Linguistically Motivated Evaluation of Machine Translation Metrics based on a Challenge Set

Eleftherios Avramidis and Vivien Macketanz
German Research Center for Artificial Intelligence (DFKI),
Speech and Language Technology, Berlin, Germany
firstname.lastname@dfki.de

Abstract

We employ a linguistically motivated challenge set in order to evaluate the state-of-the-art machine translation metrics submitted to the Metrics Shared Task of the 7th Conference for Machine Translation. The challenge set includes about 20,000 items extracted from 145 MT systems for two language directions (German⇔English), covering more than 100 linguistically-motivated phenomena organized in 14 categories. The best performing metrics are YiSi-1, BERTScore and COMET-22 for German-English, and UniTE, UniTE-ref, MetricX-XL-DA19 and MetricX-XXL-DA19 for English-German. Metrics in both directions are performing worst when it comes to named-entities & terminology and particularly measuring units. Particularly in German-English they are weak at detecting issues at punctuation, polar questions, relative clauses, dates and idioms. In English-German, they perform worst at present progressive of transitive verbs, future II progressive of intransitive verbs, simple present perfect of ditransitive verbs and focus particles.

1 Introduction

Automatic evaluation metrics have been valuable tools for Machine Translation (MT), allowing quick evaluation and suggesting directions for further development. Many metrics have been suggested throughout the years, which in turn sets the requirement for their evaluation.

Whereas MT metrics so far have been evaluated based on the agreement of their scores with human judgments on test sets drawn from broad text, little research has taken place on investigating whether the performance of the metrics generalizes enough when evaluating particular cases. A more target way of evaluating metrics is using *challenge sets*. These are targeted test sets, which have been devised in such a way, so that they benchmark the ability of metrics to score particular translation phenomena.

In this paper we present empirical results on the performance of MT metrics, using an extensive challenge set, which includes thousands of test items aiming to test the performance over more than one hundred linguistically-motivated phenomena in two language directions. It is based on thousands of manually created test items, their translation outputs from dozens of MT systems and semi-automatically evaluated with the supervision of linguists. Through this analysis we attempt to reveal strengths and weaknesses of several state-of-the-art MT metrics considering their background methods with regards to linguistic aspects.

The rest of the paper is structured as follows. In Section 2 related work is briefly described. In Section 3 we describe the construction of the challenge set and the evaluation protocol. The empirical results are outlined in Section 4, followed by a conclusion is Section 5.

2 Related work

The need for a thorough evaluation of Natural Language Processing (NLP) tools has lately received increased interest in the research community, indicated by a big amount of publications, among them several which received best paper awards (Ribeiro et al., 2020; Avelino et al., 2022; Campolungo et al., 2022). When focusing on MT, first efforts were made in the 1990s with the introduction of test suites (King and Falkedal, 1990), which were revived after the latest advances in the field (Guillou and Hardmeier, 2016). To the best of our knowledge, the first efforts relevant to the application of challenge sets on MT metrics was presented as an analysis at the Findings paper of the Metrics shared task of the 6th Conference of Machine Translation (Freitag et al., 2021), based on our test suite (Macketanz et al., 2022) that we are using on this paper.

Hereby we are advancing as to that preliminary analysis by (a) increasing the number of challenge

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 514–529
Abu Dhabi, December 7–8, 2022.

items to about 9,000-10,000, including outputs from state-of-the-art systems from 2021, (b) adding a second language direction (English-German) (c) presenting a more fine-grained analysis, not only in the category level but also on the phenomenon level. This way we can get more confident and more generalisable empirical conclusions.

3 Method

3.1 Test suite for MT systems

The challenge set is based on our test suite (Macketanz et al., 2022), a manually devised test suite for MT for German-English and its recently developed extension for English-German (Macketanz et al., 2021).[1] The German-English side consists of 5,540 German test sentences covering 107 linguistically motivated phenomena, organized in 14 categories. The English-German side consists of 4,438 English test sentences covering 105 phenomena, organized in 12 categories.

The chosen phenomena do not follow a particular linguistic theory but their definition has been inspired by observing linguistic aspects which are relevant for MT. Each phenomenon is represented by at least 20 source test sentences to guarantee a balanced test set. The test suite is used to evaluate MT systems with regard to their performance on the phenonmenon-targeting test sentences. The evaluation operates semi-automatically and it occurs based on a set of handwritten rules which contain regular expressions and fixed string tokens.

The above described test suite has been used to evaluate the outputs of 116 German-English and 29 English-German systems, submitted at the translation task of the Conference of Machine Translation (WMT) for four consequent years (2018-2021; Macketanz et al., 2018; Avramidis et al., 2019, 2020; Macketanz et al., 2021), including a preliminary system comparison in 2017 (Burchardt et al., 2017).

3.2 Challenge set for MT metrics

Here we describe how the aforementioned test suite, along with inputs from previous shared tasks, is used in order to evaluate MT metrics. A challenge set for metrics requires contrastive pairs of correct/incorrect translations and a reference, whereas our original test suite contained only source sentences and handwritten rules for the outputs, but no reference translations. We therefore use the collected MT outputs to construct the challenge items for the metrics task in order to create the required challenge sets as following. For every source sentence of the test suite we create a tuple including:

- one correct translation, to be given to the metrics as reference translation; and a pair of
- another correct translation and
- one incorrect translation, the latter two intended to be given to the metrics for scoring.

In order to generate these tuples we perform random combinations of correct and wrong translations from the WMT outputs. Also, before collecting MT outputs, we filter out a part of the original test items, to be reserved for future evaluations.

The above process resulted into a metrics challenge set with 10,402 items for German-English and 8,945 items for English-German. The fact that the correct and incorrect translations have been sampled from real MT system outputs of the last 4 years, implies that these challenge set is closer to the real MT system ecosystem, as compared to artificially created challenge sets, which may contain translations that would never be produced by state-of-the-art MT.

3.3 Evaluation of metrics

As explained, the challenge set consists of subsets of challenge items, where every subset has been deliberately created so that it can detect the metrics' performance to a particular phenomenon. For every challenge item, the two MT outputs (correct/incorrect) are given unlabelled to the metrics as two separate MT hypotheses so that they score them against the aforementioned references and/or the source. The item is considered correctly scored, if the metric gives to the correct MT output a higher score than the incorrect MT output. Then the following statistics are calculated:

Accuracy per phenomenon is given by the ratio of all correctly-scored challenge items per phenomenon to the total number of challenge items for this phenomenon

Accuracy per category is given by the ratio of all correctly-scored challenge items per category to the total number of challenge items for this category (after aggregating the underlying phenomena of this category in one set).

Significant tests for comparisons: the highest metric accuracy for every phenomenon is compared to all other metric accuracies of the same

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

phenomenon. For this, a one-tailed Z-test with $\alpha = 0.95$ is calculated. The metrics whose accuracies that are not significantly worse than the highest accuracy, are considered to share the winning position for this phenomenon. The best accuracies per category are calculated in the same way, after aggregating the challenge items from the underlying phenomena of every category.

Statistics for metric categories: We repeat this significance testing in two levels: one for all metrics participating in the shared task, and then separately for each one of the three metric categories (baseline, QE as a metric, reference-based). The significantly best systems per phenomenon over all metrics are indicated with a gray background, whereas the significantly best systems per metrics category are indicated with boldface.

Finally, we report three kinds of average scores:
Micro-average treats all items equally, aggregating all test items to compute the average percentages;
Category macro-average treats all categories equally by computing the percentages independently for each category and then averaging them
Phenomenon macro-average treats all phenomena equally, by computing the percentages independently for each phenomenon and then averaging them

4 Results

The results are displayed in detail in Tables 1 and 3 in the category level and in Tables 4 and 5 for the phenomenon level, for both language directions German-English and English-German respectively.

4.1 Metric performance analysis

Here we are observing the statistics with a focus on comparing the performance of various metrics on the challenge set.

German-English The best performing metrics for German-English are YiSi-1 (Lo, 2019), BERTScore (Zhang et al., 2020) and COMET-22 (Rei et al., 2022), achieving the significantly highest micro- and macro-average accuracies (84-85%), whereas for the macro-average, UniTE-ref (Wan et al., 2022) is also included in the first significance cluster. The two QE based metrics of HWTSC (Liu et al., 2022) get the lowest accuracies, together with the baseline BLEU (Papineni et al., 2002).

When considering the systems performance with regards to particular categories, one can see that different metrics win in different combinations of categories. Most reference-based metrics perform best for at least four categories, apart from MS-COMET which only gets two.

Interestingly enough, one QE method is outperforming reference-based metrics for one category: HWTSC-TLM is the best performing system for *punctuation*. Additionally, UNITE-src performs equally well to reference-based metrics for coordination and ellipsis.

English-German UniTE and UniTE-ref are the winning metrics based on the macro-average (82%), whereas the former seems to be stronger than the latter, winning 5 categories. MetricX-XL-DA19 and MetricX-xxl-DA19 are the winning metrics when it comes to micro-average (78%). Their average accuracies are close to 80%, which raises concerns, as this indicates that 2 out of 10 challenge items in average are not scored correctly in this language direction, even for the best performing metrics. The lowest scoring metric is MATESE (Perrella et al., 2022) in both QE and reference-based versions, very close to REUSE (Mukherjee and Shrivastava, 2022).

Also in this direction, QE methods manage to outperform submitted reference-based metrics in a few categories. REUSE is the best performing metric for *false friends* and UNITE-src for *function words*. COMET-kiwi (Rei et al., 2022) and UniTE-src are on par with reference-aware metrics when it comes to *subordination* and Cross-QE (Liu et al., 2022) for *verb tense/aspect/mood*.

4.2 Linguistically motivated analysis

Here we are looking closer to the results for particular phenomena or categories.

4.2.1 German-English

Category-level The overall average accuracy of all metrics with regards to the linguistically motivated categories is at 78% for German-English. This indicates that the metrics failed in average to predict properly the scores for about one out of four challenge items that we provided. Even for the best categories, the accuracy achieved by most metrics is considerably below the acceptable limit of 90%.

The best performing category in *negation* with 86% average accuracy. For the rest of the categories, the average accuracy is less than 82%. The worst performing categories in average are *named entity and terminology* and *punctuation* with only 67% accuracy, whereas *subordination* comes next

		baselines								QE as a metric							ref. based metrics								
ling. category	#	BERTScore	BLEU	BLEURT-20	COMET-20	YiSi-1	chrF	f101spBLEU	f200spBLEU	COMETKiwi	Cross-QE	HWTSC-TLM	HWTSC-TS	KG-BERT	MS-COMET-QE	UniTE-src	COMET-22	MS-COMET	UniTE-ref	UniTE	XL-DA19	XL-MQM20	XXL-DA19	XXL-MQM20	avg
Ambiguity	298	**90**	71	88	86	89	80	81	79	**82**	73	60	65	67	**82**	80	87	85	**89**	89	88	**90**	83	86	81
Composition	252	88	65	87	85	**90**	74	70	71	76	**77**	73	76	59	72	75	83	86	82	83	86	82	**87**	82	79
Coordination & ellipsis	316	**79**	74	**79**	77	**80**	77	72	73	**82**	78	69	72	78	69	**83**	**84**	75	79	80	79	**83**	78	78	77
False friends	90	91	64	**93**	82	**92**	78	69	70	88	74	81	**91**	87	63	44	**91**	88	**92**	**92**	90	90	87	88	82
Function word	586	**83**	72	**83**	78	81	73	73	73	**81**	77	78	**81**	70	68	77	83	81	**86**	84	84	79	83	82	79
LDD & interrogatives	1014	**85**	75	**84**	**85**	**85**	76	74	74	**84**	**83**	72	75	63	81	**82**	**86**	83	84	85	**85**	82	**85**	82	80
MWE	610	**85**	73	**85**	**85**	**86**	78	74	75	**76**	**76**	70	60	56	60	73	86	82	**89**	**90**	88	88	87	81	78
Named entity & termin.	861	74	62	68	68	**76**	67	70	71	65	**71**	64	61	55	61	61	70	66	67	64	67	**75**	70	72	67
Negation	76	**95**	84	88	92	91	88	83	80	**93**	78	62	74	87	88	92	91	88	**93**	**93**	89	78	88	83	86
Non-verbal agreement	419	77	74	**83**	81	76	75	75	76	75	72	66	63	62	**78**	73	**84**	77	84	**85**	83	81	**85**	83	77
Punctuation	293	74	77	70	68	73	69	78	**80**	55	75	**81**	73	62	61	69	**68**	65	65	61	61	53	59	47	67
Subordination	679	76	69	**77**	**77**	74	69	68	69	72	**75**	59	62	65	64	73	**80**	77	77	78	75	70	78	74	72
Verb tense/aspect/mood	4697	**88**	69	85	86	**89**	77	71	71	81	**87**	63	71	78	81	82	**86**	83	**85**	**85**	84	79	**85**	81	80
Verb valency	211	**91**	70	88	88	**90**	72	69	69	**86**	72	64	64	62	75	82	**94**	88	91	91	91	88	91	88	81
macro avg.	10402	**84**	71	83	81	**84**	75	73	74	**78**	76	69	70	68	72	75	**84**	80	**83**	**83**	82	80	82	79	78
micro avg.	10402	**84**	70	82	82	**85**	75	72	72	78	**81**	66	70	70	75	78	**84**	80	**83**	82	82	79	82	79	78

Table 1: Accuracy of the metrics (%) with regards to the 14 linguistically motivated categories for German-English. The significantly best systems per phenomenon over all metrics are indicated with a gray background, whereas the significantly best systems per metrics category are indicated with boldface.

with 72%. The lowest performing score for all systems and all categories is achieved by MetricX-XL-MQM20, which can only score correctly almost half of the punctuation challenge items (53%).

Phenomenon-level The best accuracy for this language pair is achieved for *Transitive, future I* where the metrics get an accuracy of 95%-100%. Another 13 phenomena score more than 85%. Four of them also refer to the future tenses of the transitive, in particular future I and future II in both the plain and their subjunctive form. Additionally, one can see good performance in *Intransitive-present, Modal-future I, pied-piping, comma, negation, passive voice,* and the *negated modal for future I subjunctive II*.

The lowest accuracy of all metrics in average is given for *polar questions* (61%), followed by *quotation marks* (63%). An average accuracy of less than 65% is given for some more phenomena, such as the ones including *measuring units, relative clauses, dates* and *idioms*.

The lowest phenomenon accuracies are given by QE methods, and particularly when it comes to *idioms*, where HWTSC-TLM achieves the lowest performance of 17%. This is explainable by the fact that idioms require resolving rather rare semantic relations between the source and the MT output (used for QE), but can be easily resolved with lexical matching on the reference (used by reference-aware metrics). Idioms have shown to be a particular challenge for MT systems as well.

4.2.2 English-German

Category-level The overall average accuracy of all metrics (Table 3) with regards to the linguistically motivated categories is at 69-72% for English-German. This is 6% lower than the respective average accuracy for German-English, indicating that the metrics for this MT language direction perform worse.

The category where all metrics perform best in average is *negation* (86%), whereas the one where they perform worse is *Named entity & terminology* (59%). The rest of the categories lie in rather mediocre accuracies, between 66% and 82%. The performance of metrics in English-German is worse than German-English in all categories apart from *function words, punctuation* and *subordination*, although the comparisons between the language directions have to be taken with a grain of salt, due to the fact that the two directions consist of different items.

Phenomenon-level The English-German phenomena, where metrics perform best in average are

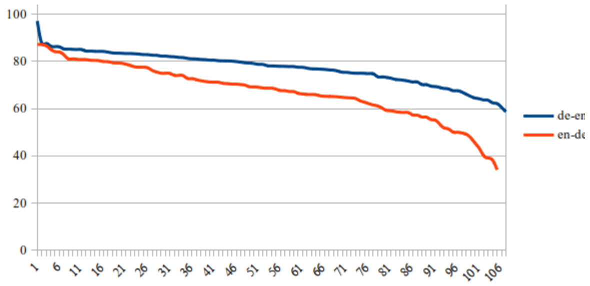

Figure 1: Plot of the accuracy of all phenomena per language direction. The accuracy percentage is shown on the vertical axis and the phenomena on the horizontal

the *Contact clause, Negation, Ditransitive - present progressive* and *question tags*, achieving more than 85% of accuracy. The most difficult phenomena to score are the *Intransitive - future II progressive* and the *Transitive - present progressive*, as they achieve less then 40% average accuracy, followed by *Ditransitive - present perfect simple*, *measuring units* and *focus particles*.

Interestingly enough, in this language direction there are metrics which scored zero accuracies in several phenomena, something that we didn't see in the opposite language direction.[2] These zero accuracies are mostly relevant to rare verb-related phenomena (e.g. intransitive constructions). A comparative plot of the accuracies for all phenomena for both language directions can be seen in Figure 1. It is very clear that English-German lacks considerably, with its lowest scored phenomena having an accuracy at half of the lower-scored phenomena of the opposite direction.

Finally, some examples of incorrectly scored challenge items from the phenomena that have the lowest accuracies can be seen in Table 2. Whereas is hard to know why each metric score in a wrong way, in many cases we may assume that it was misled by a part of the sentence which seemed distant to reference (or the source for QE), but it was correct.

5 Conclusion

In this paper we analyzed the performance of several state-of-the-art metrics with regards to particular linguistically-motivated phenomena for two language pairs, German-English and English-German. The analysis gave a multitude of observations, regarding both the performance of the metrics and the corresponding linguistic observations.

In an effort to draw conclusions after averaging accuracies, we conclude that the best performing metrics are YiSi-1, BERTScore and COMET-22 for German-English, and UniTE, UniTE-ref, MetricX-XL-DA19 and MetricX-xxl-DA19 for English-German.

The metrics are particularly good at scoring the German-English verb tense *Transitive, future I* and the category of *negation*. Concerning English-German, the best performing phenomena are *contact clause* and *negation*.

On the contrary, metrics in both directions are performing worst when it comes to *named-entities & terminology*. Particularly in German-English they are weak at detecting issues at *punctuation (quotation marks), polar questions, measuring units, relative clauses, dates* and *idioms*. In English-German at *present progressive of transitive verbs*, *future II progressive of intransitive verbs*, *present perfect of ditransitive verbs, measuring units* and *focus particles*.

We believe that further investigation on particular phenomena or categories can provide explanations for the relevant observations and possibly lead to suggestions for technical improvements in the development of the metrics in the future. For example, many observations are also relevant to whether the metrics take into account for scoring the reference translation or the source (QE as a metric). Additionally, having seen several low accuracies regarding punctuation, we note that this issue is often handled via pre-processing scripts. The low percentages of scoring punctuation issues, show that the metrics should improve their engineering on that direction.

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 (grant num. d01IW20002). We would like to thank Hans Uszkoreit, Aljoscha Burchardt, Ursula Strohriegel, Renlong Ai and He Wang for their prior contributions for the creation of the test suite.

References

Mariana Avelino, Vivien Macketanz, Eleftherios Avramidis, and Sebastian Möller. 2022. A Test Suite

[2] again this should take into consideration that English-German set has a participation of less systems and therefore less diversity than German-English

for the Evaluation of Portuguese-English Machine Translation. In *Computational Processing of the Portuguese Language*, pages 15–25, Cham. Springer International Publishing.

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

Eleftherios Avramidis, 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. Conference on Machine Translation (WMT-2019)*, pages 644–653, Florence, Italy. 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.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352, Dublin, Ireland. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej 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. 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).

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.

Yilun Liu, Xiaosong Qiao, Zhanglin Wu, Su Chang, Min Zhang, Yanqing Zhao, shimin tao Song Peng, Hao Yang, Ying Qin, Jiaxin Guo, Minghan Wang, Yinglu Li, Peng Li, and Xiaofeng Zhao. 2022. Partial Could Be Better Than Whole: HW-TSC 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. 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 (WMT18)*, pages 578–587, Brussels, Belgium. 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 Thirteenth Language Resources and Evaluation Conference*, pages 936–947, Marseille, France. European Language Resources Association.

Vivien Macketanz, Eleftherios Avramidis, Shushen Manakhimova, and Sebastian Möller. 2021. Linguistic evaluation for the 2021 state-of-the-art Machine Translation systems for German to English and English to German. In *Proceedings of the Sixth Conference on Machine Translation. (WMT21)*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ananya Mukherjee and Manish Shrivastava. 2022. REUSE: REference-free UnSupervised quality Estimation Metric. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. Machine Translation Evaluation as a Sequence Tagging Problem. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Yu Wan, Keqin Bao, Dayiheng Liu, Baosong Yang, Derek F. Wong, Lidia S. Chao, Wenqiang Lei, and Jun Xie. 2022. Alibaba-Translate China's Submission for WMT2022 Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Appendix

		German-English
idiom	src	Ich glaube, Tim hat ein Auge auf Lena geworfen.
	ref	I think Tim has a crush on Lena.
	✓	I think Tim has cast an eye on Lena.
	✗	I think Tim has an eye on Lena.
polar question	src	Willst du mit mir ins Kino gehen?
	ref	Do you want to go to a movie with me?
	✓	Do you want to go with me into the cinema?
	✗	You want to go to the cinema with me?
measuring unit	src	Ein ausgewachsener Afrikanischer Elefant wiegt etwa sechs Tonnen.
	ref	An adult African elephant weighs about six tons.
	✓	A fully grown African elephant weighs about six tons.
	✗	An adult African elephant weighs about six tonnes.
comma	src	Er fragte sich, welches Auto er kaufen sollte.
	ref	He wondered what car to buy.
	✓	He wondered which car to buy.
	✗	He asked himself, which car he should buy.
quotation marks	src	"Wann sollen wir uns treffen?", wollten sie wissen.
	ref	"When are we supposed to meet?" they asked.
	✓	"When shall we meet?" they wanted to know.
	✗	When are we going to meet? They wanted to know.
		English-German
Intransitive . future II progr	src	They will have been running.
	ref	Sie werden gelaufen sein.
	✓	Sie werden gerannt sein.
	✗	Sie würden gelaufen sein.
Focus particle	src	He even drank four bottles of wine.
	ref	Er habe sogar vier Flaschen Wein getrunken.
	✓	Er trank sogar vier Flaschen Wein.
	✗	Er trank noch vier Flaschen Wein.
Transitive present progr.	src	They are playing the piano.
	ref	Sie spielen auf dem Klavier.
	✓	Sie spielen Klavier.
	✗	Sie spielen das Klavier.
measuring unit	src	Potatoes are sold in hundredweights.
	ref	Kartoffeln werden in Zentnergewichten verkauft.
	✓	Kartoffeln werden in Zentner verkauft.
	✗	Kartoffeln werden in Hundertgewichten verkauft.

Table 2: Indicative examples of incorrectly scored challenge items for the phenomena that have the lowest accuracies

ling. category	#	baselines									QE as a metric									ref. based metrics												avg
		BERTScore	BLEU	BLEURT-20	COMET-20	YiSi-1	chrF	f101spBLEU	f200spBLEU		COMETKiwi	Cross-QE	HWTSC-TLM	HWTSC-TS	KG-BERT	MATESE-QE	MS-COMET-QE	REUSE	UniTE-src	COMET-22	MATESE	MEE	MEE2	MEE4	MS-COMET	UniTE-ref	UniTE	XL-DA19	XL-MQM20	XXL-DA19	XXL-MQM20	
Ambiguity	146	87	71	**90**	82	87	**89**	87	88		55	47	**81**	47	47	25	38	15	36	84	40	73	88	91	78	**97**	93	94	88	95	87	72
Coordination & ellipsis	836	69	61	**80**	76	73	61	64	62		**76**	71	72	70	60	33	70	38	74	79	37	59	62	62	78	79	78	81	**83**	81	80	68
False friends	225	66	63	70	**73**	67	72	66	67		67	60	64	73	68	52	73	**89**	64	69	35	69	79	**88**	76	71	71	71	71	68	69	69
Function word	200	90	76	90	**94**	78	72	74	73		91	92	78	90	90	66	92	66	**94**	**90**	28	78	80	85	90	**90**	**91**	78	82	84	82	82
MWE	829	79	72	**87**	82	85	77	74	73		78	**81**	79	**82**	**82**	37	71	32	78	**86**	46	69	76	78	81	**87**	**87**	86	**86**	79	77	75
Named entity & termin.	1272	58	55	**66**	63	**64**	61	63	64		55	**59**	56	53	54	21	55	43	53	61	30	59	63	63	62	69	68	**73**	70	**73**	**72**	59
Negation	174	87	83	89	90	**93**	85	82	84		**92**	86	87	**91**	**91**	43	**92**	78	90	91	79	84	92	92	90	**94**	**94**	82	81	82	78	86
Non-verbal agreement	372	75	72	81	**84**	78	70	74	75		77	70	59	63	59	34	**79**	39	72	**90**	48	61	73	76	84	87	86	88	**90**	**90**	**90**	73
Punctuation	336	70	**79**	76	77	77	74	71	68		68	72	70	51	51	50	68	46	**79**	79	51	64	75	74	73	**81**	**81**	67	60	72	68	69
Subordination	994	77	74	80	**83**	78	74	75	73		**86**	82	81	84	82	47	83	48	**85**	**84**	53	73	77	78	82	**85**	**85**	**85**	84	82	79	77
Verb tense/aspect/mood	3081	67	62	**70**	69	**69**	69	64	64		70	**77**	51	58	59	41	61	54	70	**77**	43	71	71	69	64	70	72	**78**	74	76	73	66
Verb valency	480	73	64	**84**	74	76	71	66	70		**82**	74	65	69	68	30	70	48	72	82	42	62	70	70	76	79	80	79	78	**85**	81	70
macro avg.	8945	75	69	**80**	**79**	77	73	72	72		**75**	73	70	69	68	40	71	50	72	81	44	69	75	77	78	**82**	**82**	80	79	80	78	72
micro avg.	8945	70	65	**76**	74	73	69	68	68		**73**	**74**	63	65	64	38	67	48	71	78	42	68	71	72	72	77	77	**79**	77	**78**	76	69

Table 3: Accuracy of the metrics (%) with regards to the 12 linguistically motivated categories for English-German

ling. category	ling. phenomenon	#	baselines								QE as a metric							ref. based metrics								avg
			BERTScore	BLEU	BLEURT-20	COMET-20	YiSi-1	chrF	f101spBLEU	f200spBLEU	COMETKiwi	Cross-QE	HWTSC-TLM	HWTSC-TS	KG-BERT	MS-COMET-QE	UniTE-src	COMET-22	MS-COMET	UniTE-ref	UniTE	XL-DA19	XL-MQM20	XXL-DA19	XXL-MQM20	
Ambiguity	Lexical ambiguity	129	91	74	**95**	**94**	88	87	82	82	81	65	56	60	57	**82**	**83**	93	81	**97**	**97**	95	93	89	88	83
	Structural ambiguity	169	**89**	69	83	80	**89**	75	80	76	**82**	79	64	69	75	**82**	78	83	**88**	83	82	82	**88**	78	84	80
Composition	Compound	129	86	64	**90**	83	**91**	74	68	70	**71**	69	64	**70**	45	64	67	81	87	82	83	90	88	**93**	88	77
	Phrasal verb	123	**91**	66	85	86	89	74	72	72	81	**86**	83	82	74	80	85	**84**	**85**	82	82	81	76	81	75	80
Coordination & ellipsis	Gapping	51	71	76	**82**	78	71	76	73	75	**100**	98	59	75	75	84	88	98	86	94	90	80	**100**	88	94	83
	Right node raising	67	90	70	76	75	**91**	75	70	67	78	**84**	64	55	82	72	72	**82**	75	78	76	76	79	75	78	76

Table 4: Accuracy of the metrics(%) with regards to the linguistically-motivated phenomena for German-English

(Continued on next page)

			baselines								QE as a metric							ref. based metrics								
ling. category	ling. phenomenon	#	BERTScore	BLEU	BLEURT-20	COMET-20	YiSi-1	chrF	f101spBLEU	f200spBLEU	COMETKiwi	Cross-QE	HWTSC-TLM	HWTSC-TS	KG-BERT	MS-COMET-QE	UniTE-src	COMET-22	MS-COMET	UniTE-ref	UniTE	XL-DA	XL-MQM	xxl-DA19	xxl-MQM20	avg
	Sluicing	128	**80**	75	77	77	78	**79**	73	75	80	66	77	79	76	59	**86**	**80**	71	73	78	**81**	76	79	70	76
	Stripping	70	76	74	**84**	79	80	76	73	73	77	80	67	71	**83**	73	**83**	81	76	81	80	77	**89**	71	81	78
False friends	False friends	90	91	64	**93**	82	**92**	78	69	70	88	74	81	**91**	87	63	44	**91**	88	**92**	**92**	90	90	87	88	82
Function word	Focus particle	64	86	75	83	**89**	88	75	72	77	83	70	70	84	**88**	81	75	84	86	86	**88**	**88**	67	**88**	81	81
	Modal particle	166	**87**	79	85	83	**86**	77	80	81	**82**	75	69	81	**83**	67	**83**	**89**	82	**89**	**88**	**89**	83	87	83	82
	Question tag	356	80	69	**82**	74	78	71	69	69	81	79	**84**	80	61	65	75	81	79	**84**	81	81	79	81	81	77
LDD & interrogatives	Extended adjective construction	320	**87**	80	**88**	87	**88**	80	80	80	90	**93**	79	82	61	91	88	**90**	87	88	**89**	**89**	86	88	84	85
	Extraposition	92	73	74	75	**82**	77	**83**	72	73	67	74	65	**79**	62	63	75	76	74	67	77	80	79	**84**	78	74
	Multiple connectors	87	74	**79**	63	72	76	76	**80**	**79**	**70**	68	67	63	64	**69**	**70**	68	**79**	61	63	66	57	66	53	69
	Pied-piping	162	94	78	93	**96**	93	77	75	75	**96**	90	73	74	70	79	94	**95**	**94**	**94**	**94**	**94**	89	**94**	90	87
	Polar question	51	**71**	43	63	61	67	45	45	47	69	49	49	53	61	69	**78**	67	55	65	71	61	**75**	61	**75**	61
	Scrambling	144	**90**	72	**90**	87	88	74	69	69	**90**	88	82	81	51	**90**	81	**98**	90	93	90	96	92	95	95	85
	Topicalization	61	85	85	**87**	84	**87**	84	**87**	**87**	77	69	66	70	77	**82**	74	82	74	80	80	82	70	**85**	79	80
	Wh-movement	97	79	62	**85**	81	77	69	63	63	72	**75**	56	64	66	**75**	**74**	81	73	**86**	84	80	73	78	75	74
MWE	Collocation	190	87	72	**91**	89	88	79	74	74	**84**	82	82	65	67	73	79	89	79	92	**93**	90	91	91	89	83
	Idiom	133	82	67	76	**85**	83	69	67	65	44	**55**	36	17	20	31	33	75	77	87	**88**	**89**	86	86	75	65
	Prepositional MWE	146	84	79	**85**	84	**86**	84	79	81	82	**84**	82	84	72	71	**85**	85	78	84	**86**	**86**	**86**	85	77	82
	Verbal MWE	141	86	74	**87**	80	84	77	77	77	89	81	77	68	57	60	**91**	92	**95**	93	91	87	87	84	82	82
Named entity & termin.	Date	203	**67**	50	65	65	66	58	58	57	**70**	**70**	63	68	68	61	66	67	63	67	63	69	**74**	68	72	65
	Domainspecific term	214	71	63	71	64	**74**	71	68	68	67	**77**	63	57	59	66	60	72	64	72	71	68	**75**	71	70	68
	Location	181	78	65	70	75	**82**	66	71	74	62	57	**76**	64	38	56	54	75	71	66	61	68	**80**	70	78	68
	Measuring unit	203	75	67	61	64	77	72	81	**84**	57	**73**	54	51	56	56	55	63	62	59	55	62	**67**	66	66	64
	Proper name	60	90	75	85	87	**92**	73	77	77	78	**88**	72	70	50	70	83	85	**90**	83	83	78	85	**90**	88	80
Negation	Negation	76	**95**	84	88	92	91	88	83	80	**93**	78	62	74	87	88	92	91	88	**93**	**93**	89	78	88	83	86
Non-verbal agreement	Coreference	251	74	68	**90**	85	75	72	71	71	81	77	73	69	66	**84**	78	**91**	82	90	90	**91**	88	**92**	**91**	80
	External possessor	104	84	**88**	75	76	82	**88**	85	86	70	68	50	51	58	68	**74**	**76**	73	75	**77**	71	71	75	70	73
	Internal possessor	64	80	80	72	72	72	67	78	**83**	61	59	62	58	52	**67**	53	69	61	73	**77**	72	69	72	72	69
Punctuation	Comma	46	91	91	**93**	85	89	87	91	91	85	**91**	83	85	87	80	80	89	85	87	83	89	80	**91**	83	87
	Quotation marks	247	71	75	66	64	70	65	76	**77**	49	72	**81**	71	57	57	67	**64**	61	60	57	56	48	53	40	63
Subordination	Adverbial clause	87	71	70	**82**	75	72	67	66	67	70	**74**	66	68	70	69	66	**77**	67	74	72	70	74	75	68	71
	Cleft sentence	109	**73**	**73**	67	71	66	66	70	69	66	69	48	64	62	55	**71**	72	**75**	66	69	66	64	65	61	66
	Free relative clause	70	63	67	**77**	71	67	71	74	71	60	70	50	56	63	69	**77**	77	**83**	80	81	74	54	76	67	70
	Indirect speech	119	76	64	**81**	**80**	71	70	62	63	**80**	75	58	58	57	62	70	**87**	**87**	**86**	**86**	83	65	84	82	73
	Infinitive clause	64	**78**	77	77	72	**78**	77	75	73	**73**	70	62	67	**73**	72	70	70	67	73	70	75	66	**80**	67	72
	Object clause	54	85	74	85	**91**	89	81	72	72	76	89	69	69	**94**	67	80	**93**	87	91	91	89	80	87	85	82

Table 4: Accuracy of the metrics(%) with regards to the linguistically-motivated phenomena for German-English

(Continued on next page)

			baselines								QE as a metric							ref. based metrics								
ling. category	ling. phenomenon	#	BERTScore	BLEU	BLEURT-20	COMET-20	YiSi-1	chrF	f101spBLEU	f200spBLEU	COMETKiwi	Cross-QE	HWTSC-TLM	HWTSC-TS	KG-BERT	MS-COMET-QE	UniTE-src	COMET-22	MS-COMET	UniTE-ref	UniTE	XL-DA	XL-MQM	xxl-DA19	xxl-MQM20	avg
	Pseudo-cleft sentence	25	**96**	72	68	88	92	60	72	72	80	**100**	48	32	60	72	96	88	80	**92**	**92**	80	88	88	68	78
	Relative clause	71	**70**	63	65	**70**	66	66	63	68	63	61	59	59	48	51	**66**	**77**	65	63	70	62	66	73	73	65
	Subject clause	80	**85**	66	**85**	**86**	**86**	65	62	71	**86**	**86**	68	70	62	70	80	86	84	82	82	85	88	85	**92**	79
Verb tense/aspect/mood	Conditional	50	**80**	**80**	76	76	**80**	**80**	78	76	80	80	82	78	74	80	**90**	82	78	82	80	82	**88**	76	78	80
	Ditransitive - future I	121	87	72	**92**	89	88	71	70	70	**93**	**92**	58	68	85	89	91	**94**	80	**94**	**94**	92	85	92	92	84
	Ditransitive - future I subjunctive II	84	90	63	89	93	**95**	75	68	68	90	**94**	50	65	92	**94**	**93**	**92**	90	**92**	**92**	**93**	88	**93**	90	85
	Ditransitive - future II	97	**94**	60	82	73	**94**	71	67	67	**98**	**98**	58	69	69	96	93	**93**	66	88	85	85	78	88	80	80
	Ditransitive - future II subjunctive II	88	93	73	86	88	**97**	78	69	69	88	**99**	65	75	89	97	93	89	77	**92**	90	83	84	85	84	84
	Ditransitive - perfect	72	**93**	62	81	78	**93**	72	67	67	93	**96**	46	58	75	88	**96**	86	71	**88**	81	81	82	**88**	81	79
	Ditransitive - pluperfect	86	83	67	83	77	**88**	79	71	71	81	**83**	57	71	74	**84**	**83**	**86**	62	**86**	**86**	79	69	83	62	77
	Ditransitive - pluperfect subjunctive II	107	**94**	71	79	86	92	88	67	67	64	69	65	66	75	68	**92**	86	75	**90**	86	82	78	77	71	78
	Ditransitive - present	90	82	61	**91**	86	81	77	66	64	70	**99**	56	60	83	81	78	**88**	**89**	**89**	**89**	83	86	**89**	**89**	80
	Ditransitive - preterite	117	84	62	85	**88**	**89**	76	68	68	84	**91**	62	61	87	84	84	**95**	86	91	90	90	92	**94**	93	83
	Ditransitive - preterite subjunctive II	110	87	61	**95**	93	90	85	65	65	85	**87**	60	61	85	85	85	**96**	89	95	95	95	**96**	**97**	**97**	85
	Imperative	98	88	78	**95**	92	89	79	81	76	**88**	84	78	74	57	78	86	88	86	**92**	**92**	90	87	**91**	90	84
	Intransitive - future I	32	84	53	88	**91**	**91**	72	59	59	84	97	69	91	**100**	94	97	84	**94**	88	88	88	88	88	91	84
	Intransitive - future I subjunctive II	56	**93**	61	**93**	91	89	70	73	71	93	95	55	68	**100**	86	89	95	88	98	**100**	98	84	96	98	86
	Intransitive - future II	62	87	60	**90**	84	89	77	65	69	79	**87**	45	58	69	63	60	90	92	94	94	95	94	**97**	92	80
	Intransitive - future II subjunctive II	94	**97**	72	94	91	**98**	89	76	74	80	**100**	63	82	86	84	71	91	86	**94**	**93**	**94**	85	**93**	87	86
	Intransitive - perfect	61	85	56	84	72	**87**	59	56	54	62	**69**	66	66	64	59	59	72	79	69	70	75	67	**82**	72	69
	Intransitive - pluperfect	85	85	79	85	80	**87**	85	81	76	68	86	46	55	**88**	64	61	78	78	**81**	**81**	**82**	71	**81**	69	76
	Intransitive - pluperfect subjunctive II	79	**100**	87	97	96	**100**	90	80	80	78	**94**	56	76	**95**	73	71	**96**	91	**97**	**97**	**96**	95	**97**	95	89
	Intransitive - present	54	96	69	91	94	**98**	74	69	72	96	**98**	65	76	94	94	91	**94**	93	93	93	89	93	87	87	87
	Intransitive - preterite	46	70	46	**89**	78	74	63	46	52	**93**	**93**	74	83	91	85	85	87	76	85	80	85	80	**89**	74	77
	Intransitive - preterite subjunctive II	100	81	43	**86**	79	79	51	60	61	79	89	58	67	**91**	83	77	**88**	81	83	84	87	**89**	82	80	76
	Modal - future I	42	**98**	90	88	95	95	95	90	90	83	**98**	76	83	74	90	74	**88**	**88**	**88**	**88**	**88**	76	86	81	87
	Modal - future I subjunctive II	86	**97**	94	81	93	**97**	94	93	93	**79**	**78**	**78**	**79**	67	**78**	62	**85**	80	**85**	**86**	**85**	64	**86**	65	83
	Modal - perfect	149	**85**	72	74	79	**85**	72	74	74	**85**	81	67	77	47	60	72	70	**78**	67	65	66	57	70	61	71
	Modal - pluperfect	75	**100**	**100**	84	95	**100**	99	**100**	**100**	69	89	83	**91**	47	49	75	76	**85**	72	72	69	44	71	48	79
	Modal - pluperfect subjunctive II	61	87	72	79	89	**90**	80	69	69	85	**90**	69	79	85	87	87	84	**87**	84	80	80	74	82	77	81
	Modal - present	30	83	57	**93**	87	80	73	63	63	**90**	80	53	80	80	83	83	80	80	80	80	73	73	**83**	80	77
	Modal - preterite	72	86	61	**88**	**88**	**88**	74	67	68	90	92	54	78	**93**	92	86	89	86	86	86	89	81	**94**	90	83
	Modal - preterite subjunctive II	30	**87**	80	83	83	83	77	80	80	**93**	87	43	73	87	**93**	83	80	**87**	83	**87**	83	70	**87**	83	81
	Modal negated - future I	43	95	93	81	93	**100**	88	93	93	86	**93**	86	91	65	86	58	81	**98**	77	79	79	74	79	70	84
	Modal negated - future I subjunctive II	73	96	92	86	90	**97**	96	90	90	79	92	79	88	77	**95**	84	86	**97**	79	82	88	75	86	67	87

Table 4: Accuracy of the metrics(%) with regards to the linguistically-motivated phenomena for German-English

(Continued on next page)

ling. category	ling. phenomenon	#	BERTScore	BLEU	BLEURT-20	COMET-20	YiSi-1	chrF	f101spBLEU	f200spBLEU	COMETKiwi	Cross-QE	HWTSC-TLM	HWTSC-TS	KG-BERT	MS-COMET-QE	UniTE-src	COMET-22	MS-COMET	UniTE-ref	UniTE	XL-DA	XL-MQM	xxl-DA19	xxl-MQM20	avg
			baselines								QE as a metric							ref. based metrics								
	Modal negated - perfect	126	71	50	66	**73**	**72**	62	52	52	73	**88**	60	71	63	85	72	63	**70**	63	66	63	60	63	51	66
	Modal negated - pluperfect	126	94	87	90	96	94	**99**	87	90	74	**95**	83	93	55	75	79	**88**	85	84	85	**88**	75	81	76	85
	Modal negated - pluperfect subjunctive II	81	75	65	73	72	**78**	74	69	68	59	**84**	64	79	**84**	**84**	73	73	**86**	74	72	79	75	74	69	74
	Modal negated - present	33	70	**79**	73	70	70	64	45	45	64	**88**	48	67	64	61	58	67	67	70	73	79	76	**82**	79	68
	Modal negated - preterite	61	90	66	90	**92**	89	87	56	56	90	82	38	75	**95**	**95**	90	85	**92**	85	87	80	79	80	84	81
	Modal negated - preterite subjunctive II	77	88	66	**91**	87	86	83	65	65	91	**95**	47	75	86	88	84	84	**91**	87	87	83	78	83	84	82
	Progressive	76	**84**	66	71	75	75	67	67	67	75	67	50	64	64	70	**76**	75	76	75	78	67	68	**79**	76	71
	Reflexive - future I	85	81	76	**89**	87	82	80	74	74	86	85	84	81	75	78	**88**	89	89	**92**	88	87	81	88	87	84
	Reflexive - future I subjunctive II	96	82	70	79	77	**84**	66	66	65	78	**89**	71	79	80	74	85	**85**	73	**86**	**85**	**85**	80	84	82	79
	Reflexive - future II	116	**97**	83	77	81	**97**	85	81	80	67	73	40	43	72	**75**	67	**87**	69	83	84	79	74	82	79	76
	Reflexive - future II subjunctive II	107	**93**	74	81	89	**93**	77	71	70	79	**92**	66	77	**91**	82	87	**89**	76	87	86	85	78	79	75	82
	Reflexive - perfect	188	81	64	81	**84**	**82**	62	69	68	86	85	53	54	78	72	**88**	86	80	**87**	85	82	78	82	83	77
	Reflexive - pluperfect	109	85	63	83	**88**	**87**	55	63	62	80	**83**	54	47	75	78	82	85	86	85	85	83	81	88	**92**	77
	Reflexive - pluperfect subjunctive II	90	**98**	76	79	87	**97**	80	78	78	70	81	66	70	**88**	76	81	**81**	76	**80**	**80**	74	64	77	71	79
	Reflexive - present	125	81	59	**90**	86	80	74	70	70	88	92	72	75	74	**94**	**94**	86	**92**	88	87	85	85	89	85	82
	Reflexive - preterite	117	86	69	85	83	**88**	75	70	71	76	83	54	56	66	**85**	83	88	76	**90**	**90**	**91**	85	85	83	79
	Reflexive - preterite subjunctive II	124	**92**	77	86	85	**91**	70	75	75	72	**83**	54	55	65	79	81	**89**	78	**89**	**88**	**89**	**88**	**88**	87	80
	Transitive - future I	43	98	95	**100**	**100**	**100**	95	95	95	**100**	95	86	**100**	**100**	95	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	98
	Transitive - future I subjunctive II	37	**100**	81	95	**100**	**100**	84	86	86	92	86	54	89	**100**	95	97	95	95	95	**97**	95	84	**97**	**97**	91
	Transitive - future II	33	**100**	76	94	94	**100**	94	79	79	88	64	70	**94**	88	**94**	76	94	88	94	94	**97**	85	**97**	85	88
	Transitive - future II subjunctive II	50	**100**	84	88	94	**100**	88	82	80	92	90	90	**98**	**98**	**98**	94	**92**	**92**	90	90	**92**	76	90	84	91
	Transitive - perfect	99	85	64	81	**88**	**88**	80	67	74	79	76	73	86	78	71	**93**	81	**90**	80	80	75	81	87	86	80
	Transitive - pluperfect	22	**91**	73	82	**91**	**91**	82	73	73	73	77	73	77	68	77	**91**	**91**	86	86	82	86	73	77	64	80
	Transitive - pluperfect subjunctive II	39	**100**	85	64	85	**100**	97	87	87	49	54	69	67	**92**	87	54	72	**92**	74	74	74	62	72	67	77
	Transitive - present	33	**94**	58	**94**	85	91	73	58	61	88	**94**	82	79	88	**94**	88	91	91	91	88	**94**	91	**94**	91	85
	Transitive - preterite	57	82	51	**86**	**86**	82	63	68	67	**95**	91	67	68	93	93	**95**	**100**	89	89	86	91	**100**	95	**100**	84
	Transitive - preterite subjunctive II	97	82	40	**86**	80	84	60	57	54	73	80	73	74	**86**	79	**85**	**85**	**86**	84	84	84	84	84	**85**	77
Verb valency	Case government	80	**89**	65	**88**	86	**89**	62	64	64	**94**	75	71	66	52	82	85	**95**	86	92	91	92	92	92	92	81
	Mediopassive voice	50	82	64	82	**84**	80	66	62	60	**74**	66	50	50	64	60	68	**90**	82	88	86	88	88	86	82	74
	Passive voice	33	**94**	85	91	**94**	**94**	82	82	79	**94**	79	64	61	64	82	91	**94**	**94**	**94**	**94**	91	91	88	91	86
	Resultative predicates	48	**100**	73	94	90	98	85	77	79	81	67	69	75	73	73	**83**	**96**	92	92	94	94	79	94	85	84
macro avg.		10402	**86**	71	83	83	**86**	76	72	72	79	**82**	65	71	73	77	79	**84**	82	**84**	**84**	83	79	**84**	80	79
micro avg.		10402	**84**	70	82	82	**85**	75	72	72	78	**81**	66	70	70	75	78	**84**	80	**83**	82	82	79	82	79	78

Table 4: Accuracy of the metrics(%) with regards to the linguistically-motivated phenomena for German-English

			baselines								QE as a metric									ref. based metrics												
ling. category	ling. phenomenon	#	BERTScore	BLEU	BLEURT-20	COMET-20	YiSi-1	chrF	f101spBLEU	f200spBLEU	COMETKiwi	Cross-QE	HWTSC-TLM	HWTSC-TS	KG-BERT	MATESE-QE	MS-COMET-QE	REUSE	UniTE-src	COMET-22	MATESE	MEE	MEE2	MEE4	MS-COMET	UniTE-ref	UniTE	XL-DA	XL-MQM20	XXL-DA19	XXL-MQM20	avg
Ambiguity	Lexical ambiguity	146	87	71	**90**	82	87	**89**	87	88	55	47	**81**	47	47	25	38	15	36	84	40	73	88	91	78	**97**	93	94	88	95	87	72
Coordination	Gapping	163	72	58	**84**	**84**	74	68	67	64	**80**	71	71	72	42	61	78	58	**81**	79	58	62	66	64	78	81	77	79	88	**91**	86	72
& ellipsis	Pseudogapping	201	82	77	**97**	87	82	67	76	74	93	92	81	78	78	41	75	26	**94**	96	53	65	73	74	89	**97**	**97**	**97**	93	95	93	80
	Right node raising	47	83	64	87	**91**	83	72	72	70	**87**	81	81	83	34	4	**87**	9	79	94	4	62	66	66	91	89	94	91	**96**	94	89	73
	Sluicing	169	54	56	**63**	57	59	54	59	58	56	47	61	51	46	14	**72**	51	53	62	17	63	59	56	59	57	55	**67**	**67**	56	65	55
	Stripping	139	66	58	65	60	**68**	58	58	60	**63**	53	53	55	55	22	41	38	55	68	29	45	58	58	71	71	68	67	**83**	73	73	58
	VP-ellipsis	117	59	47	85	**87**	75	51	46	46	84	85	**95**	91	92	29	78	26	82	82	30	54	50	48	**91**	85	84	85	80	80	76	69
False friends	False friends	225	66	63	70	**73**	67	72	66	67	67	60	64	73	68	52	73	**89**	64	69	35	69	79	**88**	76	71	71	71	71	68	69	69
Function	Focus particle	20	45	30	80	**90**	45	35	30	30	45	50	15	25	25	5	**70**	55	65	75	5	30	30	35	50	70	70	60	55	**85**	75	48
word	Question tag	180	**94**	82	91	**95**	82	76	79	78	96	**97**	84	**98**	**98**	72	94	68	**97**	92	31	83	86	91	**94**	92	**93**	80	85	83	83	85
MWE	Collocation	112	73	61	**92**	79	88	76	62	61	**88**	86	**89**	86	86	46	75	46	86	**95**	50	68	79	80	85	**95**	93	93	**96**	93	**96**	79
	Compound	63	75	51	70	78	**87**	84	71	71	98	**100**	89	94	94	21	57	3	86	**97**	22	57	70	67	73	**97**	90	90	**97**	81	84	74
	Idiom	266	86	82	**95**	**95**	92	75	80	81	86	**92**	**93**	85	86	62	75	22	82	**98**	67	77	82	84	96	**98**	**98**	97	97	97	95	85
	Nominal MWE	288	81	71	**84**	72	81	78	78	74	62	72	66	**78**	**78**	6	71	39	74	69	28	65	72	73	68	74	**76**	72	74	55	48	67
	Prepositional MWE	35	69	**86**	71	83	**86**	83	74	80	66	77	60	80	80	60	**89**	83	80	80	66	77	80	83	**89**	80	80	86	86	80	83	78
	Verbal MWE	65	66	71	**89**	78	62	74	58	58	**83**	58	69	65	65	57	55	23	58	86	42	62	74	75	72	78	80	85	75	**92**	88	69
Named entity	Date	234	55	53	**74**	66	68	60	61	65	62	**93**	80	79	79	48	80	31	50	65	54	62	60	57	64	72	**75**	67	**76**	70	73	65
& termin.	Domainspecific term	312	73	56	**89**	**90**	86	76	69	71	73	62	**78**	67	67	7	58	33	71	78	41	67	73	76	93	86	85	**97**	94	**96**	92	73
	Location	12	67	83	75	**100**	50	58	58	83	**100**	92	**100**	92	**100**	0	92	83	92	75	17	92	83	92	83	83	92	**100**	**100**	**100**	**100**	81
	Measuring unit	389	54	48	53	50	54	**55**	**57**	53	28	31	21	20	20	6	37	**43**	35	46	12	51	57	59	41	57	57	**69**	58	**69**	68	45
	Proper name	325	50	61	52	51	53	54	66	**69**	**64**	**64**	58	58	61	34	53	62	58	59	23	58	61	61	56	**64**	59	58	54	58	55	56
Negation	Negation	174	87	83	89	90	**93**	85	82	84	**92**	86	87	**91**	**91**	43	**92**	78	90	91	79	84	92	92	90	**94**	**94**	82	81	82	78	86
Non-verbal	Coreference	81	85	86	**95**	84	75	86	84	89	**77**	73	33	67	51	26	**77**	41	73	96	80	81	89	88	95	94	93	96	**100**	**99**	**99**	80
agreement	Genitive	206	76	73	73	**83**	82	68	77	76	71	63	83	56	56	44	**90**	22	62	84	42	57	70	72	81	82	82	85	85	**87**	**87**	71
	Possession	85	61	55	**86**	**85**	74	58	58	60	**93**	86	26	76	76	19	53	78	**93**	**96**	33	51	62	72	79	93	92	88	91	87	89	71
Punctuation	Quotation marks	336	70	**79**	76	77	77	74	71	68	68	72	70	51	51	50	68	46	**79**	79	51	64	75	74	73	**81**	**81**	67	60	72	68	69
Subordination	Adverbial clause	193	72	**81**	**81**	**81**	67	73	79	77	88	77	79	87	87	34	72	65	**90**	82	31	72	78	77	71	84	82	**86**	**85**	82	**85**	76
	Cleft sentence	179	66	63	60	**69**	63	57	62	60	74	59	72	**82**	**82**	45	74	45	67	71	46	59	65	68	**73**	70	**73**	**73**	66	65	63	65
	Contact clause	150	83	75	**94**	**94**	88	74	74	73	**98**	**97**	**99**	97	97	65	92	53	96	**98**	64	79	76	79	92	**97**	96	**97**	**97**	95	93	87
	Indirect speech	38	58	42	63	**66**	50	47	47	42	**95**	63	58	58	50	42	55	24	55	**76**	47	39	42	42	74	71	71	63	68	63	61	56
	Infinitive clause	85	67	55	86	87	**95**	80	66	66	95	**99**	78	79	68	66	**98**	40	**98**	93	65	54	67	71	**95**	93	92	88	91	**94**	87	80
	Object clause	16	75	38	**88**	**88**	62	56	38	38	81	62	56	81	81	62	**88**	31	81	88	50	44	50	50	94	**100**	**100**	**100**	**100**	88	75	70
	Pseudo-cleft sentence	73	**90**	88	66	70	**90**	**89**	82	82	68	85	81	**89**	**89**	62	73	70	78	75	60	**93**	**92**	**93**	58	71	77	85	86	68	68	79
	Relative clause	112	89	83	90	**94**	82	84	88	81	88	93	78	65	65	22	96	36	**100**	91	52	85	87	86	96	**98**	96	94	89	84	79	82

Table 5: Accuracy of the metrics(%) with regards to the linguistically-motivated phenomena for German-English

(Continued on next page)

ling. category	ling. phenomenon	#	BERTScore	BLEU	BLEURT-20	COMET-20	YiSi-1	chrF	f101spBLEU	f200spBLEU	COMETKiwi	Cross-QE	HWTSC-TLM	HWTSC-TS	KG-BERT	MATESE-QE	MS-COMET-QE	REUSE	UniTE-src	COMET-22	MATESE	MEE	MEE2	MEE4	MS-COMET	UniTE-ref	UniTE	XL-DA	XL-MQM	xxl-DA19	xxl-MQM20	avg
			baselines								QE as a metric									ref. based metrics												
	Subject clause	148	86	**90**	89	**91**	**91**	**90**	**91**	**91**	**89**	89	87	**89**	**89**	47	**91**	33	85	86	71	87	**92**	**93**	89	89	88	84	82	89	85	85
Verb tense	Conditional	106	74	77	**94**	90	91	70	75	75	**92**	87	86	89	89	31	81	18	**92**	87	52	75	83	87	88	**92**	**92**	**92**	84	89	89	80
/aspect/mood	Ditransitive - conditional I progr.	72	65	49	**93**	89	83	61	56	57	**99**	**99**	94	**99**	79	50	92	74	**100**	92	47	64	60	62	81	90	92	93	81	**100**	89	79
	" - conditional I simple	34	94	74	65	85	**97**	94	74	79	**100**	97	41	41	44	26	91	91	**100**	**100**	18	**100**	94	97	85	**100**	97	97	94	**100**	91	82
	" - conditional II progr.	51	75	78	**88**	78	80	82	82	82	65	67	51	55	49	24	59	63	**86**	90	27	86	82	78	82	88	86	84	82	86	**92**	73
	" - conditional II simple	59	71	64	76	**78**	66	68	64	73	69	56	53	49	47	36	63	59	**78**	73	25	78	71	75	78	78	76	80	78	**81**	78	67
	" - future I progr.	61	52	51	**62**	57	57	**62**	51	51	92	51	84	75	49	11	80	90	**97**	66	8	59	59	61	33	59	66	**79**	75	57	66	61
	" - future I simple	88	**60**	51	56	55	56	**60**	50	45	66	50	52	53	48	40	70	**90**	85	58	38	56	57	60	53	56	60	**65**	**64**	51	60	57
	" - future II progr.	91	**70**	64	66	57	47	60	65	62	71	45	84	**91**	**91**	11	78	56	77	82	14	89	73	76	54	86	77	65	62	**95**	92	68
	" - future II simple	49	71	**94**	86	**94**	65	**94**	92	92	**100**	92	76	71	65	8	65	88	92	92	18	96	94	94	65	**100**	98	86	39	88	76	79
	" - past perfect progr.	91	60	44	60	**67**	**66**	58	53	48	65	**75**	51	59	65	11	**75**	37	60	71	33	62	59	63	52	67	75	**78**	73	73	67	60
	" - past perfect simple	112	63	62	65	56	**72**	71	61	61	56	**79**	37	37	54	8	58	43	46	70	39	56	64	64	53	62	68	**71**	57	58	49	57
	" - past progr.	83	58	57	**70**	58	59	61	57	57	61	37	39	37	37	12	42	**71**	33	70	12	39	55	60	42	66	64	**72**	67	**72**	69	53
	" - present perfect progr.	48	85	54	85	75	**92**	88	56	60	85	94	90	**100**	**100**	21	77	52	**100**	92	35	75	81	81	71	**94**	**94**	79	73	92	77	78
	" - present perfect simple	54	**65**	37	56	43	30	41	37	44	33	33	31	26	**35**	28	33	33	31	48	22	44	44	48	33	57	56	65	59	**70**	69	43
	" - present progr.	72	76	38	94	**97**	90	68	36	49	**100**	**100**	**99**	**99**	**99**	94	88	35	**99**	**96**	71	72	86	83	**97**	**97**	**97**	88	88	92	88	83
	" - simple past	77	77	56	77	**83**	56	66	56	57	**97**	94	69	75	88	36	73	82	82	**94**	45	73	78	84	87	82	83	79	84	82	78	75
	" - simple present	54	72	30	**83**	70	**83**	56	41	41	67	**70**	67	67	67	54	**70**	28	**70**	59	48	56	67	69	65	80	80	81	83	89	**94**	66
	Gerund	161	92	85	**96**	**96**	92	80	83	82	97	**99**	58	87	87	19	97	78	**99**	**97**	25	83	85	88	**98**	96	96	96	96	**97**	87	85
	Imperative	50	70	50	**96**	94	70	70	58	64	**100**	92	78	86	86	80	94	82	88	**96**	60	70	70	76	94	92	92	**96**	90	94	92	82
	Intransitive - conditional I progr.	9	56	89	89	**100**	**100**	78	78	89	**100**	44	0	22	22	67	44	**100**	**100**	89	56	78	78	89	78	**100**	**100**	33	56	89	78	72
	" - conditional I simple	3	**100**	0	67	**100**	**100**	33	0	33	**100**	**100**	33	33	33	**100**	33	**100**	67	**100**	**100**	0	67	**100**	67	67	67	67	**100**	**100**	67	67
	" - future I progr.	7	71	86	**100**	**100**	57	**100**	86	86	57	57	0	29	29	**86**	57	71	71	86	0	71	86	**100**	57	**100**	**100**	71	**100**	**100**	**100**	73
	" - future I simple	24	67	**75**	**75**	71	50	67	67	71	96	**100**	71	46	46	29	92	96	**100**	62	42	58	**67**	**67**	**67**	58	62	58	58	**67**	**67**	67
	" - future II progr.	4	25	**50**	25	25	**50**	**50**	**50**	**50**	**75**	0	**75**	25	25	0	50	25	0	50	0	75	50	25	25	25	25	50	75	50	**100**	40
	" - future II simple	7	71	**100**	86	**100**	**100**	**100**	**100**	**100**	**100**	**100**	57	71	71	0	43	**100**	71	**100**	14	71	86	86	**100**	86	86	43	43	71	57	76
	" - past perfect progr.	16	56	50	38	62	69	62	**81**	69	50	69	38	44	44	0	**75**	38	44	50	6	56	62	62	**69**	31	31	56	38	62	44	50
	" - past perfect simple	18	78	72	**89**	72	61	78	61	61	**94**	50	89	78	78	0	56	44	39	83	17	**89**	78	78	78	83	72	78	67	**89**	78	69
	" - past progr.	28	43	57	**71**	**71**	54	57	54	54	**68**	50	46	36	36	29	61	46	57	50	25	50	**61**	54	50	57	57	57	54	**61**	57	52
	" - present perfect simple	2	**100**	50	**100**	**100**	**100**	**100**	50	50	**100**	**100**	**100**	**100**	**100**	0	**100**	**100**	**100**	**100**	0	50	**100**	**100**	**100**	**100**	**100**	**100**	50	**100**	**100**	84
	" - present progr.	5	80	**100**	80	80	80	**100**	80	80	**80**	**80**	0	0	0	20	**80**	60	**80**	80	60	80	**100**	**100**	80	80	80	**100**	80	80	80	72
	" - simple past	24	58	38	**62**	58	58	46	38	38	**100**	**100**	96	**100**	**100**	46	71	96	88	71	46	38	62	71	67	83	**88**	62	58	79	79	69
	" - simple present	10	40	30	**50**	40	40	40	40	40	**70**	**70**	40	60	60	**70**	**70**	40	**70**	**70**	50	20	30	30	**70**	50	50	60	50	50	40	50

Table 5: Accuracy of the metrics(%) with regards to the linguistically-motivated phenomena for German-English

(Continued on next page)

ling. category	ling. phenomenon	#	baselines: BERTScore	BLEU	BLEURT-20	COMET-20	YiSi-1	chrF	f101spBLEU	f200spBLEU	QE as a metric: COMETKiwi	Cross-QE	HWTSC-TLM	HWTSC-TS	KG-BERT	MATESE-QE	MS-COMET-QE	REUSE	UniTE-src	ref. based metrics: COMET-22	MATESE	MEE	MEE2	MEE4	MS-COMET	UniTE-ref	UniTE	XL-DA	XL-MQM	xxl-DA19	xxl-MQM20	avg
	Modal	20	**70**	60	40	60	45	55	60	55	10	15	50	60	60	0	**100**	**100**	90	25	0	55	70	**75**	65	40	40	35	35	20	20	49
	Modal negated	20	35	65	70	**75**	65	60	65	70	65	**95**	50	**95**	**95**	0	65	70	60	85	0	55	70	70	60	80	80	**95**	80	90	85	67
	Reflexive - conditional I progr.	65	66	52	48	45	45	46	**71**	**71**	38	63	15	23	23	**71**	46	28	63	52	58	74	63	54	37	51	54	83	**85**	60	58	53
	" - conditional I simple	112	**76**	70	48	67	58	70	72	72	32	**100**	9	27	27	76	43	37	60	64	80	**90**	80	66	48	57	62	86	**89**	78	72	63
	" - conditional II progr.	97	**71**	**72**	66	67	61	69	**71**	**71**	64	**80**	10	20	20	76	24	49	55	84	77	86	84	67	61	72	73	87	**89**	78	86	65
	" - conditional II simple	109	61	**68**	52	55	54	61	58	59	50	**92**	11	21	27	81	16	28	57	78	86	78	67	47	59	53	49	83	91	84	**93**	59
	" - future I progr.	70	67	67	70	54	**84**	79	66	66	59	66	60	**77**	**77**	47	69	64	63	79	40	77	74	66	56	60	67	**80**	76	70	66	67
	" - future I simple	83	69	67	71	54	76	**86**	77	77	61	61	49	61	61	45	**78**	63	66	76	33	**78**	**78**	75	45	66	71	**78**	72	65	54	66
	" - future II progr.	81	65	56	64	73	75	**80**	57	57	73	**88**	54	63	63	81	51	53	65	83	62	73	79	72	58	70	73	**85**	80	68	60	68
	" - future II simple	56	71	66	77	61	**88**	**88**	64	64	79	**98**	61	59	59	55	39	68	75	88	54	**89**	80	75	55	62	68	79	71	79	70	70
	" - past perfect progr.	98	60	50	67	63	**71**	66	60	51	66	**82**	33	46	46	44	52	51	79	**76**	42	71	67	67	55	63	62	71	73	71	66	61
	" - past perfect simple	53	62	47	68	62	**74**	55	57	57	64	**98**	25	34	34	66	17	43	57	**87**	66	81	68	62	58	51	62	79	85	85	66	61
	" - past progr.	5	**100**	**100**	40	**100**	**100**	**100**	**100**	**100**	80	60	20	20	20	40	**100**	40	40	80	20	**100**	**100**	**100**	**100**	80	**100**	80	80	80	80	74
	" - present perfect progr.	33	76	48	**88**	82	76	76	48	48	**100**	**100**	64	61	61	**100**	45	24	82	**100**	**100**	97	82	79	58	79	82	97	**100**	91	85	77
	" - present perfect simple	39	59	46	67	69	69	**72**	44	44	74	**92**	79	72	72	21	31	54	69	85	74	77	69	69	51	54	69	**87**	85	74	77	66
	" - present progr.	99	62	51	54	54	**67**	56	60	62	36	**77**	27	26	26	61	40	46	45	58	48	68	61	57	40	53	56	62	**70**	54	63	53
	" - simple past	119	71	70	73	76	73	**77**	71	71	**89**	83	37	69	76	40	46	53	76	**91**	39	81	75	71	73	74	76	82	76	83	81	71
	" - simple present	138	65	65	67	62	**88**	63	68	67	44	**89**	39	54	62	62	47	32	49	62	46	**78**	76	69	69	54	57	71	69	68	67	62
	Transitive - future II progr.	11	73	**82**	73	73	64	**82**	**82**	**82**	73	55	82	**91**	**91**	9	**91**	73	**91**	82	9	73	73	73	82	73	82	91	82	**100**	**100**	75
	" - conditional I progr.	11	55	**91**	45	73	36	82	**91**	**91**	55	18	36	45	45	0	**82**	**82**	27	45	0	55	55	55	**73**	45	45	45	27	27	18	50
	" - conditional I simple	9	67	**100**	89	89	56	**100**	**100**	**100**	**100**	67	56	67	67	0	89	67	67	**100**	33	67	67	67	**100**	78	56	78	44	67	67	72
	" - conditional II progr.	20	70	55	75	70	**80**	55	60	60	**75**	40	35	50	50	0	40	60	35	85	0	55	65	75	75	60	60	70	65	**100**	**100**	59
	" - conditional II simple	2	50	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	50	50	50	50	0	**100**	50	**100**	**100**	0	**100**	**100**	**100**	**100**	**100**	**100**	**100**	50	**100**	**100**	81
	" - future I progr.	12	42	**83**	75	67	25	50	75	75	**75**	50	50	42	42	42	58	50	42	67	25	50	50	50	**83**	42	42	42	17	50	58	52
	" - future I simple	22	64	**95**	64	64	59	77	91	**95**	36	50	41	18	18	18	**82**	50	50	55	23	68	68	68	68	45	45	59	36	64	**82**	57
	" - future II simple	39	62	**92**	59	72	67	85	90	90	**82**	**82**	64	72	72	3	69	46	79	69	10	77	**82**	**82**	69	74	67	72	38	67	54	67
	" - past perfect progr.	16	50	69	50	56	**81**	**81**	69	69	62	**75**	38	38	38	6	**75**	56	62	**75**	25	44	62	56	69	31	44	62	38	62	38	55
	" - past perfect simple	9	44	78	**89**	78	33	**89**	78	78	**100**	56	89	78	78	0	56	44	**100**	**89**	67	33	44	44	**89**	67	44	78	78	78	44	66
	" - present perfect progr.	5	20	**80**	**80**	**80**	20	**80**	**80**	**80**	40	40	**100**	60	60	0	**100**	20	20	60	60	20	20	20	**100**	60	60	60	20	60	20	52
	" - present perfect simple	9	33	67	**78**	56	44	**78**	67	67	78	33	**100**	78	78	0	**100**	44	89	**78**	44	22	33	33	**78**	67	44	67	22	67	44	58
	" - present progr.	10	30	**70**	20	30	30	40	50	50	**50**	40	40	40	40	20	40	0	20	30	20	40	40	40	40	40	40	**50**	**50**	30	40	37
	" - simple past	23	61	43	**96**	78	35	57	48	52	87	52	61	57	57	13	**91**	61	78	87	52	30	57	65	78	87	78	**91**	70	83	**91**	65
	" - simple present	16	31	62	38	44	**69**	62	56	56	94	44	31	31	31	44	**100**	50	62	**81**	31	31	38	38	69	50	44	50	25	62	62	51
Verb valency	Case government	57	**82**	67	75	79	**82**	70	70	75	**86**	79	72	77	68	75	72	44	74	**82**	65	63	77	77	63	81	**82**	77	75	81	79	74

Table 5: Accuracy of the metrics(%) with regards to the linguistically-motivated phenomena for German-English

(Continued on next page)

ling. category	ling. phenomenon	#	baselines								QE as a metric									ref. based metrics												avg
			BERTScore	BLEU	BLEURT-20	COMET-20	YiSi-1	chrF	f101spBLEU	f200spBLEU	COMETKiwi	Cross-QE	HWTSC-TLM	HWTSC-TS	KG-BERT	MATESE-QE	MS-COMET-QE	REUSE	UniTE-src	COMET-22	MATESE	MEE	MEE2	MEE4	MS-COMET	UniTE-ref	UniTE	XL-DA	XL-MQM	xxl-DA19	xxl-MQM20	
	Catenative verb	177	69	58	**86**	61	70	62	60	60	**77**	67	71	71	71	25	60	28	60	76	29	62	64	62	65	68	70	67	72	**89**	**89**	64
	Middle voice	29	90	69	**93**	**93**	79	83	83	83	79	76	**90**	83	83	21	48	31	62	83	45	83	90	**97**	**97**	**97**	**97**	**97**	93	93	86	79
	Passive voice	70	64	51	67	**74**	66	71	53	61	**87**	74	76	71	71	21	70	47	70	**87**	43	50	61	63	86	71	70	79	71	76	77	67
	Resultative	147	76	74	**90**	85	86	80	73	80	84	80	45	61	59	24	84	76	**88**	88	48	63	73	76	87	**92**	**91**	89	87	84	73	76
macro avg.		8945	67	65	**74**	**74**	70	70	66	67	**75**	72	60	62	61	35	69	53	71	**79**	39	65	70	70	72	74	75	**77**	73	**78**	74	68
micro avg.		8945	70	65	**76**	74	73	69	68	68	**73**	**74**	63	65	64	38	67	48	71	78	42	68	71	72	72	77	77	**79**	77	**78**	76	69

Table 5: Accuracy of the metrics(%) with regards to the linguistically-motivated phenomena for English-German