Linguistically motivated Evaluation of the 2022 State-of-the-art Machine Translation Systems for three Language Directions

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

[1]German Research Center for Artificial Intelligence (DFKI)
firstname.lastname@dfki.de
[2]University of Hildesheim, lapshinovakoltun@uni-hildesheim.de
[3]Saarland University, s8sebagd@stud.uni-saarland.de

Abstract

This document describes a fine-grained linguistically motivated analysis of 29 machine translation systems submitted at the Shared Task of the 7th Conference of Machine Translation (WMT22). This submission expands the test suite work of previous years by adding the language direction of English–Russian. As a result, evaluation takes place for the language directions of German–English, English–German, and English–Russian. We find that the German–English systems suffer in translating idioms, some tenses of modal verbs, and resultative predicates, the English–German ones in idioms, transitive-past progressive, and middle voice, whereas the English–Russian ones in pseudo-gapping and idioms.

1 Introduction

Neural Machine Translation has seen enormous progress and reached a quality that is helpful for many everyday use cases. However, several methods for evaluating MT suggest that there is still plenty of room for improvement. An evaluation method for revealing the translation flaws in a more structured way refers to the use of *test suites* or *challenge sets*. Contrary to the classical evaluation, where test sets are drawn from random everyday texts, test suites consist of manually devised or selected sentences that focus on testing the ability of the MT systems to translate a particular phenomenon. Here, we are presenting test suite results while analyzing the state-of-the-art systems with regard to many linguistically-motivated phenomena. The test suites[1] were applied to the MT systems submitted at the 7th Conference of Machine Translation (WMT22) for the language directions German–English, English–German, and English–Russian. The test suites for the first two language directions have also been showcased during the previous years, whereas English–Russian is published for the first time.

This paper is structured as follows: Section 2 goes through related papers, 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 first test suites were introduced as early as the first MT systems in the 1990s (King and Falkedal, 1990; Way, 1991; Heid and Hildenbrand, 1991). Recent years saw the rise of Deep Learning and the drastic improvement of the quality of MT outputs, which has led to the current revival of test suites. Most of these test suites, however, focus on evaluating specific linguistic phenomena, e.g., Guillou and Hardmeier (2016), or on the comparison of different MT technologies (Isabelle et al., 2017; Burchardt et al., 2017), and Quality Estimation methods (Avramidis et al., 2018).

Over the last few years, several test suites for multiple language directions have emerged as a part of the Conference on Machine Translation test suite track. These test suites, however, focus on one or a few different phenomena, including the works of Popović (2019) Cinkova and Bojar (2018), Bojar et al. (2018), Rysová et al. (2019), Vojtěchová et al. (2019), Kocmi et al. (2020), Zouhar et al. (2020), Burlot et al. (2018), Guillou et al. (2018), Rios et al. (2018), Raganato et al. (2019), Scherrer et al. (2020). Our test suite, on the other hand, performs a systematic evaluation of more than one hundred phenomena per language direction (Macketanz et al., 2022). Similar to our work, the test suite approach and human evaluation are also used to evaluate MT quality metrics (Freitag et al., 2021; Avramidis and Macketanz, 2022).

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

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 432–449
Abu Dhabi, December 7–8, 2022.

Test set	Test sentences	Categories	Phenomena
De–En	∼5,500	14	106
En–De	∼4,400	13	110
En–Ru	∼300	12	51

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

3 Method

We have created a large-scale test suite with the goal of testing and comparing the performance of MT systems. Currently, the test suite covers four different language pairs. We will present three in this paper: German to English, English to German, and English to Russian (the fourth language pair being Portuguese to English). The test suite is based on a number of linguistic categories which are in turn divided into more fine-grained linguistic phenomena. The categories and phenomena are language-specific; however, there is a significant overlap between many of the categories and phenomena across the different language pairs. Each linguistic phenomenon in the test suite is represented by multiple test sentences. All categories, phenomena, and test sentences are the result of extensive research and knowledge of the syntax and morphology of the languages under inspection. The categories and phenomena do not follow a specific linguistic theory, but they were created by linguistic experts who are native speakers or highly proficient speakers of the languages. Furthermore, the set of categories and phenomena was reviewed internally by linguists and experienced translators to achieve objectivity in the classification.

The number of test sentences, categories, and phenomena for each language pair can be found in Table 1. As can be seen in the table, the English–Russian test set is considerably smaller than the other two test sets. This is due to the fact that we started creating the English–Russian test set only recently. However, we are currently working on expanding the test set by creating more phenomena and test sentences.

In order to allow for a semi-automatic evaluation of the test sentences, we have created a set of rules which determine whether a test sentence is translated correctly or incorrectly. The rules consist of hand-crafted regular expressions and fixed strings of translation outputs. They can be applied with the help of an internal evaluation tool (Macketanz et al., 2022). The workflow of the preparation and application of the test suite is depicted in Figure 1.

3.1 Application of the test suite

The thorough building and application of the test suite can be found in the previous test suite track papers (Macketanz et al., 2018; Avramidis et al., 2019, 2020; Macketanz et al., 2021). This paper gives a quick overview of the whole system. As shown in Figure 1, the building of the test suite follows steps a to c. Once the test sentences are fed as input to the MT systems, begins the application of the test suite (step d). The MT outputs are then automatically evaluated by the test suite tool with the help of the rules defined earlier by linguists and annotators (step e). The rules combine pre-set regular expressions and fixed strings (correct and incorrect translations from earlier MT system outputs). The rules are designed to evaluate each phenomenon in question's correct and incorrect translations. Note that only the phenomenon under inspection is being evaluated, meaning that all translation errors that are unrelated to the phenomenon are being ignored. The test sentence is marked with a warning if the output cannot be automatically sorted as correct or incorrect with the predefined rules. These warnings are then manually reviewed by human linguist annotators who decide on the translation's correctness and adapt the rules accordingly (step e). After that, the phenomenon-specific translation accuracy is calculated by dividing the number of correctly translated test sentences of a phenomenon by the total number of test sentences of that phenomenon:

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

Since this evaluation aims to compare the systems fairly, only the test items that do not contain any warnings for any systems are included in the calculation. If a test item has an unresolved warning for any MT systems, we exclude them from the calculation. Unfortunately, this reduces the number of test items. We see great importance in the extensive manual evaluation and human annotators designing rules with good coverage.

To define which system(s) perform better for a particular phenomenon (or category), we first identify the best 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 $\alpha = 0.95$. The systems that do not differ significantly from the best system are considered in the first performance cluster and indicated with boldface in the tables.

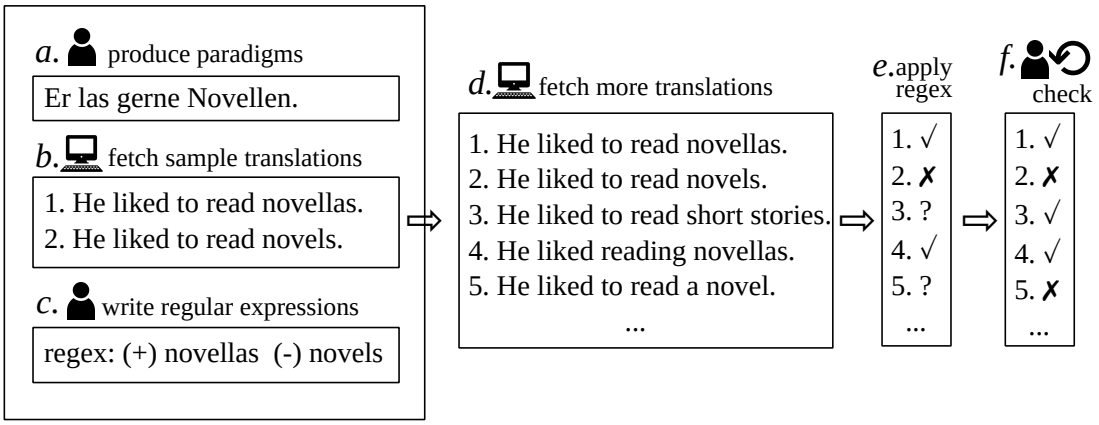

Figure 1: Example of the preparation and application of the test suite for one test sentence

The boldfaces, therefore, have a meaning only for the respective row of the table.

The average scores are computed in three ways as each category or phenomenon has a different number of test items. Micro-average aggregates the contributions of all test items to compute the average percentages. Category macro-average computes the percentages independently for each category and then averages them (i.e., treating all categories equally). Phenomenon macro-average computes the percentages independently for each phenomenon and then takes the average (i.e., treating all phenomena equally).

4 Experiment Setup

In this paper, we present the evaluation of 29 systems with our test suite. The systems are part of the *news translation task* of the Seventh Conference on Machine Translation (WMT22). The systems cover three different language pairs: nine systems for German–English, nine systems for English–German, and 11 systems for English–Russian.

This year is the second time that the English–German systems are being evaluated and the first time that the English–Russian systems are being evaluated with our test suite. 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). This year, there were on average 7.8 % of warnings for German–English, 9.7 % for English–German, and 20,6 % for English–Russian. It is not surprising that the English–Russian test set had a comparably bigger amount of warnings as this was the first time the test set was evaluated and therefore, the database of evaluation rules for this language pair was still rather small. It was also expected that English–German would have a higher amount of warnings than German–English as the German–English test set has the largest rules database since this language pair has been evaluated five years in a row.

Two annotators with extensive linguistic knowledge of the three languages under investigation conducted the manual evaluation of the warnings. No inter-annotator agreement was calculated; however, problematic cases were discussed with several linguistic experts to exclude subjectivity. The manual evaluation took around four weeks and involved around 50 person-hours. After the manual evaluation, there were on average 1.2 % of warnings left for German–English, 3.2 % for English–German, and 0.7 % 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 a result, our analysis was conducted on 5049 (91 %) test sentences for German–English, 3723 (83 %) test sentences for English–German, and 300 (97 %) test sentences for English–Russian.

5 Results

All result tables can be found in the Appendix.

5.1 System comparison

For **German–English**, two systems have the highest micro-average (85 %), Online-W and Online-A, whereas when considering the macro-average, three more systems also achieve the highest scores (89-90 %), Online-B, Land-Bridge, and JDExplore-Academy.

For **English–German**, two systems have the highest micro-average (97 %), Online-B and Lan-

Bridge. However, on the macro-average, a different system displays the highest score (94 %), JDExploreAcademy. The system with the lowest micro- and macro-average, Online-Y, still achieves scores of 84 % for both averages.

For **English–Russian**, the same four systems achieve the highest scores on both the micro- (78-81 %) and the macro-average (82-85 %), Online-W, Online-G, Online-B, and JDExploreAcademy. The average scores of English–Russian on the category level are comparably smaller than the scores of German–English and English–German. One plausible explanation is that English and Russian are more distant from a typological perspective than English and German.

5.2 Category-level analysis

For **German–English**, the categories with the highest average by all systems (> 90 %) are *composition*, *coordination & ellipses*, *named entity & terminology*, *negation*, and *non-verbal agreement*. The category with the lowest average score (77.2 %) is *false friends*.

For **English–German**, the categories with the highest average scores (> 96 %) are *function words*, *negation*, *non-verbal agreement*, *subordination*, and *verb tense/aspect/mood*. The category with the lowest average score (77.8 %) is *punctuation*.

For **English–Russian**, the category with the highest average score (92 %) is *punctuation*, with seven of the 11 systems achieving 100 % of accuracy, followed by *ambiguity*, *function words*, *negation*, and *subordination* (all > 80 %). The category with the lowest accuracies is *coordination & ellipsis*, followed by *false friends*.

5.3 Phenomenon-level analysis

For **German–English**, there are many phenomena that reach an average of 90-100 %, while the phenomenon macro-average reaches 85 %. Phenomena that reach more than 95 % of accuracy are *gapping*, *sluicing*, *polar question*, *verbal MWE*, *date*, *measuring unit*, *negation*, *internal possessor*, *comma*, *infinitive clause*, *object clause*, several verb tenses in *ditransitive*, *intransitive*, *transitive*, and *modal verbs*, and *passive voice*.

Yet there are some phenomena with a very low accuracy: The phenomena *idiom*, *modal pluperfect*, *modal pluperfect subjunctive II modal negated pluperfect*, *modal negated pluperfect subjunctive II*, and *resultative predicates* are the phenomena with the lowest averages, ranging only between 20-57 % accuracy. This result goes hand in hand with last year's result where the phenomena *modal pluperfect*, *resultative predicates*, and *idioms* reached the lowest accuracy.

Idiom	
Er macht aus einer Mücke immer gleich einen Elefanten.	
It always makes out of a mosquito an elephant.	fail
He always turns a gnat into an elephant.	fail
He always makes a mountain out of a molehill.	pass
Modal negated pluperfect	
Ich hatte nicht lesen sollen.	
I wasn't supposed to read.	fail
I shouldn't have read.	fail
I didn't want to read.	fail
Right node raising	
Lena soll und Tim will den Vertrag kündigen.	
Lena will and Tim will terminate the contract.	fail
L. should and T. want to terminate the contract.	fail
L. should and T. wants to terminate the contract.	pass

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

Table 2 contains example outputs from three different phenomena for German–English. The first example is from the phenomenon *idiom*. Idioms are multiword expressions whose meaning goes beyond the meaning of their separate elements. This also means that a simple literal translation into another language is usually incorrect. In our example at hand, the German idiom "aus einer Mücke einen Elefanten machen" means "to blow something out of proportion". A literal translation like the first and second outputs leads to an incorrect English meaning. What is further interesting about the incorrect outputs is that while the second one ("turns a gnat into an elephant") is at least grammatically correct, the first one ("makes out of a mosquito an elephant") is also grammatically incorrect. The translation of "Mücke" ("mosquito") as the term "gnat" is also unexpected. Only the third translation "makes a mountain out of a molehill" is a correct translation of this idiom.

The second example contains a *negated modal verb* in the *pluperfect tense*. The German sentence "Ich hatte nicht lesen sollen." can only be correctly translated as "I had not been supposed to read". This year, all systems failed to produce this correct output. Instead, there were different incorrect outputs with incorrect tenses ("I wasn't supposed to read.", "I shouldn't have read.") or incorrect translations of the modal verb ("I didn't want to read.").

The third example sentence contains an elliptical

right node raising construction. *Right node raising constructions* often consist of parallel coordinate sentences (sentences joined by "and") in which two conjuncts share some material on the right side of the structure. In the example sentence, the two conjuncts "Lena soll" ("Lena should") and "Tim will" ("Tim wants to") are sharing the material "den Vertrag kündigen" ("terminate the contract") on the right side of the construction. In the first incorrect example, the verbs "soll" and "will" are both translated as "will" which is an incorrect translation for both verbs. In the second incorrect output, the verbs are translated correctly, however, the verb "want" is incorrectly conjugated, missing the third person singular ending. Surprisingly, there were multiple systems that created this incorrectly conjugated translation.

At this point, it is also interesting to mention that there was one system that often incorrectly conjugated the verb "to sleep" in the past tense: Instead of "slept", the outputs by that particular system often contained the non-existing conjugation "sleeped".

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 100 % accuracy are *question tag*, *compound*, *prepositional MWE*, *subject clause*, *intransitive - present perfect progressive, present perfect simple, simple present*, and *transitive - future I progressive*.

The phenomena with the lowest accuracies, ranging between 35-61 %, are *idioms*, *transitive - past progressive*, and *middle voice*. These results are more in line with last year's results, as *idioms* and *middle voice* were also among the lowest accuracy phenomena.

Table 3 contains correct and incorrect translation examples from English–German. The first example contains a *coreference*. While many English nouns are gender-neutral, the same German nouns are in most cases gender specific. This can lead to translation errors if the context of a sentence clarifies the gender in English yet the German translation contains the incorrect gender. The test sentence at hand provides a clear context of the nurse being male. Yet, many systems incorrectly translated "nurse" as the female "Krankenschwester' instead of the male "Krankenpfleger". [2]

Coreference	
My brother is a nurse in the local hospital.	
Mein Bruder ist Krankenschwester im örtlichen Krankenhaus.	fail
Mein Bruder ist Krankenpfleger im örtlichen Krankenhaus.	pass
Verbal MWE	
She takes after her mother.	
Sie nimmt nach ihrer Mutter.	fail
Sie hinterlässt ihre Mutter.	fail
Sie kommt nach ihrer Mutter.	pass
Transitive future II progressive	
I will have been playing the piano.	
Ich würde Klavier gespielt haben.	fail
Ich habe Klavier gespielt.	fail
Ich werde Klavier gespielt haben.	pass

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

The second example contains the *verbal multiword expression* "to take after somebody". As explained above, multiword expressions cannot be translated literally as their meaning goes beyond their separate elements. The first incorrect output "Sie nimmt nach ihrer Mutter." is, however, a literal translation of this multiword expression. The second incorrect output "Sie hinterlässt ihre Mutter." is not a literal translation, yet still incorrect as it means "She leaves her mother behind". Only the translation "Sie kommt nach ihrer Mutter.", which is the German equivalent of this multiword expression, is correct.

The third example output contains a *transitive verb* in the tense *future II progressive*. The future II tense was often mistranslated as a conditional II tense "würde gespielt haben" ("would have played") instead of the correct form "werde". The second incorrect output contains a completely incorrect tense, the present perfect "habe gespielt" ("have played").

For **English–Russian**, the phenomenon level macro-average accuracy lies at 76 %. Also for this language pair, there are some phenomena which reach 100 % accuracy for all systems, like *nominal MWE*, *prepositional MWE*, *contact clause*, *indirect speech*, and passive voice. On the other hand, there are quite a few phenomena that reach a very low accuracy, ranging between 30-50 %: *gapping*, *pseudogapping*, *idioms*, *verbal MWE*, *anaphora agreement*, *intransitive verbs*, and *middle voice*. The low accuracies of *idioms* and *verbal MWEs* are

[2]We are aware that genders and their translation are a large topic on their own which we can only scratch on the surface within the scope of our test suite. We would like to point the interested reader to the following research: (Hardmeier et al., 2022)

Collocation	
She is careful to eat light and exercise often.	
Она старается есть легкую пищу и часто занимается спортом.	pass
Она старается есть свет и часто тренируется.	fail
Она осторожно ест легко и часто занимается спортом.	fail
Она следит за тем, чтобы есть мало и часто заниматься спортом.	pass
Pseudogapping	
I don't know that and don't think you do.	
Я этого не знаю и не думаю, что вы это делаете.	fail
Я этого не знаю и не думаю, что знаешь.	fail
Я не знаю этого и не думаю, что вы знаете.	pass
Resultative	
He read the children to sleep.	
Он зачитывал детей спать.	fail
Он читал детям спать.	fail
Он читал детям перед сном.	pass

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

not surprising as multiword expressions generally tend to cause translation errors across all language pairs. What is interesting is that the accuracy of *intransitive verbs* is considerably lower than the accuracies of the other verb types. One potential reason might be that in our small-scale English–Russian test suite the intransitives are presented by the verb of motion "to go", which has a number of equivalents in Russian that can convey various aspects such as tense, frequency, or incompleteness. This ambiguity increases the overall number of equivalents in the training data which could lead to faulty results when analyzing the translations with respect to specific phenomena.

Table 4 covers example translations of some low-accuracy phenomena for English–Russian. The first example contains the *collocation* "to eat light" that does not have an exact equivalent in Russian. The word "light" poses some extra difficulty, as it is lexically and semantically ambiguous in both languages. In different contexts, it could function as an adverb, adjective, or noun. This year, a typical incorrect output is "есть свет" (*est' svet*) meaning to consume light as in electromagnetic radiation, and "есть/питаться легко" (*est'/pitat'sya legko*), a combination of the verb to eat with an ill-passing adverb. Some possible translations would be Russian equivalents "to eat light food" or "to eat little" that we see in the first and fourth translations.

The second example is taken from the phenomenon of *pseudogapping*. *Pseudogapping* is an ellipsis mechanism in which a part of the verb phrase is omitted. In the example at hand, the non-finite verb part "know" is omitted in the second conjunct of the construction. Instead, the auxiliary verb "do" is used as a substitute for the full verb. Verbal substitution is not common in Russian. Moreover, Russian does not employ auxiliary verbs (such as "to do" or "to be") to form parallel elliptical constructions standard in English. The verb "does" in the second part of the sentence is translated as "сделает" (*sdelaet*) in the first incorrect Russian translation, leading to an impossible Russian phrasing. The second translation leaves out the subject "you" or "ты" (*ty*) in the conjunct resulting in a syntactically incorrect construction.

The last example contains a *resultative predicate*. *Resultatives* contain a verb with an adjective describing the result of the verb action. *Resultative predicates* usually require a significant construction change to get an equivalent translation in the target language. "He read the children to sleep" would be transformed in Russian as "on chital detyam pered snom" meaning "he read to the children before they were going to bed," as in the third translation in the table or as "on chital detyam, chtobi oni spali" meaning "he read to the children so that they would sleep".

5.4 Comparison with previous years

The progress of the systems' accuracy for particular categories through the last years can be seen in Table 6 for German-English (since 2018) and Table 9 for English-German (since 2021). The calculation has been done based on the common test items without warnings over all these years, which is 4307 items for German-English and 3616 items for English-German. The general trend of this year suggests small but steady improvements for most systems and categories. In a few cases where the accuracies deteriorated, this is only for very few percentage points.

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 29 machine translation systems in three different language pairs: German–English, English–German, and (for the first time) English–Russian. Altogether, almost 7,000 test sen-

tences, structured in various linguistic categories and phenomena, were evaluated altogether across the three language pairs. Additionally, a comparison to the evaluation in previous years for the language pairs German–English and English–German was drawn.

The average accuracy for most categories and phenomena is relatively high for German–English and English–German, with only about 5 % room for improvement. As compared to last year, this is an improvement of around 5 %. For English–Russian, the average accuracy is not as high, yet still around 80 %.

The high average accuracies do not necessarily mean that the respective categories and phenomena no longer pose difficulties for MT. Instead, it could mean that the difficulty of the test sentences has become too easy over the past few years and should thus be increased. Therefore, we are currently constructing more complex test sentences for German–English and English–German. Further work also includes expanding the English–Russian test suite with more phenomena and more test sentences.

Acknowledgements

This research was supported by the Deutsche Forschungsgemeinschaft (DFG) through the project TextQ, by the German Federal Ministry of Education through the project SocialWear (grant num. 01IW20002) and by the project European Language Equality 2, which has received funding from the European Union (grant agreement no. LC-01884166 – 101075356 ELE2). We would like to thank Hans Uszkoreit, Aljoscha Burchardt, Ursula Strohriegel, Renlong Ai and He Wang for their prior contributions on 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*, 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.

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.

Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen, editors. 2022. *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Association for Computational Linguistics, United States.

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.

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.

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, 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, Eleftherios Avramidis, Aljoscha Burchardt, and Hans Uszkoreit. 2018. Fine-grained evaluation of German-English Machine Translation based on a Test Suite. In *Proceedings of the Third Conference on Machine Translation, 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. 2022. A Linguistically Motivated Test Suite to Semi-Automatically Evaluate German–English Machine Translation Output. In *Proceedings of the Language Resources and Evaluation Conference*, pages 936–947, Marseille, France. European Language Resources Association.

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.

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.

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.

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.

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 German–English

category	count	Onl-B	Onl-W	LanBr	Onl-A	JDExp	Onl-Y	PROMT	Onl-G	LT22	avg
Ambiguity	81	**93.8**	84.0	**90.1**	84.0	**92.6**	84.0	77.8	**87.7**	43.2	81.9
Composition	49	93.9	**95.9**	93.9	**95.9**	**98.0**	**100.0**	**98.0**	**98.0**	65.3	93.2
Coordination & ellipsis	56	**92.9**	**94.6**	**91.1**	**94.6**	**94.6**	**94.6**	**92.9**	**94.6**	67.9	90.9
False friends	36	**77.8**	**80.6**	**83.3**	**83.3**	**66.7**	**77.8**	**80.6**	**83.3**	61.1	77.2
Function word	69	**91.3**	**89.9**	**89.9**	**88.4**	**92.8**	**84.1**	**89.9**	**91.3**	52.2	85.5
LDD & interrogatives	149	**89.3**	**86.6**	**89.3**	**88.6**	**91.9**	77.9	**88.6**	**90.6**	61.1	84.9
MWE	76	**81.6**	**89.5**	**78.9**	**82.9**	**81.6**	**82.9**	**78.9**	**80.3**	48.7	78.4
Named entitiy & terminology	87	**94.3**	**95.4**	**95.4**	**90.8**	**90.8**	**94.3**	**88.5**	**94.3**	78.2	91.3
Negation	19	**100.0**	**94.7**	**100.0**	**100.0**	**94.7**	**100.0**	**100.0**	**100.0**	68.4	95.3
Non-verbal agreement	60	**96.7**	**96.7**	**98.3**	**91.7**	**96.7**	**93.3**	88.3	88.3	61.7	90.2
Punctuation	59	**91.5**	**98.3**	89.8	**98.3**	89.8	**98.3**	**91.5**	66.1	67.8	87.9
Subordination	167	**93.4**	87.4	**92.8**	**88.0**	**92.2**	**92.8**	**90.4**	**91.6**	74.3	89.2
Verb tense/aspect/mood	4058	81.3	83.3	81.6	**84.9**	81.5	83.1	80.4	83.3	57.3	79.6
Verb valency	83	**84.3**	**83.1**	**81.9**	**83.1**	**88.0**	**81.9**	**81.9**	77.1	50.6	79.1
micro-average	5049	83.1	**84.6**	83.2	**85.7**	83.3	84.1	81.8	84.2	58.2	80.9
macro-average	5049	**90.1**	**90.0**	**89.7**	**89.6**	**89.4**	88.9	87.7	87.6	61.3	86.0

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

		Onl-B					Onl-Y				PROMT			Onl-A					Onl-W		Onl-G				
category	count	2018	2019	2020	2021	2022	2018	2019	2021	2022	2019	2020	2022	2018	2019	2020	2021	2022	2021	2022	2018	2019	2020	2021	2022
Ambiguity	76	76.3	77.6	78.9	85.5	93.4	67.1	78.9	82.9	84.2	50.0	65.8	77.6	68.4	69.7	77.6	81.6	84.2	85.5	84.2	72.4	75.0	84.2	85.5	88.2
Composition	47	97.9	97.9	95.7	100.0	95.7	89.4	91.5	91.5	100.0	78.7	89.4	97.9	80.9	91.5	93.6	95.7	95.7	95.7	95.7	70.2	83.0	95.7	97.9	97.9
Coordination & ellipsis	33	87.9	87.9	90.9	90.9	93.9	87.9	87.9	90.9	90.9	81.8	87.9	87.9	87.9	87.9	87.9	90.9	90.9	90.9	90.9	51.5	66.7	75.8	90.9	90.9
False friends	36	75.0	77.8	80.6	75.0	77.8	66.7	91.7	75.0	77.8	72.2	72.2	80.6	72.2	72.2	69.4	83.3	83.3	86.1	80.6	72.2	72.2	77.8	80.6	83.3
Function word	61	78.7	78.7	91.8	88.5	93.4	90.2	90.2	83.6	85.2	85.2	91.8	90.2	83.6	88.5	86.9	90.2	90.2	93.4	90.2	50.8	91.8	91.8	93.4	91.8
LDD & interrogatives	73	83.6	83.6	89.0	94.5	91.8	83.6	79.5	90.4	84.9	74.0	83.6	89.0	76.7	75.3	82.2	89.0	89.0	90.4	91.8	64.4	72.6	90.4	89.0	90.4
MWE	65	72.3	72.3	76.9	76.9	81.5	69.2	70.8	73.8	81.5	56.9	69.2	76.9	64.6	66.2	70.8	80.0	81.5	87.7	87.7	64.6	67.7	78.5	78.5	78.5
Named entity & term.	58	91.4	91.4	87.9	91.4	96.6	91.4	89.7	93.1	94.8	84.5	91.4	94.8	89.7	89.7	94.8	94.8	94.8	96.6	98.3	89.7	87.9	91.4	94.8	96.6
Negation	16	93.8	93.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	93.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	62.5	100.0	100.0	100.0	100.0
Non-verbal agreement	55	87.3	87.3	87.3	98.2	96.4	80.0	81.8	85.5	92.7	70.9	81.8	89.1	78.2	83.6	83.6	92.7	92.7	96.4	96.4	58.2	81.8	90.9	90.9	89.1
Punctuation	35	97.1	97.1	94.3	100.0	94.3	100.0	100.0	100.0	100.0	85.7	97.1	94.3	100.0	100.0	100.0	100.0	100.0	97.1	97.1	82.9	82.9	85.7	85.7	85.7
Subordination	90	87.8	88.9	93.3	94.4	95.6	92.2	92.2	92.2	93.3	86.7	93.3	92.2	94.4	78.9	93.3	92.2	92.2	91.1	91.1	81.1	90.0	92.2	91.1	93.3
Verb tense/aspect/mood	3604	77.1	77.3	79.5	78.6	81.3	73.8	75.6	76.7	83.3	78.3	78.0	80.5	75.4	86.4	80.8	85.3	85.2	86.3	84.2	49.4	69.2	83.5	79.2	83.7
Verb valency	58	81.0	81.0	89.7	89.7	87.9	79.3	81.0	81.0	86.2	70.7	82.8	87.9	77.6	81.0	86.2	86.2	86.2	87.9	87.9	69.0	77.6	86.2	86.2	84.5
micro-average	4307	78.2	78.5	80.9	80.6	83.0	75.3	77.2	78.3	84.3	77.6	78.9	81.7	76.3	85.6	81.6	**86.0**	**86.0**	**87.2**	85.3	52.6	71.0	84.2	80.8	84.7
macro-average	4307	84.8	85.2	88.3	90.3	**91.4**	83.6	86.5	86.9	89.6	76.4	84.6	88.5	82.1	83.6	86.2	90.1	90.4	**91.8**	**91.1**	67.1	79.9	87.4	88.8	89.6

Table 6: Comparisons of the accuracy (%) of several German–English systems through the years.

categ	count	Onl-B	Onl-W	LanBr	Onl-A	JDExp	Onl-Y	PROMT	Onl-G	LT22	avg
Ambiguity	81	**93.8**	84.0	**90.1**	84.0	**92.6**	84.0	77.8	**87.7**	43.2	81.9
Lexical ambiguity	63	**95.2**	**88.9**	**92.1**	84.1	**90.5**	85.7	79.4	**88.9**	47.6	83.6
Structural ambiguity	18	**88.9**	66.7	83.3	83.3	**100.0**	77.8	72.2	83.3	27.8	75.9
Composition	49	93.9	**95.9**	93.9	**95.9**	**98.0**	**100.0**	**98.0**	**98.0**	65.3	93.2
Compound	29	**96.6**	**96.6**	**96.6**	**93.1**	**96.6**	**100.0**	**96.6**	**96.6**	75.9	94.3
Phrasal verb	20	**90.0**	**95.0**	**90.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	50.0	91.7
Coordination & ellipsis	56	**92.9**	**94.6**	**91.1**	**94.6**	**94.6**	**94.6**	**92.9**	**94.6**	67.9	90.9
Gapping	19	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.7**	**100.0**	68.4	95.9
Right node raising	19	**84.2**	**84.2**	**78.9**	**84.2**	**84.2**	**84.2**	**84.2**	**84.2**	42.1	78.9
Sluicing	18	94.4	100.0	94.4	100.0	100.0	100.0	100.0	100.0	94.4	98.1
False friends	36	**77.8**	**80.6**	**83.3**	**83.3**	**66.7**	**77.8**	**80.6**	**83.3**	61.1	77.2
Function word	69	**91.3**	**89.9**	**89.9**	**88.4**	**92.8**	**84.1**	**89.9**	**91.3**	52.2	85.5
Focus particle	24	**100.0**	**95.8**	**95.8**	**95.8**	**95.8**	**100.0**	87.5	**95.8**	70.8	93.1
Modal particle	25	**76.0**	**76.0**	**76.0**	**80.0**	**84.0**	**72.0**	**84.0**	**80.0**	56.0	76.0
Question tag	20	**100.0**	**100.0**	**100.0**	**90.0**	**100.0**	80.0	**100.0**	**100.0**	25.0	88.3
LDD & interrogatives	149	**89.3**	**86.6**	**89.3**	**88.6**	**91.9**	77.9	**88.6**	**90.6**	61.1	84.9
Extended adjective construction	14	**100.0**	**100.0**	**100.0**	**100.0**	**92.9**	**92.9**	**100.0**	**100.0**	64.3	94.4
Extraposition	17	64.7	64.7	64.7	64.7	76.5	70.6	64.7	52.9	52.9	64.1
Multiple connectors	19	84.2	94.7	84.2	89.5	78.9	78.9	84.2	89.5	94.7	86.5
Pied-piping	18	**88.9**	**83.3**	**88.9**	**88.9**	**88.9**	**88.9**	**88.9**	**88.9**	55.6	84.6
Polar question	18	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	72.2	**100.0**	**100.0**	83.3	95.1
Scrambling	17	**94.1**	**88.2**	**94.1**	**82.4**	**94.1**	**94.1**	**76.5**	**94.1**	29.4	83.0
Topicalization	17	82.4	52.9	82.4	82.4	**100.0**	**88.2**	**88.2**	**94.1**	58.8	81.0
Wh-movement	29	**96.6**	**100.0**	**96.6**	**96.6**	**100.0**	55.2	**100.0**	**100.0**	51.7	88.5
MWE	76	**81.6**	**89.5**	**78.9**	**82.9**	**81.6**	**82.9**	**78.9**	**80.3**	48.7	78.4
Collocation	19	**100.0**	**100.0**	**100.0**	**100.0**	**94.7**	**100.0**	**100.0**	**100.0**	52.6	94.2
Idiom	18	**38.9**	**61.1**	**22.2**	**27.8**	**33.3**	**27.8**	**16.7**	**27.8**	0.0	28.4
Prepositional MWE	20	**90.0**	**95.0**	**90.0**	**100.0**	**95.0**	**100.0**	**95.0**	**95.0**	65.0	91.7
Verbal MWE	19	**94.7**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.7**	73.7	95.9
Named entitiy & terminology	87	**94.3**	**95.4**	**95.4**	**90.8**	**90.8**	**94.3**	**88.5**	**94.3**	78.2	91.3
Date	19	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	94.7	99.4
Domainspecific term	20	**85.0**	**85.0**	**90.0**	**75.0**	**80.0**	**85.0**	**65.0**	**85.0**	50.0	77.8
Location	20	**95.0**	**100.0**	**95.0**	**95.0**	**95.0**	**95.0**	**95.0**	**95.0**	85.0	94.4
Measuring unit	19	**100.0**	**100.0**	**100.0**	**100.0**	**89.5**	**100.0**	**100.0**	**100.0**	84.2	97.1
Proper name	9	88.9	88.9	88.9	77.8	88.9	88.9	77.8	88.9	77.8	85.2
Negation	19	**100.0**	**94.7**	**100.0**	**100.0**	**94.7**	**100.0**	**100.0**	**100.0**	68.4	95.3
Non-verbal agreement	60	**96.7**	**96.7**	**98.3**	**91.7**	**96.7**	**93.3**	88.3	88.3	61.7	90.2
Coreference	19	**89.5**	**100.0**	**94.7**	84.2	**94.7**	84.2	78.9	78.9	57.9	84.8
External possessor	21	**100.0**	**95.2**	**100.0**	**90.5**	**95.2**	**95.2**	**90.5**	**90.5**	42.9	88.9
Internal possessor	20	**100.0**	**95.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.0**	**95.0**	85.0	96.7
Punctuation	59	**91.5**	**98.3**	89.8	**98.3**	89.8	**98.3**	**91.5**	66.1	67.8	87.9
Comma	20	100.0	95.0	95.0	100.0	100.0	100.0	100.0	100.0	100.0	98.9
Quotation marks	39	87.2	**100.0**	87.2	**97.4**	84.6	**97.4**	87.2	48.7	51.3	82.3

categ	count	Onl-B	Onl-W	LanBr	Onl-A	JDExp	Onl-Y	PROMT	Onl-G	LT22	avg
Subordination	167	**93.4**	87.4	**92.8**	**88.0**	**92.2**	**92.8**	**90.4**	**91.6**	74.3	89.2
Adverbial clause	20	**100.0**	85.0	**95.0**	**90.0**	**90.0**	**95.0**	85.0	**90.0**	85.0	90.6
Cleft sentence	20	**95.0**	**95.0**	**95.0**	**95.0**	**95.0**	**95.0**	**95.0**	**95.0**	60.0	91.1
Free relative clause	17	88.2	88.2	88.2	82.4	88.2	94.1	94.1	94.1	76.5	88.2
Indirect speech	15	**93.3**	73.3	**93.3**	**93.3**	**100.0**	**100.0**	**93.3**	**100.0**	60.0	89.6
Infinitive clause	19	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	84.2	98.2
Object clause	18	**100.0**	**100.0**	**94.4**	**94.4**	**94.4**	**100.0**	**100.0**	**100.0**	83.3	96.3
Pseudo-cleft sentence	20	85.0	65.0	85.0	70.0	85.0	70.0	75.0	75.0	60.0	74.4
Relative clause	18	83.3	94.4	83.3	77.8	83.3	83.3	83.3	77.8	88.9	84.0
Subject clause	20	**95.0**	85.0	**100.0**	**90.0**	**95.0**	**100.0**	**90.0**	**95.0**	70.0	91.1
Verb tense/aspect/mood	4058	81.3	83.3	81.6	**84.9**	81.5	83.1	80.4	83.3	57.3	79.6
Conditional	20	**95.0**	**90.0**	**95.0**	**95.0**	85.0	**90.0**	**100.0**	**95.0**	75.0	91.1
Ditransitive - future I	36	**100.0**	**94.4**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	86.1	97.8
Ditransitive - future I subjunctive II	36	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	86.1	98.5
Ditransitive - future II	36	**97.2**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	52.8	94.4
Ditransitive - future II subjunctive II	36	**100.0**	88.9	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.4**	98.1
Ditransitive - perfect	36	100.0	97.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.7
Ditransitive - pluperfect	36	61.1	**97.2**	61.1	**88.9**	**97.2**	**86.1**	52.8	50.0	75.0	74.4
Ditransitive - pluperfect subjunctive II	36	**100.0**	**100.0**	**100.0**	**100.0**	83.3	**100.0**	**100.0**	**100.0**	69.4	94.8
Ditransitive - present	36	**100.0**	86.1	**100.0**	91.7	**94.4**	**94.4**	80.6	88.9	91.7	92.0
Ditransitive - preterite	36	**91.7**	**94.4**	**88.9**	75.0	**83.3**	**88.9**	80.6	80.6	77.8	84.6
Ditransitive - preterite subjunctive II	35	71.4	68.6	71.4	68.6	71.4	80.0	71.4	71.4	65.7	71.1
Imperative	19	**100.0**	**94.7**	**100.0**	**89.5**	**100.0**	**100.0**	**100.0**	**94.7**	63.2	93.6
Intransitive - future I	35	100.0	100.0	97.1	100.0	100.0	100.0	100.0	100.0	97.1	99.4
Intransitive - future I subjunctive II	36	100.0	97.2	100.0	100.0	100.0	97.2	100.0	100.0	97.2	99.1
Intransitive - future II	37	**100.0**	**100.0**	**100.0**	91.9	78.4	89.2	**100.0**	**97.3**	40.5	88.6
Intransitive - future II subjunctive II	35	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	40.0	93.3
Intransitive - perfect	83	**100.0**	90.4	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	60.2	94.5
Intransitive - pluperfect	35	**82.9**	**85.7**	**88.6**	**94.3**	57.1	**88.6**	57.1	77.1	37.1	74.3
Intransitive - pluperfect subjunctive II	29	**100.0**	**100.0**	**100.0**	**93.1**	**100.0**	**96.6**	**100.0**	**100.0**	37.9	92.0
Intransitive - present	35	**100.0**	65.7	**100.0**	**100.0**	**100.0**	**97.1**	**100.0**	**100.0**	71.4	92.7
Intransitive - preterite	66	**92.4**	**84.8**	**92.4**	**93.9**	**92.4**	**90.9**	**93.9**	**90.9**	65.2	88.6
Intransitive - preterite subjunctive II	36	**80.6**	63.9	**75.0**	**69.4**	66.7	**69.4**	**86.1**	**77.8**	36.1	69.4
Modal - future I	160	86.9	**88.8**	86.9	**93.1**	**89.4**	69.4	85.0	**89.4**	58.1	83.0
Modal - future I subjunctive II	151	**86.8**	**87.4**	**87.4**	**90.1**	82.8	**88.1**	80.8	**90.1**	49.0	82.5
Modal - perfect	165	73.9	70.9	73.9	80.0	**93.9**	79.4	82.4	75.2	3.6	70.4
Modal - pluperfect	146	2.7	**55.5**	**0.7**	37.0	14.4	34.2	13.0	25.3	1.4	20.5
Modal - pluperfect subjunctive II	144	**56.3**	**60.4**	**56.3**	**57.6**	41.7	**57.6**	**56.3**	**54.9**	31.9	52.5
Modal - present	173	**87.3**	**89.6**	**87.9**	**91.9**	**87.9**	83.2	79.8	**87.9**	79.2	86.1
Modal - preterite	169	**100.0**	92.9	**100.0**	**99.4**	98.2	97.6	95.3	97.6	89.9	96.8
Modal - preterite subjunctive II	146	80.8	78.8	80.1	82.2	82.2	74.0	79.5	76.7	73.3	78.6
Modal negated - future I	151	**95.4**	**95.4**	**94.7**	92.1	**94.7**	**96.7**	**93.4**	**97.4**	70.2	92.2
Modal negated - future I subjunctive II	171	83.0	85.4	81.9	84.2	83.6	91.8	79.5	**98.8**	51.5	82.2

categ	count	Onl-B	Onl-W	LanBr	Onl-A	JDExp	Onl-Y	PROMT	Onl-G	LT22	avg
Modal negated - perfect	165	77.0	84.8	73.3	78.8	**93.9**	76.4	82.4	70.9	14.5	72.5
Modal negated - pluperfect	131	13.7	**67.9**	14.5	24.4	8.4	28.2	4.6	26.0	0.0	20.9
Modal negated - pluperfect subjunctive II	153	**56.2**	**67.3**	**60.1**	**62.1**	41.2	**66.0**	**62.7**	**62.1**	32.7	56.7
Modal negated - present	157	89.8	92.4	**94.3**	**96.2**	**97.5**	87.9	84.7	87.3	88.5	90.9
Modal negated - preterite	169	**98.8**	94.7	**99.4**	**100.0**	97.6	**99.4**	94.1	97.0	79.3	95.6
Modal negated - preterite subjunctive II	141	**84.4**	**86.5**	**87.2**	**80.9**	**83.7**	**85.8**	**87.9**	**87.9**	75.2	84.4
Progressive	18	**94.4**	83.3	**94.4**	83.3	**94.4**	**100.0**	**88.9**	**88.9**	50.0	86.4
Reflexive - future I	35	**88.6**	80.0	**88.6**	**97.1**	**85.7**	**94.3**	**88.6**	**91.4**	57.1	85.7
Reflexive - future I subjunctive II	34	**85.3**	76.5	**85.3**	**97.1**	82.4	**94.1**	79.4	**85.3**	41.2	80.7
Reflexive - future II	36	75.0	66.7	**83.3**	**91.7**	58.3	**83.3**	**80.6**	**91.7**	41.7	74.7
Reflexive - future II subjunctive II	36	**83.3**	66.7	**86.1**	**88.9**	**80.6**	**88.9**	**72.2**	**86.1**	36.1	76.5
Reflexive - perfect	36	**91.7**	58.3	**86.1**	**97.2**	**94.4**	**91.7**	**91.7**	**94.4**	41.7	83.0
Reflexive - pluperfect	35	77.1	71.4	77.1	**94.3**	**82.9**	80.0	**88.6**	**85.7**	37.1	77.1
Reflexive - pluperfect subjunctive II	32	**84.4**	**75.0**	**87.5**	**90.6**	**78.1**	**87.5**	**78.1**	**84.4**	43.8	78.8
Reflexive - present	36	**97.2**	63.9	**94.4**	**97.2**	**88.9**	**94.4**	**86.1**	**91.7**	50.0	84.9
Reflexive - preterite	34	**91.2**	52.9	**91.2**	79.4	**88.2**	**82.4**	73.5	**94.1**	47.1	77.8
Reflexive - preterite subjunctive II	28	**100.0**	71.4	**100.0**	89.3	**96.4**	82.1	85.7	**100.0**	57.1	86.9
Transitive - future I	41	97.6	97.6	97.6	97.6	97.6	97.6	97.6	97.6	92.7	97.0
Transitive - future I subjunctive II	36	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	94.4	99.4
Transitive - future II	36	**100.0**	**97.2**	**100.0**	**100.0**	**97.2**	**97.2**	**100.0**	**100.0**	86.1	97.5
Transitive - future II subjunctive II	35	**100.0**	**97.1**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	85.7	98.1
Transitive - perfect	42	97.6	95.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.2
Transitive - pluperfect	36	47.2	**100.0**	50.0	**94.4**	83.3	**94.4**	**94.4**	83.3	88.9	81.8
Transitive - pluperfect subjunctive II	36	**97.2**	**100.0**	**100.0**	**100.0**	86.1	**100.0**	**100.0**	**97.2**	80.6	95.7
Transitive - present	48	**100.0**	89.6	**100.0**	**100.0**	**100.0**	**100.0**	**97.9**	**100.0**	91.7	97.7
Transitive - preterite	36	**97.2**	86.1	**94.4**	**100.0**	**97.2**	83.3	88.9	**100.0**	80.6	92.0
Transitive - preterite subjunctive II	35	68.6	60.0	68.6	62.9	71.4	60.0	80.0	60.0	65.7	66.3
Verb valency	83	**84.3**	**83.1**	**81.9**	**83.1**	**88.0**	**81.9**	**81.9**	77.1	50.6	79.1
Case government	28	**92.9**	**92.9**	**92.9**	**92.9**	**85.7**	**89.3**	**92.9**	**89.3**	50.0	86.5
Mediopassive voice	20	**90.0**	**90.0**	**85.0**	**90.0**	**95.0**	**85.0**	**85.0**	65.0	50.0	81.7
Passive voice	19	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	73.7	97.1
Resultative predicates	16	43.8	37.5	37.5	37.5	68.8	43.8	37.5	43.8	25.0	41.7
micro-average	5049	83.1	**84.6**	83.2	**85.7**	83.3	84.1	81.8	84.2	58.2	80.9
phen. macro-average	5049	**88.4**	86.5	**88.2**	**88.9**	**88.1**	**88.1**	86.5	**88.1**	62.8	85.1
categ. macro-average	5049	**90.1**	**90.0**	**89.7**	**89.6**	**89.4**	88.9	87.7	87.6	61.3	86.0

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

B English–German

category	count	JDExp	Onl-A	Onl-W	Onl-B	LanBr	Onl-G	PROMT	Onl-Y	OpenN	avg
Ambiguity	24	**91.7**	**87.5**	**95.8**	**91.7**	**83.3**	**83.3**	**79.2**	**79.2**	62.5	83.8
Coordination & ellipsis	74	78.4	79.7	67.6	**91.9**	**90.5**	**85.1**	**83.8**	79.7	66.2	80.3
False friends	38	92.1	84.2	89.5	86.8	84.2	89.5	84.2	84.2	78.9	86.0
Function word	42	97.6	97.6	100.0	97.6	97.6	97.6	97.6	97.6	95.2	97.6
MWE	110	**90.9**	84.5	**93.6**	**90.0**	83.6	85.5	80.0	82.7	80.9	85.8
Named entitiy & terminology	74	94.6	93.2	90.5	94.6	93.2	90.5	90.5	90.5	89.2	91.9
Negation	17	100.0	100.0	100.0	100.0	100.0	94.1	94.1	100.0	94.1	98.0
Non-verbal agreement	71	98.6	95.8	98.6	97.2	98.6	95.8	95.8	95.8	93.0	96.6
Punctuation	19	**100.0**	**94.7**	84.2	63.2	63.2	63.2	68.4	63.2	**100.0**	77.8
Subordination	162	**100.0**	**99.4**	98.1	**99.4**	**99.4**	**100.0**	**98.8**	93.2	98.1	98.5
Verb tense/aspect/mood	3009	97.4	98.1	96.2	**98.7**	**99.2**	97.6	**98.7**	95.2	83.6	96.1
Verb valency	83	**91.6**	**84.3**	**84.3**	**86.7**	**85.5**	**84.3**	**83.1**	**84.3**	77.1	84.6
micro-average	3723	96.7	96.7	95.2	**97.6**	**97.7**	96.3	96.9	93.8	84.1	95.0
macro-average	3723	**94.4**	91.6	91.5	91.5	89.9	88.9	87.9	87.1	84.9	89.7

Table 8: Accuracies (%) of successful translations on the category level for English–German. Boldface indicates the significantly best performing systems per row.

		Onl-A		Onl-W		Onl-B		Onl-G		Onl-Y	
category	count	2021	2022	2021	2022	2021	2022	2021	2022	2021	2022
Ambiguity	24	92	88	96	96	92	92	75	83	71	79
Coordination & ellipsis	75	69	80	65	64	84	91	73	84	68	75
False friends	39	85	85	87	90	82	87	82	90	85	85
Function word	41	98	98	100	100	100	98	98	98	98	98
MWE	109	83	85	93	95	89	90	79	86	80	83
Named entitiy & terminology	72	93	93	96	93	93	97	76	93	92	93
Negation	16	94	100	100	100	94	100	94	94	100	100
Non-verbal agreement	71	97	97	96	97	94	97	92	96	92	96
Punctuation	36	97	97	97	92	78	78	69	78	78	78
Subordination	161	99	99	98	98	99	99	95	99	94	93
Verb tense/aspect/mood	2885	96	98	97	96	99	99	95	98	92	95
Verb valency	87	83	86	89	85	86	89	75	86	79	86
micro-avg	3616	95	97	96	95	97	98	93	96	90	94
macro-avg	3616	90	92	93	92	91	93	84	90	86	88

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

category/phenomenon	count	JDExp	Onl-A	Onl-W	Onl-B	LanBr	Onl-G	PROMT	Onl-Y	OpenN	avg
Ambiguity	24	**91.7**	**87.5**	**95.8**	**91.7**	**83.3**	**83.3**	**79.2**	**79.2**	62.5	83.8
Lexical ambiguity	24	**91.7**	**87.5**	**95.8**	**91.7**	**83.3**	**83.3**	**79.2**	**79.2**	62.5	83.8
Coordination & ellipsis	74	78.4	79.7	67.6	**91.9**	**90.5**	**85.1**	**83.8**	79.7	66.2	80.3
Gapping	14	78.6	85.7	64.3	92.9	92.9	78.6	85.7	71.4	64.3	79.4
Pseudogapping	6	83.3	50.0	66.7	83.3	83.3	83.3	66.7	50.0	50.0	68.5
Right node raising	11	90.9	100.0	90.9	90.9	90.9	90.9	100.0	81.8	90.9	91.9
Sluicing	14	**100.0**	78.6	**100.0**	**92.9**	78.6	78.6	78.6	78.6	78.6	84.9
Stripping	19	52.6	78.9	36.8	**94.7**	**100.0**	**89.5**	78.9	**89.5**	47.4	74.3
VP-ellipsis	10	80.0	70.0	60.0	90.0	90.0	90.0	90.0	90.0	70.0	81.1
False friends	38	92.1	84.2	89.5	86.8	84.2	89.5	84.2	84.2	78.9	86.0
Function word	42	97.6	97.6	100.0	97.6	97.6	97.6	97.6	97.6	95.2	97.6
Focus particle	23	95.7	95.7	100.0	95.7	95.7	95.7	95.7	95.7	91.3	95.7
Question tag	19	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
MWE	110	**90.9**	84.5	**93.6**	**90.0**	83.6	85.5	80.0	82.7	80.9	85.8
Collocation	17	100.0	100.0	100.0	94.1	88.2	100.0	94.1	94.1	88.2	95.4
Compound	19	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Idiom	18	**55.6**	**33.3**	**72.2**	**50.0**	**33.3**	22.2	**11.1**	22.2	16.7	35.2
Nominal MWE	18	94.4	88.9	94.4	100.0	94.4	100.0	88.9	88.9	94.4	93.8
Prepositional MWE	14	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Verbal MWE	24	95.8	87.5	95.8	95.8	87.5	91.7	87.5	91.7	87.5	91.2
Named entitiy & terminology	74	94.6	93.2	90.5	94.6	93.2	90.5	90.5	90.5	89.2	91.9
Date	16	100.0	100.0	100.0	100.0	100.0	100.0	100.0	93.8	100.0	99.3
Domainspecific term	9	77.8	88.9	77.8	77.8	77.8	77.8	77.8	77.8	88.9	80.2
Location	17	88.2	88.2	94.1	88.2	88.2	88.2	88.2	88.2	88.2	88.9
Measuring unit	18	**100.0**	**94.4**	83.3	**100.0**	**100.0**	**94.4**	**94.4**	**100.0**	**88.9**	95.1
Proper name	14	**100.0**	**92.9**	**92.9**	**100.0**	**92.9**	**85.7**	**85.7**	**85.7**	78.6	90.5
Negation	17	100.0	100.0	100.0	100.0	100.0	94.1	94.1	100.0	94.1	98.0
Non-verbal agreement	71	98.6	95.8	98.6	97.2	98.6	95.8	95.8	95.8	93.0	96.6
Coreference	25	96.0	88.0	96.0	92.0	96.0	96.0	88.0	88.0	88.0	92.0
Genitive	17	100.0	100.0	100.0	100.0	100.0	94.1	100.0	100.0	94.1	98.7
Possession	29	100.0	100.0	100.0	100.0	100.0	96.6	100.0	100.0	96.6	99.2
Punctuation	19	**100.0**	**94.7**	84.2	63.2	63.2	63.2	68.4	63.2	**100.0**	77.8
Quotation marks	19	**100.0**	**94.7**	84.2	63.2	63.2	63.2	68.4	63.2	**100.0**	77.8
Subordination	162	**100.0**	**99.4**	98.1	**99.4**	**99.4**	**100.0**	**98.8**	93.2	98.1	98.5
Adverbial clause	15	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	93.3	99.3
Cleft sentence	17	**100.0**	**94.1**	**100.0**	**94.1**	**100.0**	**100.0**	**94.1**	82.4	**94.1**	95.4
Contact clause	23	100.0	100.0	95.7	100.0	100.0	100.0	100.0	100.0	100.0	99.5
Indirect speech	13	100.0	100.0	84.6	100.0	100.0	100.0	100.0	100.0	100.0	98.3
Infinitive clause	18	100.0	100.0	100.0	100.0	100.0	100.0	100.0	94.4	94.4	98.8
Object clause	13	100.0	100.0	100.0	100.0	100.0	100.0	100.0	84.6	100.0	98.3
Pseudo-cleft sentence	16	**100.0**	**100.0**	**100.0**	**100.0**	**93.8**	**100.0**	**100.0**	75.0	**100.0**	96.5
Relative clause	34	100.0	100.0	100.0	100.0	100.0	100.0	97.1	97.1	100.0	99.3
Subject clause	13	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

category/phenomenon	count	JDExp	Onl-A	Onl-W	Onl-B	LanBr	Onl-G	PROMT	Onl-Y	OpenN	avg
Verb tense/aspect/mood	3009	97.4	98.1	96.2	**98.7**	**99.2**	97.6	**98.7**	95.2	83.6	96.1
Conditional	17	94.1	94.1	94.1	88.2	88.2	94.1	94.1	94.1	94.1	92.8
Ditransitive - conditional I progressive	56	**100.0**	**100.0**	**96.4**	**100.0**	**100.0**	**100.0**	**98.2**	**100.0**	58.9	94.8
Ditransitive - conditional I simple	58	77.6	82.8	**98.3**	**100.0**	**100.0**	**98.3**	**96.6**	**100.0**	50.0	89.3
Ditransitive - conditional II progressive	56	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.4	96.4	99.2
Ditransitive - conditional II simple	57	**100.0**	**100.0**	**96.5**	**100.0**	**100.0**	**100.0**	**100.0**	94.7	94.7	98.4
Ditransitive - future I progressive	47	**97.9**	**97.9**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	89.4	98.3
Ditransitive - future I simple	99	94.9	**99.0**	**99.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	96.0	98.8
Ditransitive - future II progressive	51	94.1	**100.0**	92.2	**100.0**	**100.0**	92.2	**100.0**	58.8	37.3	86.1
Ditransitive - future II simple	57	84.2	94.7	**100.0**	**100.0**	**100.0**	93.0	**96.5**	77.2	36.8	86.9
Ditransitive - past perfect progressive	55	**96.4**	**100.0**	**98.2**	**100.0**	**100.0**	92.7	92.7	90.9	61.8	92.5
Ditransitive - past perfect simple	56	**98.2**	**100.0**	**96.4**	**100.0**	**100.0**	89.3	**98.2**	94.6	75.0	94.6
Ditransitive - past progressive	44	**100.0**	**100.0**	79.5	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**97.7**	97.5
Ditransitive - present perfect progressive	56	**100.0**	**100.0**	**98.2**	**100.0**	**100.0**	**100.0**	**100.0**	94.6	**96.4**	98.8
Ditransitive - present perfect simple	55	**100.0**	**100.0**	**98.2**	**100.0**	**100.0**	**100.0**	**100.0**	**96.4**	94.5	98.8
Ditransitive - present progressive	44	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	95.5	99.5
Ditransitive - simple past	73	**100.0**	**100.0**	**98.6**	**100.0**	**100.0**	**100.0**	**100.0**	**97.3**	95.9	99.1
Ditransitive - simple present	52	**100.0**	**100.0**	94.2	**100.0**	**100.0**	**98.1**	**100.0**	**100.0**	90.4	98.1
Gerund	21	100.0	100.0	95.2	100.0	100.0	100.0	100.0	100.0	95.2	98.9
Imperative	13	**100.0**	**100.0**	**92.3**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	61.5	94.9
Intransitive - conditional I progressive	27	**96.3**	**100.0**	85.2	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**92.6**	97.1
Intransitive - conditional I simple	29	**96.6**	**100.0**	**96.6**	**93.1**	**100.0**	**100.0**	**100.0**	**100.0**	89.7	97.3
Intransitive - conditional II progressive	22	**100.0**	**100.0**	81.8	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	98.0
Intransitive - conditional II simple	21	100.0	100.0	95.2	100.0	100.0	100.0	100.0	100.0	100.0	99.5
Intransitive - future I progressive	24	91.7	100.0	91.7	100.0	100.0	100.0	100.0	100.0	91.7	97.2
Intransitive - future I simple	64	95.3	**100.0**	89.1	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	93.8	97.6
Intransitive - future II progressive	24	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	20.8	62.5	87.0
Intransitive - future II simple	35	**97.1**	**100.0**	**100.0**	**100.0**	**100.0**	**97.1**	**97.1**	**94.3**	88.6	97.1
Intransitive - past perfect progressive	25	**92.0**	**100.0**	**96.0**	**100.0**	**100.0**	**96.0**	**96.0**	**100.0**	76.0	95.1
Intransitive - past perfect simple	33	**97.0**	**100.0**	**97.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	78.8	97.0
Intransitive - past progressive	28	100.0	100.0	100.0	92.9	100.0	100.0	100.0	96.4	100.0	98.8
Intransitive - present perfect progressive	2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Intransitive - present perfect simple	27	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Intransitive - present progressive	55	98.2	100.0	96.4	100.0	100.0	100.0	100.0	100.0	98.2	99.2
Intransitive - simple past	38	100.0	97.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.7
Intransitive - simple present	33	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Modal	271	**100.0**	**100.0**	98.9	98.5	**100.0**	**100.0**	**99.6**	98.5	96.3	99.1
Modal negated	270	**98.5**	**98.9**	97.4	**98.1**	**99.6**	**98.9**	**99.3**	**98.9**	95.2	98.3
Reflexive - conditional I progressive	34	**97.1**	**97.1**	91.2	**100.0**	**100.0**	91.2	**100.0**	**100.0**	58.8	92.8
Reflexive - conditional I simple	28	**92.9**	**92.9**	**92.9**	**100.0**	**100.0**	**100.0**	**100.0**	**92.9**	64.3	92.9
Reflexive - conditional II progressive	30	**100.0**	**100.0**	83.3	**96.7**	**96.7**	**100.0**	**100.0**	90.0	70.0	93.0
Reflexive - conditional II simple	31	**100.0**	**100.0**	87.1	**96.8**	**100.0**	**100.0**	**100.0**	90.3	90.3	96.1
Reflexive - future I progressive	32	**100.0**	**100.0**	**96.9**	**96.9**	**96.9**	**96.9**	**100.0**	**100.0**	75.0	95.8

category/phenomenon	count	JDExp	Onl-A	Onl-W	Onl-B	LanBr	Onl-G	PROMT	Onl-Y	OpenN	avg
Reflexive - future I simple	50	**100.0**	**100.0**	**96.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	90.0	98.4
Reflexive - future II progressive	27	**100.0**	**92.6**	88.9	**100.0**	**100.0**	**100.0**	**100.0**	48.1	55.6	87.2
Reflexive - future II simple	28	**100.0**	**96.4**	**92.9**	**100.0**	**100.0**	**100.0**	**100.0**	**96.4**	75.0	95.6
Reflexive - past perfect progressive	28	**100.0**	**92.9**	85.7	**92.9**	**96.4**	67.9	75.0	71.4	35.7	79.8
Reflexive - past perfect simple	28	**100.0**	**100.0**	**92.9**	**100.0**	**100.0**	85.7	**96.4**	82.1	53.6	90.1
Reflexive - past progressive	34	**100.0**	**100.0**	**100.0**	**97.1**	**97.1**	85.3	**100.0**	**97.1**	76.5	94.8
Reflexive - present perfect progressive	30	**100.0**	**100.0**	**100.0**	**96.7**	**100.0**	**100.0**	**100.0**	**100.0**	80.0	97.4
Reflexive - present perfect simple	31	100.0	100.0	93.5	100.0	100.0	100.0	100.0	100.0	96.8	98.9
Reflexive - present progressive	34	**94.1**	**94.1**	**97.1**	**94.1**	**97.1**	**85.3**	**91.2**	**97.1**	70.6	91.2
Reflexive - simple past	32	100.0	100.0	96.9	96.9	100.0	96.9	100.0	100.0	93.8	98.3
Reflexive - simple present	27	**100.0**	85.2	**100.0**	**100.0**	**96.3**	88.9	**100.0**	**96.3**	81.5	94.2
Transitive - future II progressive	29	**96.6**	**100.0**	**96.6**	**100.0**	**100.0**	**100.0**	**100.0**	37.9	44.8	86.2
Transitive - conditional I progressive	29	**100.0**	**100.0**	89.7	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	69.0	95.4
Transitive - conditional I simple	30	**93.3**	83.3	**100.0**	**100.0**	**100.0**	**100.0**	**93.3**	**100.0**	76.7	94.1
Transitive - conditional II progressive	29	**100.0**	**100.0**	**96.6**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	79.3	97.3
Transitive - conditional II simple	28	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	92.9	99.2
Transitive - future I progressive	26	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Transitive - future I simple	50	100.0	98.0	100.0	100.0	100.0	98.0	98.0	98.0	100.0	99.1
Transitive - future II simple	32	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**96.9**	46.9	93.8
Transitive - past perfect progressive	28	**100.0**	**100.0**	71.4	**100.0**	**100.0**	**96.4**	**100.0**	**100.0**	53.6	91.3
Transitive - past perfect simple	28	**100.0**	**100.0**	**96.4**	**100.0**	**100.0**	**96.4**	**100.0**	**100.0**	67.9	95.6
Transitive - past progressive	28	39.3	39.3	**92.9**	57.1	42.9	60.7	71.4	**96.4**	57.1	61.9
Transitive - present perfect progressive	30	**100.0**	**100.0**	**96.7**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	83.3	97.8
Transitive - present perfect simple	35	**100.0**	**100.0**	**97.1**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	85.7	98.1
Transitive - present progressive	35	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	88.6	98.7
Transitive - simple past	38	**97.4**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	92.1	98.8
Transitive - simple present	35	100.0	100.0	100.0	97.1	100.0	100.0	100.0	100.0	100.0	99.7
Verb valency	83	**91.6**	**84.3**	**84.3**	**86.7**	**85.5**	**84.3**	**83.1**	**84.3**	77.1	84.6
Case government	18	94.4	94.4	88.9	94.4	94.4	94.4	94.4	88.9	94.4	93.2
Catenative verb	18	100.0	100.0	94.4	100.0	100.0	100.0	100.0	100.0	88.9	98.1
Middle voice	15	73.3	53.3	73.3	60.0	60.0	53.3	46.7	53.3	26.7	55.6
Passive voice	15	100.0	100.0	93.3	100.0	100.0	100.0	100.0	100.0	100.0	99.3
Resultative	17	88.2	70.6	70.6	76.5	70.6	70.6	70.6	76.5	70.6	73.9
micro-average	3723	96.7	96.7	95.2	**97.6**	**97.7**	96.3	96.9	93.8	84.1	95.0
phen. macro-average	3723	**95.8**	94.8	93.4	**96.3**	**95.9**	94.5	94.8	91.4	82.4	93.3
categ. macro-average	3723	**94.4**	91.6	91.5	91.5	89.9	88.9	87.9	87.1	84.9	89.7

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

C English–Russian

category	count	Onl-W	Onl-G	Onl-B	JDExp	LanBr	Huawe	Onl-A	PROMT	Onl-Y	SRPOL	eTran	avg
Ambiguity	11	**100.0**	**90.9**	**90.9**	**90.9**	**81.8**	**81.8**	72.7	63.6	**81.8**	72.7	63.6	81.0
Coordination & ellipsis	27	**70.4**	**77.8**	**55.6**	**74.1**	48.1	**55.6**	44.4	**59.3**	**51.9**	**55.6**	**51.9**	58.6
False friends	5	80.0	80.0	80.0	80.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	67.3
Function word	10	**100.0**	**90.0**	**90.0**	70.0	**80.0**	**80.0**	**80.0**	**80.0**	**90.0**	**80.0**	**80.0**	83.6
MWE	39	76.9	74.4	76.9	74.4	66.7	64.1	66.7	64.1	66.7	59.0	61.5	68.3
Named entitiy & terminology	26	**73.1**	**88.5**	**84.6**	**84.6**	**84.6**	65.4	**69.2**	**73.1**	**73.1**	65.4	65.4	75.2
Negation	5	100.0	80.0	100.0	80.0	100.0	80.0	80.0	80.0	80.0	100.0	80.0	87.3
Non-verbal agreement	23	73.9	78.3	73.9	82.6	69.6	73.9	69.6	73.9	69.6	69.6	65.2	72.7
Punctuation	5	100.0	100.0	100.0	100.0	80.0	100.0	100.0	100.0	80.0	80.0	80.0	92.7
Subordination	49	**91.8**	**89.8**	**91.8**	**93.9**	**89.8**	79.6	81.6	81.6	79.6	81.6	75.5	85.2
Verb tense/aspect/mood	68	70.6	72.1	72.1	76.5	75.0	67.6	73.5	73.5	66.2	70.6	67.6	71.4
Verb valency	32	87.5	87.5	78.1	84.4	75.0	75.0	84.4	71.9	78.1	75.0	68.8	78.7
micro-average	300	**80.3**	**81.3**	**78.7**	**81.7**	75.0	70.7	72.3	72.3	71.0	70.3	67.0	74.6
macro-average	300	**85.4**	**84.1**	**82.8**	**82.6**	75.9	73.6	73.5	73.4	73.1	72.5	68.3	76.8

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

category/phenomenon	count	Onl-W	Onl-G	Onl-B	JDExp	LanBr	Huawe	Onl-A	PROMT	Onl-Y	SRPOL	eTran	avg
Ambiguity	11	**100.0**	**90.9**	**90.9**	**90.9**	**81.8**	**81.8**	72.7	63.6	**81.8**	72.7	63.6	81.0
Lexical ambiguity	11	**100.0**	**90.9**	**90.9**	**90.9**	**81.8**	**81.8**	72.7	63.6	**81.8**	72.7	63.6	81.0
Coordination & ellipsis	27	**70.4**	**77.8**	**55.6**	**74.1**	48.1	**55.6**	44.4	**59.3**	**51.9**	**55.6**	**51.9**	58.6
Gapping	5	**40.0**	**80.0**	**20.0**	**80.0**	**20.0**	**60.0**	0.0	**60.0**	**40.0**	**60.0**	**40.0**	45.5
Pseudogapping	6	**50.0**	**83.3**	**50.0**	**66.7**	**16.7**	0.0	0.0	**16.7**	**16.7**	0.0	**16.7**	28.8
Right node raising	5	100.0	80.0	80.0	100.0	80.0	100.0	80.0	80.0	80.0	80.0	80.0	85.5
Sluicing	3	**100.0**	**66.7**	**66.7**	**33.3**	**66.7**	**33.3**	**66.7**	**66.7**	**66.7**	**100.0**	0.0	60.6
Stripping	5	80.0	80.0	40.0	60.0	40.0	60.0	60.0	80.0	60.0	80.0	80.0	65.5
VP-ellipsis	3	66.7	66.7	100.0	100.0	100.0	100.0	100.0	66.7	66.7	33.3	100.0	81.8
False friends	5	80.0	80.0	80.0	80.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	67.3
Function word	10	**100.0**	**90.0**	**90.0**	70.0	**80.0**	**80.0**	**80.0**	**80.0**	**90.0**	**80.0**	**80.0**	83.6
Focus particle	5	100.0	80.0	80.0	80.0	80.0	100.0	100.0	80.0	100.0	100.0	100.0	90.9
Question tag	5	100.0	100.0	100.0	60.0	80.0	60.0	60.0	80.0	80.0	60.0	60.0	76.4
MWE	39	76.9	74.4	76.9	74.4	66.7	64.1	66.7	64.1	66.7	59.0	61.5	68.3
Collocation	8	75.0	62.5	62.5	87.5	62.5	50.0	50.0	62.5	62.5	62.5	37.5	61.4
Compound Adjectives	6	100.0	100.0	100.0	83.3	100.0	100.0	100.0	66.7	83.3	66.7	100.0	90.9
Idiom	8	25.0	50.0	50.0	37.5	25.0	25.0	37.5	37.5	25.0	12.5	12.5	30.7
Nominal MWE	6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Prepositional MWE	5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Verbal MWE	6	83.3	50.0	66.7	50.0	33.3	33.3	33.3	33.3	50.0	33.3	50.0	47.0
Named entitiy & terminology	26	**73.1**	**88.5**	**84.6**	**84.6**	**84.6**	65.4	**69.2**	**73.1**	**73.1**	65.4	65.4	75.2
Date	5	100.0	100.0	100.0	100.0	100.0	80.0	80.0	80.0	100.0	100.0	100.0	94.5

category/phenomenon	count	Onl-W	Onl-G	Onl-B	JDExp	LanBr	Huawe	Onl-A	PROMT	Onl-Y	SRPOL	eTran	avg
Domainspecific Term	5	80.0	100.0	100.0	80.0	100.0	40.0	60.0	60.0	40.0	40.0	40.0	67.3
Location	5	40.0	60.0	80.0	80.0	80.0	60.0	80.0	60.0	80.0	60.0	60.0	67.3
Measuring unit	5	80.0	80.0	80.0	80.0	80.0	80.0	60.0	100.0	80.0	60.0	60.0	76.4
Proper name	6	66.7	100.0	66.7	83.3	66.7	66.7	66.7	66.7	66.7	66.7	66.7	71.2
Negation	5	100.0	80.0	100.0	80.0	100.0	80.0	80.0	80.0	80.0	100.0	80.0	87.3
Non-verbal agreement	23	73.9	78.3	73.9	82.6	69.6	73.9	69.6	73.9	69.6	69.6	65.2	72.7
Anaphora agreement	7	57.1	71.4	42.9	71.4	42.9	42.9	42.9	57.1	42.9	42.9	42.9	50.6
Coreference	5	80.0	80.0	100.0	100.0	80.0	100.0	80.0	80.0	60.0	100.0	80.0	85.5
Genitive	6	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	83.3	66.7	66.7	80.3
Possession	5	80.0	80.0	80.0	80.0	80.0	80.0	80.0	80.0	100.0	80.0	80.0	81.8
Punctuation	5	100.0	100.0	100.0	100.0	80.0	100.0	100.0	100.0	80.0	80.0	80.0	92.7
Direct Speech	5	100.0	100.0	100.0	100.0	80.0	100.0	100.0	100.0	80.0	80.0	80.0	92.7
Subordination	49	**91.8**	**89.8**	**91.8**	**93.9**	**89.8**	79.6	81.6	81.6	79.6	81.6	75.5	85.2
Adverbial clause	5	80.0	80.0	80.0	80.0	80.0	80.0	80.0	80.0	40.0	60.0	80.0	74.5
Cleft sentence	5	80.0	80.0	80.0	80.0	60.0	40.0	40.0	80.0	60.0	60.0	40.0	63.6
Contact clause	5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Indirect speech	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
Infinitive clause	5	100.0	80.0	80.0	80.0	80.0	80.0	80.0	80.0	80.0	60.0	80.0	80.0
Object clause	5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	80.0	80.0	100.0	100.0	96.4
Participle clause	5	100.0	100.0	100.0	100.0	100.0	80.0	100.0	100.0	60.0	100.0	100.0	94.5
Pseudo-cleft sentence	5	80.0	80.0	100.0	100.0	100.0	60.0	80.0	80.0	100.0	100.0	20.0	81.8
Relative clause	5	80.0	80.0	80.0	100.0	80.0	80.0	40.0	60.0	80.0	60.0	60.0	72.7
Subject clause	5	100.0	100.0	100.0	100.0	100.0	80.0	100.0	60.0	100.0	80.0	80.0	90.9
Verb tense/aspect/mood	68	70.6	72.1	72.1	76.5	75.0	67.6	73.5	73.5	66.2	70.6	67.6	71.4
Conditional	5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	80.0	98.2
Ditransitive	16	75.0	81.3	81.3	93.8	93.8	87.5	87.5	87.5	68.8	81.3	81.3	83.5
Gerund	5	80.0	80.0	100.0	80.0	80.0	80.0	100.0	80.0	100.0	80.0	60.0	83.6
Imperative	5	80.0	100.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	65.5
Intransitive	16	43.8	43.8	43.8	50.0	50.0	43.8	43.8	56.3	50.0	43.8	43.8	46.6
Reflexive	5	100.0	100.0	80.0	100.0	80.0	80.0	80.0	80.0	80.0	80.0	80.0	85.5
Transitive	16	68.8	62.5	75.0	75.0	75.0	56.3	75.0	68.8	56.3	75.0	75.0	69.3
Verb valency	32	87.5	87.5	78.1	84.4	75.0	75.0	84.4	71.9	78.1	75.0	68.8	78.7
Case government	5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	80.0	80.0	96.4
Catenative verb	7	85.7	85.7	71.4	71.4	57.1	71.4	71.4	57.1	71.4	71.4	57.1	70.1
Impersonal Subject	5	100.0	100.0	100.0	100.0	100.0	80.0	100.0	100.0	100.0	100.0	100.0	98.2
Middle voice	5	40.0	60.0	40.0	40.0	40.0	40.0	60.0	40.0	40.0	40.0	40.0	43.6
Passive voice	5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Resultative	5	100.0	80.0	60.0	100.0	60.0	60.0	80.0	40.0	60.0	60.0	40.0	67.3
micro-average	300	**80.3**	**81.3**	**78.7**	**81.7**	75.0	70.7	72.3	72.3	71.0	70.3	67.0	74.6
phen. macro-average	300	**83.2**	**84.0**	**81.0**	**83.2**	76.7	72.7	74.2	73.8	73.4	72.2	68.3	76.6
categ. macro-average	300	**85.4**	**84.1**	**82.8**	**82.6**	75.9	73.6	73.5	73.4	73.1	72.5	68.3	76.8

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