	**82.1**	42.9	25.0	66.7
MWE	95	**95.8**	**97.9**	**96.8**	91.6	**95.8**	85.3	89.5	86.3	86.3	92.6	78.9	68.4	88.8
Named entity & terminology	73	**95.9**	**95.9**	**95.9**	**97.3**	**97.3**	83.6	**94.5**	87.7	**94.5**	87.7	78.1	90.4	91.6
Negation	13	100.0	100.0	100.0	100.0	100.0	92.3	100.0	92.3	100.0	100.0	100.0	100.0	98.7
Non-verbal agreement	90	**97.8**	**94.4**	90.0	88.9	92.2	**94.4**	**93.3**	**95.6**	**95.6**	92.2	87.8	74.4	91.4
Punctuation	36	83.3	**97.2**	80.6	**88.9**	77.8	80.6	80.6	**86.1**	83.3	61.1	80.6	72.2	81.0
Subordination	136	**99.3**	**97.1**	**97.8**	**97.8**	**96.3**	**97.8**	**97.8**	**97.8**	**97.8**	**97.1**	**99.3**	92.6	97.4
Verb semantics	4	75.0	75.0	75.0	50.0	50.0	100.0	50.0	75.0	50.0	50.0	50.0	25.0	60.4
Verb tense/aspect/mood	2237	99.1	98.4	98.7	99.0	**99.6**	97.0	99.1	97.1	98.4	99.2	97.2	91.6	97.9
Verb valency	94	**86.2**	**86.2**	**88.3**	**86.2**	**79.8**	77.7	76.6	**80.9**	**78.7**	**86.2**	72.3	59.6	79.9
micro-average	3109	**97.8**	97.0	**97.2**	97.0	**97.4**	94.4	96.6	94.7	95.9	95.9	93.5	87.1	95.4

categ	count	ChatG	Onl-W	Onl-B	Onl-A	Onl-Y	NLLBG	Onl-G	NLLBM	Onl-M	ZengH	LanBr	AIRC	avg
macro-average	3109	**92.9**	**92.6**	**92.0**	89.0	88.1	88.0	87.1	86.8	86.5	84.8	81.2	71.0	86.7

Table 6: Accuracies (%) of successful translations on the category level for English–German. Boldface indicates the significantly best performing systems per row.

categ	count	Onl-G	Onl-W	Onl-B	ChatG	Onl-Y	Onl-A	NLLBM	NLLBG	Onl-M	PROMT	ZengH	LanBr	avg
Ambiguity	20	**70.0**	**60.0**	50.0	**85.0**	55.0	45.0	50.0	50.0	35.0	30.0	45.0	25.0	50.0
Coordination & ellipsis	89	**82.0**	**83.1**	67.4	**77.5**	68.5	65.2	62.9	66.3	67.4	58.4	50.6	49.4	66.6
False friends	14	85.7	85.7	78.6	64.3	85.7	71.4	57.1	64.3	64.3	57.1	71.4	50.0	69.6
Function word	29	**96.6**	**96.6**	**96.6**	**93.1**	**82.8**	**82.8**	**96.6**	**93.1**	**96.6**	**86.2**	37.9	75.9	86.2
LDD & interrogatives	61	**95.1**	**95.1**	**91.8**	**88.5**	**93.4**	**91.8**	**88.5**	**88.5**	85.2	85.2	73.8	78.7	88.0
Lexical Morphology	29	**86.2**	**86.2**	**75.9**	**86.2**	**65.5**	62.1	62.1	**65.5**	41.4	51.7	58.6	55.2	66.4
MWE	71	**76.1**	**73.2**	**76.1**	**70.4**	59.2	**69.0**	**67.6**	**66.2**	**60.6**	**60.6**	**69.0**	54.9	66.9
Named entity & terminology	71	**87.3**	**77.5**	**81.7**	73.2	69.0	**76.1**	63.4	63.4	69.0	59.2	**80.3**	60.6	71.7
Negation	4	75.0	100.0	100.0	75.0	100.0	75.0	75.0	75.0	100.0	75.0	100.0	50.0	83.3
Non-verbal agreement	80	**76.3**	**86.3**	**75.0**	**82.5**	73.8	72.5	**81.3**	**81.3**	73.8	**75.0**	66.3	65.0	75.7
Punctuation	12	**100.0**	**83.3**	**91.7**	66.7	75.0	**100.0**	**83.3**	**83.3**	66.7	**91.7**	0.0	**91.7**	77.8
Subordination	130	**93.8**	**96.9**	**93.8**	**93.8**	**93.8**	90.0	86.9	88.5	**93.8**	89.2	68.5	83.1	89.4
Verb semantics	17	**94.1**	**82.4**	**76.5**	47.1	58.8	**76.5**	52.9	47.1	58.8	58.8	58.8	41.2	62.7
Verb tense/aspect/mood	156	**91.7**	**94.2**	85.9	85.9	87.2	87.8	84.0	82.1	83.3	84.0	66.7	75.0	84.0
Verb valency	126	**84.9**	**81.7**	**79.4**	**78.6**	**77.0**	72.2	68.3	64.3	73.0	70.6	69.8	60.3	73.3
micro-average	909	**86.9**	**86.8**	81.7	81.7	78.3	78.0	75.2	74.8	75.4	72.9	65.0	65.7	76.9
macro-average	909	**86.3**	**85.5**	81.3	77.9	76.3	75.8	72.0	71.9	71.3	68.9	61.1	61.1	74.1

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

B Comparison through the years

		Lan-Bridge		online-A						online-B						online-G						online-W			online-Y				
category	count	22	23	18	19	20	21	22	23	18	19	20	21	22	23	18	19	20	21	22	23	21	22	23	18	19	21	22	23
Ambiguity	75	89	80	69	71	77	83	84	88	76	77	79	85	93	93	72	75	84	85	88	88	85	84	87	68	79	83	84	88
Composition	45	98	91	82	93	96	96	96	100	98	98	98	100	98	98	73	87	98	98	98	96	96	96	100	91	93	93	100	100
Coordination & ellipsis	28	86	71	86	86	86	89	89	89	86	86	89	89	93	89	50	64	75	89	89	89	89	89	89	86	86	89	89	89
False friends	36	83	75	72	72	69	83	83	86	75	78	81	75	78	78	72	72	78	81	83	81	86	81	92	67	92	75	78	69
Function word	58	95	81	84	90	88	91	91	95	78	78	93	88	95	95	50	93	93	95	93	91	95	93	90	91	91	84	88	91
LDD & interrogatives	72	92	81	79	75	85	89	89	93	85	85	89	94	92	92	64	72	92	88	89	90	89	90	92	83	79	90	85	89
MWE	64	77	73	67	67	73	81	83	83	73	73	78	78	80	80	67	69	81	81	81	81	89	89	91	72	73	75	83	81
Named entity & terminology	9	100	100	100	100	100	100	100	100	100	100	78	100	100	100	100	100	100	100	100	100	100	100	89	89	100	100	100	100
Negation	16	100	100	100	100	100	100	100	100	94	94	100	100	100	100	63	100	100	100	100	100	100	100	100	100	100	100	100	100

category	count	Lan-Bridge		online-A						online-B						online-G						online-W			online-Y				
		22	23	18	19	20	21	22	23	18	19	20	21	22	23	18	19	20	21	22	23	21	22	23	18	19	21	22	23
Non-verbal agreement	55	98	85	78	84	84	93	93	91	87	87	87	98	96	96	58	82	91	91	89	89	96	96	95	80	82	85	93	96
Punctuation	33	94	94	100	100	100	100	100	100	97	97	97	100	94	94	88	88	91	91	91	91	97	100	100	100	100	100	100	97
Subordination	87	94	76	94	79	94	94	94	95	87	89	94	95	97	97	82	90	93	91	94	93	93	92	95	93	93	93	94	95
Verb tense/aspect/mood	2775	87	79	80	89	86	90	90	90	82	82	84	83	86	86	55	74	88	84	89	90	90	88	92	78	80	81	88	89
Verb valency	56	88	84	79	84	88	88	88	88	82	82	91	89	88	88	71	79	88	88	84	84	89	89	88	80	82	82	86	86
micro-average	3409	88	79	80	88	86	90	90	90	83	83	85	85	87	87	57	76	88	85	89	90	90	88	92	79	81	82	89	89
macro-average	3409	91	84	84	85	88	91	91	93	86	86	88	91	92	92	69	82	89	90	91	90	93	92	93	84	88	88	91	91

Table 8: Comparisons of the accuracy (%) of several German–English systems through the years.

category	count	Lan-Bridge		online-A			online-B			online-G			online-W			online-Y		
		22	23	21	22	23	21	22	23	21	22	23	21	22	23	21	22	23
Ambiguity	24	83.3	75.0	91.7	87.5	87.5	91.7	91.7	91.7	75.0	83.3	87.5	95.8	95.8	95.8	70.8	79.2	83.3
Coordination & ellipsis	68	86.8	63.2	70.6	82.4	79.4	82.4	88.2	88.2	73.5	85.3	88.2	66.2	67.6	69.1	67.6	76.5	86.8
False friends	36	86.1	86.1	86.1	86.1	88.9	83.3	88.9	91.7	83.3	91.7	83.3	88.9	91.7	91.7	86.1	86.1	86.1
Function word	39	97.4	97.4	97.4	97.4	97.4	100.0	97.4	97.4	97.4	97.4	97.4	100.0	100.0	97.4	97.4	97.4	97.4
MWE	96	87.5	81.3	86.5	89.6	91.7	92.7	95.8	96.9	81.3	89.6	90.6	93.8	97.9	97.9	80.2	85.4	95.8
Named entity & terminology	64	98.4	79.7	96.9	96.9	96.9	93.8	100.0	96.9	81.3	93.8	95.3	98.4	95.3	96.9	93.8	96.9	98.4
Negation	15	100.0	93.3	93.3	100.0	100.0	93.3	100.0	93.3	93.3	93.3	93.3	100.0	100.0	100.0	100.0	100.0	100.0
Non-verbal agreement	64	98.4	96.9	96.9	96.9	96.9	96.9	96.9	96.9	95.3	98.4	98.4	95.3	96.9	95.3	95.3	95.3	98.4
Punctuation	18	66.7	66.7	94.4	94.4	83.3	66.7	66.7	66.7	50.0	66.7	66.7	94.4	88.9	94.4	66.7	66.7	66.7
Subordination	129	98.4	99.2	98.4	98.4	97.7	97.7	98.4	97.7	93.8	99.2	97.7	97.7	97.7	98.4	94.6	93.8	96.1
Verb tense/aspect/mood	2526	99.3	97.3	96.1	98.6	98.7	99.1	98.8	98.4	94.9	97.6	98.9	96.8	96.6	98.1	93.0	96.6	99.6
Verb valency	76	88.2	78.9	84.2	86.8	93.4	88.2	89.5	92.1	77.6	88.2	85.5	89.5	88.2	90.8	80.3	85.5	86.8
micro-average	3155	97.9	94.9	94.9	97.4	97.5	97.7	97.9	97.6	92.8	96.5	97.4	95.9	95.8	97.1	91.6	95.0	98.2
macro-average	3155	90.9	84.6	91.0	92.9	92.6	90.5	92.7	92.3	83.1	90.4	90.2	93.1	93.1	93.8	85.5	88.3	91.3

Table 9: Comparisons of the accuracy (%) of several English–German systems through the years.

category	count	Lan-Bridge		online-A		online-B		online-G		online-W		online-Y		PROMT	
		22	23	22	23	22	23	22	23	22	23	22	23	22	23
Ambiguity	7	71.0	57.0	71.0	86.0	86.0	86.0	86.0	86.0	100.0	86.0	71.0	86.0	57.0	71.0
Coordination & ellipsis	30	50.0	40.0	43.0	60.0	57.0	57.0	80.0	80.0	73.0	77.0	53.0	60.0	60.0	53.0
False friends	5	60.0	80.0	60.0	60.0	80.0	80.0	80.0	80.0	80.0	80.0	60.0	80.0	60.0	60.0
Function word	10	80.0	60.0	80.0	90.0	90.0	90.0	90.0	90.0	90.0	100.0	90.0	70.0	80.0	80.0
MWE	32	63.0	59.0	63.0	66.0	75.0	75.0	69.0	72.0	75.0	75.0	66.0	66.0	63.0	66.0
Named entity & terminology	22	82.0	64.0	68.0	77.0	86.0	86.0	91.0	95.0	77.0	68.0	73.0	77.0	73.0	68.0

		Lan-Bridge		online-A		online-B		online-G		online-W		online-Y		PROMT	
category	count	22	23	22	23	22	23	22	23	22	23	22	23	22	23
Negation	4	100.0	50.0	75.0	75.0	100.0	100.0	75.0	75.0	100.0	100.0	100.0	100.0	75.0	75.0
Non-verbal agreement	10	80.0	50.0	80.0	80.0	90.0	90.0	80.0	80.0	80.0	90.0	70.0	80.0	80.0	80.0
Punctuation	5	80.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	80.0	80.0	100.0	80.0
Subordination	48	90.0	81.0	81.0	83.0	92.0	92.0	90.0	92.0	92.0	98.0	79.0	96.0	81.0	85.0
Verb tense/aspect/mood	61	77.0	75.0	77.0	82.0	75.0	79.0	77.0	84.0	75.0	89.0	70.0	74.0	77.0	82.0
Verb valency	30	73.0	53.0	83.0	80.0	77.0	77.0	87.0	83.0	90.0	83.0	77.0	77.0	73.0	80.0
micro-average	264	75.0	65.0	72.0	77.0	80.0	80.0	82.0	84.0	82.0	86.0	72.0	77.0	73.0	75.0
macro-average	264	75.0	64.0	74.0	78.0	84.0	84.0	84.0	85.0	86.0	87.0	74.0	79.0	73.0	73.0

Table 10: Comparisons of the accuracy (%) of several German–English systems through the years.

C Detailed analysis on a phenomenon-level

categ	count	Onl-W	Onl-A	Onl-B	ChatG	Onl-M	Onl-Y	NLLBM	NLLBG	Onl-G	LanBr	GTCOM	ZengH	AIRC	avg
Ambiguity	78	**85.9**	**88.5**	**93.6**	**91.0**	84.6	**87.2**	**87.2**	84.6	**87.2**	78.2	75.6	**88.5**	62.8	84.2
Lexical ambiguity	62	**91.9**	**93.5**	**95.2**	**90.3**	85.5	**87.1**	**90.3**	85.5	**88.7**	80.6	79.0	**90.3**	67.7	86.6
Structural ambiguity	16	62.5	**68.8**	**87.5**	**93.8**	**81.3**	**87.5**	**75.0**	**81.3**	**81.3**	**68.8**	62.5	**81.3**	43.8	75.0
Composition	45	**100.0**	**100.0**	**97.8**	**100.0**	**97.8**	**100.0**	93.3	**95.6**	**95.6**	91.1	**95.6**	**95.6**	77.8	95.4
Compound	26	**100.0**	**100.0**	**100.0**	**100.0**	**96.2**	**100.0**	88.5	**92.3**	**96.2**	84.6	**92.3**	**96.2**	76.9	94.1
Phrasal verb	19	**100.0**	**100.0**	**94.7**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.7**	**100.0**	**100.0**	**94.7**	78.9	97.2
Coordination & ellipsis	49	**93.9**	**93.9**	**91.8**	**89.8**	77.6	**91.8**	**85.7**	**83.7**	**93.9**	77.6	**91.8**	**87.8**	81.6	87.8
Gapping	19	**100.0**	**100.0**	**100.0**	**89.5**	**94.7**	**94.7**	**89.5**	**89.5**	**100.0**	73.7	**94.7**	**94.7**	**89.5**	93.1
Right node raising	18	**83.3**	**83.3**	**83.3**	**83.3**	50.0	**83.3**	**72.2**	**66.7**	**83.3**	**66.7**	**83.3**	**88.9**	**61.1**	76.1
Sluicing	12	**100.0**	**100.0**	**91.7**	**100.0**	**91.7**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	75.0	**100.0**	96.8
False friends	36	**91.7**	**86.1**	**77.8**	**83.3**	**83.3**	69.4	**83.3**	**80.6**	**80.6**	75.0	75.0	72.2	52.8	77.8
Function word	61	**90.2**	**93.4**	**93.4**	**91.8**	**91.8**	**88.5**	**95.1**	**91.8**	**90.2**	78.7	83.6	52.5	65.6	85.1
Focus particle	22	**95.5**	**100.0**	**100.0**	**100.0**	**100.0**	**95.5**	**95.5**	**90.9**	**100.0**	**90.9**	**90.9**	**95.5**	81.8	95.1
Modal particle	20	**80.0**	**85.0**	**80.0**	**75.0**	**75.0**	**75.0**	**90.0**	**85.0**	**70.0**	50.0	**70.0**	40.0	**70.0**	72.7
Question tag	19	**94.7**	**94.7**	**100.0**	**100.0**	**100.0**	**94.7**	**100.0**	**100.0**	**100.0**	**94.7**	**89.5**	15.8	42.1	86.6
LDD & interrogatives	154	87.0	90.3	88.3	**95.5**	87.7	87.7	87.0	89.6	90.3	79.2	85.1	72.1	66.2	85.1
Extended adjective construction	14	**100.0**	**92.9**	**100.0**	**85.7**	**92.9**	**92.9**	78.6	78.6	**100.0**	**92.9**	**92.9**	**92.9**	78.6	90.7
Extraposition	18	72.2	83.3	61.1	83.3	77.8	77.8	83.3	88.9	66.7	72.2	72.2	72.2	66.7	75.2
Multiple connectors	19	84.2	78.9	**89.5**	**100.0**	73.7	78.9	78.9	78.9	84.2	**89.5**	**89.5**	**89.5**	78.9	84.2
Pied-piping	20	85.0	**90.0**	**90.0**	**100.0**	**95.0**	**90.0**	**90.0**	**95.0**	**90.0**	75.0	80.0	80.0	60.0	86.2
Polar question	20	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	70.0	**100.0**	25.0	75.0	90.0
Scrambling	15	**86.7**	**93.3**	**93.3**	**93.3**	**93.3**	**86.7**	**93.3**	**93.3**	**93.3**	**66.7**	60.0	**86.7**	33.3	82.6
Topicalization	17	58.8	**76.5**	**70.6**	**94.1**	**76.5**	**76.5**	**76.5**	**82.4**	**88.2**	**70.6**	**82.4**	64.7	41.2	73.8
Wh-movement	31	**100.0**	**100.0**	**96.8**	**100.0**	90.3	**93.5**	90.3	**93.5**	**96.8**	90.3	**93.5**	74.2	80.6	92.3
MWE	76	**90.8**	**82.9**	**82.9**	**88.2**	77.6	80.3	**81.6**	**82.9**	80.3	71.1	76.3	**84.2**	53.9	79.5

categ	count	Onl-W	Onl-A	Onl-B	ChatG	Onl-M	Onl-Y	NLLBM	NLLBG	Onl-G	LanBr	GTCOM	ZengH	AIRC	avg
Collocation	19	**100.0**	**100.0**	**100.0**	**100.0**	**94.7**	**100.0**	**100.0**	**94.7**	**100.0**	84.2	**89.5**	**100.0**	57.9	93.9
Idiom	19	**63.2**	**31.6**	**42.1**	**57.9**	**15.8**	**21.1**	**31.6**	**36.8**	**26.3**	**10.5**	**15.8**	**36.8**	0.0	30.0
Prepositional MWE	19	**100.0**	**100.0**	**89.5**	**94.7**	**100.0**	**100.0**	**94.7**	**100.0**	**100.0**	**94.7**	**100.0**	**100.0**	68.4	95.5
Verbal MWE	19	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	94.7	94.7	100.0	100.0	89.5	98.4
Named entity & terminology	20	**95.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.0**	50.0	0.0	**90.0**	86.9
Date	20	**95.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.0**	50.0	0.0	**90.0**	86.9
Negation	19	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	94.7	99.6
Non-verbal agreement	60	**93.3**	90.0	**96.7**	**93.3**	**95.0**	**98.3**	**96.7**	**96.7**	88.3	86.7	81.7	**95.0**	71.7	91.0
Coreference	19	**94.7**	84.2	**89.5**	**94.7**	**89.5**	**94.7**	**94.7**	**94.7**	78.9	78.9	68.4	**100.0**	63.2	86.6
External possessor	21	**90.5**	**90.5**	**100.0**	**90.5**	**95.2**	**100.0**	**95.2**	**95.2**	**90.5**	**95.2**	81.0	**90.5**	57.1	90.1
Internal possessor	20	**95.0**	**95.0**	**100.0**	**95.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.0**	85.0	**95.0**	**95.0**	**95.0**	96.2
Punctuation	50	**100.0**	**100.0**	94.0	76.0	**100.0**	74.0	74.0	74.0	64.0	84.0	70.0	50.0	94.0	81.1
Comma	19	100.0	100.0	100.0	94.7	100.0	94.7	100.0	100.0	100.0	94.7	94.7	100.0	94.7	98.0
Quotation marks	31	**100.0**	**100.0**	90.3	64.5	**100.0**	61.3	58.1	58.1	41.9	77.4	54.8	19.4	**93.5**	70.7
Subordination	158	**91.1**	89.2	**92.4**	**91.8**	**92.4**	**93.7**	**94.9**	**93.0**	**92.4**	75.9	86.7	85.4	76.6	88.9
Adverbial clause	20	**90.0**	**90.0**	**100.0**	**90.0**	**90.0**	**95.0**	**95.0**	**90.0**	**90.0**	75.0	85.0	**90.0**	**90.0**	90.0
Cleft sentence	20	**95.0**	**95.0**	**95.0**	**95.0**	**95.0**	**100.0**	**95.0**	**95.0**	**100.0**	60.0	**90.0**	**95.0**	70.0	90.8
Free relative clause	14	100.0	92.9	92.9	100.0	85.7	100.0	100.0	85.7	100.0	100.0	100.0	85.7	92.9	95.1
Indirect speech	15	**86.7**	80.0	**93.3**	**86.7**	**100.0**	**93.3**	**100.0**	**100.0**	**93.3**	60.0	66.7	80.0	66.7	85.1
Infinitive clause	19	100.0	94.7	94.7	100.0	100.0	94.7	100.0	100.0	100.0	89.5	100.0	94.7	89.5	96.8
Object clause	16	**100.0**	**100.0**	**93.8**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**87.5**	**93.8**	**93.8**	81.3	96.2
Pseudo-cleft sentence	18	**77.8**	**83.3**	**83.3**	**83.3**	**72.2**	**83.3**	**72.2**	**72.2**	**72.2**	**66.7**	**66.7**	**61.1**	27.8	70.9
Relative clause	18	83.3	77.8	83.3	83.3	**88.9**	77.8	**100.0**	**100.0**	83.3	83.3	83.3	83.3	83.3	85.5
Subject clause	18	**88.9**	**88.9**	**94.4**	**88.9**	**100.0**	**100.0**	**94.4**	**94.4**	**94.4**	66.7	**94.4**	83.3	**88.9**	90.6
Verb tense/aspect/mood	2347	**93.7**	**94.0**	91.4	**92.9**	88.0	**94.0**	84.3	84.4	**93.1**	81.8	**93.8**	**93.8**	86.6	90.1
Conditional	16	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	93.8	99.5
Ditransitive - future I	36	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	91.7	**100.0**	**100.0**	**100.0**	99.4
Ditransitive - future I subjunctive II	24	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	87.5	**100.0**	**100.0**	**100.0**	99.0
Ditransitive - future II	31	**100.0**	**96.8**	**100.0**	83.9	51.6	**100.0**	32.3	25.8	**100.0**	80.6	**100.0**	**100.0**	67.7	79.9
Ditransitive - future II subjunctive II	31	**93.5**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**96.8**	**100.0**	**100.0**	87.1	98.3
Ditransitive - perfect	35	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**97.1**	**97.1**	**100.0**	88.6	**100.0**	**100.0**	**97.1**	98.5
Ditransitive - pluperfect	29	**100.0**	89.7	58.6	75.9	10.3	**93.1**	31.0	34.5	65.5	75.9	**93.1**	**96.6**	75.9	69.2
Ditransitive - pluperfect subjunctive II	25	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
Ditransitive - present	24	**91.7**	**95.8**	**100.0**	**100.0**	**100.0**	**95.8**	**91.7**	87.5	87.5	83.3	**100.0**	**100.0**	**95.8**	94.6
Ditransitive - preterite	31	**100.0**	**93.5**	**93.5**	**96.8**	90.3	90.3	**96.8**	**96.8**	83.9	74.2	87.1	**96.8**	77.4	90.6
Ditransitive - preterite subjunctive II	26	**92.3**	88.5	80.8	88.5	**96.2**	84.6	**100.0**	**100.0**	84.6	80.8	**96.2**	80.8	76.9	88.5
Imperative	19	**100.0**	**100.0**	**100.0**	**100.0**	**89.5**	**94.7**	**100.0**	**94.7**	**94.7**	63.2	**89.5**	**89.5**	78.9	91.9
Intransitive - future I	32	**96.9**	**96.9**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	68.8	**96.9**	87.5	**96.9**	95.7
Intransitive - future I subjunctive II	29	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	89.7	**100.0**	**100.0**	**100.0**	99.2
Intransitive - future II	31	**100.0**	90.3	**96.8**	74.2	61.3	**100.0**	51.6	54.8	**96.8**	58.1	29.0	**100.0**	90.3	77.2
Intransitive - future II subjunctive II	7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	14.3	100.0	100.0	93.4
Intransitive - perfect	76	**100.0**	**100.0**	**100.0**	**100.0**	**97.4**	**100.0**	94.7	92.1	**100.0**	60.5	**100.0**	**98.7**	92.1	95.0
Intransitive - pluperfect	32	**90.6**	**90.6**	**84.4**	**96.9**	28.1	**96.9**	25.0	25.0	68.8	37.5	**93.8**	**96.9**	**84.4**	70.7

categ	count	Onl-W	Onl-A	Onl-B	ChatG	Onl-M	Onl-Y	NLLBM	NLLBG	Onl-G	LanBr	GTCOM	ZengH	AIRC	avg
Intransitive - pluperfect subjunctive II	15	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	60.0	**100.0**	**100.0**	80.0	95.4
Intransitive - present	31	90.3	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	54.8	**100.0**	**100.0**	**100.0**	95.8
Intransitive - preterite	55	92.7	94.5	94.5	**100.0**	92.7	**100.0**	**96.4**	94.5	94.5	52.7	**96.4**	85.5	83.6	90.6
Intransitive - preterite subjunctive II	19	**57.9**	**63.2**	**78.9**	**84.2**	**63.2**	**78.9**	**68.4**	**68.4**	**68.4**	21.1	**73.7**	**57.9**	**57.9**	64.8
Modal - future I	95	**100.0**	**100.0**	**100.0**	**98.9**	**100.0**	89.5	96.8	93.7	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	98.4
Modal - future I subjunctive II	59	91.5	94.9	88.1	64.4	64.4	88.1	57.6	59.3	93.2	91.5	**100.0**	89.8	83.1	82.0
Modal - perfect	78	**78.2**	**78.2**	**76.9**	69.2	**74.4**	**82.1**	**84.6**	**78.2**	**79.5**	55.1	**83.3**	**83.3**	41.0	74.2
Modal - pluperfect	37	**86.5**	45.9	16.2	32.4	10.8	67.6	8.1	10.8	56.8	40.5	**75.7**	51.4	54.1	42.8
Modal - pluperfect subjunctive II	46	**73.9**	**71.7**	**73.9**	**76.1**	**69.6**	**71.7**	45.7	52.2	**69.6**	**73.9**	**76.1**	**73.9**	54.3	67.9
Modal - present	109	93.6	94.5	92.7	**100.0**	85.3	89.9	78.9	84.4	89.9	95.4	**100.0**	84.4	96.3	91.2
Modal - preterite	111	**100.0**	**99.1**	**100.0**	**98.2**	**98.2**	**100.0**	97.3	97.3	**99.1**	91.9	**100.0**	**99.1**	**100.0**	98.5
Modal - preterite subjunctive II	78	**88.5**	**89.7**	84.6	73.1	84.6	83.3	78.2	76.9	84.6	**89.7**	**88.5**	**89.7**	**93.6**	85.0
Modal negated - future I	82	98.8	98.8	100.0	100.0	100.0	100.0	100.0	97.6	98.8	100.0	100.0	100.0	98.8	99.4
Modal negated - future I subjunctive II	76	**100.0**	**100.0**	**100.0**	**98.7**	**98.7**	**100.0**	96.1	96.1	**100.0**	**100.0**	93.4	**100.0**	**98.7**	98.6
Modal negated - perfect	71	**98.6**	**98.6**	**98.6**	**98.6**	**100.0**	**97.2**	**97.2**	95.8	**100.0**	81.7	**100.0**	**98.6**	91.5	96.6
Modal negated - pluperfect	8	62.5	37.5	12.5	12.5	62.5	37.5	12.5	12.5	37.5	37.5	**100.0**	12.5	50.0	37.5
Modal negated - pluperfect subjunctive II	62	95.2	91.9	88.7	93.5	**100.0**	95.2	80.6	83.9	95.2	90.3	95.2	95.2	83.9	91.4
Modal negated - present	93	91.4	**98.9**	92.5	**100.0**	**98.9**	94.6	88.2	89.2	91.4	92.5	**100.0**	95.7	**100.0**	94.9
Modal negated - preterite	101	**100.0**	**100.0**	**100.0**	**99.0**	**100.0**	**98.0**	94.1	93.1	**99.0**	88.1	**99.0**	**99.0**	**100.0**	97.6
Modal negated - preterite subjunctive II	62	**98.4**	**98.4**	**98.4**	**100.0**	**100.0**	**100.0**	**98.4**	95.2	**98.4**	**100.0**	**100.0**	**98.4**	**96.8**	98.6
Progressive	19	**89.5**	**89.5**	**89.5**	**89.5**	**94.7**	**89.5**	**84.2**	**89.5**	**78.9**	47.4	68.4	**78.9**	26.3	78.1
Reflexive - future I	23	82.6	**100.0**	87.0	**100.0**	87.0	**100.0**	87.0	**91.3**	**100.0**	**95.7**	**100.0**	82.6	78.3	91.6
Reflexive - future I subjunctive II	25	80.0	**100.0**	80.0	**100.0**	88.0	**92.0**	88.0	**92.0**	**96.0**	80.0	**92.0**	**100.0**	72.0	89.2
Reflexive - future II	9	66.7	**88.9**	44.4	44.4	66.7	**88.9**	11.1	22.2	**100.0**	44.4	44.4	**100.0**	55.6	59.8
Reflexive - future II subjunctive II	10	**80.0**	**80.0**	**100.0**	**80.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	40.0	40.0	**100.0**	50.0	82.3
Reflexive - perfect	15	80.0	**93.3**	**86.7**	**100.0**	**100.0**	**93.3**	**86.7**	80.0	**93.3**	**86.7**	**93.3**	**100.0**	66.7	89.2
Reflexive - pluperfect	20	**75.0**	**85.0**	70.0	**95.0**	70.0	**90.0**	60.0	60.0	**90.0**	70.0	**80.0**	**95.0**	70.0	77.7
Reflexive - pluperfect subjunctive II	24	66.7	**91.7**	83.3	**91.7**	83.3	**95.8**	79.2	87.5	87.5	58.3	66.7	**100.0**	62.5	81.1
Reflexive - present	23	**91.3**	**100.0**	**95.7**	**100.0**	**100.0**	**95.7**	82.6	**91.3**	**95.7**	60.9	82.6	**100.0**	73.9	90.0
Reflexive - preterite	19	**89.5**	84.2	**100.0**	**100.0**	78.9	**94.7**	78.9	78.9	**100.0**	63.2	**89.5**	**89.5**	47.4	84.2
Reflexive - preterite subjunctive II	18	**94.4**	**94.4**	**94.4**	**100.0**	**94.4**	**94.4**	**88.9**	83.3	**94.4**	66.7	**88.9**	**100.0**	50.0	88.0
Transitive - future I	39	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	92.3	**100.0**	**100.0**	**100.0**	99.4
Transitive - future I subjunctive II	34	100.0	97.1	100.0	97.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.5
Transitive - future II	29	**100.0**	**96.6**	**100.0**	86.2	62.1	**100.0**	44.8	41.4	**100.0**	89.7	**100.0**	**100.0**	82.8	84.9
Transitive - future II subjunctive II	17	94.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	88.2	100.0	94.1	98.2
Transitive - perfect	41	**97.6**	**100.0**	**97.6**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	92.7	**100.0**	**100.0**	87.8	98.1
Transitive - pluperfect	31	**100.0**	**96.8**	45.2	**96.8**	22.6	**100.0**	45.2	54.8	**93.5**	**93.5**	**100.0**	**100.0**	80.6	79.2
Transitive - pluperfect subjunctive II	26	100.0	100.0	96.2	100.0	100.0	100.0	96.2	96.2	100.0	100.0	100.0	100.0	100.0	99.1
Transitive - present	43	**97.7**	**97.7**	**100.0**	**100.0**	**100.0**	93.0	**100.0**	**100.0**	**100.0**	**97.7**	**100.0**	**100.0**	**100.0**	98.9
Transitive - preterite	31	87.1	90.3	**96.8**	**100.0**	**100.0**	87.1	90.3	90.3	**100.0**	71.0	**93.5**	90.3	87.1	91.1
Transitive - preterite subjunctive II	29	72.4	69.0	69.0	**93.1**	**89.7**	69.0	69.0	69.0	65.5	44.8	**86.2**	72.4	62.1	71.6
Verb valency	81	**84.0**	**84.0**	**85.2**	**88.9**	**85.2**	**84.0**	**85.2**	**87.7**	77.8	77.8	**82.7**	**81.5**	65.4	82.2
Case government	28	**96.4**	**89.3**	**92.9**	**89.3**	**96.4**	**89.3**	**89.3**	**96.4**	**89.3**	78.6	**85.7**	**85.7**	71.4	88.5

categ	count	Onl-W	Onl-A	Onl-B	ChatG	Onl-M	Onl-Y	NLLBM	NLLBG	Onl-G	LanBr	GTCOM	ZengH	AIRC	avg
Mediopassive voice	18	83.3	**94.4**	**88.9**	**100.0**	**88.9**	**94.4**	83.3	83.3	72.2	**94.4**	**94.4**	**88.9**	77.8	88.0
Passive voice	19	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	94.7	100.0	100.0	89.5	98.8
Resultative predicates	16	43.8	43.8	50.0	62.5	43.8	43.8	62.5	62.5	37.5	37.5	43.8	43.8	12.5	45.2
micro-average	3234	**92.9**	**93.0**	91.2	**92.5**	88.3	**92.5**	85.6	85.6	91.5	81.2	90.7	89.4	82.2	89.0
phen. macro-average	3234	**91.0**	**91.3**	89.5	**91.5**	87.0	**91.6**	84.6	84.9	90.3	78.1	86.8	86.5	77.0	86.9
categ. macro-average	3234	**92.6**	**92.3**	**91.8**	**91.6**	90.1	89.2	89.2	88.9	88.1	82.3	82.0	75.6	74.3	86.8

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

categ	count	ChatG	Onl-W	Onl-B	Onl-A	Onl-Y	NLLBG	Onl-G	NLLBM	Onl-M	ZengH	LanBr	AIRC	avg
Ambiguity	24	**95.8**	**95.8**	**91.7**	**87.5**	**83.3**	**83.3**	**87.5**	**83.3**	**87.5**	**91.7**	75.0	50.0	84.4
Lexical ambiguity	24	**95.8**	**95.8**	**91.7**	**87.5**	**83.3**	**83.3**	**87.5**	**83.3**	**87.5**	**91.7**	75.0	50.0	84.4
Coordination & ellipsis	74	**90.5**	78.4	**93.2**	**85.1**	**93.2**	70.3	**90.5**	67.6	82.4	74.3	71.6	63.5	80.1
Gapping	12	**100.0**	75.0	**100.0**	**100.0**	**100.0**	58.3	**100.0**	50.0	**91.7**	**91.7**	75.0	75.0	84.7
Pseudogapping	7	**100.0**	**85.7**	**100.0**	**100.0**	**71.4**	**85.7**	**85.7**	**85.7**	**85.7**	**100.0**	**71.4**	42.9	84.5
Right node raising	15	**100.0**	**93.3**	80.0	80.0	**86.7**	**86.7**	**86.7**	**86.7**	80.0	73.3	**86.7**	80.0	85.0
Sluicing	14	**100.0**	**100.0**	**92.9**	**85.7**	**92.9**	**92.9**	**85.7**	**85.7**	**92.9**	78.6	**92.9**	78.6	89.9
Stripping	17	58.8	47.1	**94.1**	70.6	**100.0**	41.2	**94.1**	41.2	70.6	41.2	41.2	47.1	62.3
VP-ellipsis	9	**100.0**	**77.8**	**100.0**	**88.9**	**100.0**	66.7	**88.9**	66.7	**77.8**	**88.9**	66.7	44.4	80.6
False friends	33	**93.9**	**93.9**	**97.0**	**93.9**	**93.9**	**97.0**	**90.9**	**97.0**	**93.9**	**93.9**	**93.9**	81.8	93.4
Function word	41	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	75.6	**97.6**	85.4	94.7
Focus particle	22	95.5	95.5	95.5	95.5	95.5	95.5	95.5	95.5	95.5	95.5	95.5	90.9	95.1
Question tag	19	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	52.6	**100.0**	78.9	94.3
LDD & interrogatives	131	**96.9**	**96.2**	**95.4**	**96.9**	**96.9**	**94.7**	**93.9**	**93.9**	**93.9**	88.5	**92.4**	84.7	93.7
Extraposition	14	**85.7**	**85.7**	**78.6**	**85.7**	**78.6**	**78.6**	**64.3**	**71.4**	**71.4**	**78.6**	**57.1**	42.9	73.2
Inversion	13	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	92.3	100.0	100.0	100.0	99.4
Multiple connectors	17	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
Negative inversion	17	**94.1**	**94.1**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.1**	**100.0**	76.5	**100.0**	**94.1**	96.1
Pied-piping	11	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	90.9	100.0	90.9	98.5
Polar question	8	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**87.5**	62.5	95.8
Preposition stranding	7	85.7	100.0	100.0	100.0	100.0	85.7	100.0	100.0	100.0	100.0	100.0	100.0	97.6
Split infinitive	11	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
Topicalization	12	**100.0**	**83.3**	**83.3**	**91.7**	**91.7**	**91.7**	**83.3**	**91.7**	**83.3**	50.0	75.0	50.0	81.3
Wh-movement	21	100.0	100.0	95.2	95.2	100.0	90.5	95.2	90.5	95.2	95.2	100.0	95.2	96.0
Lexical Morphology	28	**85.7**	**85.7**	**82.1**	**75.0**	**67.9**	**67.9**	**64.3**	**64.3**	57.1	**82.1**	42.9	25.0	66.7
Functional shift	14	**92.9**	**85.7**	**78.6**	**85.7**	**64.3**	**85.7**	**71.4**	**71.4**	50.0	**78.6**	50.0	28.6	70.2
Noun formation (er)	14	**78.6**	**85.7**	**85.7**	**64.3**	**71.4**	**50.0**	**57.1**	**57.1**	**64.3**	**85.7**	35.7	21.4	63.1
MWE	95	**95.8**	**97.9**	**96.8**	91.6	**95.8**	85.3	89.5	86.3	86.3	92.6	78.9	68.4	88.8
Collocation	13	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**92.3**	**92.3**	**92.3**	**92.3**	**100.0**	**84.6**	69.2	93.6
Compound	17	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	94.1	99.5

categ	count	ChatG	Onl-W	Onl-B	Onl-A	Onl-Y	NLLBG	Onl-G	NLLBM	Onl-M	ZengH	LanBr	AIRC	avg
Idiom	12	**66.7**	**91.7**	**75.0**	50.0	**91.7**	41.7	33.3	41.7	25.0	50.0	16.7	0.0	48.6
Nominal MWE	20	**100.0**	**95.0**	**100.0**	**90.0**	85.0	85.0	**95.0**	**90.0**	85.0	**95.0**	70.0	70.0	88.3
Prepositional MWE	14	100.0	100.0	100.0	100.0	100.0	92.9	100.0	92.9	100.0	100.0	100.0	100.0	98.8
Verbal MWE	19	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**89.5**	**100.0**	**89.5**	**100.0**	**100.0**	**89.5**	63.2	94.3
Named entity & terminology	73	**95.9**	**95.9**	**95.9**	**97.3**	**97.3**	83.6	**94.5**	87.7	**94.5**	87.7	78.1	90.4	91.6
Date	13	**92.3**	**100.0**	**100.0**	**100.0**	**92.3**	**92.3**	**100.0**	**92.3**	**100.0**	69.2	61.5	**92.3**	91.0
Domainspecific Term	6	100.0	83.3	83.3	100.0	83.3	83.3	83.3	83.3	100.0	83.3	83.3	83.3	87.5
Location	17	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.1**	**100.0**	**94.1**	**100.0**	**100.0**	64.7	**100.0**	96.1
Measuring unit	18	**100.0**	**94.4**	**100.0**	**100.0**	**100.0**	66.7	**100.0**	77.8	83.3	**94.4**	83.3	**88.9**	90.7
Proper name	19	**89.5**	**94.7**	**89.5**	**89.5**	**100.0**	84.2	84.2	**89.5**	**94.7**	84.2	**94.7**	84.2	89.9
Negation	13	100.0	100.0	100.0	100.0	100.0	92.3	100.0	92.3	100.0	100.0	100.0	100.0	98.7
Non-verbal agreement	90	**97.8**	**94.4**	90.0	88.9	92.2	**94.4**	**93.3**	**95.6**	**95.6**	92.2	87.8	74.4	91.4
Coreference	29	**100.0**	**96.6**	86.2	86.2	89.7	**93.1**	89.7	**96.6**	**93.1**	89.7	79.3	58.6	88.2
Genitive	19	**100.0**	**94.7**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**89.5**	**100.0**	**94.7**	57.9	94.7
Personal Pronoun Coreference	12	**100.0**	**83.3**	58.3	58.3	66.7	**91.7**	75.0	**91.7**	**100.0**	66.7	66.7	**83.3**	78.5
Possession	27	92.6	96.3	100.0	96.3	100.0	96.3	100.0	96.3	100.0	100.0	100.0	96.3	97.8
Substitution	3	100.0	100.0	100.0	100.0	100.0	66.7	100.0	66.7	100.0	100.0	100.0	100.0	94.4
Punctuation	36	83.3	**97.2**	80.6	**88.9**	77.8	80.6	80.6	**86.1**	83.3	61.1	80.6	72.2	81.0
Quotation marks	36	83.3	**97.2**	80.6	**88.9**	77.8	80.6	80.6	**86.1**	83.3	61.1	80.6	72.2	81.0
Subordination	136	**99.3**	**97.1**	**97.8**	**97.8**	**96.3**	**97.8**	**97.8**	**97.8**	**97.8**	**97.1**	**99.3**	92.6	97.4
Adverbial clause	11	100.0	100.0	100.0	100.0	100.0	90.9	100.0	100.0	100.0	100.0	100.0	100.0	99.2
Cleft sentence	10	100.0	100.0	90.0	100.0	90.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.3
Contact clause	23	**100.0**	**95.7**	**95.7**	**100.0**	**100.0**	**100.0**	**95.7**	**100.0**	**95.7**	**100.0**	**100.0**	78.3	96.7
Indirect speech	12	91.7	83.3	100.0	91.7	91.7	100.0	100.0	100.0	91.7	91.7	100.0	100.0	95.1
Infinitive clause	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	100.0
Object clause	8	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
Pseudo-cleft sentence	14	100.0	100.0	100.0	100.0	92.9	100.0	92.9	100.0	100.0	100.0	100.0	92.9	98.2
Relative clause	34	**100.0**	**97.1**	**97.1**	**97.1**	**97.1**	**97.1**	**97.1**	**97.1**	**97.1**	91.2	**100.0**	**94.1**	96.8
Subject clause	14	100.0	100.0	100.0	92.9	92.9	92.9	100.0	85.7	100.0	100.0	92.9	85.7	95.2
Verb semantics	4	75.0	75.0	75.0	50.0	50.0	100.0	50.0	75.0	50.0	50.0	50.0	25.0	60.4
Verb tense/aspect/mood	2237	99.1	98.4	98.7	99.0	**99.6**	97.0	99.1	97.1	98.4	99.2	97.2	91.6	97.9
Conditional	19	94.7	89.5	94.7	94.7	94.7	94.7	94.7	94.7	89.5	94.7	84.2	78.9	91.7
Ditransitive - conditional I progressive	36	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**97.2**	91.7	99.1
Ditransitive - conditional I simple	42	**100.0**	**100.0**	**100.0**	**95.2**	**100.0**	73.8	**100.0**	76.2	92.9	**100.0**	73.8	66.7	89.9
Ditransitive - conditional II progressive	42	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	92.9	99.4
Ditransitive - conditional II simple	39	100.0	100.0	97.4	97.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.6
Ditransitive - future I progressive	39	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.4	100.0	99.8
Ditransitive - future I simple	81	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.5	99.8
Ditransitive - future II progressive	7	**85.7**	**100.0**	**100.0**	**100.0**	**100.0**	**85.7**	**100.0**	**100.0**	**100.0**	**100.0**	**71.4**	0.0	86.9
Ditransitive - future II simple	21	**100.0**	**100.0**	**100.0**	**95.2**	**100.0**	**90.5**	**100.0**	85.7	**90.5**	**100.0**	71.4	9.5	86.9
Ditransitive - past perfect progressive	35	**100.0**	**97.1**	**94.3**	**100.0**	**100.0**	88.6	**100.0**	85.7	**94.3**	**100.0**	**100.0**	**97.1**	96.4
Ditransitive - past perfect simple	34	**97.1**	**94.1**	**97.1**	**100.0**	**100.0**	85.3	**100.0**	85.3	**94.1**	**100.0**	**100.0**	**100.0**	96.1
Ditransitive - past progressive	30	**96.7**	86.7	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	98.6

categ	count	ChatG	Onl-W	Onl-B	Onl-A	Onl-Y	NLLBG	Onl-G	NLLBM	Onl-M	ZengH	LanBr	AIRC	avg
Ditransitive - present perfect progressive	38	94.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.4	99.3
Ditransitive - present perfect simple	43	97.7	97.7	100.0	100.0	100.0	95.3	100.0	95.3	100.0	100.0	100.0	100.0	98.8
Ditransitive - present progressive	40	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	87.5	92.5	98.3
Ditransitive - simple past	48	100.0	100.0	100.0	100.0	100.0	95.8	100.0	97.9	100.0	100.0	100.0	95.8	99.1
Ditransitive - simple present	43	**100.0**	**97.7**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	90.7	90.7	98.3
Gerund	19	**94.7**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.7**	**100.0**	**94.7**	**94.7**	84.2	96.9
Imperative	9	88.9	100.0	88.9	100.0	100.0	100.0	100.0	88.9	100.0	77.8	100.0	77.8	93.5
Intransitive - conditional I progressive	24	100.0	95.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.7
Intransitive - conditional I simple	25	100.0	100.0	92.0	96.0	100.0	96.0	100.0	96.0	100.0	100.0	100.0	100.0	98.3
Intransitive - conditional II progressive	9	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
Intransitive - conditional II simple	20	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
Intransitive - future I progressive	24	100.0	91.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	95.8	100.0	100.0	99.0
Intransitive - future I simple	56	**100.0**	87.5	**98.2**	**100.0**	**100.0**	**98.2**	**98.2**	**98.2**	**98.2**	**100.0**	**100.0**	**100.0**	98.2
Intransitive - future II progressive	4	100.0	100.0	100.0	75.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	75.0	95.8
Intransitive - future II simple	24	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.8**	87.5	**100.0**	**100.0**	37.5	93.4
Intransitive - past perfect progressive	12	100.0	100.0	100.0	100.0	100.0	91.7	100.0	91.7	91.7	100.0	100.0	91.7	97.2
Intransitive - past perfect simple	25	100.0	96.0	100.0	100.0	100.0	96.0	100.0	96.0	100.0	100.0	100.0	100.0	99.0
Intransitive - past progressive	22	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
Intransitive - present perfect progressive	4	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
Intransitive - present perfect simple	25	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.0	99.7
Intransitive - present progressive	49	**100.0**	93.9	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**98.0**	99.3
Intransitive - simple past	32	**100.0**	**100.0**	**100.0**	**96.9**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	84.4	98.4
Intransitive - simple present	33	97.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	93.9	99.2
Modal	283	98.9	**100.0**	98.9	**99.6**	**99.6**	98.6	**100.0**	98.9	**99.6**	**100.0**	98.9	**99.6**	99.4
Modal negated	251	**100.0**	**99.6**	98.8	**99.6**	**99.6**	98.8	**99.6**	98.8	**99.2**	**99.6**	**99.2**	98.4	99.3
Reflexive - conditional I progressive	23	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.7**	**100.0**	**100.0**	**100.0**	**95.7**	87.0	98.2
Reflexive - conditional I simple	22	100.0	100.0	100.0	90.9	100.0	90.9	100.0	90.9	95.5	95.5	95.5	90.9	95.8
Reflexive - conditional II progressive	12	100.0	91.7	100.0	100.0	100.0	100.0	91.7	100.0	100.0	100.0	91.7	100.0	97.9
Reflexive - conditional II simple	19	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	94.7	100.0	99.6
Reflexive - future I progressive	11	100.0	100.0	90.9	100.0	100.0	100.0	100.0	100.0	90.9	90.9	100.0	81.8	96.2
Reflexive - future I simple	33	100.0	100.0	100.0	100.0	100.0	100.0	97.0	100.0	100.0	100.0	100.0	100.0	99.7
Reflexive - future II progressive	5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	20.0	93.3
Reflexive - future II simple	11	**100.0**	**100.0**	**81.8**	**100.0**	**100.0**	**81.8**	**100.0**	**90.9**	**100.0**	**100.0**	**100.0**	63.6	93.2
Reflexive - past perfect progressive	12	**100.0**	75.0	**91.7**	**100.0**	**100.0**	66.7	**100.0**	66.7	66.7	**100.0**	**100.0**	**91.7**	88.2
Reflexive - past perfect simple	22	**95.5**	**86.4**	**95.5**	**95.5**	**95.5**	72.7	**95.5**	72.7	**86.4**	**95.5**	**95.5**	**90.9**	89.8
Reflexive - past progressive	25	**100.0**	**100.0**	**100.0**	**100.0**	**96.0**	**100.0**	88.0	**100.0**	**100.0**	**100.0**	88.0	88.0	96.7
Reflexive - present perfect progressive	11	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
Reflexive - present perfect simple	24	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	87.5	99.0
Reflexive - present progressive	20	100.0	100.0	90.0	95.0	90.0	95.0	90.0	100.0	95.0	95.0	95.0	90.0	94.6
Reflexive - simple past	25	**100.0**	**100.0**	**92.0**	**100.0**	**100.0**	**100.0**	**96.0**	**100.0**	**100.0**	**100.0**	**92.0**	88.0	97.3
Reflexive - simple present	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	100.0
Transitive - future II progressive	5	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	0.0	91.7
Transitive - conditional I progressive	19	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	84.2	98.7

categ	count	ChatG	Onl-W	Onl-B	Onl-A	Onl-Y	NLLBG	Onl-G	NLLBM	Onl-M	ZengH	LanBr	AIRC	avg
Transitive - conditional I simple	12	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	75.0	**100.0**	75.0	**91.7**	**100.0**	**83.3**	**83.3**	92.4
Transitive - conditional II progressive	22	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	90.9	99.2
Transitive - conditional II simple	26	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	84.6	98.7
Transitive - future I progressive	23	**100.0**	**95.7**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	87.0	98.6
Transitive - future I simple	48	**100.0**	**100.0**	**100.0**	**97.9**	**97.9**	**100.0**	**100.0**	**100.0**	**97.9**	**97.9**	**100.0**	91.7	98.6
Transitive - future II simple	15	**93.3**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	6.7	91.7
Transitive - past perfect progressive	17	100.0	94.1	100.0	100.0	100.0	88.2	100.0	88.2	88.2	100.0	100.0	88.2	95.6
Transitive - past perfect simple	18	100.0	100.0	100.0	100.0	100.0	94.4	100.0	94.4	100.0	100.0	100.0	88.9	98.1
Transitive - past progressive	16	68.8	**93.8**	68.8	50.0	**100.0**	62.5	62.5	68.8	81.3	56.3	43.8	56.3	67.7
Transitive - present perfect progressive	20	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	90.0	99.2
Transitive - present perfect simple	29	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	93.1	99.4
Transitive - present progressive	28	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.4	92.9	99.1
Transitive - simple past	30	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**96.7**	**100.0**	**100.0**	**100.0**	**100.0**	90.0	98.9
Transitive - simple present	33	**100.0**	**100.0**	**97.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	90.9	99.0
Verb valency	94	**86.2**	**86.2**	**88.3**	**86.2**	**79.8**	77.7	76.6	**80.9**	**78.7**	**86.2**	72.3	59.6	79.9
Case government	20	90.0	90.0	95.0	95.0	90.0	90.0	90.0	95.0	90.0	90.0	85.0	75.0	89.6
Catenative verb	15	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**93.3**	80.0	97.8
Mediopassive voice	15	**80.0**	**73.3**	**73.3**	**66.7**	**53.3**	**46.7**	**40.0**	**60.0**	33.3	**73.3**	**40.0**	20.0	55.0
Passive voice	14	92.9	92.9	92.9	92.9	92.9	92.9	92.9	92.9	92.9	92.9	92.9	71.4	91.1
Resultative	16	87.5	87.5	87.5	93.8	87.5	75.0	87.5	81.3	93.8	93.8	75.0	68.8	84.9
Semantic roles	14	64.3	71.4	78.6	64.3	50.0	57.1	42.9	50.0	57.1	64.3	42.9	35.7	56.5
micro-average	3109	**97.8**	97.0	**97.2**	97.0	**97.4**	94.4	96.6	94.7	95.9	95.9	93.5	87.1	95.4
phen. macro-average	3109	**96.7**	95.6	**96.0**	95.3	95.8	91.9	94.5	92.0	93.7	93.6	90.3	80.3	93.0
categ. macro-average	3109	**92.9**	**92.6**	**92.0**	89.0	88.1	88.0	87.1	86.8	86.5	84.8	81.2	71.0	86.7

Table 12: Accuracies (%) of successful translations on the phenomenon level for English–German. Boldface indicates the significantly best-performing systems per row.

categ	count	Onl-G	Onl-W	Onl-B	ChatG	Onl-Y	Onl-A	NLLBM	NLLBG	Onl-M	PROMT	ZengH	LanBr	avg
Ambiguity	20	**70.0**	**60.0**	50.0	**85.0**	55.0	45.0	50.0	50.0	35.0	30.0	45.0	25.0	50.0
Lexical ambiguity	20	**70.0**	**60.0**	50.0	**85.0**	55.0	45.0	50.0	50.0	35.0	30.0	45.0	25.0	50.0
Coordination & ellipsis	89	**82.0**	**83.1**	67.4	**77.5**	68.5	65.2	62.9	66.3	67.4	58.4	50.6	49.4	66.6
Gapping	17	**88.2**	**76.5**	29.4	**64.7**	**76.5**	41.2	52.9	**64.7**	**70.6**	29.4	17.6	29.4	53.4
Pseudogapping	14	**78.6**	**78.6**	**57.1**	**64.3**	**35.7**	**42.9**	28.6	28.6	**42.9**	**50.0**	**50.0**	14.3	47.6
Right node raising	16	75.0	81.3	81.3	68.8	75.0	68.8	75.0	68.8	68.8	56.3	68.8	68.8	71.4
Sluicing	12	83.3	83.3	75.0	83.3	66.7	83.3	58.3	58.3	75.0	66.7	50.0	41.7	68.7
Stripping	16	**93.8**	**93.8**	81.3	**100.0**	68.8	**87.5**	**93.8**	**93.8**	**87.5**	**87.5**	75.0	75.0	86.5
VP-ellipsis	14	**71.4**	**85.7**	**85.7**	**85.7**	**85.7**	**71.4**	**64.3**	**78.6**	**57.1**	**64.3**	42.9	**64.3**	71.4
False friends	14	85.7	85.7	78.6	64.3	85.7	71.4	57.1	64.3	64.3	57.1	71.4	50.0	69.6
Function word	29	**96.6**	**96.6**	**96.6**	**93.1**	**82.8**	**82.8**	**96.6**	**93.1**	**96.6**	**86.2**	37.9	75.9	86.2
Focus particle	11	90.9	90.9	90.9	81.8	81.8	90.9	90.9	81.8	90.9	90.9	81.8	72.7	86.4

categ	count	Onl-G	Onl-W	Onl-B	ChatG	Onl-Y	Onl-A	NLLBM	NLLBG	Onl-M	PROMT	ZengH	LanBr	avg
Question tag	18	**100.0**	**100.0**	**100.0**	**100.0**	83.3	77.8	**100.0**	**100.0**	**100.0**	83.3	11.1	77.8	86.1
LDD & interrogatives	61	**95.1**	**95.1**	**91.8**	**88.5**	**93.4**	**91.8**	**88.5**	**88.5**	85.2	85.2	73.8	78.7	88.0
Inversion	13	**100.0**	**100.0**	**92.3**	**92.3**	**100.0**	**92.3**	**92.3**	**92.3**	76.9	**92.3**	**84.6**	**100.0**	92.9
Modifying Comparison	5	60.0	80.0	80.0	80.0	80.0	60.0	80.0	80.0	100.0	60.0	60.0	60.0	73.3
Multiple connectors	13	**100.0**	**100.0**	**92.3**	**92.3**	**92.3**	**100.0**	**92.3**	**92.3**	**84.6**	**92.3**	76.9	**84.6**	91.7
Pied-piping	7	100.0	100.0	100.0	100.0	85.7	85.7	100.0	100.0	85.7	100.0	85.7	85.7	94.0
Preposition stranding	9	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**88.9**	**88.9**	**88.9**	**88.9**	66.7	**88.9**	92.6
Topicalization	9	**100.0**	**77.8**	**88.9**	**88.9**	**88.9**	**88.9**	**88.9**	**77.8**	**77.8**	55.6	**88.9**	44.4	80.6
Wh-movement	5	80.0	100.0	80.0	40.0	100.0	100.0	60.0	80.0	100.0	100.0	20.0	60.0	76.7
Lexical Morphology	29	**86.2**	**86.2**	**75.9**	**86.2**	**65.5**	62.1	62.1	**65.5**	41.4	51.7	58.6	55.2	66.4
Functional shift	15	**86.7**	**100.0**	**93.3**	**93.3**	73.3	**86.7**	73.3	80.0	53.3	66.7	**86.7**	73.3	80.6
Noun formation (er)	14	**85.7**	**71.4**	**57.1**	**78.6**	**57.1**	35.7	**50.0**	**50.0**	28.6	35.7	28.6	35.7	51.2
MWE	71	**76.1**	**73.2**	**76.1**	**70.4**	59.2	**69.0**	**67.6**	**66.2**	**60.6**	**60.6**	**69.0**	54.9	66.9
Collocation	8	75.0	87.5	87.5	75.0	75.0	75.0	62.5	62.5	75.0	62.5	87.5	50.0	72.9
Compound	4	75.0	25.0	75.0	50.0	25.0	75.0	50.0	25.0	50.0	50.0	50.0	50.0	50.0
Idiom	14	50.0	50.0	50.0	57.1	21.4	35.7	21.4	28.6	14.3	28.6	28.6	28.6	34.5
Nominal MWE	17	88.2	88.2	88.2	82.4	82.4	94.1	88.2	88.2	88.2	82.4	82.4	76.5	85.8
Prepositional MWE	8	100.0	100.0	100.0	100.0	100.0	100.0	87.5	100.0	100.0	100.0	100.0	100.0	99.0
Verbal MWE	20	**75.0**	**70.0**	**70.0**	**60.0**	**50.0**	**55.0**	**80.0**	**70.0**	**50.0**	**50.0**	**70.0**	40.0	61.7
Named entity & terminology	71	**87.3**	**77.5**	**81.7**	73.2	69.0	**76.1**	63.4	63.4	69.0	59.2	**80.3**	60.6	71.7
Date	19	**94.7**	78.9	**100.0**	**89.5**	84.2	84.2	73.7	73.7	84.2	68.4	84.2	73.7	82.5
Domainspecific Term	9	77.8	66.7	77.8	55.6	55.6	66.7	22.2	22.2	33.3	44.4	77.8	33.3	52.8
Measuring unit	13	**92.3**	76.9	**92.3**	**92.3**	**84.6**	**84.6**	**92.3**	**100.0**	76.9	**92.3**	**92.3**	**92.3**	89.1
Onomatopeia	11	**72.7**	**90.9**	**63.6**	**54.5**	36.4	**63.6**	36.4	27.3	27.3	18.2	**81.8**	27.3	50.0
Proper name	6	100.0	66.7	66.7	66.7	66.7	66.7	83.3	83.3	100.0	66.7	66.7	66.7	75.0
Proper Name & Location	13	84.6	76.9	69.2	61.5	69.2	76.9	61.5	61.5	84.6	53.8	69.2	53.8	68.6
Negation	4	75.0	100.0	100.0	75.0	100.0	75.0	75.0	75.0	100.0	75.0	100.0	50.0	83.3
Non-verbal agreement	80	**76.3**	**86.3**	**75.0**	**82.5**	73.8	72.5	**81.3**	**81.3**	73.8	**75.0**	66.3	65.0	75.7
Coreference	23	**52.2**	**73.9**	**52.2**	**60.9**	**47.8**	**47.8**	**73.9**	**65.2**	**52.2**	**47.8**	39.1	34.8	54.0
Genitive	13	84.6	84.6	92.3	61.5	92.3	84.6	84.6	76.9	76.9	84.6	69.2	69.2	80.1
Personal Pronoun Coreference	17	82.4	**88.2**	70.6	**100.0**	64.7	76.5	**94.1**	**100.0**	**88.2**	82.4	76.5	82.4	83.8
Possessive Pronouns	16	**87.5**	**93.8**	**93.8**	**100.0**	**87.5**	**93.8**	81.3	81.3	81.3	**87.5**	**87.5**	**87.5**	88.5
Substitution	11	**90.9**	**100.0**	**81.8**	**100.0**	**100.0**	72.7	72.7	**90.9**	**81.8**	**90.9**	72.7	63.6	84.8
Punctuation	12	**100.0**	**83.3**	**91.7**	66.7	75.0	**100.0**	**83.3**	**83.3**	66.7	**91.7**	0.0	**91.7**	77.8
Quotation marks	12	**100.0**	**83.3**	**91.7**	66.7	75.0	**100.0**	**83.3**	**83.3**	66.7	**91.7**	0.0	**91.7**	77.8
Subordination	130	**93.8**	**96.9**	**93.8**	**93.8**	**93.8**	90.0	86.9	88.5	**93.8**	89.2	68.5	83.1	89.4
Adverbial clause	11	**81.8**	**100.0**	**81.8**	**100.0**	**81.8**	**81.8**	63.6	63.6	**81.8**	72.7	45.5	**90.9**	78.8
Cleft sentence	12	**100.0**	**100.0**	**91.7**	**91.7**	**91.7**	**91.7**	66.7	75.0	**91.7**	**91.7**	75.0	66.7	86.1
Complex object	18	94.4	94.4	94.4	94.4	88.9	94.4	88.9	94.4	94.4	94.4	77.8	77.8	90.7
Contact clause	10	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	90.0	80.0	97.5
Indirect speech	4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	75.0	100.0	100.0	75.0	100.0	95.8
Infinitive clause	21	**95.2**	**90.5**	**95.2**	**90.5**	**95.2**	**90.5**	**100.0**	**90.5**	**95.2**	**90.5**	66.7	**90.5**	90.9
Object clause	5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	80.0	100.0	98.3

categ	count	Onl-G	Onl-W	Onl-B	ChatG	Onl-Y	Onl-A	NLLBM	NLLBG	Onl-M	PROMT	ZengH	LanBr	avg
Participle clause	20	85.0	95.0	90.0	90.0	90.0	85.0	80.0	85.0	95.0	85.0	85.0	80.0	87.1
Pseudo-cleft sentence	5	100.0	100.0	100.0	100.0	100.0	80.0	80.0	100.0	80.0	100.0	80.0	80.0	91.7
Relative clause	4	75.0	100.0	75.0	100.0	100.0	25.0	75.0	75.0	100.0	75.0	50.0	50.0	75.0
Subject clause	20	**100.0**	**100.0**	**100.0**	**90.0**	**100.0**	**100.0**	**95.0**	**100.0**	**95.0**	85.0	40.0	**90.0**	91.3
Verb semantics	17	**94.1**	**82.4**	**76.5**	47.1	58.8	**76.5**	52.9	47.1	58.8	58.8	58.8	41.2	62.7
Verb tense/aspect/mood	156	**91.7**	**94.2**	85.9	85.9	87.2	87.8	84.0	82.1	83.3	84.0	66.7	75.0	84.0
Conditional	24	**100.0**	**100.0**	**100.0**	**100.0**	**95.8**	**100.0**	**91.7**	87.5	87.5	**91.7**	45.8	87.5	90.6
Ditransitive	30	93.3	96.7	93.3	90.0	93.3	96.7	90.0	83.3	96.7	96.7	90.0	83.3	91.9
Gerund	15	**86.7**	**86.7**	**86.7**	66.7	**100.0**	80.0	73.3	53.3	80.0	73.3	53.3	53.3	74.4
Imperative	24	**87.5**	**95.8**	66.7	75.0	62.5	66.7	**83.3**	**95.8**	75.0	70.8	45.8	54.2	73.3
Intransitive	25	88.0	92.0	80.0	88.0	88.0	84.0	84.0	84.0	80.0	84.0	80.0	84.0	84.7
Reflexive	19	**89.5**	**89.5**	**78.9**	**84.2**	**89.5**	**89.5**	**68.4**	**68.4**	**73.7**	**78.9**	**78.9**	57.9	78.9
Transitive	19	**94.7**	**94.7**	**94.7**	**89.5**	**84.2**	**94.7**	**89.5**	**89.5**	**84.2**	**84.2**	63.2	**94.7**	88.2
Verb valency	126	**84.9**	**81.7**	**79.4**	**78.6**	**77.0**	72.2	68.3	64.3	73.0	70.6	69.8	60.3	73.3
Case government	25	96.0	96.0	92.0	96.0	92.0	88.0	84.0	84.0	96.0	92.0	84.0	84.0	90.3
Catenative verb	21	81.0	90.5	95.2	85.7	90.5	95.2	81.0	76.2	90.5	90.5	76.2	81.0	86.1
Impersonal Subject	5	100.0	100.0	100.0	80.0	100.0	100.0	100.0	60.0	100.0	100.0	100.0	80.0	93.3
Mediopassive voice	19	**84.2**	**63.2**	**73.7**	**73.7**	**57.9**	47.4	52.6	52.6	**68.4**	52.6	52.6	36.8	59.6
Passive voice	25	96.0	92.0	92.0	96.0	96.0	92.0	92.0	88.0	96.0	92.0	88.0	80.0	91.7
Resultative	16	**68.8**	**75.0**	**62.5**	**62.5**	**62.5**	**37.5**	25.0	18.8	18.8	**37.5**	**50.0**	18.8	44.8
Semantic roles	15	66.7	53.3	33.3	33.3	33.3	40.0	40.0	40.0	26.7	20.0	40.0	26.7	37.8
micro-average	909	**86.9**	**86.8**	81.7	81.7	78.3	78.0	75.2	74.8	75.4	72.9	65.0	65.7	76.9
phen. macro-average	909	**87.0**	**86.8**	82.7	81.2	79.1	78.0	74.7	73.9	76.3	73.5	65.9	65.5	77.0
categ. macro-average	909	**86.3**	**85.5**	81.3	77.9	76.3	75.8	72.0	71.9	71.3	68.9	61.1	61.1	74.1

Table 13: Accuracies (%) of successful translations on the phenomenon level for English–Russian. The boldface indicates the significantly best-performing systems per row.