Linguistic evaluation for the 2021 state-of-the-art Machine Translation systems for German to English and English to German

Vivien Macketanz, Eleftherios Avramidis, Shushen Manakhimova, Sebastian Möller
German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
firstname.lastname@dfki.de

Abstract

We are using a semi-automated test suite in order to provide a fine-grained linguistic evaluation for state-of-the-art machine translation systems. The evaluation includes 18 German to English and 18 English to German systems, submitted to the Translation Shared Task of the 2021 Conference on Machine Translation. Our submission adds up to the submissions of the previous years by creating and applying a wide-range test suite for English to German as a new language pair. The fine-grained evaluation allows spotting significant differences between systems that cannot be distinguished by the direct assessment of the human evaluation campaign. We find that most of the systems achieve good accuracies in the majority of linguistic phenomena but there are few phenomena with lower accuracy, such as the idioms, the modal pluperfect and the German resultative predicates. Two systems have significantly better test suite accuracy in macro-average in every language direction, Online-W and Facebook-AI for German to English and VolcTrans and Online-W for English to German. The systems show a steady improvement as compared to previous years.

1 Introduction

Evaluation in NLP and particularly in Machine Translation (MT) is an essential process for identifying flaws and leading further system improvements. Nevertheless, the exact method of evaluation to be used varies, given the quality requirements of the particular use case. Whereas the vast majority of the evaluation methods reside on metrics or direct assessment by humans to produce a single quality score given an entire test set, a recent trend has opted to evaluating the details of the produced translations, with major focus on their correctness from a linguistic perspective. For this reason, the translation systems are not tested based on generic test-sets, but they are given input which is particularly crafted to trial their performance. Most commonly, this is done with the help of a test suite (cf. Guillou and Hardmeier, 2016; Isabelle et al., 2017b; Burchardt et al., 2017).

The paper at hand describes the use of a test suite in order to evaluate 18 German to English and 18 English to German MT systems that participated at the Shared Task of the Sixth Conference on Machine Translation (WMT21)[1]. The evaluation is performed by an extensive test suite that tests a wide range of linguistically motivated phenomena. In addition to our contributions in the previous years, which focused only on German to English, this year we are presenting for the first time results with an extensive test suite with a similar logic for the opposite direction English to German. Our German to English test set contains 5,560 test sentences, covering 107 linguistic phenomena that are organized in 14 categories. The English to German test set contains 4,443 test sentences, covering 111 linguistic phenomena that are organized in 12 categories.

2 Related Work

Test suites have already been used since the beginnings of MT in the 1990s (King and Falkedal, 1990; Way, 1991; Heid and Hildenbrand, 1991). With the rise of deep learning, the quality of MT outputs has improved significantly, which in turn lead to a recent revival of test suites that focus on the evaluation of specific linguistic phenomena (e.g., pronoun translation (Guillou and Hardmeier, 2016), or on the comparison of different MT technologies (Isabelle et al., 2017a; Burchardt et al., 2017), and Quality Estimation methods (Avramidis et al., 2018).

Within the scope of the test suite track of the Conference on Machine Translation, several test suites for multiple language directions have

[1] http://statmt.org/wmt21/

Lexical Ambiguity	
Er las gerne Novellen.	
He liked to read novels.	fail
He liked to read novellas.	pass
Phrasal verb	
Warum starben die Dinosaurier aus?	
Why did the dinosaurs die?	fail
Why did the dinosaurs die out?	pass
Why did the dinosaurs become extinct?	pass
Ditransitive Perfect	
Ich habe Tim einen Kuchen gebacken.	
I have baked a cake.	fail
I baked Tim a cake.	pass

Table 1: Examples of passing and failing MT outputs

been introduced. These test suites focus on one or multiple different phenomena, such as conjunctions (Popović, 2019), grammatical contrasts (Cinkova and Bojar, 2018), discourse (Bojar et al., 2018; Rysová et al., 2019), domain-specific translations (Vojtěchová et al., 2019), gender coreference (Kocmi et al., 2020), markables (Zouhar et al., 2020), morphology (Burlot et al., 2018), pronouns (Guillou et al., 2018), or word sense disambiguation (Rios et al., 2018; Raganato et al., 2019; Scherrer et al., 2020). In contrast to the majority of these test suites, our test suite does not focus on a single phenomenon but performs a systematic evaluation of more than one hundred phenomena per language direction.

3 Methods

Our test suite consists of two test sets (one per language direction) that have been created manually with the aim of testing the performance of MT systems. They cover a wide variety of linguistic phenomena which are grouped in different categories. While there is a big overlap between the linguistic categories and phenomena in the two test sets, there are also many differences as the categories and phenomena are language-specific. Some exemplary test sentences can be seen in Table 1.[2] Each linguistic phenomenon in the test suite is represented by multiple test sentences. Each test sentence is tied to a number of rules that determine whether a translation of the sentence would be deemed correct or incorrect. The performance of an MT system with regard to the linguistic phenomena is then evaluated by observing the amount of test sentences that are translated correctly.

[2]A larger set of exemplary test sentences can be found in the GitHub repository: https://github.com/DFKI-NLP/TQ_AutoTest.

3.1 Application of the test suite

The construction of the test suite has been described in detail in the papers for the test suite track from the previous years. Figure 1 depicts the preparation and application of the test suite with steps *a* to *c* representing the construction. The application starts with step *d*: The test sentences are given as input to the MT systems. The MT outputs are then evaluated by the set of rules which define whether the phenomenon under inspection is translated correctly or not (step *e*). The rules consist of regular expressions and fixed strings. When the rules cannot be applied to a translation to automatically determine whether it is correct or incorrect, the test sentence is marked with a warning. Those warnings are consequently inspected manually by a human annotator with linguistic knowledge who decides on the correctness of the translation and adapts the set of rules accordingly (step *e*).

Thereafter, 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 the aim of this evaluation is to compare the systems in a fair way, we include only the test items that do not contain any warnings for any of the systems in the calculation. Test items that have an unresolved warning for at least one system are excluded from the calculation. Unfortunately, this reduces the amount of the test items by removing properly validated ones, and this is where we see the importance of the extensive manual evaluation and the creation of rules with good coverage.

To define which systems perform better for a particular phenomenon (or category), we compare all systems to the one with the highest accuracy. When we compare the highest scoring system with the rest, we confirm the significance of the comparison with a one-tailed Z-test with $\alpha = 0.95$. The systems which do not differ significantly from the best system are considered to be in the first performance cluster and indicated with boldface in the tables. The boldfaces therefore have a meaning only for the respective row of the table.

The average scores are computed in three different ways, because each category or phenomenon has a different amount of test items. *Micro-average* aggregates the contributions of all test items to

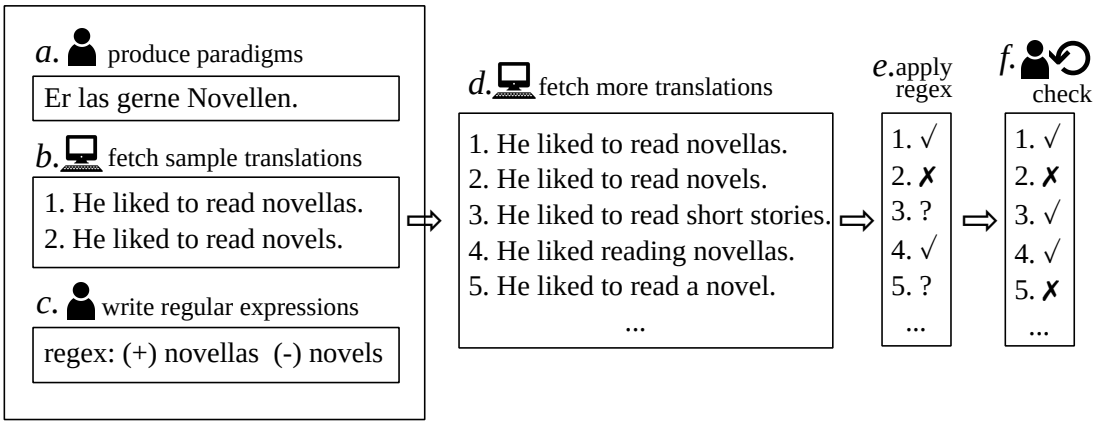

Figure 1: Example of the preparation and application of the test suite for one test sentence

compute the average percentages, *category macro-average* computes the percentages independently for each category and then averages them (i.e. treating all categories equally), and *phenomenon macro-average* computes the percentages independently for each phenomenon and then takes the average (i.e. treating all phenomena equally).

3.2 Experiment setup

In the evaluation presented in this paper, we obtained translations of our test suite by 36 systems that are part of the *news translation task* of the Sixth Conference on Machine Translation (WMT21). In previous years, we solely applied our test suite to the German to English MT outputs. However, this year, we did not only analyse the MT outputs from 18 German to English systems, but also from 18 English to German systems.

While there were already many rules for the evaluation of German to English MT output in our test suite, very few rules were available for the other language direction when we received the translations. Therefore, a significantly bigger amount of manual work was involved in the evaluation this year. For German to English there were on average 5.76% of warnings when we received the translations, while for English to German there were on average 84.21% of warnings. The manual evaluation process was conducted by three annotators with linguistic knowledge over the course of seven weeks and involved around 80 person hours. After the extensive manual evaluation, there were on average 3.04% of warnings for German to English and 4.87% for English to German.

As we explained previously, in order to have a fair comparison between the systems we excluded items where at least one system has an unresolved warning. Therefore, in the results that we are presenting in this paper we can only use 3,806 out of the 5,560 (68.4%) test items for German to English and 3,096 out of the 4,443 (69.7%) test items for English to German for the systems comparison.

4 Results

The accuracies resulting from the application of the test suite on the system outputs can be seen in the tables in the Appendix. We first present the results aggregated in categories (Tables 4 and 5) so one can have a broad overview of the systems performance, whereas afterwards a yearly comparison with last years (Table 6) and the detailed phenomenon-level results (Tables 7 and 8) are shown. The systems are ordered based on their macro-average accuracy, from high to low.[3]

4.1 Comparison between systems

For German to English, two systems have the highest category macro-averaged accuracy, Online-W and FacebookAI, whereas when considering the phenomenon macro-averaged accuracy, the significantly best systems are FacebookAI and Online-A. UEdin, Online-A and borderline compete with the best systems when the micro-average is considered, mainly because of their good accuracies on phenomena related to *verb tense/aspect/mood*, where there are many individual phenomena with a lot of test items in one category. Overall, the average accuracies are very high, with the lowest system (happyface) having a micro-average of 72.3%. Despite the high accuracies there is definitely room

[3]For German-English the two VolcTrans system variations appear as one system, since they delivered the same output. This is not the case for the English-German direction where they appear separately.

for improvement.

For English to German, based on the category macro-average, FacebookAI and VolcTransAT share the first position. Based on the micro-average and the phenomenon macro-average however, FacebookAI, Online-B and VolcTrans-GLAT share the first position. The accuracies for this language direction in overall are much higher on the micro-average, but not on the macro-average. However, due to the fact that the test items are different in their nature and in the amount, we cannot make a direct comparison between the two language directions.

4.2 Categories

For some categories, the accuracies have reached very high numbers, which is the case for *negation* and *punctuation*, both having a 100% for German to English. Concerning punctuation, in the previous years we had seen individual systems with considerable punctuation errors, which seem to not appear this year. However, the high scores do not necessarily mean that all problems for these phenomena are solved. It could rather mean that our test suite does not cover the current edge cases, a consideration that is subject to further research. Other categories such as *composition*, *subordination* and *named entities & terminology* reach an average of more than 90% accuracy in German to English. The worst performing category in German to English is *false friends*, where all systems perform 64-86%. *Ambiguity*, *verb tense/aspect/mood* and *multi-word expressions* (MWE) also perform relatively low, with accuracies less than 85%.

For English to German, there are no categories for which all systems reach an accuracy of 100%. However, there are several categories with average accuracies above 95%, that is *function words*, *negation*, *verb tense/aspect/mood*, and *subordination*. The category with the lowest average is *coordination & ellipsis*, with an average accuracy of only 70.8%. The individual systems reach a wide range of 58.6% to 81.6% accuracy for this category while for most other systems, the range is not as big for the systems. There are two more categories with a relatively low accuracy on average (below 85%), namely *verb valency* (81.4 % accuracy) and *ambiguity* (83.3% accuracy).

4.3 Phenomena

For **German to English**, the most difficult phenomena this year remain the *modal pluperfect* (*negated* and *non-negated*), the *resultative predicates* and the *idioms*. Online-W does impressively well with *idioms*, achieving almost 60%, with another two systems, FacebookAI and Online-A, reaching 33.3%. These numbers were significantly lower in the previous years, which indicates an improvement in this direction. There are some phenomena for which all systems reached 100% accuracy, such as *negation*, *internal possessor*, *comma*, *ditransitive perfect*, and *intransitive future I*.

Idiom	
Er redet um den heißen Brei herum.	
He's talking around the hot porridge.	fail
He's talking around the bush.	fail
He's beating around the bush.	pass
Modal pluperfect	
Sie hatten lesen wollen.	
They wanted to read.	fail
They had to read.	fail
They had wanted to read.	pass
Resultative predicate	
Lisa fuhr das Auto kaputt.	
Lisa drove the car broken.	fail
Lisa broke the car.	pass
Lisa crashed the car.	pass

Table 2: Examples of De-En linguistic phenomena with low accuracy with passing and failing MT outputs

Table 2 contains translation examples from linguistic phenomena with the lowest accuracy for German to English. *Idioms* are types of multiword expressions. The meaning of an idiom goes beyond the meanings of its individual elements. Most idioms are very language-specific and therefore difficult to translate. For the German idiom "um den heißen Brei herumreden", there is the equivalent English idiom "to beat about the bush". The first incorrect translation contains a direct translation of all the individual elements of the German idiom. The second incorrect translation, which was produced by several MT systems, is very interesting because it does indeed contain the "bush" of the English idiom. However, it still contains the wrong verb as the verb "is talking" is simply a translation of the German "redet". Therefore, the second translation is still incorrect. Only the third translation which contains the full English idiom is correct.

The second example contains a test sentence from the phenomenon *modal pluperfect*. Modal verbs can usually have several meanings which often leads to translation errors. Furthermore, the tense pluperfect is often mistranslated as preterite, as in the first incorrect translation. The second in-

Idiom	
The mafia boss has spilled the beans.	
Der Mafiaboss hat die Bohnen verschüttet.	fail
Der Mafiaboss hat sich verplappert.	pass
Der Mafiaboss hat es ausgeplaudert.	pass
Pseudogapping	
Jackie likes the doctor but she doesn't the nurse.	
Jackie mag den Arzt, aber sie nicht die Krankenschwester.	fail
Jackie mag den Arzt, aber sie ist nicht die Krankenschwester.	fail
Jackie mag den Arzt, aber nicht die Krankenschwester.	pass
Middle Voice	
This car drives easily.	
Dieses Auto fährt leicht.	fail
Dieses Auto fährt sich leicht.	pass
Das Auto ist leicht zu fahren.	pass

Table 3: Examples of En-De linguistic phenomena with low accuracy with passing and failing MT outputs

correct translation additionally leaves out the German modal verb "wollen" ("to want") which completely changes the meaning of the translation.

Resultative predicates contain a verb and an adjective which describes the result of the verb action. *Resultative predicates* do not exist that way in English, which makes them hard to translate. In the example at hand, the meaning of the German sentence is that Lena drove the car which resulted in the car being broken. A literal translation like in the first translation is ungrammatical. The second and third translation are possible correct translations – even though the "driving" part is left out, these translations are still deemed best options to translate this phenomenon.

In **English to German**, *idioms* show even more difficulties as in German to English (average accuracy only 14.6%, the lowest average accuracy on any phenomenon for this language direction). Here, 9 systems totally fail to translate any idiom, whereas the system with the highest accuracy is an unconstrained system, which may attributed to the fact that additional data led to better coverage of such cases. Furthermore, *middle voice* (45.9%), *pseudogapping* (60.5%), and *stripping* (57.0%) and also have a relatively low accuracy. On the other hand, there were also many phenomena which reached (nearly) 100% accuracy, such as *internal possessor*, *comma*, *indirect speech*, *infinitive clause*, *object clause*, *subject clause*, *passive voice*, and *ditransitive*, *intransitive* and *transitive verbs* in many tenses.

Table 3 covers example translation from low accuracy phenomena for English to German. The first example again contains an *idiom*. The English idiom "to spill the beans" does not have an equivalent idiomatic translation in German. Therefore, the first translation, which is a literal translation of the separate idiom elements, is incorrect. The second and third translation are possible correct translations.

The second example sentence is taken from the phenomenon *pseudogapping*. Put simply, in *pseudogapping*, part of the verb phrase is omitted. In the example at hand, the non-finite verb part "like" is omitted in the second conjunct of the construction. In the first incorrect German translation, the verb has been completely left out in the second conjunct (while the subject "sie" persists). In the second incorrect translation, the second conjunct contains the auxiliary verb 'ist" which also leads to ungrammaticality. The third translation leaves out the non-finite verb part "like" as well as the subject which results in a grammatical German construction.

The third example contains a sentence in *middle voice*. In middle voice, the subject of the verb is neither agent nor patient. A sentence in active voice would be: "I am driving the car.", with the subject ("I") being the agent. A sentence in passive voice would be: "The car is driven by me." with the subject ("the car") being the patient. The subject of the example sentence in Table 3 ("This car") is neither agent nor patient. As middle voice does not exist in German, such sentences have to be translated in other constructions. A literal translation like the first example translation is incorrect. Possible correct translations can be seen in the second and third translation.

5 Comparison with previous years

The progress of the systems performance through the last four years for German-English can be seen in Table 6. The calculation is done based on the common test items without warnings over all these years (4,366 test items), this is why the scores differ slightly from the ones in Table 4. In the first columns of Table 6 the best systems of every year are compared. One can see that the best system of 2021 has significantly better macro-averaged accuracy as compared to the best system of 2020, but when the micro-averaged accuracy is considered, there has been no significant improvement or deterioration. This year's best system also seems

to perform better in a few categories, with most impressive improvements at *false friends* (+14%) and the *non-verbal agreement* (+5%).

Individual systems show some small improvements in general, but the fine-grained evaluation is able to indicate some significant deterioration in particular categories. For example, Online-B, Online-G and VolcTrans, despite their overall improvement, show a significant deterioration regarding *verb tense/aspect/mood*, which reaches a -9% in the case of VolcTrans. Other deteriorations occur for several systems regarding *false friends* and *function words*. This shows that the overall improvement in translation quality may occur at the expense of particular qualitative aspects.

6 Conclusions and Further Work

We presented the result of applying a fine-grained linguistically motivated test suite on the outputs of 36 state-of-the-art machine translation systems, as submitted in the Sixth Conference on Machine Translation. We presented detailed accuracies of translations of 18 German to English as well as 18 English to German MT systems based on more than 3,000 test items each, organized in various linguistic categories and fine-grained phenomena. Additionally, we drew a comparison to previous years' evaluations.

In both language directions, the systems achieve good accuracies in most phenomena or categories and there is some advancement as compared to last year, although there is space for about 10% improvement on the average accuracy. A few phenomena still suffer considerably, such as the *idioms*, the *modal pluperfect* and the German *resultative predicates*, although there is notable improvement as compared to previous years.

As discussed, the very high accuracies for some categories or phenomena raise the question whether the difficulty of the respective test items should be increased. In future work, we plan to investigate this by constructing more test items. Further work includes the development of similar test suites for other language pairs.

Acknowledgements

This research was supported by the Deutsche Forschungsgemeinschaft (DFG) through the project TextQ, and by the German Federal Ministry of Education through the project SocialWear. We thank our colleague Tatjana Zeen for her valuable help with the evaluation.

References

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.

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(1):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.

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.

Pierre Isabelle, Colin Cherry, and George Foster. 2017a. 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.

Pierre Isabelle, Colin Cherry, and George Foster. 2017b. A Challenge Set Approach to Evaluating Machine Translation.

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.

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.

Appendix

category	count	Onl-W	Faceb	Onl-B	VolcT	Onl-A	SMU	Onl-G	Huawe	borde	Nemo	uedin	Water	P3AI	ICL	Onl-Y	Manif	happy	avg
Ambiguity	74	**87.8**	**90.5**	**86.5**	**86.5**	**81.1**	**83.8**	**85.1**	**89.2**	**83.8**	**83.8**	**83.8**	75.7	79.7	**86.5**	**82.4**	**81.1**	60.8	82.8
Composition	43	**97.7**	**97.7**	**100.0**	**100.0**	**97.7**	**95.3**	**97.7**	**95.3**	**95.3**	93.0	**97.7**	**97.7**	**97.7**	**95.3**	93.0	93.0	74.4	95.2
Coordination & ellipsis	57	89.5	89.5	89.5	89.5	87.7	86.0	86.0	87.7	89.5	87.7	87.7	86.0	87.7	77.2	87.7	89.5	80.7	87.0
False friends	36	**86.1**	**80.6**	**75.0**	**75.0**	**83.3**	**83.3**	**80.6**	63.9	**77.8**	**72.2**	66.7	**80.6**	**80.6**	**72.2**	**75.0**	**69.4**	63.9	75.7
Function word	40	**92.5**	**92.5**	**92.5**	**92.5**	**90.0**	**85.0**	**95.0**	**92.5**	**85.0**	**92.5**	**92.5**	**92.5**	**85.0**	**87.5**	**90.0**	72.5	80.0	88.8
LDD & interrogatives	103	**91.3**	**91.3**	**91.3**	**91.3**	**91.3**	**91.3**	**90.3**	**93.2**	**90.3**	**91.3**	**89.3**	**92.2**	**89.3**	**91.3**	**88.3**	57.3	74.8	87.9
MWE	66	**90.9**	**86.4**	**83.3**	**83.3**	**86.4**	**86.4**	**84.8**	**86.4**	**86.4**	**86.4**	**83.3**	**80.3**	**84.8**	**86.4**	**81.8**	**84.8**	69.7	84.2
Named entity & terminology	71	**95.8**	**94.4**	**93.0**	**93.0**	**94.4**	**93.0**	**94.4**	**95.8**	**91.5**	**91.5**	**95.8**	**88.7**	**90.1**	**91.5**	**93.0**	**90.1**	83.1	92.3
Negation	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	100.0	100.0	100.0	100.0	100.0
Non-verbal agreement	57	**98.2**	**94.7**	**98.2**	**98.2**	**93.0**	**91.2**	89.5	**93.0**	**93.0**	**93.0**	89.5	89.5	**91.2**	**93.0**	84.2	**93.0**	73.7	91.5
Punctuation	18	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	100.0	100.0	100.0	100.0
Subordination	115	**92.2**	**93.9**	**95.7**	**95.7**	**93.9**	**92.2**	**93.0**	**92.2**	**92.2**	**93.9**	**93.9**	**94.8**	**93.0**	**93.9**	**93.9**	**93.9**	87.0	93.2
Verb tense/aspect/mood	3058	**87.3**	**87.3**	79.6	79.6	**86.4**	**85.8**	80.5	82.7	**86.5**	83.9	**86.9**	84.1	81.3	82.6	77.7	84.1	71.1	82.8
Verb valency	54	88.9	90.7	92.6	92.6	90.7	90.7	87.0	90.7	90.7	90.7	88.9	88.9	88.9	90.7	85.2	90.7	81.5	89.4
micro-average	3806	**88.3**	**88.2**	82.0	81.9	**87.3**	86.6	82.4	84.3	**87.1**	85.1	**87.4**	85.0	82.8	83.9	79.7	84.0	72.3	84.0
macro-average	3806	**92.7**	**92.1**	91.2	91.2	91.1	90.3	90.3	90.2	90.1	90.0	89.7	89.3	89.2	89.2	88.0	85.7	78.6	89.4

Table 4: Accuracies (%) of successful translations on a category level for German-English. Boldface indicates the significantly best performing systems in each row.

categ	count	Faceb	VolcA	Onl-W	Onl-A	Huawe	Nemo	Onl-B	VolcG	uedin	P3AI	eTran	happy	nucle	Onl-Y	Manif	BUPT	ICL	Onl-G	avg
Ambiguity	23	**91.3**	**95.7**	**95.7**	**91.3**	**87.0**	**87.0**	**91.3**	**91.3**	**82.6**	**82.6**	**78.3**	73.9	73.9	69.6	**82.6**	69.6	**82.6**	73.9	83.3
Coordination & ellipsis	87	**81.6**	**71.3**	**71.3**	**73.6**	**77.0**	**75.9**	**80.5**	**79.3**	69.0	69.0	66.7	64.4	65.5	**71.3**	63.2	63.2	58.6	**72.4**	70.8
False friends	38	92.1	92.1	89.5	86.8	86.8	84.2	84.2	84.2	86.8	86.8	86.8	86.8	81.6	86.8	86.8	86.8	84.2	84.2	86.5
Function word	35	**97.1**	**97.1**	**100.0**	**97.1**	**97.1**	**94.3**	**100.0**	**100.0**	**100.0**	**97.1**	**94.3**	**97.1**	**97.1**	**97.1**	65.7	**97.1**	**100.0**	**97.1**	95.9
MWE	98	**89.8**	**93.9**	**91.8**	85.7	**87.8**	**88.8**	**90.8**	**90.8**	82.7	85.7	84.7	**89.8**	82.7	82.7	83.7	81.6	80.6	81.6	86.4
Named entity & terminology	82	**93.9**	**97.6**	**93.9**	**93.9**	**93.9**	89.0	**93.9**	**93.9**	**92.7**	89.0	**93.9**	90.2	90.2	**92.7**	89.0	**92.7**	81.7	80.5	91.3
Negation	15	100.0	100.0	100.0	93.3	93.3	100.0	93.3	93.3	100.0	100.0	93.3	93.3	86.7	100.0	100.0	93.3	93.3	93.3	95.9
Non-verbal agreement	68	**100.0**	**98.5**	**97.1**	95.6	95.6	92.6	92.6	92.6	92.6	89.7	91.2	92.6	88.2	89.7	92.6	88.2	88.2	89.7	92.6
Punctuation	37	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	78.4	78.4	91.9	81.1	78.4	81.1	86.5	75.7	78.4	78.4	78.4	70.3	86.5
Subordination	161	**99.4**	**98.1**	**98.1**	**99.4**	95.7	**99.4**	**98.1**	**98.1**	**98.1**	**98.8**	**98.1**	**96.9**	**97.5**	93.8	**96.9**	94.4	92.5	96.3	97.2
Verb tense/aspect/mood	2366	98.6	97.9	97.3	96.9	96.1	97.4	**99.0**	**99.1**	**99.2**	97.4	98.4	96.7	97.3	90.7	98.6	94.8	95.2	94.7	97.0
Verb valency	96	**90.6**	81.3	**85.4**	81.3	**84.4**	81.3	**83.3**	**83.3**	81.3	**83.3**	**84.4**	80.2	80.2	77.1	81.3	77.1	75.0	74.0	81.4
micro-average	3106	**97.4**	96.5	95.9	95.3	94.7	95.6	**96.9**	**96.9**	96.5	95.1	95.8	94.4	94.5	89.4	95.2	92.3	92.1	92.0	94.8
macro-average	3106	**94.5**	**93.6**	93.3	91.2	91.2	90.8	90.5	90.4	89.7	88.4	87.4	86.9	85.6	85.6	84.9	84.8	84.2	84.0	88.7

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

category	count	best				Faceb		Onl-B				Volc		Onl-A				Onl-G				uedin				Onl-Y		
		2018	2019	2020	2021	2019	2021	2018	2019	2020	2021	2020	2021	2018	2019	2020	2021	2018	2019	2020	2021	2018	2019	2020	2021	2018	2019	2021
Ambiguity	76	76	92	83	86	92	89	76	78	79	86	78	86	68	70	78	82	72	75	84	86	50	62	75	84	67	79	83
Composition	45	98	98	98	96	98	98	98	98	98	100	98	100	80	93	93	96	71	82	96	98	76	84	93	96	89	91	93
Coordination & ellipsis	43	88	88	88	91	88	91	88	88	91	88	88	88	86	86	86	88	49	60	77	88	81	81	86	88	77	86	88
False friends	36	75	75	72	86	75	81	75	78	81	75	81	75	72	72	69	83	72	72	78	81	53	67	72	67	67	92	75
Function word	52	81	92	94	96	92	96	81	81	94	88	85	88	88	90	90	92	50	96	96	98	83	90	92	96	92	94	88
LDD & interrogatives	73	85	90	92	92	90	92	85	85	89	96	90	96	81	79	85	92	67	77	90	90	75	77	90	86	85	82	90
MWE	64	75	84	84	89	84	84	75	75	81	81	78	81	69	69	75	83	67	72	83	83	58	63	73	81	73	75	78
Named entity & terminology	57	91	91	96	96	91	93	91	91	88	91	91	91	91	91	95	93	88	86	91	93	82	91	95	96	91	89	93
Negation	17	94	100	100	100	100	100	94	94	100	100	100	100	100	100	100	100	65	100	100	100	100	94	100	100	100	100	100
Non-verbal agreement	56	88	91	91	96	91	95	88	88	88	98	89	98	77	84	84	91	57	80	91	91	71	84	89	89	79	82	86
Punctuation	35	97	97	97	97	97	100	97	97	94	100	100	100	100	100	100	100	83	83	86	86	94	91	100	100	100	100	100
Subordination	83	88	93	95	94	93	93	88	89	96	98	94	98	96	86	95	95	84	93	96	93	92	89	96	95	94	94	95
Verb tense/aspect/mood	3676	77	82	88	86	82	86	77	77	79	78	87	78	75	87	81	85	49	69	83	79	79	84	83	85	74	75	76
Verb valency	53	83	89	89	87	89	91	83	83	92	91	91	91	79	83	87	87	72	79	89	89	74	77	85	89	81	83	83
micro-avg	4366	78	83	88	87	83	87	78	78	80	80	88	80	76	86	82	86	52	71	84	80	78	83	84	85	75	77	78
macro-avg	4366	85	90	91	92	90	92	85	86	89	91	89	91	83	85	87	90	68	80	89	90	76	81	88	89	83	87	88

Table 6: Accuracies (%) of the German to English systems that were submitted also in previous years.

phenomenon	count	Onl-W	Faceb	Onl-B	VolcT	Onl-A	SMU	Onl-G	Huawe	borde	Nemo	uedin	Water	P3AI	ICL	Onl-Y	Manif	happy	avg
Ambiguity	74	**87.8**	**90.5**	**86.5**	**86.5**	**81.1**	**83.8**	**85.1**	**89.2**	**83.8**	**83.8**	**83.8**	75.7	79.7	**86.5**	**82.4**	**81.1**	60.8	82.8
Lexical ambiguity	61	**91.8**	**91.8**	**88.5**	**88.5**	**82.0**	**86.9**	**86.9**	**88.5**	**86.9**	**85.2**	**85.2**	78.7	**82.0**	**88.5**	**85.2**	**82.0**	62.3	84.8
Structural ambiguity	13	**69.2**	**84.6**	**76.9**	**76.9**	**76.9**	**69.2**	**76.9**	**92.3**	**69.2**	**76.9**	**76.9**	**61.5**	**69.2**	**76.9**	**69.2**	**76.9**	53.8	73.8
Composition	43	**97.7**	**97.7**	**100.0**	**100.0**	**97.7**	**95.3**	**97.7**	**95.3**	**95.3**	93.0	**97.7**	**97.7**	**97.7**	**95.3**	93.0	93.0	74.4	95.2
Compound	25	**96.0**	**96.0**	**100.0**	**100.0**	**96.0**	**92.0**	**96.0**	**96.0**	**92.0**	88.0	**96.0**	**96.0**	**96.0**	**92.0**	88.0	**92.0**	84.0	93.9
Phrasal verb	18	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.4**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.4**	61.1	97.1
Coordination & ellipsis	57	89.5	89.5	89.5	89.5	87.7	86.0	86.0	87.7	89.5	87.7	87.7	86.0	87.7	77.2	87.7	89.5	80.7	87.0
Gapping	15	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	93.3	93.3	93.3	93.3	100.0	100.0	86.7	97.6
Right node raising	15	80.0	80.0	80.0	80.0	80.0	73.3	80.0	80.0	80.0	80.0	80.0	80.0	80.0	80.0	73.3	80.0	73.3	78.8
Sluicing	13	100.0	100.0	92.3	92.3	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	99.1
Stripping	14	**78.6**	**78.6**	**85.7**	**85.7**	**71.4**	**71.4**	**64.3**	**71.4**	**78.6**	**71.4**	**78.6**	**71.4**	**78.6**	35.7	**78.6**	**78.6**	**64.3**	73.1
False friends	36	**86.1**	**80.6**	**75.0**	**75.0**	**83.3**	**83.3**	**80.6**	63.9	**77.8**	**72.2**	66.7	**80.6**	**80.6**	**72.2**	**75.0**	**69.4**	63.9	75.7
Function word	40	**92.5**	**92.5**	**92.5**	**92.5**	**90.0**	**85.0**	**95.0**	**92.5**	**85.0**	**92.5**	**92.5**	**92.5**	**85.0**	**87.5**	**90.0**	72.5	80.0	88.8
Focus particle	21	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**90.5**	**100.0**	**100.0**	**90.5**	**100.0**	**100.0**	**100.0**	**90.5**	**95.2**	**100.0**	**95.2**	85.7	96.9
Modal particle	14	78.6	78.6	78.6	78.6	78.6	71.4	85.7	78.6	71.4	78.6	78.6	85.7	71.4	71.4	71.4	64.3	78.6	76.5
Question tag	5	**100.0**	**100.0**	**100.0**	**100.0**	**80.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**80.0**	**100.0**	**100.0**	**100.0**	0.0	**60.0**	89.4

phenomenon	count	Onl-W	Faceb	Onl-B	VolcT	Onl-A	SMU	Onl-G	Huawe	borde	Nemo	uedin	Water	P3AI	ICL	Onl-Y	Manif	happy	avg
LDD & interrogatives	103	**91.3**	**91.3**	**91.3**	**91.3**	**91.3**	**91.3**	**90.3**	**93.2**	**90.3**	**91.3**	**89.3**	**92.2**	**89.3**	**91.3**	**88.3**	57.3	74.8	87.9
Extended adj. construction	9	**88.9**	**88.9**	**77.8**	**77.8**	**100.0**	**77.8**	**88.9**	**88.9**	66.7	**100.0**	**77.8**	**88.9**	66.7	**88.9**	55.6	66.7	**77.8**	81.0
Extraposition	11	90.9	90.9	90.9	90.9	81.8	100.0	81.8	90.9	100.0	90.9	90.9	90.9	90.9	90.9	90.9	100.0	81.8	90.9
Multiple connectors	13	**84.6**	**84.6**	**84.6**	**84.6**	**84.6**	**92.3**	76.9	**100.0**	**84.6**	**84.6**	**84.6**	**84.6**	**84.6**	**84.6**	76.9	**84.6**	**84.6**	85.1
Pied-piping	14	**92.9**	**85.7**	**85.7**	**85.7**	**85.7**	**92.9**	**85.7**	**85.7**	**92.9**	**85.7**	**85.7**	**92.9**	**92.9**	**85.7**	**85.7**	**92.9**	50.0	86.1
Polar question	12	**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**	**100.0**	0.0	**91.7**	93.6
Scrambling	9	88.9	88.9	88.9	88.9	88.9	88.9	88.9	88.9	88.9	88.9	77.8	77.8	77.8	88.9	88.9	77.8	44.4	83.7
Topicalization	10	70.0	**90.0**	**90.0**	**90.0**	**90.0**	**100.0**	**90.0**	**90.0**	**90.0**	**90.0**	**90.0**	**100.0**	**90.0**	**90.0**	**100.0**	**90.0**	**80.0**	90.0
Wh-movement	25	**100.0**	**96.0**	**100.0**	**100.0**	**96.0**	84.0	**100.0**	**96.0**	**92.0**	**92.0**	**96.0**	**96.0**	**96.0**	**96.0**	**96.0**	8.0	80.0	89.6
MWE	66	**90.9**	**86.4**	**83.3**	**83.3**	**86.4**	**86.4**	**84.8**	**86.4**	**86.4**	**86.4**	**83.3**	**80.3**	**84.8**	**86.4**	**81.8**	**84.8**	69.7	84.2
Collocation	16	**100.0**	**100.0**	**93.8**	**93.8**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**87.5**	**100.0**	**100.0**	**93.8**	**100.0**	81.3	97.1
Idiom	12	58.3	33.3	25.0	25.0	33.3	25.0	25.0	25.0	25.0	25.0	16.7	16.7	16.7	25.0	16.7	16.7	16.7	25.0
Prepositional MWE	19	**94.7**	**94.7**	**94.7**	**94.7**	**100.0**	**100.0**	**94.7**	**100.0**	**100.0**	**100.0**	**94.7**	**94.7**	**100.0**	**100.0**	**94.7**	**100.0**	78.9	96.3
Verbal MWE	19	**100.0**	**100.0**	**100.0**	**100.0**	**94.7**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	84.2	98.8
Named entity & terminology	71	**95.8**	**94.4**	**93.0**	**93.0**	**94.4**	**93.0**	**94.4**	**95.8**	**91.5**	**91.5**	**95.8**	**88.7**	**90.1**	**91.5**	**93.0**	**90.1**	83.1	92.3
Date	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	94.1	100.0	100.0	94.1	100.0	99.3
Domainspecific term	10	80.0	80.0	70.0	70.0	70.0	70.0	80.0	80.0	70.0	70.0	80.0	60.0	70.0	70.0	70.0	70.0	60.0	71.8
Location	19	94.7	94.7	94.7	94.7	94.7	94.7	94.7	94.7	89.5	89.5	94.7	89.5	94.7	94.7	94.7	94.7	78.9	92.9
Measuring unit	19	**100.0**	**94.7**	**94.7**	**94.7**	**100.0**	**94.7**	**100.0**	**100.0**	**94.7**	**94.7**	**100.0**	**89.5**	**89.5**	**89.5**	**94.7**	**89.5**	78.9	94.1
Proper name	6	100.0	100.0	100.0	100.0	100.0	100.0	83.3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.0
Negation	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	100.0	100.0	100.0	100.0	100.0
Non-verbal agreement	57	**98.2**	**94.7**	**98.2**	**98.2**	**93.0**	**91.2**	89.5	**93.0**	**93.0**	**93.0**	89.5	89.5	**91.2**	**93.0**	84.2	**93.0**	73.7	91.5
Coreference	19	**100.0**	**89.5**	**94.7**	**94.7**	84.2	78.9	73.7	78.9	78.9	78.9	73.7	78.9	73.7	78.9	73.7	78.9	52.6	80.2
External possessor	20	**95.0**	**95.0**	**100.0**	**100.0**	**95.0**	**95.0**	**95.0**	**100.0**	**100.0**	**100.0**	**95.0**	**90.0**	**100.0**	**100.0**	80.0	**100.0**	70.0	94.7
Internal possessor	18	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	100.0	100.0	100.0	100.0
Punctuation	18	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	100.0	100.0	100.0	100.0
Comma	18	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	100.0	100.0	100.0	100.0
Subordination	115	**92.2**	**93.9**	**95.7**	**95.7**	**93.9**	**92.2**	**93.0**	**92.2**	**92.2**	**93.9**	**93.9**	**94.8**	**93.0**	**93.9**	**93.9**	**93.9**	87.0	93.2
Adverbial clause	17	82.4	**88.2**	**100.0**	**100.0**	**94.1**	**94.1**	**88.2**	**94.1**	**94.1**	**88.2**	**94.1**	**94.1**	**94.1**	**94.1**	**94.1**	**94.1**	**100.0**	93.4
Cleft sentence	14	92.9	92.9	92.9	92.9	92.9	85.7	85.7	92.9	85.7	92.9	92.9	92.9	92.9	92.9	92.9	92.9	85.7	91.2
Free relative clause	12	91.7	83.3	83.3	83.3	83.3	83.3	91.7	75.0	83.3	91.7	91.7	91.7	83.3	83.3	83.3	83.3	83.3	85.3
Indirect speech	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	100.0	100.0	100.0	77.8	98.7
Infinitive clause	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	100.0	100.0	100.0	94.1	99.7
Object clause	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	100.0	100.0	100.0	85.7	99.2
Pseudo-cleft sentence	9	66.7	88.9	77.8	77.8	77.8	77.8	77.8	66.7	77.8	77.8	66.7	77.8	66.7	88.9	77.8	88.9	55.6	75.8
Relative clause	13	92.3	92.3	100.0	100.0	92.3	84.6	92.3	92.3	84.6	92.3	92.3	92.3	92.3	84.6	100.0	84.6	92.3	91.9
Subject 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	100.0	90.0	100.0	90.0	98.8
Verb tense/aspect/mood	3058	**87.3**	**87.3**	79.6	79.6	**86.4**	**85.8**	80.5	82.7	**86.5**	83.9	**86.9**	84.1	81.3	82.6	77.7	84.1	71.1	82.8
Conditional	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	100.0	92.9	100.0	92.9	99.2
Ditransitive - future I	23	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	100.0	100.0	91.3	99.5
Ditransitive - future I subjunct. II	28	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	100.0	100.0	96.4	99.8
Ditransitive - future II	14	100.0	100.0	100.0	100.0	100.0	100.0	92.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	92.9	99.2
Ditransitive - future II subjunct. II	27	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	100.0	100.0	96.3	99.8

phenomenon	count	Onl-W	Faceb	Onl-B	VolcT	Onl-A	SMU	Onl-G	Huawe	borde	Nemo	uedin	Water	P3AI	ICL	Onl-Y	Manif	happy	avg
Ditransitive - perfect	23	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	100.0	100.0	100.0	100.0
Ditransitive - pluperfect	27	**100.0**	**92.6**	63.0	63.0	**92.6**	**92.6**	48.1	77.8	**96.3**	85.2	**100.0**	85.2	85.2	85.2	7.4	**92.6**	**92.6**	80.0
Ditransitive - pluperf. subjunct. II	29	**100.0**	**96.6**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**93.1**	**100.0**	**96.6**	**100.0**	**100.0**	62.1	97.0
Ditransitive - present	26	84.6	**96.2**	88.5	88.5	88.5	**100.0**	76.9	**92.3**	**96.2**	**96.2**	**96.2**	**92.3**	**96.2**	88.5	80.8	**96.2**	76.9	90.3
Ditransitive - preterite	21	**100.0**	**90.5**	**100.0**	**100.0**	85.7	**90.5**	85.7	**90.5**	85.7	**95.2**	**95.2**	**90.5**	**95.2**	**95.2**	85.7	**90.5**	81.0	91.6
Ditransitive - preterite subjunct. II	17	100.0	100.0	100.0	100.0	94.1	100.0	94.1	94.1	100.0	100.0	100.0	88.2	100.0	100.0	100.0	100.0	94.1	97.9
Imperative	15	**93.3**	**93.3**	**86.7**	**86.7**	**93.3**	**93.3**	**93.3**	**86.7**	**86.7**	**86.7**	**86.7**	**80.0**	**93.3**	**93.3**	**86.7**	**86.7**	60.0	87.5
Intransitive - future I	31	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	100.0	100.0	100.0	100.0
Intransitive - future I subjunct. II	30	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.7	100.0	96.7	100.0	100.0	99.6
Intransitive - future II	34	91.2	**97.1**	**94.1**	**94.1**	91.2	91.2	85.3	**100.0**	**100.0**	**94.1**	**97.1**	79.4	85.3	**97.1**	79.4	**97.1**	55.9	90.0
Intransitive - future II subjunct. II	35	**94.3**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**97.1**	**100.0**	**100.0**	**100.0**	**100.0**	**94.3**	65.7	**94.3**	**94.3**	82.9	57.1	92.9
Intransitive - perfect	72	**100.0**	**100.0**	**98.6**	**98.6**	**100.0**	95.8	**100.0**	**100.0**	95.8	**100.0**	**100.0**	91.7	**97.2**	94.4	**100.0**	**100.0**	91.7	97.9
Intransitive - pluperfect	31	**87.1**	71.0	19.4	19.4	**93.5**	77.4	41.9	**87.1**	77.4	**87.1**	**87.1**	67.7	**96.8**	71.0	41.9	**83.9**	74.2	69.6
Intransitive - pluperf. subjunct. II	35	**97.1**	**100.0**	**100.0**	**100.0**	**94.3**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.3**	85.7	**94.3**	**100.0**	**94.3**	**100.0**	51.4	94.8
Intransitive - present	25	100.0	100.0	96.0	96.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.0	100.0	92.0	98.8
Intransitive - preterite	49	**95.9**	**95.9**	**98.0**	**98.0**	**95.9**	91.8	**95.9**	**98.0**	93.9	**95.9**	**95.9**	93.9	**98.0**	91.8	91.8	**100.0**	91.8	95.4
Intransitive - preterite subjunct. II	23	**82.6**	**87.0**	**91.3**	**91.3**	**78.3**	**82.6**	**82.6**	**91.3**	**82.6**	**95.7**	**82.6**	69.6	**82.6**	**82.6**	**91.3**	**95.7**	**87.0**	85.7
Modal - future I	115	**93.0**	**93.9**	84.3	84.3	**95.7**	**94.8**	**93.9**	89.6	**93.9**	**93.0**	**91.3**	**93.9**	**93.0**	**91.3**	79.1	**90.4**	83.5	90.5
Modal - future I subjunct. II	111	**86.5**	**84.7**	82.0	82.0	**91.9**	81.1	82.0	83.8	80.2	**85.6**	**85.6**	83.8	80.2	73.0	**88.3**	68.5	75.7	82.0
Modal - perfect	113	72.6	**89.4**	68.1	68.1	75.2	**85.8**	73.5	53.1	**85.0**	68.1	78.8	**86.7**	57.5	74.3	70.8	72.6	44.2	72.0
Modal - pluperfect	124	**60.5**	38.7	4.8	4.8	41.1	40.3	17.7	11.3	37.1	29.0	41.1	10.5	25.0	11.3	4.0	13.7	9.7	23.6
Modal - pluperf. subjunct. II	137	**61.3**	**59.1**	**58.4**	**58.4**	**59.9**	**60.6**	**54.0**	**56.2**	**60.6**	**55.5**	**60.6**	**54.0**	**59.9**	**60.6**	**58.4**	**59.9**	39.4	57.4
Modal - present	102	**98.0**	92.2	89.2	89.2	**94.1**	**95.1**	81.4	92.2	**95.1**	**95.1**	**96.1**	**96.1**	86.3	93.1	72.5	93.1	88.2	91.0
Modal - preterite	123	97.6	**99.2**	**98.4**	**98.4**	**100.0**	96.7	97.6	96.7	97.6	91.1	**100.0**	**100.0**	89.4	91.9	96.7	95.9	90.2	96.3
Modal - preterite subjunct. II	111	**82.9**	**86.5**	**79.3**	**78.4**	**87.4**	**85.6**	**87.4**	**85.6**	**85.6**	77.5	**87.4**	**84.7**	73.9	**82.0**	**82.9**	**84.7**	76.6	82.8
Modal neg. - future I	97	**94.8**	**95.9**	**95.9**	**95.9**	**92.8**	**92.8**	**95.9**	**90.7**	87.6	**95.9**	**91.8**	**93.8**	85.6	**91.8**	**95.9**	**89.7**	74.2	91.8
Modal neg. - future I subjunct. II	125	**95.2**	**95.2**	91.2	91.2	**96.0**	**94.4**	**95.2**	**93.6**	**95.2**	**95.2**	**95.2**	**96.0**	91.2	**93.6**	**97.6**	92.0	83.2	93.6
Modal neg. - perfect	87	80.5	**86.2**	71.3	71.3	77.0	**89.7**	79.3	64.4	**90.8**	72.4	**86.2**	**88.5**	70.1	79.3	78.2	79.3	48.3	77.2
Modal neg. - pluperfect	102	**66.7**	38.2	3.9	3.9	27.5	8.8	3.9	7.8	6.9	17.6	31.4	18.6	20.6	**1.0**	7.8	17.6	18.6	17.7
Modal neg. - pluperf. subjunct. II	122	**70.5**	64.8	55.7	55.7	66.4	**72.1**	53.3	66.4	**75.4**	50.0	**80.3**	69.7	**71.3**	**71.3**	50.8	**76.2**	47.5	64.6
Modal neg. - present	125	92.0	96.8	91.2	91.2	96.0	93.6	76.0	96.0	95.2	**98.4**	**100.0**	97.6	86.4	94.4	73.6	92.8	96.8	92.2
Modal neg. - preterite	128	**99.2**	**99.2**	96.9	96.9	**100.0**	**98.4**	**100.0**	**98.4**	**100.0**	**100.0**	**100.0**	**100.0**	97.7	**98.4**	**98.4**	**98.4**	89.8	98.3
Modal neg. - preterite subjunct. II	118	**93.2**	**91.5**	66.9	66.9	83.1	85.6	**93.2**	**91.5**	**94.1**	83.1	83.9	75.4	**89.8**	82.2	73.7	**90.7**	83.9	84.0
Progressive	11	**90.9**	**90.9**	72.7	72.7	72.7	**81.8**	**81.8**	**100.0**	**100.0**	**81.8**	**90.9**	**81.8**	**90.9**	**100.0**	**90.9**	**90.9**	**81.8**	86.6
Reflexive - future I	21	**76.2**	**95.2**	**90.5**	**90.5**	**95.2**	**95.2**	**76.2**	**90.5**	**95.2**	**90.5**	**81.0**	**95.2**	**90.5**	**95.2**	**81.0**	**95.2**	66.7	88.2
Reflexive - future I subjunct. II	32	71.9	**87.5**	**93.8**	**93.8**	**96.9**	**96.9**	71.9	**93.8**	**93.8**	**93.8**	68.8	**96.9**	**93.8**	**93.8**	75.0	**96.9**	62.5	87.1
Reflexive - future II	24	83.3	**95.8**	**95.8**	**95.8**	**91.7**	**95.8**	87.5	**91.7**	**95.8**	**95.8**	87.5	**100.0**	87.5	**91.7**	87.5	**100.0**	62.5	90.9
Reflexive - future II subjunct. II	29	79.3	**89.7**	**96.6**	**96.6**	**86.2**	75.9	79.3	**96.6**	**96.6**	**96.6**	72.4	**96.6**	44.8	**96.6**	**82.8**	**82.8**	44.8	83.2
Reflexive - perfect	27	77.8	**100.0**	**92.6**	**92.6**	**96.3**	**96.3**	**92.6**	**92.6**	**96.3**	**96.3**	**92.6**	**100.0**	81.5	**96.3**	81.5	**96.3**	55.6	90.4
Reflexive - pluperfect	28	71.4	**89.3**	**82.1**	**82.1**	**96.4**	**92.9**	78.6	**96.4**	**92.9**	**89.3**	**92.9**	**96.4**	**92.9**	**92.9**	78.6	**96.4**	53.6	86.8
Reflexive - pluperf. subjunct. II	29	72.4	**82.8**	**93.1**	**93.1**	**89.7**	**82.8**	79.3	75.9	**86.2**	**89.7**	72.4	**86.2**	79.3	**96.6**	75.9	79.3	44.8	81.1
Reflexive - present	23	69.6	**91.3**	**95.7**	**95.7**	**95.7**	**95.7**	78.3	**91.3**	**91.3**	**91.3**	**91.3**	**100.0**	87.0	**91.3**	87.0	**95.7**	73.9	89.5
Reflexive - preterite	17	76.5	**94.1**	**94.1**	**94.1**	82.4	**88.2**	82.4	82.4	**88.2**	**94.1**	**88.2**	**100.0**	**88.2**	82.4	82.4	**94.1**	35.3	85.1

phenomenon	count	Onl-W	Faceb	Onl-B	VolcT	Onl-A	SMU	Onl-G	Huawe	borde	Nemo	uedin	Water	P3AI	ICL	Onl-Y	Manif	happy	avg
Reflexive - preterite subjunct. II	15	80.0	**100.0**	**100.0**	**100.0**	**86.7**	73.3	**86.7**	**93.3**	73.3	**93.3**	**93.3**	**93.3**	**86.7**	73.3	80.0	**93.3**	46.7	85.5
Transitive - future I	40	97.5	97.5	97.5	97.5	97.5	97.5	97.5	97.5	97.5	97.5	97.5	97.5	97.5	97.5	97.5	97.5	97.5	97.5
Transitive - future I subjunct. II	35	100.0	100.0	97.1	97.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.1	99.5
Transitive - future II	35	100.0	97.1	94.3	94.3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.1	100.0	100.0	97.1	98.8
Transitive - future II subjunct. II	35	97.1	100.0	97.1	97.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.1	100.0	100.0	100.0	94.3	99.0
Transitive - perfect	40	100.0	100.0	95.0	95.0	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	99.4
Transitive - pluperfect	31	**100.0**	90.3	58.1	58.1	**93.5**	**100.0**	80.6	**100.0**	**100.0**	**100.0**	**100.0**	80.6	**93.5**	**100.0**	58.1	**96.8**	**93.5**	88.4
Transitive - pluperf. subjunct. II	32	**100.0**	**96.9**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**96.9**	**96.9**	**93.8**	**100.0**	**96.9**	**100.0**	**96.9**	50.0	95.8
Transitive - present	36	**97.2**	**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**	91.7	**100.0**	**94.4**	99.0
Transitive - preterite	26	**92.3**	**100.0**	**96.2**	**96.2**	**100.0**	**96.2**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**92.3**	**92.3**	**100.0**	**100.0**	88.5	97.3
Transitive - preterite subjunct. II	23	**82.6**	**82.6**	**82.6**	**82.6**	**82.6**	**78.3**	**73.9**	**82.6**	**87.0**	**87.0**	**73.9**	**82.6**	**73.9**	**78.3**	**91.3**	**87.0**	56.5	80.3
Verb valency	54	88.9	90.7	92.6	92.6	90.7	90.7	87.0	90.7	90.7	90.7	88.9	88.9	88.9	90.7	85.2	90.7	81.5	89.4
Case government	17	94.1	94.1	94.1	94.1	94.1	94.1	94.1	94.1	94.1	94.1	88.2	88.2	94.1	94.1	82.4	94.1	88.2	92.4
Mediopassive voice	13	100.0	100.0	100.0	100.0	100.0	100.0	84.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	92.3	100.0	84.6	97.7
Passive voice	15	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	100.0	100.0	86.7	99.2
Resultative predicates	9	44.4	55.6	66.7	66.7	55.6	55.6	55.6	55.6	55.6	55.6	55.6	55.6	44.4	55.6	55.6	55.6	55.6	55.6
micro-average	3806	**88.3**	**88.2**	82.0	81.9	**87.3**	86.6	82.4	84.3	**87.1**	85.1	**87.4**	85.0	82.8	83.9	79.7	84.0	72.3	84.0
phen. macro-average	3806	90.3	**91.6**	88.4	88.4	**90.5**	90.0	87.0	89.9	90.4	90.3	90.0	88.9	87.6	89.0	85.2	87.1	75.8	88.3
categ. macro-average	3806	**92.7**	**92.1**	91.2	91.2	91.1	90.3	90.3	90.2	90.1	90.0	89.7	89.3	89.2	89.2	88.0	85.7	78.6	89.4

Table 7: Accuracies (%) of successful translations on a phenomenon-level granularity for German-English, organized in categories. Boldface indicates the best scoring system in each row, including all systems which are not significantly inferior than the best scoring system. Grey rows average the accuracies of the phenomena per category.

categ	count	Faceb	VolcA	Onl-W	Onl-A	Huawe	Nemo	Onl-B	VolcG	uedin	P3AI	eTran	happy	nucle	Onl-Y	Manif	BUPT	ICL	Onl-G	avg
Ambiguity	23	**91.3**	**95.7**	**95.7**	**91.3**	**87.0**	**87.0**	**91.3**	**91.3**	**82.6**	**82.6**	**78.3**	73.9	73.9	69.6	**82.6**	69.6	**82.6**	73.9	83.3
Lexical ambiguity	23	**91.3**	**95.7**	**95.7**	**91.3**	**87.0**	**87.0**	**91.3**	**91.3**	**82.6**	**82.6**	**78.3**	73.9	73.9	69.6	**82.6**	69.6	**82.6**	73.9	83.3
Coordination & ellipsis	87	**81.6**	**71.3**	**71.3**	**73.6**	**77.0**	**75.9**	**80.5**	**79.3**	69.0	69.0	66.7	64.4	65.5	**71.3**	63.2	63.2	58.6	**72.4**	70.8
Gapping	16	81.3	68.8	56.3	75.0	87.5	81.3	87.5	87.5	68.8	81.3	68.8	62.5	56.3	81.3	62.5	68.8	56.3	75.0	72.6
Pseudogapping	9	88.9	77.8	77.8	55.6	55.6	55.6	55.6	55.6	66.7	55.6	55.6	66.7	44.4	44.4	55.6	55.6	55.6	66.7	60.5
Right node raising	14	92.9	78.6	92.9	78.6	85.7	85.7	71.4	71.4	92.9	92.9	85.7	92.9	92.9	64.3	92.9	92.9	85.7	85.7	85.3
Sluicing	19	**94.7**	**100.0**	**94.7**	73.7	84.2	84.2	78.9	78.9	78.9	78.9	73.7	78.9	78.9	73.7	84.2	73.7	68.4	68.4	80.4
Stripping	19	**73.7**	36.8	42.1	68.4	63.2	**73.7**	**94.7**	**89.5**	42.1	47.4	42.1	42.1	52.6	**73.7**	36.8	42.1	36.8	68.4	57.0
VP-ellipsis	10	**50.0**	**70.0**	**70.0**	**90.0**	**80.0**	**60.0**	**80.0**	**80.0**	**70.0**	**50.0**	**80.0**	40.0	**60.0**	**80.0**	40.0	40.0	**50.0**	**70.0**	64.4
False friends	38	92.1	92.1	89.5	86.8	86.8	84.2	84.2	84.2	86.8	86.8	86.8	86.8	81.6	86.8	86.8	86.8	84.2	84.2	86.5
Function word	35	**97.1**	**97.1**	**100.0**	**97.1**	**97.1**	**94.3**	**100.0**	**100.0**	**100.0**	**97.1**	**94.3**	**97.1**	**97.1**	**97.1**	65.7	**97.1**	**100.0**	**97.1**	95.9
Focus particle	23	95.7	95.7	100.0	95.7	95.7	91.3	100.0	100.0	100.0	95.7	91.3	95.7	95.7	95.7	100.0	95.7	100.0	95.7	96.6
Question tag	12	**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**	0.0	**100.0**	**100.0**	**100.0**	94.4
MWE	98	**89.8**	**93.9**	**91.8**	85.7	**87.8**	**88.8**	**90.8**	**90.8**	82.7	85.7	84.7	**89.8**	82.7	82.7	83.7	81.6	80.6	81.6	86.4
Collocation	16	**100.0**	**100.0**	**100.0**	**87.5**	**93.8**	**100.0**	**100.0**	**100.0**	**87.5**	**100.0**	**93.8**	**93.8**	**87.5**	**87.5**	**87.5**	**87.5**	81.3	**87.5**	93.1
Compound	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	100.0	100.0	94.1	100.0	100.0	99.7
Idiom	11	**36.4**	**63.6**	**45.5**	0.0	**18.2**	**9.1**	**27.3**	**27.3**	0.0	0.0	**9.1**	**27.3**	0.0	0.0	0.0	0.0	0.0	0.0	14.6

categ	count	Faceb	VolcA	Onl-W	Onl-A	Huawe	Nemo	Onl-B	VolcG	uedin	P3AI	eTran	happy	nucle	Onl-Y	Manif	BUPT	ICL	Onl-G	avg
Nominal MWE	18	**88.9**	**94.4**	**88.9**	**94.4**	**94.4**	**94.4**	**100.0**	**100.0**	83.3	**94.4**	83.3	**100.0**	83.3	**88.9**	**88.9**	83.3	83.3	**88.9**	90.7
Prepositional MWE	15	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	100.0	100.0	100.0	100.0	100.0
Verbal MWE	21	**95.2**	**95.2**	**100.0**	**100.0**	**95.2**	**100.0**	**95.2**	**95.2**	**95.2**	**90.5**	**95.2**	**95.2**	**95.2**	**90.5**	**95.2**	**95.2**	**90.5**	85.7	94.7
Named entity & terminology	82	**93.9**	**97.6**	**93.9**	**93.9**	**93.9**	89.0	**93.9**	**93.9**	**92.7**	89.0	**93.9**	90.2	90.2	**92.7**	89.0	**92.7**	81.7	80.5	91.3
Date	16	**100.0**	**100.0**	**100.0**	**93.8**	**100.0**	**100.0**	**100.0**	**100.0**	**93.8**	**93.8**	**93.8**	**93.8**	**93.8**	**100.0**	**100.0**	**87.5**	81.3	**87.5**	95.5
Domainspecific term	11	**90.9**	**90.9**	72.7	**100.0**	**90.9**	**100.0**	**90.9**	**90.9**	**100.0**	**90.9**	**100.0**	**81.8**	**90.9**	**100.0**	**81.8**	**90.9**	**81.8**	**90.9**	90.9
Location	19	94.7	94.7	94.7	94.7	94.7	94.7	94.7	94.7	94.7	94.7	94.7	89.5	94.7	89.5	94.7	94.7	84.2	89.5	93.3
Measuring unit	17	82.4	**100.0**	**94.1**	**88.2**	**94.1**	70.6	**88.2**	**88.2**	**94.1**	**94.1**	**94.1**	**94.1**	**88.2**	**94.1**	**94.1**	**100.0**	**94.1**	82.4	90.8
Proper name	19	**100.0**	**100.0**	**100.0**	**94.7**	**89.5**	84.2	**94.7**	**94.7**	84.2	73.7	**89.5**	**89.5**	84.2	84.2	73.7	**89.5**	68.4	57.9	86.3
Negation	15	100.0	100.0	100.0	93.3	93.3	100.0	93.3	93.3	100.0	100.0	93.3	93.3	86.7	100.0	100.0	93.3	93.3	93.3	95.9
Non-verbal agreement	68	**100.0**	**98.5**	**97.1**	95.6	95.6	92.6	92.6	92.6	92.6	89.7	91.2	92.6	88.2	89.7	92.6	88.2	88.2	89.7	92.6
Coreference	26	**100.0**	**96.2**	**96.2**	88.5	**92.3**	88.5	88.5	88.5	84.6	80.8	84.6	**92.3**	76.9	88.5	84.6	84.6	80.8	80.8	87.6
Genitive	15	**100.0**	**100.0**	**93.3**	**100.0**	**93.3**	**93.3**	**93.3**	**93.3**	**93.3**	**86.7**	**86.7**	80.0	**86.7**	**86.7**	**93.3**	73.3	80.0	**86.7**	90.0
Possession	27	100.0	100.0	100.0	100.0	100.0	96.3	96.3	96.3	100.0	100.0	100.0	100.0	100.0	92.6	100.0	100.0	100.0	100.0	99.0
Punctuation	37	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	78.4	78.4	91.9	81.1	78.4	81.1	86.5	75.7	78.4	78.4	78.4	70.3	86.5
Quotation marks	37	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	78.4	78.4	91.9	81.1	78.4	81.1	86.5	75.7	78.4	78.4	78.4	70.3	86.5
Subordination	161	**99.4**	**98.1**	**98.1**	**99.4**	95.7	**99.4**	**98.1**	**98.1**	**98.1**	**98.8**	**98.1**	**96.9**	**97.5**	93.8	**96.9**	94.4	92.5	96.3	97.2
Adverbial clause	14	100.0	100.0	100.0	100.0	92.9	100.0	100.0	100.0	92.9	100.0	100.0	92.9	92.9	92.9	100.0	92.9	92.9	85.7	96.4
Cleft sentence	16	**100.0**	**93.8**	**87.5**	**93.8**	**87.5**	**93.8**	**93.8**	**93.8**	**93.8**	**93.8**	**93.8**	**93.8**	**93.8**	**87.5**	**93.8**	**93.8**	81.3	**93.8**	92.4
Contact clause	24	**95.8**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.8**	**95.8**	**100.0**	**100.0**	**95.8**	**91.7**	**95.8**	**95.8**	**91.7**	**91.7**	87.5	**95.8**	96.3
Indirect speech	10	100.0	80.0	90.0	100.0	100.0	100.0	90.0	100.0	100.0	100.0	100.0	100.0	100.0	90.0	90.0	90.0	100.0	90.0	95.6
Infinitive clause	16	100.0	100.0	100.0	100.0	87.5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	93.8	100.0	99.0
Object clause	15	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	93.3	93.3	100.0	93.3	100.0	100.0	98.9
Pseudo-cleft sentence	18	100.0	100.0	100.0	100.0	100.0	100.0	100.0	94.4	100.0	100.0	100.0	100.0	100.0	88.9	100.0	100.0	94.4	94.4	98.5
Relative clause	36	**100.0**	**100.0**	**100.0**	**100.0**	**97.2**	**100.0**	**100.0**	**100.0**	**97.2**	**97.2**	**97.2**	**97.2**	**100.0**	**94.4**	**97.2**	91.7	91.7	**100.0**	97.8
Subject clause	12	100.0	100.0	100.0	100.0	91.7	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	99.5
Verb tense/aspect/mood	2366	98.6	97.9	97.3	96.9	96.1	97.4	**99.0**	**99.1**	**99.2**	97.4	98.4	96.7	97.3	90.7	98.6	94.8	95.2	94.7	97.0
Conditional	15	93.3	86.7	93.3	93.3	93.3	86.7	80.0	80.0	93.3	93.3	93.3	93.3	93.3	93.3	93.3	93.3	86.7	86.7	90.4
Ditransitive - conditional I progr.	57	**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**	**98.2**	93.0	**98.2**	**96.5**	**100.0**	99.2
Ditransitive - conditional I simple	55	**96.4**	90.9	**96.4**	81.8	**100.0**	94.5	**100.0**	**100.0**	**98.2**	**98.2**	**96.4**	**96.4**	**98.2**	**96.4**	**96.4**	89.1	92.7	**96.4**	95.5
Ditransitive - conditional II progr.	14	**100.0**	**100.0**	**100.0**	**100.0**	**85.7**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**85.7**	**85.7**	**100.0**	**100.0**	78.6	**100.0**	96.4
Ditransitive - conditional II simple	15	100.0	100.0	93.3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	86.7	100.0	100.0	100.0	93.3	100.0	98.5
Ditransitive - future I progr.	39	97.4	100.0	100.0	94.9	100.0	97.4	97.4	97.4	97.4	100.0	97.4	100.0	100.0	100.0	100.0	100.0	100.0	94.9	98.6
Ditransitive - future I simple	67	88.1	**100.0**	95.5	95.5	**100.0**	**97.0**	**100.0**	**100.0**	**100.0**	**100.0**	95.5	**100.0**	**98.5**	91.0	**97.0**	94.0	**97.0**	94.0	96.8
Ditransitive - future II progr.	54	**96.3**	**98.1**	**96.3**	94.4	**98.1**	**96.3**	**98.1**	**98.1**	**98.1**	88.9	**100.0**	70.4	**98.1**	33.3	92.6	88.9	66.7	88.9	89.0
Ditransitive - future II simple	44	88.6	**100.0**	**100.0**	90.9	**100.0**	90.9	90.9	90.9	**97.7**	**100.0**	**100.0**	**100.0**	**95.5**	65.9	**100.0**	77.3	93.2	**95.5**	93.2
Ditransitive - past perf. progr.	47	**95.7**	**97.9**	93.6	83.0	**100.0**	87.2	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	74.5	**100.0**	87.2	**95.7**	78.7	94.1
Ditransitive - past perf. simple	49	**98.0**	**98.0**	**100.0**	**95.9**	**100.0**	91.8	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.9**	**95.9**	**98.0**	93.9	81.6	97.2
Ditransitive - past progr.	30	**93.3**	76.7	90.0	**100.0**	**100.0**	**93.3**	**96.7**	**96.7**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**93.3**	**100.0**	**100.0**	96.7
Ditransitive - present perf. progr.	38	**100.0**	**100.0**	**100.0**	89.5	**100.0**	**100.0**	**97.4**	**97.4**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**97.4**	**94.7**	**100.0**	**94.7**	**100.0**	98.4
Ditransitive - present perf. simple	44	**100.0**	90.9	**95.5**	93.2	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**97.7**	93.2	93.2	**100.0**	**95.5**	**100.0**	97.7
Ditransitive - present progr.	38	**100.0**	**100.0**	**97.4**	92.1	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**97.4**	**97.4**	**100.0**	**97.4**	**94.7**	**100.0**	**97.4**	**94.7**	98.2
Ditransitive - simple past	53	100.0	98.1	98.1	96.2	100.0	98.1	98.1	98.1	100.0	100.0	100.0	98.1	100.0	100.0	100.0	100.0	98.1	100.0	99.1

categ	count	Faceb	VolcA	Onl-W	Onl-A	Huawe	Nemo	Onl-B	VolcG	uedin	P3AI	eTran	happy	nucle	Onl-Y	Manif	BUPT	ICL	Onl-G	avg
Ditransitive - simple present	36	100.0	100.0	100.0	97.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.2	100.0	100.0	100.0	100.0	99.7
Gerund	21	**100.0**	**95.2**	**95.2**	**100.0**	**100.0**	**95.2**	**95.2**	**100.0**	**95.2**	**100.0**	**95.2**	**100.0**	85.7	**90.5**	**100.0**	81.0	**95.2**	**95.2**	95.5
Imperative	9	100.0	100.0	100.0	88.9	100.0	100.0	100.0	100.0	100.0	100.0	88.9	100.0	88.9	100.0	88.9	77.8	88.9	88.9	95.1
Intransitive - conditional I progr.	24	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.8**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	87.5	**100.0**	99.1
Intransitive - conditional I simple	28	100.0	100.0	100.0	96.4	100.0	100.0	92.9	92.9	100.0	100.0	92.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.6
Intransitive - future I progr.	27	**100.0**	**100.0**	**96.3**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	59.3	**100.0**	77.8	**100.0**	**96.3**	**100.0**	74.1	40.7	22.2	87.0
Intransitive - future I simple	46	100.0	95.7	97.8	100.0	100.0	97.8	97.8	97.8	100.0	100.0	100.0	100.0	97.8	100.0	100.0	97.8	100.0	97.8	98.9
Intransitive - future II progr.	10	**100.0**	**100.0**	**100.0**	**100.0**	**90.0**	**100.0**	**100.0**	**100.0**	**90.0**	20.0	**100.0**	50.0	**80.0**	30.0	**90.0**	0.0	20.0	50.0	73.3
Intransitive - future II simple	24	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	87.5	**100.0**	**100.0**	**95.8**	**100.0**	70.8	**100.0**	**91.7**	**100.0**	**100.0**	97.0
Intransitive - past perf. progr.	18	**94.4**	**88.9**	**94.4**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.4**	**100.0**	**100.0**	**100.0**	72.2	**100.0**	**88.9**	**100.0**	72.2	94.8
Intransitive - past perf. simple	30	**100.0**	**100.0**	**96.7**	**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**	83.3	98.9
Intransitive - past progr.	11	100.0	100.0	100.0	100.0	81.8	100.0	90.9	90.9	100.0	100.0	100.0	90.9	100.0	81.8	100.0	100.0	100.0	100.0	96.5
Intransitive - present perf. simple	25	100.0	100.0	92.0	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	100.0	100.0	99.6
Intransitive - present progr.	50	**98.0**	**98.0**	**98.0**	**100.0**	**100.0**	**98.0**	**100.0**	**100.0**	**100.0**	**96.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**98.0**	94.0	98.9
Intransitive - simple past	30	**100.0**	**100.0**	**100.0**	86.7	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**96.7**	**100.0**	**100.0**	**100.0**	99.1
Intransitive - simple present	24	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	95.8	100.0	95.8	100.0	100.0	100.0	99.5
Modal	226	**100.0**	**99.6**	**99.6**	**100.0**	**100.0**	**100.0**	**99.6**	**99.6**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	96.9	**100.0**	**100.0**	**99.6**	**100.0**	99.7
Modal negated	213	**99.5**	**99.5**	98.6	98.6	93.9	**99.1**	**100.0**	**100.0**	**99.5**	**99.1**	**99.5**	**99.5**	**99.1**	**99.5**	**100.0**	**99.1**	**99.1**	95.8	98.9
Reflexive - conditional I progr.	25	**96.0**	**100.0**	**100.0**	**100.0**	84.0	**96.0**	**100.0**	**100.0**	**100.0**	**100.0**	**96.0**	**100.0**	**96.0**	88.0	**100.0**	**100.0**	**100.0**	**100.0**	97.6
Reflexive - conditional I simple	22	**95.5**	**100.0**	**95.5**	86.4	77.3	**90.9**	**100.0**	**100.0**	**95.5**	**100.0**	86.4	**95.5**	**95.5**	86.4	**100.0**	**95.5**	**100.0**	**100.0**	94.4
Reflexive - conditional II progr.	6	100.0	100.0	83.3	100.0	66.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	83.3	83.3	100.0	100.0	100.0	100.0	95.4
Reflexive - conditional II simple	23	**100.0**	**100.0**	**95.7**	**100.0**	73.9	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.7**	78.3	**100.0**	**100.0**	**100.0**	87.0	96.1
Reflexive - future I progr.	20	**100.0**	**100.0**	**100.0**	**100.0**	85.0	**95.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**90.0**	**90.0**	**90.0**	**100.0**	**90.0**	**95.0**	**100.0**	96.4
Reflexive - future I simple	40	**100.0**	**97.5**	**100.0**	**100.0**	77.5	**95.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	87.5	**100.0**	**100.0**	**100.0**	**100.0**	97.6
Reflexive - future II progr.	24	**100.0**	**95.8**	**100.0**	**100.0**	87.5	**100.0**	**100.0**	**100.0**	**100.0**	75.0	**95.8**	62.5	**95.8**	54.2	**95.8**	66.7	70.8	83.3	88.0
Reflexive - future II simple	30	**100.0**	**100.0**	90.0	**100.0**	83.3	**100.0**	**96.7**	**96.7**	**100.0**	**100.0**	**100.0**	**100.0**	90.0	63.3	**100.0**	**96.7**	**100.0**	**100.0**	95.4
Reflexive - past perf. progr.	28	**100.0**	**100.0**	**100.0**	**92.9**	82.1	85.7	**100.0**	**100.0**	**100.0**	**100.0**	89.3	**96.4**	**96.4**	71.4	**100.0**	85.7	**100.0**	**100.0**	94.4
Reflexive - past perf. simple	31	**100.0**	**100.0**	**100.0**	**100.0**	87.1	**96.8**	**100.0**	**100.0**	**100.0**	**100.0**	87.1	**100.0**	**100.0**	90.3	**100.0**	**96.8**	**100.0**	**96.8**	97.5
Reflexive - past progr.	4	100.0	50.0	50.0	100.0	75.0	50.0	75.0	75.0	100.0	100.0	75.0	50.0	75.0	50.0	100.0	75.0	100.0	50.0	75.0
Reflexive - present perf. progr.	23	**100.0**	**100.0**	**100.0**	**100.0**	82.6	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.7**	**100.0**	**100.0**	**100.0**	**100.0**	98.8
Reflexive - present perf. simple	30	**100.0**	**100.0**	**100.0**	**100.0**	90.0	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**93.3**	**100.0**	**100.0**	**100.0**	**100.0**	99.1
Reflexive - present progr.	25	**96.0**	**100.0**	**100.0**	**96.0**	84.0	**96.0**	**100.0**	**100.0**	**96.0**	**96.0**	88.0	**92.0**	**92.0**	**92.0**	**96.0**	**96.0**	**92.0**	**92.0**	94.7
Reflexive - simple past	32	**100.0**	**100.0**	**96.9**	**100.0**	84.4	**100.0**	**96.9**	**96.9**	**100.0**	**100.0**	90.6	**96.9**	87.5	84.4	**100.0**	90.6	**96.9**	**93.8**	95.3
Reflexive - simple present	25	**96.0**	**100.0**	**100.0**	**96.0**	80.0	**96.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**96.0**	**96.0**	88.0	**100.0**	**100.0**	**96.0**	**100.0**	96.9
Transitive - future II progr.	27	**100.0**	**96.3**	**96.3**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	44.4	**100.0**	40.7	**100.0**	14.8	**92.6**	51.9	51.9	77.8	81.5
Transitive - conditional I progr.	26	**100.0**	84.6	88.5	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	98.5
Transitive - conditional I simple	30	**100.0**	**100.0**	90.0	83.3	**100.0**	**100.0**	**100.0**	**100.0**	86.7	**96.7**	**93.3**	**93.3**	**93.3**	**100.0**	**100.0**	86.7	**100.0**	**93.3**	95.4
Transitive - conditional II progr.	28	**100.0**	89.3	**92.9**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**96.4**	**100.0**	**96.4**	**100.0**	98.6
Transitive - conditional II simple	29	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.6	100.0	100.0	100.0	96.6	100.0	99.6
Transitive - future I progr.	21	**100.0**	85.7	81.0	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.2**	**100.0**	**100.0**	**95.2**	**100.0**	**100.0**	**95.2**	**95.2**	**100.0**	97.1
Transitive - future I simple	41	**100.0**	**100.0**	**100.0**	**97.6**	**100.0**	85.4	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	90.2	**95.1**	**100.0**	92.7	**100.0**	**97.6**	97.7
Transitive - future II simple	35	**100.0**	**97.1**	**97.1**	**100.0**	**100.0**	**100.0**	**97.1**	**97.1**	**100.0**	**100.0**	**100.0**	**100.0**	91.4	82.9	**100.0**	88.6	**94.3**	**100.0**	97.0
Transitive - past perf. progr.	27	**100.0**	**92.6**	74.1	**100.0**	**100.0**	**92.6**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**96.3**	**96.3**	77.8	**100.0**	**96.3**	**100.0**	**92.6**	95.5

categ	count	Faceb	VolcA	Onl-W	Onl-A	Huawe	Nemo	Onl-B	VolcG	uedin	P3AI	eTran	happy	nucle	Onl-Y	Manif	BUPT	ICL	Onl-G	avg
Transitive - past perf. simple	30	100.0	100.0	96.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.7	100.0	96.7	99.4
Transitive - past progr.	4	25.0	100.0	100.0	25.0	25.0	25.0	100.0	100.0	25.0	25.0	50.0	25.0	25.0	100.0	25.0	25.0	25.0	75.0	50.0
Transitive - present perf. progr.	21	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	95.2	95.2	95.2	95.2	100.0	100.0	95.2	100.0	100.0	98.7
Transitive - present perf. simple	31	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.8	100.0	100.0	100.0	100.0	100.0	99.8
Transitive - present progr.	37	**100.0**	**97.3**	**97.3**	**100.0**	**100.0**	91.9	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.6**	**97.3**	**100.0**	**100.0**	**97.3**	**94.6**	98.3
Transitive - simple past	40	**100.0**	**100.0**	**100.0**	92.5	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.0**	**100.0**	**100.0**	**97.5**	**100.0**	92.5	98.8
Transitive - simple present	40	**100.0**	**100.0**	**97.5**	**100.0**	**97.5**	**100.0**	**97.5**	**97.5**	**100.0**	**100.0**	**100.0**	**100.0**	**95.0**	**100.0**	**100.0**	**97.5**	**100.0**	90.0	98.5
Verb valency	96	**90.6**	81.3	**85.4**	81.3	**84.4**	81.3	**83.3**	**83.3**	81.3	**83.3**	**84.4**	80.2	80.2	77.1	81.3	77.1	75.0	74.0	81.4
Case government	20	90.0	90.0	85.0	90.0	95.0	90.0	95.0	95.0	95.0	95.0	95.0	95.0	95.0	85.0	95.0	90.0	90.0	90.0	91.9
Catenative verb	20	**100.0**	**90.0**	**95.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.0**	**95.0**	**100.0**	**100.0**	**90.0**	**90.0**	85.0	96.7
Middle voice	19	68.4	63.2	63.2	52.6	42.1	36.8	47.4	47.4	47.4	42.1	52.6	42.1	36.8	42.1	42.1	36.8	31.6	31.6	45.9
Passive voice	17	100.0	94.1	88.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	88.2	100.0	100.0	100.0	94.1	98.0
Resultative	20	**95.0**	70.0	**95.0**	65.0	**85.0**	**80.0**	**75.0**	**75.0**	65.0	**80.0**	**75.0**	70.0	**75.0**	70.0	70.0	70.0	65.0	70.0	75.0
micro-average	3106	**97.4**	96.5	95.9	95.3	94.7	95.6	**96.9**	**96.9**	96.5	95.1	95.8	94.4	94.5	89.4	95.2	92.3	92.1	92.0	94.8
phen. macro-average	3106	**95.7**	94.6	93.9	93.3	91.7	93.0	**95.1**	**95.1**	93.8	91.8	93.1	90.8	90.8	86.8	91.7	88.3	88.2	89.1	92.1
categ. macro-average	3106	**94.5**	**93.6**	93.3	91.2	91.2	90.8	90.5	90.4	89.7	88.4	87.4	86.9	85.6	85.6	84.9	84.8	84.2	84.0	88.7

Table 8: Accuracies (%) of successful translations on a phenomenon-level granularity for English-German, organized in categories. Boldface indicates the best scoring system in each row, including all systems which are not significantly inferior than the best scoring system. Grey rows average the accuracies of the phenomena per category.