Challenging the state-of-the-art Machine Translation Metrics from a Linguistic Perspective

Eleftherios Avramidis, Shushana Manakhimova, Vivien Macketanz and Sebastian Möller
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 8th Conference for Machine Translation. The challenge set includes about 21,000 items extracted from 155 machine translation systems for three language directions (German⇔English, English→Russian), covering more than 100 linguistically-motivated phenomena organized in 14 categories. The metrics that have the best performance with regard to our linguistically motivated analysis are the COMETOID22-WMT23 (a trained metric based on distillation) for German-English and METRICX-23-C (based on a fine-tuned mT5 encoder-decoder language model) for English-German and English-Russian. Some of the most difficult phenomena are *passive voice* for German-English, *named entities, terminology* and *measurement units* for English-German and *focus particles, adverbial clause* and *stripping* for English-Russian.

1 Introduction

Most NLP evaluation has relied for years on testing the system performance on randomly picked test sets and producing a single generic score. Yet, machine learned systems learn to make abstractions and due to these, phenomena who are on the long tail of the training and test data may be overlooked hidden behind a very high generic score. Additionally, generic scores are often helpful to show relative improvement and reflect overall quality, but cannot explain the performance in a comprehensive way.

For example, old-style machine translation (MT) metrics measuring lexical overlap would equally penalize the omission of an article and the omission of the particle forming the negation in a sentence, although negation is more crucial for its meaning. While the evaluation of so obvious errors has been addressed by the trained MT metrics, their evaluation relies on correlations with human judgments on randomly picked test-sets. In this case, a single correlation score may not be able to explain the strengths and weaknesses of the metrics with regard to the functioning of language.

Motivated by these considerations, we employ a multifold test set with linguistically-motivated challenges that will allow us to understand the metric performance from a linguistic perspective. These challenges are organized in smaller sets, one set per phenomenon, whereas the phenomena are organized in broader categories. By measuring the ability of the metrics to detect the errors in these challenge sets, we can get scores that indicate different aspects of linguistic performance.

This paper describes the application of such a challenge set on the evaluation of the MT metrics submitted at the relevant shared task of the 8th Conference of Machine Translation (Freitag et al., 2023). The rest of the paper is structured as following: Section 2 describes related work, and section 3 describes the way the challenges were selected. In Section 4 the results are presented and described, first from the perspective of metric comparison and then focusing on the performance for particular linguistically-motivated categories and phenomena per language direction. Some conclusions are given in Section 5.

2 Related work

There has been a growing interest for more fine-grained evaluation of Natural Language Processing (NLP) tools, as shown by the increasing number of publications many of whom have received distinctions (Ribeiro et al., 2020; Avelino et al., 2022; Campolungo et al., 2022). Concerning machine translation (MT), initial efforts were made in the 1990s with the introduction of test suites (King and Falkedal, 1990), and these efforts have been revitalized in light of recent advancements in the

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 713–729
December 6–7, 2023.

field (Guillou and Hardmeier, 2016). To the best of our knowledge, the first endeavours related to the use of challenge sets in a meta-level in order to evaluate MT metrics were applied to Quality Estimation metrics (Avramidis et al., 2018), based on the first version of our linguistically-motivated test suite (Macketanz et al., 2018). The analysis was broadened to cover a broader range of MT metrics, including reference-based ones, as appeared in the Findings paper of the Metrics shared task of the 6th Conference on Machine Translation (Freitag et al., 2021), which was based on a later version of our test suite on German-English (Avramidis et al., 2019, 2020; Macketanz et al., 2021, 2022a), a resource also employed in this paper.

With the occasion of the first challenge set subtask for the metrics shared task of the 7th Conference on Machine Translation (Freitag et al., 2022), a few more challenge sets emerged. ACES (Amrhein et al., 2022) for example, focuses on 68 accuracy errors. Similarly, Alves et al. (2022) evaluate the robustness of MT metrics by generating translations with critical errors. In a more linguistic direction, Chen et al. (2022) examine the capability of the metrics to correlate synonyms in different areas and to discern catastrophic errors at both word- and sentence-levels.

Our submission at that sub-task (Avramidis and Macketanz, 2022) augmented the preliminary analysis appearing at Freitag et al. (2021) by adding the language direction of English-German and presenting a more fine-grained analysis, not only in the category level but also on the phenomenon level. This year's submission, explained on our paper, includes that same challenge set as last year, whereas English-Russian has been added as an additional language direction.

3 Method

3.1 Test suite for MT systems

Here, we are going to explain how we created the pool of MT sentences that were used for the challenge set. The selection was based on a linguistically-motivated test suite (Macketanz et al., 2022a)[1]. The test suite contains a set of source sentences focusing on particular phenomena, each of them accompanied by some rules or regular expressions that can detect which translations would be accepted for these source sentences. This allows a semi-automatic evaluation when new translations are provided, whereas a human annotator resolves cases not covered by the rules.

For this experiment, we employed the test suite on three language directions: German-English (Avramidis et al., 2020), English-German (Macketanz et al., 2021) and English-Russian (Macketanz et al., 2022b). The German-English side consists of 5,539 German test sentences covering 107 linguistically motivated phenomena, the English-German side consists of 4,782 English test sentences covering 126 phenomena, and the English-Russian side consists of 1,225 English test sentences covering 64 phenomena. All language directions are organized in 14 categories, which nevertheless differ among the directions.

The above described test suite has been used to evaluate the outputs of 116 German-English, 29 English-German systems and 10 English-Russian systems submitted at the translation task of the Conference of Machine Translation (WMT). German-English outputs were collected from systems submitted in the years 2018-2021, English-German outputs in the years 2020-2021 and English-Russian in 2022.

3.2 Challenge set for MT metrics

The sentences selected with the help of the test suite are consequently used to create the challenge set. The source sentences and the system outputs have to be organized in contrastive pairs of correct/incorrect translations and a reference. In order to achieve this, for every source sentence from the test suite selection we create a challenge item including:

- one correct translation to be used as a reference translation,
- another correct translation to be used as the first translation candidate
- one incorrect translation to be used as the contrastive translation candidate

The two candidate translations and the reference consist one challenge item. Since source and translations were collected as a result of testing for a particular phenomenon, the same phenomenon will be what the challenge item will test.

Given that we may have many correct and wrong translations for the same source, the reference and the translations of the challenge items result from random combinations of correct and wrong translations from the collected WMT outputs. Therefore,

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

the same source sentence may appear many times.

As a result, we get a challenge set with 10,402 items for German-English, 8,945 items for English-German and 1,727 items for English-Russian.

3.3 Evaluation of metrics

For each challenge item, the two machine translation (MT) outputs, are provided to the metrics as separate MT hypotheses. Which output is correct, and which is incorrect, is hidden from the metrics. These hypotheses are then evaluated against the previously mentioned reference and/or the source. An item is deemed correctly scored when the metric assigns a higher score to the correct MT output compared to the incorrect one. Following this, the statistics below are computed:

i) **Accuracy per Phenomenon**: the ratio of all correctly-scored challenge items per phenomenon to the total number of challenge items for that particular phenomenon.

ii) **Accuracy per Category**: the ratio of all correctly-scored challenge items per category to the total number of challenge items for that category, after consolidating the underlying phenomena of that category into a single set.

Significance tests are performed to compare the highest metric accuracy for each phenomenon with all other metric accuracies for the same phenomenon. This is a one-tailed Z-test, conducted with a significance level of $\alpha = 0.95$. Metrics with accuracies that are not significantly worse than the highest accuracy are considered to share the top position for that phenomenon. A similar approach is used to identify the best accuracies per category, after aggregating the challenge items from the underlying phenomena within each category.

Metric categories We conduct this significance testing in two stages: first, for all metrics involved in the shared task, and then separately for each of the three metric categories (baseline, Quality Estimation (QE) as a metric, reference-based). Systems that are significantly superior per phenomenon across all metrics are highlighted with a gray background, while those that are significantly superior per metric category are denoted in boldface.

Averaging Lastly, we provide three types of averaging scores:

i) **Micro-average:** This approach considers all items equally, aggregating all test items to compute the average percentages.

ii) **Category macro-average:** Here, all categories are treated equally, with the percentages being computed independently for each category and then averaged.

iii) **Phenomenon macro-average:** This average treats all phenomena equally, with the percentages being computed independently for each phenomenon and then averaged.

4 Results

The results are displayed in detail in Tables 1, 2 and 3 for the category level and in Tables 4, 5 and 6 for the phenomenon level, for the three language directions 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 accuracies of the metrics, as measured for several categories in German-English, can be seen in Table 1. The best performing metric for German-English is COMETOID22-WMT23 (Gowda et al., 2023), which, wins significantly based on both the micro-average (83%) and the macro-average (87%). This metric is a distilled QE model that has been trained on COMET (Rei et al., 2020) scores of WMT outputs, including the ones of WMT23. For this reason, we include it into the reference-aware metrics. We notice that its performance among the other metrics is impressive. It is the first metric in 6 categories and among them the only one who wins at *Verb tense/aspect/mood* and *function words*, achieving 93% and 91% accuracy respectively.

Another two reference-based baseline metrics, COMET and PRISMREF (Thompson and Post, 2020a,b) share the first position when the category macro-average is considered (82%). None of the other reference-aware metrics submitted this year managed to compete with the metrics with the highest accuracy mentioned above.

The lowest performing metric is the reference-less random baseline RANDOM-SYSNAME, provided by the organizers (44%), followed by XL-SIMQE (55-58%; Mukherjee and Shrivastava,

2023) and MATESE (57-58%; Perrella et al., 2022).

When considering the metric performance with regard to particular categories, one can see, again this year, that different metrics win in different combinations of categories. Here, only COMETOID22-WMT23 as mentioned above, wins 6 metrics, followed by PRISMREF and METRIC-23-C, which win 4 categories. 17 metrics do not win any category, ranging in accuracies around 75%.

English-German The accuracies of the metrics, as measured for several categories in English-German, can be seen in Table 2. The best performing metric in English-German is METRICX-23-C, which is in the first significance cluster based on both the micro-average (81%) and the category macro-average (84%). This metric uses the mT5 encoder-decoder language model, which is fine-tuned using direct assessment data, MQM (Lommel et al., 2014) data and synthetic data. The categories to which its success may be mostly attributed are the *multi-word expressions* (MWE; with 92%) and the non-verbal agreement (95%).

Another three metrics share the first position, when the micro average is considered, namely the QE version of the latter, *MetricX-23-QE-c* and also *mbr-metricx-qe* (Naskar et al., 2023) and XCOMET-Ensemble. It is impressive that QE methods manage to reach high accuracy without access to reference content.

When looking at the worse-performing metrics, MATESE here performs worse than the baseline (36-38%), followed by PARTOKENGRAM_F (55-56%; Dreano et al., 2023b).

In English-German it is even harder to say which metrics perform well in multiple categories, as only one of them, XCOMET-QE-ENSEMBLE, achieves the highest performance in 3 categories (*function words, non-verbal agreement and subordination*). The rest of the metrics show a good performance in 2 categories or fewer.

English-Russian The accuracies of the metrics, as measured for several categories in English-German, can be seen in Table 3. For this language pair, variants of the MetricX achieve significantly higher accuracies than all the other metrics. In particular, METRICX-23-C achieves the highest accuracy based on both micro-average and category macro-average, whereas METRICX-23-B and METRICX-23-QE-C achieve a slightly lower macro-average, which is nevertheless not significantly worse than the one of the former. MATESE is again by far the lowest performing metric (32/34%), lower than the random baseline. We may assume that this metric has not been optimized for this language direction.

4.2 Linguistically motivated analysis

In this section, we are focusing on the results for particular phenomena or categories.

4.2.1 German-English

Category-level The overall average accuracy of all metrics with regard to the linguistically motivated categories is at 76% for German-English, which is two percentage points lower than last year's average. It is still a fact, that the metrics in average fail to predict properly the scores for one out of four challenge items that we provided. Luckily, there has been noticeable accuracy for some categories, for example METRICX reference-based variants achieved an accuracy of 96% for *false friends*, whereas *negation* errors have been scored correctly with a 98.5

The worse performing category is *Verb valency*, where the best metrics achieved only 66% accuracy, and the rest of the metrics averaged to a mere 56%. In this category one can observe the lowest accuracy, given by an LLM-based metric, EMBED_LLAMA (Dreano et al., 2023a) with 41%.

Phenomenon-level The best accuracy for this language pair (Table 4) is achieved this year at several variations of verb tenses, i.e. *Transitive - future II, Modal negated - present, Reflexive - preterite subjunctive II* and *Intransitive - pluperfect* which get more than 85% in average.

The lowest accuracy of all metrics in average is given for *passive voice*, where the highest accuracy achieved by several metrics is only 54.5%. Errors related to *commas, domain-specific terms* and *locations* have also been scored with a less than 65% accuracy.

4.2.2 English-German

Category-level The overall average accuracy of all metrics with regard to the linguistically motivated categories is at 71-73% for English-German. The category where all metrics perform better in average is *negation* (83%), where 11 of the metrics achieve more than 90% accuracy. Negation is closely followed by *function words Non-verbal agreement* (80%).

The worse performing category in average is *named entities and terminology* (58,8%), where most metrics' accuracies are close to 50%, except for BLEURT (Yan et al., 2023) (80.3%). The rest of the categories lie in rather mediocre accuracies, between 58.8% and 80%.

Phenomenon-level The English-German phenomena, where metrics perform best in average (Table 5) are the *transitive conditional II simple, gerunds, contact clause* and the *intransitive present perfect simple*, achieving more than 85% of accuracy. The phenomena which incur the lowest average accuracies are the *transitive present progressive, measuring units*, *modals* and *intransitive - future II progressive* with less than 50% accuracy. The former and the latter were observed as the most difficult phenomena to score also last year.

4.3 English-Russian

This analysis for English-Russian occurs for the first time this year, based on the MT outputs collected at last year's shared task. For this purpose the test instances are much fewer than the other language pairs and therefore the numbers are not very conclusive. Therefore, categories and phenomena that have only a handful of samples will not be included in our analysis, although they appear in the tables.

Category-level Here, the average accuracy over all metrics is much lower than the other language directions, reaching only 66%, only 20% above the random baseline. The best performing category is *ambiguity* (86,3%), more than 13% better than the following categories. The worst performing categories are *function words* and *punctuation*, with less than 55%. The rest of the categories range in accuracies between 53 and 73%.

Phenomenon-level The good performance of the *ambiguity* category is also confirmed in the table on the phenomenon level (Table 6), as in Russian this is the only phenomenon of this category, as opposed to other language pairs where we also have examples of structural ambiguity. The most difficult phenomena to score appear to be the *focus particles, adverbial clause* and *stripping* with less than 50% average accuracy, in many cases lower than the random baseline.

5 Conclusion

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

The metrics demonstrating the best performance in average are COMETOID22-WMT23 for the German-English language pair, and METRICX-23-C for both the English-German and English-Russian language pairs. Quality estimation methods have impressively good performance in several phenomena. Some metrics that report usage of LLMs (EMBED_LLAMA) have not scored very high in overall, indicating that more work is required in this direction.

Among the various linguistic phenomena, we could identify some of the particularly challenging ones. In German-English, metrics have difficulties scoring the *passive voice* properly. In English-German *named entities and terminology* as well as specific *measurement units* pose the most difficulties. In English-Russian translation, translations with *focus particles*, *adverbial clause*, and *stripping* phenomena emerge as particularly complex challenges.

Acknowledgements

This research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft; DFG) through the project TextQ and by the German Federal Ministry of Education (BMBF) through the project SocialWear (grant num. d01IW20002). We would like to thank Hans Uszkoreit, Aljoscha Burchardt, Ursula Strohriegel, Renlong Ai, He Wang, Ekaterina Lapshinova-Koltunski and Sergei Bagdasarov for their prior contributions for the creation of the test suite.

References

Duarte Alves, Ricardo Rei, Ana C Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT evaluation with sentence-level multilingual augmentation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 469–478, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Eleftherios Avramidis and Vivien Macketanz. 2022. Linguistically motivated evaluation of machine translation metrics based on a challenge set. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 514–529, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Arle Lommel, and Hans Uszkoreit. 2018. Fine-grained evaluation of quality estimation for machine translation based on a linguistically motivated test suite. In *Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing*, pages 243–248, Boston, MA. Association for Machine Translation in the Americas.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, Aljoscha Burchardt, and Sebastian Möller. 2020. Fine-grained linguistic evaluation for state-of-the-art machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 346–356, Online. Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit. 2019. Linguistic evaluation of German-English machine translation using a test suite. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 445–454, Florence, Italy. Association for Computational Linguistics.

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.

Xiaoyu Chen, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Ting Zhu, Mengli Zhu, Ning Xie, Lizhi Lei, Shimin Tao, Hao Yang, and Ying Qin. 2022. Exploring robustness of machine translation metrics: A study of twenty-two automatic metrics in the WMT22 metric task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 530–540, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Sören Dreano, Derek Molloy, and Noel Murphy. 2023a. Embed_Llama: using LLM embeddings for the Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Sören Dreano, Derek Molloy, and Noel Murphy. 2023b. Tokengram_F, a fast and accurate token-based chrF++ derivative. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thomson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Thamme Gowda, Tom Kocmi, and Marcin Junczys-Dowmunt. 2023. Cometoid: Distilling Strong Reference-based Machine Translation Metrics into Even Stronger Quality Estimation Metrics. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. *Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

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.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 165–172, Dubrovnik, Croatia. European Association for Machine Translation.

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: Shared Task Papers*, pages 578–587, Belgium, Brussels. Association for Computational Linguistics.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, and Hans Uszkoreit. 2022a. A linguistically motivated test suite to semi-automatically evaluate German–English machine translation output. In *Proceedings of the 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*, pages 1059–1073, Online. Association for Computational Linguistics.

Vivien Macketanz, Shushen Manakhimova, Eleftherios Avramidis, Ekaterina Lapshinova-koltunski, Sergei Bagdasarov, and Sebastian Möller. 2022b. Linguistically motivated evaluation of the 2022 state-of-the-art machine translation systems for three language directions. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 432–449, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ananya Mukherjee and Manish Shrivastava. 2023. MEE4 and XLsim : IIIT HYD's Submissions' for WMT23 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Subhajit Naskar, Daniel Deutsch, and Markus Freitag. 2023. Quality Estimation using Minimum Bayes Risk. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. MaTESe: Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. 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.

Brian Thompson and Matt Post. 2020a. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020b. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

Yiming Yan, Tao Wang, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Mingxuan Wang. 2023. BLEURT has universal translations: An analysis of automatic metrics by minimum risk training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5428–5443, Toronto, Canada. Association for Computational Linguistics.

		baselines								QE as a metric													ref. based metrics															
ling. category	\#	BERTscore	BLEU	BLEURT-20	COMET	YiSi-1	chrF	prismRef	spBLEU	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	prismSrc2	KG-BERT	MEE4	MEE4_stsb_xlm	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	cometoid22-wmt23	eBLEU	embed_llama	partokengram_F	tokengram_F	avg
Ambiguity	298	87	71	89	**90**	**90**	81	**91**	79	78	**90**	85	77	82	78	86	80	46	87	61	86	85	66	85	83	88	74	87	90	90	88	89	**92**	84	**91**	74	56	48	80	81
Composition	252	**90**	66	88	87	88	74	86	71	83	87	80	81	75	85	**90**	81	44	86	40	81	81	70	80	85	86	69	85	**90**	85	87	84	83	79	87	74	65	61	75	79
Coordination & ellipsis	316	81	74	81	78	**83**	78	**82**	74	**83**	75	**84**	69	67	78	77	78	45	79	59	75	73	79	**84**	81	77	49	81	**82**	81	81	79	70	77	81	73	68	58	78	75
False friends	90	90	64	**91**	**92**	88	76	84	69	82	82	82	77	72	**89**	**90**	88	47	80	51	87	83	86	82	87	80	73	**96**	**96**	**96**	84	84	74	74	93	72	68	51	72	80
Function word	586	82	73	**83**	**83**	82	73	**84**	73	84	**87**	76	77	68	80	86	80	47	86	60	84	**87**	81	76	79	76	57	87	83	86	87	83	72	80	**91**	79	66	55	75	78
LDD & interrogatives	1014	84	74	85	**87**	84	76	**86**	74	80	81	**84**	72	81	**83**	**83**	**82**	45	80	61	81	79	72	84	81	76	53	82	83	82	84	82	72	77	**87**	76	68	54	77	77
MWE	610	82	72	84	84	**85**	76	**86**	73	74	83	77	73	59	80	**87**	80	45	83	53	72	73	65	76	80	81	67	84	86	**87**	84	84	80	75	85	75	66	54	76	76
Named entity & termin.	861	73	64	69	74	**76**	69	73	72	65	67	66	61	62	65	**70**	66	44	67	59	66	67	**69**	66	72	70	58	72	74	72	72	**75**	66	69	71	64	57	52	70	67
Negation	76	79	78	91	**93**	84	82	84	76	91	91	93	75	84	91	**95**	86	50	93	55	91	89	92	93	86	84	51	91	**95**	92	**95**	91	79	76	91	75	59	45	82	82
Non-verbal agreement	419	78	78	**85**	**84**	78	78	79	79	78	**80**	77	67	78	78	**81**	79	43	**80**	64	77	79	74	76	78	75	59	81	**84**	81	81	77	72	70	81	76	60	74	77	76
Punctuation	293	75	74	69	69	70	65	**80**	76	62	62	63	59	59	59	61	60	39	65	67	66	66	**86**	63	**76**	67	48	62	55	62	66	66	44	69	67	65	66	67	67	65
Subordination	679	75	70	76	**78**	75	70	**78**	69	**79**	**78**	**78**	64	65	77	77	**78**	43	**79**	53	75	75	68	**78**	77	75	50	77	78	77	**79**	75	64	70	**79**	73	62	71	71	72
Verb tense/aspect/mood	4697	88	69	85	88	**89**	76	**90**	71	83	84	85	84	81	85	**87**	81	43	82	60	74	75	80	85	84	84	56	82	87	83	85	83	78	78	**93**	76	62	77	77	79
Verb valency	211	**64**	54	63	**64**	64	55	**65**	52	**59**	54	**59**	45	51	**60**	58	55	39	57	33	57	**59**	48	59	62	56	44	62	63	61	62	61	53	61	**66**	55	41	56	56	56
macro avg.	10402	81	70	**81**	**82**	**81**	73	**82**	72	77	79	78	70	70	78	**81**	77	44	79	55	76	77	74	78	79	77	58	81	82	81	81	79	71	74	**83**	72	62	59	74	74
micro avg.	10402	83	70	82	**84**	**84**	75	**85**	72	79	80	80	75	74	80	**83**	78	44	80	58	75	76	75	80	81	79	57	81	83	81	82	81	74	76	**87**	74	63	67	75	76

Table 1: Accuracy of the metrics (%) with regard 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.

ling. category	#	baselines								QE as a metric																			ref. based metrics																avg
		BERTscore	BLEU	BLEURT-20	COMET	YiSi-1	chrF	prismRef	spBLEU	Calibri-COMET22-QE	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	KG-BERT	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	cometoid22-wmt23	mbr-metricx-qe0p2p1-qe	mbr-metricx-qexv1p-qe	prismSrc2	Calibri-COMET22	MEE4	MEE4_stsb_xlm	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	eBLEU	embed_llama	mt0regressor	partokengram_F	tokengram_F	
Ambiguity	146	84	71	90	84	86	88	**97**	88	42	77	60	42	77	42	39	75	85	82	46	56	38	59	58	87	86	**95**	14	84	90	92	50	88	**95**	88	82	88	62	**93**	70	62	81	62	88	73
Coordination & ellipsis	836	69	61	**80**	78	69	61	74	63	73	74	75	74	60	74	71	78	79	80	48	77	38	75	74	76	**83**	80	58	75	62	63	33	76	**81**	72	79	**81**	64	75	61	56	72	49	62	69
False friends	225	65	63	69	**74**	71	71	67	65	71	61	75	71	65	71	73	64	80	66	46	76	**82**	64	52	66	76	63	68	74	83	**87**	30	60	82	62	65	59	53	76	66	55	56	52	68	67
Function word	200	90	78	79	90	82	74	**93**	74	92	**94**	92	90	82	90	78	86	72	85	36	**95**	62	**94**	94	94	90	67	84	86	82	76	60	86	86	84	**94**	86	64	76	70	52	76	60	76	80
MWE	829	79	72	87	**90**	86	77	85	74	74	86	79	73	66	73	71	88	89	85	45	83	37	78	78	**90**	88	89	47	89	76	78	18	86	**92**	85	90	**93**	76	81	69	69	72	54	78	76
Named entity & termin.	1272	57	54	**66**	64	62	61	**67**	62	54	51	49	55	56	54	55	61	65	58	44	54	46	60	54	68	62	**80**	56	62	60	63	30	72	**74**	71	62	60	60	71	60	48	54	49	60	59
Negation	174	87	83	89	**92**	**92**	84	84	86	90	90	84	89	76	89	92	83	88	78	47	84	40	90	90	**96**	89	91	90	**91**	**91**	**91**	58	85	**91**	80	84	85	79	80	86	67	88	52	86	83
Non-verbal agreement	372	74	72	81	**88**	78	70	83	75	82	90	84	81	83	81	78	**95**	93	91	45	**95**	43	93	91	93	82	94	71	87	76	69	57	90	**95**	87	**94**	91	88	73	72	65	84	54	69	80
Punctuation	336	69	73	74	70	66	72	**77**	69	**82**	72	76	**82**	60	**82**	65	73	66	73	42	60	44	81	**82**	**81**	75	70	75	65	**74**	72	28	**74**	70	67	65	65	46	68	**74**	49	57	44	73	68
Subordination	994	78	74	81	84	78	75	**86**	74	**89**	**88**	**88**	**89**	76	**89**	80	83	85	83	42	**89**	43	83	83	86	83	84	76	83	79	78	16	83	84	83	**89**	85	79	75	75	71	75	56	76	78
Verb tense/aspect/mood	3081	68	62	**70**	**70**	**69**	69	67	64	71	72	75	75	71	75	61	82	**84**	**85**	43	83	52	62	65	75	71	77	61	70	70	74	46	77	77	78	**83**	76	71	69	72	55	71	62	69	70
Verb valency	480	73	64	**82**	79	74	70	79	69	77	81	85	77	62	77	71	86	80	86	42	87	52	79	80	77	**87**	**88**	66	77	70	70	36	**84**	81	**85**	**85**	80	75	79	66	69	65	70	70	74
macro avg.	8945	74	69	**79**	**80**	76	73	**80**	72	75	78	77	75	70	75	70	80	80	79	44	78	48	76	75	**82**	**81**	**81**	64	79	76	76	38	80	**84**	79	81	79	68	76	70	60	71	55	73	73
micro avg.	8945	70	65	**75**	**76**	72	69	74	68	73	74	75	74	68	74	67	79	**81**	80	44	78	47	71	71	78	77	**81**	62	74	71	73	36	79	**81**	78	**81**	77	70	73	69	59	69	56	70	71

Table 2: Accuracy of the metrics (%) with regard to the 12 linguistically motivated categories for English-German

		baselines								QE as a metric																				ref. based metrics											
ling. category	#	BERTscore	BLEU	BLEURT-20	COMET	YiSi-1	chrF	prismRef	spBLEU	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	KG-BERT	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	cometoid22-wmt23	prismSrc2	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	eBLEU	embed_llama	partokengram_F	tokengram_F	avg
Ambiguity	66	80	67	**100**	**97**	88	67	83	73	**100**	94	**100**	**100**	**100**	95	**100**	**100**	**100**	39	**100**	30	**100**	**100**	**100**	56	**98**	**100**	**100**	**100**	**100**	**100**	**100**	88	77	44	59	70	86
Coordination & ellipsis	203	**84**	73	68	73	76	67	77	69	73	64	68	38	68	**76**	62	**79**	64	45	64	41	70	67	69	55	21	61	69	56	67	64	50	74	**80**	60	57	67	64
False friends	6	**100**	17	**100**	**100**	**100**	17	83	50	0	0	0	**100**	0	0	0	**100**	17	17	0	17	33	33	**100**	0	17	**100**	**100**	**100**	33	33	**100**	83	**100**	50	83	33	50
Function word	116	59	**69**	**69**	59	53	38	48	**72**	53	55	50	51	50	36	66	55	48	44	**72**	39	**73**	64	54	28	9	**63**	57	**60**	**60**	**60**	22	49	**61**	52	53	38	53
MWE	122	70	64	**79**	**80**	74	64	**79**	67	72	70	77	57	77	77	75	**88**	75	40	67	35	75	70	77	23	57	87	**91**	85	70	73	69	75	66	59	59	64	69
Named entity & termin.	243	79	69	**95**	91	79	76	85	76	84	62	59	54	59	69	63	73	73	49	64	53	85	86	**95**	31	51	**89**	86	86	**91**	**91**	71	84	75	65	58	75	73
Negation	34	74	74	79	68	74	**85**	68	74	68	85	85	91	85	41	71	**100**	9	59	91	9	62	29	59	91	0	**94**	71	**94**	79	71	74	65	76	76	85	85	69
Non-verbal agreement	61	69	56	**82**	77	67	67	61	57	84	**100**	54	72	54	74	72	**98**	66	41	77	26	95	**100**	85	62	43	**92**	**92**	**92**	85	79	31	64	61	66	69	69	70
Punctuation	121	42	46	53	68	**86**	41	73	55	66	29	65	27	65	**87**	55	66	43	41	61	44	66	67	36	56	1	60	71	52	66	57	12	63	66	**84**	45	45	54
Subordination	499	62	62	61	**70**	**67**	66	63	63	72	66	66	41	66	**75**	65	66	61	45	**77**	32	68	61	65	37	18	60	**84**	58	80	70	52	63	54	57	65	65	61
Verb tense/aspect/mood	135	79	62	76	**85**	76	**84**	81	74	79	79	73	61	73	64	67	65	66	39	74	60	70	59	**85**	38	50	90	**94**	90	85	**92**	74	79	70	73	82	82	73
Verb valency	121	70	67	70	**75**	**76**	67	56	68	71	**88**	79	66	79	71	76	85	76	36	76	26	63	61	65	25	45	79	81	78	72	70	65	**84**	76	64	69	69	68
macro avg.	1727	72	60	**78**	**79**	**76**	61	71	66	69	66	65	63	65	64	64	**81**	58	41	69	34	72	66	74	42	34	**81**	**83**	79	74	72	60	73	72	62	65	64	66
micro avg.	1727	69	64	72	**76**	73	65	70	67	**74**	67	68	51	68	**71**	67	**74**	64	44	**72**	39	**73**	68	**71**	40	32	73	**82**	71	77	74	55	71	67	62	63	66	66

Table 3: Accuracy of the metrics (%) with regard to the linguistically motivated categories for English-Russian

ling. category	ling. phenomenon	#	baselines: BERTscore	BLEU	BLEURT-20	COMET	YiSi-1	chrF	prismRef	spBLEU	QE as a metric: CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	prismSrc2	ref. based metrics: KG-BERT	MEE4	MEE4_stsb_xlm	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	cometoid22-wmt23	eBLEU	embed_llama	partokengram_F	tokengram_F	avg
Ambiguity	Lexical ambiguity	129	91	75	95	**97**	89	88	95	83	81	**97**	88	79	82	81	91	84	50	91	51	84	87	63	88	84	**97**	83	91	95	95	95	93	96	94	**98**	78	73	49	86	85
	Structural ambiguity	169	85	69	84	84	**90**	75	88	76	75	84	82	75	82	76	82	77	43	84	68	**87**	84	69	82	82	80	67	83	86	85	83	86	**89**	76	86	71	44	47	75	77
Composition	Compound	129	91	64	**93**	90	**92**	71	90	71	80	85	77	79	70	81	**91**	80	43	84	29	74	74	64	77	91	90	70	87	**95**	87	88	86	89	80	86	71	61	57	72	78
	Phrasal verb	123	**89**	67	82	85	85	77	81	72	87	**89**	84	82	80	**89**	**89**	83	46	87	50	87	87	76	84	78	82	67	82	85	83	86	81	76	77	**89**	77	69	66	79	79
Coordination & ellipsis	Gapping	51	86	69	88	88	84	75	**90**	67	**96**	**96**	92	75	78	88	94	90	49	94	51	82	86	80	92	84	78	67	**96**	90	94	94	94	86	73	94	80	65	57	73	82
	Right node raising	67	81	69	72	69	**84**	73	**84**	69	75	57	**84**	76	61	75	58	69	52	66	42	55	52	76	**84**	79	69	45	81	78	73	73	72	64	70	75	72	70	55	73	69
	Sluicing	128	80	81	80	79	81	**84**	81	80	82	74	**85**	69	68	77	77	78	45	79	73	79	77	78	**85**	82	77	49	79	82	80	80	80	69	84	77	73	69	59	**84**	76
	Stripping	70	80	71	**84**	76	**84**	74	77	73	**84**	79	74	59	64	73	81	80	37	80	57	81	77	83	74	77	81	40	74	81	79	79	74	64	73	**84**	71	66	59	76	73
False friends	False friends	90	90	64	**91**	**92**	88	76	84	69	82	82	82	77	72	**89**	**90**	88	47	80	51	87	83	86	82	87	80	73	**96**	**96**	**96**	84	84	74	74	93	72	68	51	72	80
Function word	Focus particle	64	**88**	78	84	80	**88**	75	83	78	77	81	81	84	70	86	78	83	47	80	55	78	81	**88**	81	88	80	48	86	84	**89**	81	72	70	81	81	80	73	64	78	78
	Modal particle	166	84	77	84	89	89	77	**90**	81	88	92	72	72	69	84	**95**	86	43	**94**	55	92	93	69	72	86	85	51	89	89	88	91	89	78	84	**94**	81	70	54	76	80
	Question tag	356	80	70	**82**	**81**	78	71	81	69	83	**86**	76	78	67	77	83	77	49	83	63	82	**85**	85	77	74	71	61	86	81	84	86	82	70	78	**92**	78	63	54	74	76
LDD & interrogatives	Extended adjective construction	320	85	78	85	**90**	86	80	88	79	76	80	**88**	70	**88**	85	83	82	49	78	62	81	83	74	88	84	78	59	82	85	82	85	81	78	81	**92**	81	67	53	82	79
	Extraposition	92	77	75	**79**	**80**	77	**79**	78	74	68	68	70	54	66	62	67	66	48	**74**	50	**74**	**74**	49	70	**84**	78	37	76	75	80	73	65	42	75	76	82	74	51	77	69
	Multiple connectors	87	79	79	72	**84**	**84**	77	76	78	**80**	75	76	61	78	71	75	74	48	72	66	72	56	69	76	**78**	71	24	70	70	62	74	77	46	68	**79**	71	67	52	77	70
	Pied-piping	162	**90**	73	**89**	**89**	**88**	73	**89**	71	**87**	**86**	85	76	73	84	85	85	38	85	71	85	81	**87**	85	83	78	56	**89**	86	**88**	86	85	75	78	87	77	67	60	73	80
	Polar question	51	**84**	63	76	76	76	53	76	61	84	82	84	75	86	**88**	78	**88**	43	78	57	84	84	73	84	71	65	47	78	80	78	80	80	59	67	**90**	63	67	53	57	73
	Scrambling	144	81	72	87	87	85	78	**90**	72	81	89	85	78	83	**91**	**91**	86	41	87	58	81	81	67	85	75	77	69	83	**90**	88	88	86	84	79	85	72	61	56	78	79
	Topicalization	61	90	70	**97**	90	89	72	89	70	90	82	**97**	90	89	**97**	92	93	41	85	72	90	92	74	**97**	87	90	59	92	92	90	90	93	82	77	93	75	80	43	75	83
	Wh-movement	97	78	68	85	**87**	79	74	85	69	82	84	75	82	71	**86**	**86**	**85**	49	76	49	78	78	69	75	78	65	52	76	78	78	85	**87**	76	71	**87**	67	76	54	74	75
MWE	Collocation	190	80	71	**86**	**85**	**86**	76	81	74	77	85	76	84	72	87	**92**	82	45	84	55	75	74	67	76	79	79	70	84	**91**	86	86	86	74	80	83	72	62	51	76	77
	Idiom	133	84	72	83	89	88	78	**95**	72	62	77	59	64	32	73	**80**	71	41	78	43	58	59	41	59	80	93	69	80	92	86	80	84	88	75	**96**	81	75	51	77	73
	Prepositional MWE	146	82	77	86	82	85	81	**88**	77	78	85	**87**	71	67	80	84	84	52	**86**	50	79	79	72	87	86	78	66	88	84	**92**	87	79	81	74	81	78	67	57	81	78
	Verbal MWE	141	**82**	67	80	80	81	70	**82**	70	78	83	84	72	61	79	**92**	81	42	83	62	77	79	78	82	74	74	62	**86**	74	**86**	82	84	80	70	82	71	63	57	70	75
Named entity & termin.	Date	203	67	54	67	70	67	61	**77**	61	78	**80**	70	68	64	68	76	70	43	75	67	67	66	75	70	66	66	62	73	**76**	**76**	74	**76**	70	70	72	57	54	50	60	68
	Domainspecific term	214	**73**	60	70	69	**73**	70	64	66	58	63	64	51	63	66	65	62	43	61	62	66	**69**	49	65	70	72	49	72	**76**	71	69	**76**	66	71	69	60	51	50	71	64
	Location	181	72	69	70	**81**	**82**	69	72	77	54	55	60	55	61	58	66	65	48	56	42	62	62	**74**	60	**74**	69	57	65	70	68	63	72	57	70	70	58	60	52	72	64
	Measuring unit	203	79	69	65	72	80	73	76	**84**	66	65	63	62	55	64	71	63	42	70	58	62	66	**77**	64	75	69	61	72	69	68	74	74	66	63	70	**79**	61	54	76	68
	Proper name	60	77	73	83	83	**88**	77	87	75	78	82	83	82	80	82	75	85	47	83	78	85	**87**	83	83	83	83	72	85	85	85	**88**	85	83	82	83	73	65	48	80	79
Negation	Negation	76	79	78	91	**93**	84	82	84	76	91	91	93	75	84	91	**95**	86	50	93	55	91	89	92	93	86	84	51	91	**95**	92	**95**	91	79	76	91	75	59	45	82	82
Non-verbal agreement	Coreference	251	76	71	**89**	87	77	75	77	72	86	**89**	81	72	83	86	**88**	85	43	87	67	84	85	74	81	74	73	62	87	**89**	**88**	87	81	80	64	86	72	57	69	74	78
	External possessor	104	82	88	75	78	81	84	80	**90**	62	65	65	59	**68**	62	**68**	**67**	38	66	59	58	62	**67**	65	83	79	48	67	79	65	69	66	58	72	70	**88**	69	83	83	70
	Internal possessor	64	81	**88**	84	83	80	78	**88**	**88**	73	72	78	58	75	72	73	77	47	75	64	83	80	**89**	77	83	78	66	78	75	78	78	77	66	**88**	78	77	59	80	80	76
Punctuation	Comma	46	80	76	61	67	**83**	74	78	**83**	48	52	61	46	63	46	54	54	35	59	65	70	67	**85**	61	**78**	74	48	52	43	50	57	63	37	61	63	70	65	74	74	63
	Quotation marks	247	74	74	70	69	68	63	**81**	74	64	64	64	62	58	61	62	61	40	66	68	65	66	**86**	64	**76**	66	48	64	57	64	67	66	45	70	67	64	66	66	66	65
Subordination	Adverbial clause	87	62	56	**75**	**75**	69	56	**75**	54	**82**	71	78	54	63	77	69	77	44	69	51	74	74	68	**78**	72	68	49	68	74	64	74	67	49	68	**77**	62	66	56	56	66
	Cleft sentence	109	75	**83**	74	74	69	78	64	80	73	72	72	49	59	68	71	**76**	50	74	69	65	67	55	72	71	68	45	69	69	67	73	72	53	66	71	**79**	72	**78**	**78**	69
	Free relative clause	70	77	64	77	**84**	70	67	79	66	71	**90**	**90**	63	69	71	79	71	46	80	60	76	73	59	**90**	74	79	59	71	74	73	81	74	69	77	80	70	59	67	67	72
	Indirect speech	119	75	68	**79**	**79**	73	73	**80**	66	84	86	74	71	68	82	79	80	43	**89**	43	80	81	79	74	81	80	56	84	82	79	**86**	83	76	69	82	72	61	74	74	75
	Infinitive clause	64	80	80	84	86	84	81	**88**	77	78	78	78	61	62	80	73	**81**	34	73	38	73	72	64	78	**88**	78	56	80	81	80	80	75	78	67	80	77	55	80	80	74
	Object clause	54	**76**	61	56	74	74	59	74	67	69	69	80	76	72	74	**87**	78	46	69	56	76	74	72	**80**	74	72	31	70	74	74	70	57	37	72	78	72	59	63	63	68
	Pseudo-cleft sentence	25	80	60	72	60	72	64	**92**	60	76	72	80	76	20	**92**	84	84	36	84	12	72	68	60	80	80	72	52	**88**	**88**	84	84	**88**	68	64	84	80	44	64	64	70
	Relative clause	71	76	68	**82**	76	80	70	80	72	82	75	79	62	62	79	77	77	35	**83**	48	68	76	63	79	76	72	46	86	83	**92**	82	82	66	73	80	73	63	72	72	73

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	YiSi-1	chrF	prismRef	spBLEU	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	prismSrc2	KG-BERT	MEE4	MEE4_stsb_xlm	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	cometoid22-wmt23	eBLEU	embed_llama	partokengram_F	tokengram_F	
	Subject clause	80	82	78	80	**86**	84	74	**86**	76	**88**	**86**	84	76	80	85	85	84	48	84	71	**86**	**88**	82	84	82	81	49	85	81	**86**	84	80	74	75	**88**	78	59	74	74	79
Verb tense/aspect/mood	Conditional	50	90	68	**92**	**92**	88	66	84	64	90	**96**	**96**	92	86	92	88	88	34	88	64	80	80	82	**96**	74	70	52	88	94	86	92	92	86	78	90	66	40	62	62	80
	Ditransitive - future I	121	89	69	88	87	**91**	74	**90**	71	95	**97**	94	**96**	93	93	95	90	53	92	56	80	81	91	**94**	79	74	63	91	**95**	89	93	92	89	81	93	74	57	74	74	84
	Ditransitive - future I subj. II	84	**96**	62	86	90	94	75	89	68	**100**	94	**100**	**100**	95	**100**	98	95	35	95	52	75	80	94	**100**	79	82	57	80	92	86	96	88	88	63	95	65	57	75	75	83
	Ditransitive - future II	97	**98**	65	84	90	96	72	84	66	93	94	97	94	96	88	**99**	79	41	85	66	77	77	86	97	72	73	68	84	91	87	92	90	82	68	**99**	67	53	71	71	81
	Ditransitive - future II subj. II	88	90	69	88	**94**	**94**	77	91	73	98	98	98	95	92	97	**100**	89	39	95	67	91	90	92	**98**	78	78	81	93	**97**	93	93	92	91	76	93	80	55	76	76	86
	Ditransitive - perfect	72	89	64	78	88	**90**	72	79	62	88	83	83	75	79	**89**	76	83	50	75	58	72	75	72	83	68	75	42	75	81	69	75	71	60	74	**89**	76	69	72	72	75
	Ditransitive - pluperfect	86	**91**	69	72	**90**	**90**	86	78	69	55	66	64	64	72	52	**74**	43	51	55	52	49	52	**76**	64	78	76	43	73	72	71	69	66	45	71	**97**	74	67	87	87	69
	Ditransitive - pluperfect subj. II	107	82	68	**91**	87	83	80	88	70	89	89	84	93	79	93	**95**	82	40	84	55	69	64	93	84	82	79	64	**93**	**93**	**93**	92	90	91	78	92	74	64	79	79	81
	Ditransitive - present	90	87	58	87	86	**89**	77	83	64	88	90	89	**96**	82	92	**97**	92	53	82	51	69	70	69	89	80	81	69	93	88	90	**97**	93	90	79	89	70	69	79	79	81
	Ditransitive - preterite	117	88	62	**93**	**92**	89	82	**93**	66	93	88	92	90	84	**95**	**95**	**94**	48	89	59	83	84	81	92	89	85	81	96	97	96	95	**98**	88	76	**97**	73	68	85	85	85
	Ditransitive - preterite subj. II	110	83	76	93	92	89	83	**95**	76	79	88	84	86	76	90	85	**92**	45	85	65	78	79	72	84	92	89	75	**98**	88	96	92	95	85	77	95	77	70	85	85	84
	Imperative	98	84	59	**94**	93	91	74	**95**	63	**94**	90	**95**	**94**	**94**	**94**	**94**	93	44	**94**	57	**94**	93	84	**95**	90	84	73	**95**	92	90	**95**	91	92	79	**96**	76	59	77	77	85
	Intransitive - future I	32	91	66	94	91	91	75	**100**	75	97	**100**	88	97	78	**100**	97	97	47	97	66	69	81	81	88	94	91	62	**100**	94	**100**	97	88	94	75	94	75	72	78	78	86
	Intransitive - future I subj. II	56	84	61	93	88	91	73	**96**	70	82	82	77	80	62	77	**86**	80	36	**86**	32	66	68	59	77	89	**95**	59	89	**95**	91	84	80	80	77	**95**	70	55	73	73	77
	Intransitive - future II	62	92	69	89	95	**97**	89	**97**	76	92	76	92	69	82	90	**97**	87	44	85	73	71	77	68	92	95	**98**	55	85	94	82	85	82	77	89	**98**	73	52	87	87	83
	Intransitive - future II subj. II	94	88	65	**90**	82	**90**	72	84	63	82	82	83	**88**	71	85	86	82	44	83	54	80	79	73	83	80	83	62	83	**88**	84	83	83	80	79	84	65	54	71	71	77
	Intransitive - perfect	61	85	70	82	84	**89**	82	**89**	69	70	69	77	85	64	**93**	85	82	39	67	61	74	75	61	77	82	**92**	34	77	80	74	69	62	64	80	**92**	87	72	82	82	75
	Intransitive - pluperfect	85	**98**	86	**96**	95	95	88	**98**	82	**92**	86	**93**	**92**	73	88	91	88	44	87	62	79	82	61	93	96	**99**	74	92	95	94	88	89	91	94	**98**	85	62	91	91	87
	Intransitive - pluperfect subj. II	79	91	75	92	96	**99**	78	**97**	72	92	87	94	**97**	86	**97**	95	**96**	43	87	48	77	78	84	94	95	94	59	90	**96**	94	**97**	95	94	78	**97**	78	57	81	81	85
	Intransitive - present	54	72	50	85	83	74	61	**87**	57	81	74	**91**	85	85	87	78	74	46	74	76	83	89	74	91	76	74	46	87	91	83	81	80	85	70	**93**	67	63	63	63	76
	Intransitive - preterite	46	78	43	83	**85**	80	52	80	63	83	89	**93**	80	85	72	83	78	50	89	52	91	89	87	**93**	87	83	67	91	91	91	83	72	87	70	85	74	61	57	57	77
	Intransitive - preterite subj. II	100	89	69	88	92	88	77	**95**	76	85	84	**89**	81	87	85	83	79	44	84	52	68	72	72	89	**93**	89	64	85	89	86	88	79	80	81	**94**	88	69	81	81	81
	Modal - future I	42	**98**	**98**	74	93	**98**	95	95	**98**	**93**	83	79	57	69	74	76	67	50	76	62	33	38	57	79	90	98	29	67	74	74	86	81	55	98	98	**100**	83	95	95	78
	Modal - future I subj. II	86	86	77	81	78	**88**	76	**90**	79	85	79	78	65	71	86	86	81	50	85	56	44	42	**94**	78	83	**90**	31	65	79	67	84	74	72	80	**91**	79	57	77	77	75
	Modal - perfect	149	91	83	77	87	90	84	**93**	83	**85**	81	83	56	52	83	82	76	37	79	61	44	46	73	83	81	89	21	52	68	56	77	66	62	86	**98**	83	66	85	85	73
	Modal - pluperfect	75	88	80	81	92	92	84	**95**	77	79	76	80	75	81	76	**85**	67	37	73	77	83	80	84	80	85	92	47	75	85	80	77	79	67	89	**95**	81	76	83	83	79
	Modal - pluperfect subj. II	61	85	57	89	80	85	74	**95**	66	80	85	87	84	89	72	84	74	46	87	59	69	74	**92**	87	85	89	46	75	82	74	**90**	84	82	72	80	84	70	74	74	78
	Modal - present	30	93	67	90	90	93	87	**100**	73	90	83	90	77	90	**97**	**97**	**97**	43	90	40	73	80	**97**	90	93	93	53	**100**	97	**100**	97	93	97	90	93	90	63	87	87	86
	Modal - preterite	72	85	78	82	88	88	76	**94**	78	**94**	83	85	93	89	88	78	86	46	90	57	72	69	83	86	88	86	44	75	78	78	83	82	68	79	**89**	86	72	76	76	80
	Modal - preterite subj. II	30	**100**	**100**	80	93	**100**	93	93	97	87	87	**90**	87	**90**	87	87	**90**	30	80	73	63	63	73	90	93	**100**	40	70	83	80	77	67	47	87	**100**	97	77	93	93	83
	Modal negated - future I	43	95	91	88	95	**98**	93	**98**	91	91	95	**98**	88	93	93	93	95	47	91	77	56	58	77	**98**	88	95	42	79	91	81	91	84	65	95	95	91	77	93	93	86
	Modal negated - future I subj. II	73	78	60	77	74	78	73	**82**	59	85	67	84	68	88	**89**	75	75	48	71	63	59	56	78	84	79	78	33	59	71	64	75	70	64	84	**92**	63	52	75	75	71
	Modal negated - perfect	126	81	67	71	82	80	77	**87**	71	79	71	75	64	80	79	75	77	46	67	**83**	54	54	56	75	84	84	33	61	66	65	67	58	52	80	**94**	75	63	78	78	71
	Modal negated - pluperfect	126	86	77	83	85	86	87	**90**	79	72	68	75	71	**79**	75	72	70	39	63	76	72	71	67	75	84	**89**	38	74	84	81	64	67	50	**89**	**89**	79	65	87	87	75
	Modal negated - pluperfect subj. II	81	78	70	80	83	80	75	**85**	57	84	77	75	**91**	81	84	77	80	42	78	69	80	81	**90**	75	**84**	80	38	77	80	79	79	72	65	75	81	68	54	70	70	75
	Modal negated - present	33	91	64	**97**	91	91	91	94	55	94	91	97	88	97	97	**100**	94	52	94	48	82	88	**100**	97	94	91	61	91	94	91	97	**100**	91	79	**100**	67	67	91	91	87
	Modal negated - preterite	61	84	72	90	89	**92**	87	90	69	89	74	75	67	89	87	84	72	38	77	38	75	75	**90**	75	89	85	49	77	85	87	84	80	70	84	**90**	75	62	87	87	78
	Modal negated - preterite subj. II	77	77	66	82	81	78	70	**88**	68	**91**	79	86	61	79	**90**	75	77	34	84	56	70	74	82	86	84	78	57	83	**91**	84	88	81	74	69	87	74	73	69	69	76
	Progressive	76	82	66	79	**86**	79	71	82	63	76	**83**	79	78	67	75	76	74	45	82	59	72	70	68	79	68	75	57	80	**83**	78	**83**	79	74	75	82	70	57	68	68	73
	Reflexive - future I	85	**86**	76	**86**	**87**	85	79	**87**	74	72	**89**	78	85	71	87	85	**89**	36	87	54	68	69	69	78	84	86	62	87	91	87	88	88	84	78	**94**	80	45	79	79	79
	Reflexive - future I subj. II	96	85	77	69	77	86	73	**90**	72	69	79	**83**	**82**	75	72	**82**	66	42	71	66	78	73	72	83	76	80	38	76	78	75	79	79	69	74	**89**	70	62	73	73	74
	Reflexive - future II	116	94	77	83	92	**96**	79	91	75	72	80	84	**88**	78	76	85	69	49	79	53	72	76	81	84	82	86	54	77	81	76	85	84	72	77	**97**	71	53	78	78	78

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	YiSi-1	chrF	prismRef	spBLEU
			baselines							
	Reflexive - future II subj. II	107	90	67	84	**93**	92	71	**93**	70
	Reflexive - perfect	188	**85**	61	83	**85**	80	60	**85**	64
	Reflexive - pluperfect	109	**96**	74	76	91	94	67	86	71
	Reflexive - pluperfect subj. II	90	90	68	87	90	87	74	**92**	74
	Reflexive - present	125	82	55	**90**	86	82	72	88	66
	Reflexive - preterite	117	**92**	75	86	85	90	77	**92**	77
	Reflexive - preterite subj. II	124	94	89	90	93	**97**	81	**96**	85
	Transitive - future I	43	98	84	95	95	**100**	86	98	86
	Transitive - future I subj. II	37	97	81	86	95	**100**	89	95	81
	Transitive - future II	33	97	82	94	94	**100**	91	94	79
	Transitive - future II subj. II	50	**96**	62	78	80	86	82	68	74
	Transitive - perfect	99	88	73	76	91	**95**	85	**96**	77
	Transitive - pluperfect	22	82	41	86	86	86	59	**95**	50
	Transitive - pluperfect subj. II	39	90	64	**92**	90	90	72	**92**	69
	Transitive - present	33	73	52	85	**91**	82	67	**91**	70
	Transitive - preterite	57	82	39	**84**	79	77	56	77	54
	Transitive - preterite subj. II	97	81	48	**88**	**88**	**89**	59	**88**	54
Verb valency	Case government	80	89	74	86	**91**	85	72	**90**	70
	Mediopassive voice	50	94	72	92	90	**96**	78	94	72
	Passive voice	33	**55**	**55**	**55**	**55**	**55**	**55**	**55**	**55**
macro avg.		10402	84	69	82	84	**85**	75	**86**	71
micro avg.		10402	83	70	82	**84**	**84**	75	**85**	72

ling. category	ling. phenomenon	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	prismSrc2
		QE as a metric													
	Reflexive - future II subj. II	89	**93**	91	**93**	76	84	88	79	41	92	63	84	92	83
	Reflexive - perfect	82	88	**90**	80	78	83	**90**	80	46	88	58	79	84	84
	Reflexive - pluperfect	70	77	80	**84**	75	66	**84**	62	43	73	55	72	71	81
	Reflexive - pluperfect subj. II	78	87	87	**92**	88	89	83	78	38	86	64	89	84	88
	Reflexive - present	79	84	88	81	**91**	89	**92**	87	46	85	58	89	86	**91**
	Reflexive - preterite	71	68	74	**89**	83	80	80	73	39	71	61	81	80	78
	Reflexive - preterite subj. II	86	81	90	**96**	88	93	86	93	44	85	55	82	82	88
	Transitive - future I	84	**93**	77	88	91	84	81	79	44	**93**	51	84	84	**93**
	Transitive - future I subj. II	95	95	84	81	**100**	95	97	89	43	92	68	70	95	**100**
	Transitive - future II	91	94	97	91	85	94	97	82	45	94	67	76	85	**100**
	Transitive - future II subj. II	92	**98**	70	92	66	96	**98**	**98**	56	96	54	90	88	86
	Transitive - perfect	70	77	72	84	82	82	83	72	43	73	60	62	69	**86**
	Transitive - pluperfect	77	91	82	**95**	**95**	91	91	77	27	91	45	68	68	91
	Transitive - pluperfect subj. II	79	87	92	**95**	**95**	92	90	92	49	82	51	82	85	85
	Transitive - present	82	88	91	**100**	91	85	97	82	48	85	70	88	91	70
	Transitive - preterite	**84**	81	81	63	72	77	77	75	40	79	58	81	81	74
	Transitive - preterite subj. II	82	87	92	**94**	84	85	**94**	77	41	88	65	82	85	79
Verb valency	Case government	**84**	74	**84**	60	70	**82**	78	78	57	76	50	79	80	71
	Mediopassive voice	90	84	88	84	86	90	**96**	88	54	94	40	88	94	72
	Passive voice	39	39	42	18	27	**45**	39	30	27	39	27	42	39	27
macro avg.		80	81	81	76	75	81	**83**	79	43	80	57	74	75	76
micro avg.		79	80	80	75	74	80	**83**	78	44	80	58	75	76	75

ling. category	ling. phenomenon	KG-BERT	MEE4	MEE4_stsb_xlm	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	cometoid22-wmt23	eBLEU	embed_llama	partokengram_F	tokengram_F	avg
		ref. based metrics																
	Reflexive - future II subj. II	91	75	82	55	86	84	88	**94**	**93**	82	69	**93**	69	57	73	73	81
	Reflexive - perfect	90	76	74	63	84	**91**	88	86	83	82	69	**91**	70	57	60	60	77
	Reflexive - pluperfect	80	80	81	55	69	83	67	79	74	70	77	**96**	72	75	72	72	75
	Reflexive - pluperfect subj. II	87	84	86	58	79	82	79	84	81	84	76	**92**	83	69	76	76	81
	Reflexive - present	88	81	78	62	87	90	84	**92**	87	86	69	88	70	58	70	70	80
	Reflexive - preterite	74	**92**	82	64	90	**93**	86	77	82	79	71	85	85	79	80	80	79
	Reflexive - preterite subj. II	90	95	90	77	**98**	**98**	**98**	88	96	92	77	97	85	73	83	83	87
	Transitive - future I	77	95	91	60	93	**98**	93	93	86	86	86	95	65	56	86	86	85
	Transitive - future I subj. II	84	89	89	62	86	89	84	89	89	73	84	**97**	81	54	84	84	85
	Transitive - future II	**97**	85	**97**	61	**97**	**97**	**97**	94	88	79	94	**97**	73	36	94	94	87
	Transitive - future II subj. II	70	72	70	50	82	76	82	80	80	80	82	**96**	68	56	82	82	79
	Transitive - perfect	72	88	94	58	83	84	84	81	78	73	85	**96**	80	72	86	86	79
	Transitive - pluperfect	86	82	64	77	82	82	86	**91**	**91**	77	68	**91**	64	32	55	55	75
	Transitive - pluperfect subj. II	92	97	87	87	97	**100**	97	87	90	90	85	95	85	59	72	72	84
	Transitive - present	91	94	88	88	**97**	94	**97**	88	85	94	79	**97**	85	64	76	76	83
	Transitive - preterite	81	77	70	58	77	79	75	82	75	**84**	68	**84**	74	53	58	58	72
	Transitive - preterite subj. II	92	78	78	77	92	91	**95**	93	**95**	92	74	88	64	62	62	62	80
Verb valency	Case government	84	84	70	57	86	86	86	86	89	74	80	**92**	70	51	75	75	77
	Mediopassive voice	88	90	88	68	90	90	**94**	90	90	84	92	**94**	86	58	80	80	84
	Passive voice	42	**55**	**55**	36	52	**55**	39	48	36	33	**55**	**55**	**55**	52	**55**	**55**	45
macro avg.		81	81	80	56	81	84	81	83	80	73	76	**87**	74	62	69	75	76
micro avg.		80	81	79	57	81	83	81	82	81	74	76	**87**	74	63	67	75	76

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

ling. phenomenon	#	baselines								QE as a metric																				ref. based metrics																	
		BERTscore	BLEU	BLEURT-20	COMET	YiSi-1	chrF	prismRef	spBLEU	Calibri-COMET22-QE	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	KG-BERT	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	cometoid22-wmt23	mltsamp-bleurt0p2p1-qe	mltsamp-bleurtxv1p-qe	prismSrc2	Calibri-COMET22	MEE4	MEE4_stsb_xlm	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	eBLEU	embed_llama	mt0regressor	partokengram_F	tokengram_F	avg
Ambiguity																																														
Lexical ambiguity	146	84	71	90	84	86	88	**97**	88	42	77	60	42	77	42	39	75	85	82	46	56	38	59	58	87	86	**95**	14	84	90	92	50	88	**95**	88	82	88	62	**93**	70	62	81	62	88	73
Coordination & ellipsis																																														
Gapping	163	72	59	86	84	71	67	**88**	66	75	64	79	76	84	76	78	91	94	93	42	84	56	78	75	75	**98**	91	66	83	63	67	65	76	**89**	71	83	**88**	71	74	58	55	61	51	67	74
Pseudogapping	201	81	76	**98**	92	79	67	81	73	94	97	**98**	94	83	94	79	93	98	93	45	**98**	23	96	95	**99**	93	96	63	88	74	73	42	90	96	84	**98**	**98**	84	87	75	62	76	48	68	82
Right node raising	47	83	64	87	**98**	83	72	89	70	91	87	89	91	81	91	85	91	91	85	53	89	9	91	89	**94**	89	85	87	89	68	70	51	89	**98**	87	89	91	83	81	55	64	91	55	72	80
Sluicing	169	55	**58**	**57**	50	54	54	45	**58**	37	47	53	39	19	39	**69**	49	56	57	51	57	49	39	41	44	63	54	43	44	56	59	11	57	66	53	64	67	44	57	56	57	**69**	44	55	51
Stripping	139	72	58	68	**76**	68	58	**76**	61	64	73	50	64	45	64	36	65	56	69	50	55	46	65	62	68	74	**77**	45	73	63	60	29	68	66	75	56	60	40	**78**	60	58	71	50	58	61
VP-ellipsis	117	59	47	**85**	84	68	53	74	45	88	85	84	**89**	59	**89**	84	81	82	82	52	82	25	**90**	**90**	85	85	79	67	81	49	50	6	82	78	71	**85**	81	67	74	49	44	79	51	56	70
False friends																																														
False friends	225	65	63	69	**74**	71	71	67	65	71	61	75	71	65	71	73	64	80	66	46	76	**82**	64	52	66	76	63	68	74	83	**87**	30	60	82	62	65	59	53	76	66	55	56	52	68	67
Function word																																														
Focus particle	20	45	30	**70**	60	50	35	**70**	30	85	90	90	85	50	85	50	90	30	40	40	**100**	95	50	50	50	70	65	40	60	35	35	25	80	**100**	70	95	60	45	45	35	40	80	40	35	59
Question tag	180	**96**	83	80	93	86	78	**96**	79	92	95	92	91	86	91	81	86	77	90	36	94	58	**99**	**98**	**98**	93	67	89	89	88	81	64	87	84	86	**94**	89	66	79	74	53	75	63	80	83
MWE																																														
Collocation	112	73	62	**93**	90	89	77	88	60	84	96	92	80	81	80	74	**97**	93	**97**	45	**98**	15	81	73	83	**97**	**97**	57	88	82	81	30	95	96	**99**	**99**	**99**	79	80	61	68	71	51	78	79
Compound	63	73	51	56	89	95	84	**98**	68	95	97	**100**	**100**	79	**100**	56	**100**	98	**100**	48	**100**	22	86	94	81	90	98	83	81	62	65	2	**100**	**100**	92	**100**	83	79	78	51	65	40	51	84	78
Idiom	266	86	82	**97**	**98**	93	75	93	83	89	93	97	90	**99**	90	75	98	**100**	97	44	96	23	82	84	**98**	95	98	36	**98**	82	89	15	95	97	89	96	97	86	80	80	71	81	60	76	83
Nominal MWE	288	81	72	83	**86**	80	78	75	74	51	75	54	50	26	50	71	74	78	65	42	66	61	73	72	**92**	81	73	39	86	72	70	17	72	84	76	83	**93**	71	85	65	70	71	54	79	69
Prepositional MWE	35	69	**86**	71	83	83	83	83	83	83	80	80	83	71	83	**89**	80	71	83	63	83	46	86	86	**89**	71	**89**	80	83	83	83	17	71	**89**	57	83	80	69	80	77	74	83	46	83	77
Verbal MWE	65	68	72	**88**	77	66	74	78	60	69	75	71	66	66	66	55	94	86	95	43	65	37	72	66	75	86	**98**	60	77	77	74	32	**95**	88	89	69	75	57	75	74	57	66	49	71	71
Named entity & termin.																																														
Date	234	52	55	**74**	71	64	60	68	63	67	72	60	67	75	67	77	77	81	77	46	64	62	83	82	**88**	85	84	82	63	55	65	24	**85**	79	81	75	79	69	77	69	54	79	51	61	69
Domainspecific term	312	69	56	**88**	82	79	74	**88**	66	71	46	58	74	72	73	69	66	87	65	46	64	12	83	69	89	83	**98**	61	82	68	74	53	91	**98**	96	76	71	78	96	65	42	41	47	71	71
Location	12	67	83	75	67	50	58	**100**	83	67	58	67	67	83	67	92	92	75	92	58	92	67	**100**	**100**	75	75	**100**	**100**	67	92	92	67	**100**	**100**	**100**	**100**	**100**	**100**	83	83	58	**100**	58	58	81
Measuring unit	389	54	48	53	**56**	55	**56**	**57**	53	28	37	20	28	28	28	37	51	50	40	44	27	56	28	23	53	35	**77**	36	55	58	**63**	10	57	56	53	38	35	49	**62**	49	48	42	48	55	45
Proper name	325	50	57	54	50	53	54	56	**68**	57	57	**66**	58	58	58	48	56	49	58	39	**66**	54	58	56	50	57	64	55	50	58	51	35	61	**68**	59	**67**	62	49	53	61	48	60	48	56	56
Negation																																														
Negation	174	87	83	89	**92**	**92**	84	84	86	90	90	84	89	76	89	92	83	88	78	47	84	40	90	90	**96**	89	91	90	**91**	**91**	**91**	58	85	**91**	80	84	85	79	80	86	67	88	52	86	83
Non-verbal agreement																																														
Coreference	81	85	85	**95**	80	73	86	84	88	83	77	95	83	69	83	75	95	89	91	52	**98**	74	85	81	88	93	**98**	62	79	88	79	91	91	98	96	**100**	96	85	80	79	72	79	56	85	84
Genitive	206	75	73	73	**93**	82	68	80	76	77	94	78	76	83	76	90	**97**	95	90	46	96	24	**97**	**96**	**97**	75	93	69	90	72	67	53	90	**95**	81	91	87	88	76	77	65	85	52	69	79
Possession	85	62	55	**86**	**86**	74	58	**87**	60	93	93	89	93	**95**	93	53	89	92	93	39	92	58	89	87	89	91	91	86	86	72	64	34	88	**93**	**93**	**93**	**93**	91	59	54	61	86	56	54	78
Punctuation																																														
Quotation marks	336	69	73	74	70	66	72	**77**	69	**82**	72	76	**82**	60	**82**	65	73	66	73	42	60	44	81	**82**	**81**	75	70	75	65	**74**	72	28	**74**	70	67	65	65	46	68	**74**	49	57	44	73	68
Subordination																																														
Adverbial clause	193	72	80	80	82	71	73	**83**	77	91	**93**	90	91	80	91	68	83	88	87	42	88	44	79	73	76	86	83	70	80	77	75	8	75	84	74	**91**	**92**	87	67	69	67	63	51	74	76
Cleft sentence	179	67	64	61	**75**	66	59	**75**	62	76	72	74	76	58	76	75	70	72	75	41	74	45	**78**	75	**78**	64	69	63	73	68	67	14	66	65	71	**75**	69	59	63	70	66	65	54	61	66
Contact clause	150	84	75	**95**	**94**	91	75	89	75	**97**	95	**98**	**97**	93	**97**	91	87	**98**	91	47	**98**	44	94	96	**97**	91	93	91	94	79	79	47	97	97	97	**100**	**99**	89	86	79	71	91	61	77	87
Indirect speech	38	55	42	63	63	47	47	**82**	42	84	**97**	71	79	61	79	50	68	66	76	47	82	26	82	79	76	95	87	84	50	45	47	5	79	63	79	**82**	74	74	58	45	55	58	45	50	64
Infinitive clause	85	66	55	89	**92**	87	80	88	67	**98**	**99**	95	**98**	87	**98**	89	95	95	91	44	95	75	**98**	**98**	**98**	94	93	79	91	73	75	8	92	93	91	**95**	**94**	85	88	60	66	89	49	79	83
Object clause	16	56	44	**94**	69	62	62	75	50	75	**100**	88	81	81	81	88	**100**	81	**100**	31	**100**	69	81	81	81	81	**100**	81	62	50	69	38	94	81	**100**	**100**	**100**	75	50	50	50	75	56	62	75
Pseudo-cleft sentence	73	**90**	86	68	73	85	89	**92**	84	77	**82**	**82**	74	58	74	73	78	**82**	74	38	**82**	68	68	**82**	**82**	64	70	79	75	**92**	82	12	79	75	71	82	67	73	86	79	86	88	56	86	76

Accuracy of the metrics(%) with regards to the linguistically-motivated phenomena for English-German. Continued on next page

ling. phenomenon	#	baselines								QE as a metric																			ref. based metrics																
		BERTscore	BLEU	BLEURT-20	COMET	YiSi-1	chrF	prismRef	spBLEU	Calibri-COMET22-QE	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	KG-BERT	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	cometoid22-wmt23	mtsamp-bleurt0p2p1-qe	mtsamp-bleurtxv1p-qe	prismSrc2	Calibri-COMET22	MEE4	MEE4_stsb_xlm	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	eBLEU	embed_llama	mt0regressor	partokengram_F	tokengram_F	avg
Relative clause	112	91	80	88	93	77	83	92	77	100	84	98	100	79	100	94	94	85	81	39	94	40	80	80	94	97	91	84	93	87	90	5	92	97	91	92	76	74	76	88	64	62	61	81	82
Subject clause	148	92	93	91	89	90	91	89	91	89	89	89	89	80	89	88	80	86	82	43	93	14	86	90	89	85	91	70	89	96	92	11	91	89	93	93	92	86	85	92	89	83	62	93	83
Verb tense/aspect/mood																																													
cond.	106	74	77	91	88	86	72	93	74	97	85	94	97	86	97	82	95	91	96	37	92	38	99	100	100	94	97	92	87	86	85	14	97	89	90	91	92	84	82	76	60	80	57	73	83
Ditrans. - cond. I progr.	72	65	49	93	96	83	61	81	57	100	90	100	100	100	100	92	94	100	94	43	99	65	99	99	100	100	96	94	92	61	62	46	83	100	85	99	92	74	67	57	61	82	49	Q60	82
Ditrans. - cond. I simple	34	94	74	65	88	97	94	91	79	88	68	97	88	76	88	91	100	100	97	56	79	79	50	71	82	65	91	47	91	97	97	74	79	85	71	97	79	56	94	76	85	97	65	94	82
Ditrans. - cond. II progr.	51	75	78	88	82	80	82	86	82	75	43	82	75	53	75	59	96	75	94	31	75	61	86	86	92	96	90	96	82	82	80	59	96	94	92	67	65	61	82	80	49	82	61	82	77
Ditrans. - cond. II simple	59	71	64	76	75	68	68	86	73	71	85	85	69	63	69	63	86	71	76	41	92	64	76	76	76	81	85	80	73	75	80	53	86	83	83	93	86	58	64	75	58	75	41	66	73
Ditrans. - future I progr.	61	52	51	62	61	57	62	57	51	90	28	97	90	11	90	80	57	62	67	39	85	56	48	56	74	67	72	89	59	62	69	11	69	57	66	70	36	52	49	57	43	85	54	61	61
Ditrans. - future I simple	88	60	51	56	66	56	61	67	45	64	60	73	64	58	64	70	72	67	91	43	64	76	58	59	72	57	62	99	65	61	61	18	62	59	59	67	53	49	55	43	47	72	44	58	61
Ditrans. - future II progr.	91	71	64	66	66	47	60	86	62	92	30	87	92	99	92	78	81	80	100	41	90	70	64	68	64	70	99	87	67	77	84	46	96	100	95	86	70	77	80	70	56	91	59	62	75
Ditrans. - future II simple	49	71	94	86	71	65	94	86	92	98	76	92	96	100	96	65	98	100	98	41	100	55	51	55	94	47	49	41	73	94	94	18	65	63	57	98	57	59	94	82	73	92	41	94	76
Ditrans. - past perf. progr.	91	60	44	62	75	66	58	75	48	75	68	85	75	33	75	75	87	79	79	40	93	33	79	81	80	63	74	78	76	64	65	59	56	82	66	85	77	76	56	58	47	58	54	58	67
Ditrans. - past perf. simple	112	63	62	65	63	72	71	65	61	60	66	57	58	36	58	58	70	82	64	47	66	69	41	44	54	55	45	67	62	64	57	39	54	57	59	65	58	52	73	57	39	67	52	70	59
Ditrans. - past progr.	83	58	59	71	52	59	61	77	60	48	61	51	46	22	46	43	78	36	78	37	61	71	40	35	39	77	76	78	51	63	60	20	63	69	75	67	73	31	59	52	51	53	53	66	56
Ditrans. - pres. perf. progr.	48	85	54	85	85	92	88	90	60	100	71	73	100	65	100	77	56	56	92	44	79	31	54	94	100	92	88	100	90	81	81	56	75	96	79	94	79	73	94	79	81	88	58	88	79
Ditrans. - pres. perf. simple	54	65	37	56	39	30	41	61	44	41	93	61	39	24	39	33	89	78	89	52	76	57	30	30	28	67	74	48	39	48	46	28	69	44	74	87	76	35	63	65	56	59	54	37	53
Ditrans. - pres. progr.	72	76	38	94	97	90	68	79	49	92	85	85	93	75	93	88	97	97	93	54	92	75	100	97	96	89	89	60	97	83	89	89	85	89	83	96	97	93	93	79	71	89	53	68	84
Ditrans. - simple past	77	77	56	77	79	56	66	87	57	92	82	95	92	99	92	73	97	92	99	45	96	53	61	68	92	92	87	82	81	84	81	45	68	88	71	96	91	71	78	60	60	79	42	64	77
Ditrans. - simple pres.	54	72	30	83	80	83	56	72	41	67	98	70	69	61	69	70	98	74	81	46	96	43	70	67	69	93	85	81	80	69	70	65	63	78	67	98	96	69	80	59	50	61	54	56	71
Gerund	161	93	90	96	96	86	83	93	86	99	100	99	99	84	99	90	93	100	94	41	98	13	96	94	95	98	97	60	96	89	91	0	98	98	94	99	98	95	92	77	90	94	66	86	87
Imperative	50	70	50	96	96	70	70	90	64	90	80	96	90	94	90	94	100	100	88	48	100	30	94	100	98	100	100	82	98	76	76	50	98	100	94	100	94	92	74	72	66	76	60	68	83
Intrans. - cond. I progr.	9	56	89	89	100	100	78	100	89	44	44	22	100	100	100	44	100	100	100	56	44	100	100	100	100	100	67	100	100	89	89	78	56	89	56	78	100	44	67	78	100	89	44	67	80
Intrans. - cond. I simple	3	100	0	67	100	100	33	100	33	100	33	33	100	100	100	33	100	100	100	33	33	100	100	100	100	100	100	100	100	100	67	67	67	100	67	67	100	33	100	0	67	67	0	33	73
Intrans. - future I progr.	7	71	86	100	86	57	100	86	86	14	100	86	71	100	71	57	100	100	100	43	100	86	71	71	71	100	100	86	86	100	100	43	100	86	86	100	100	86	100	100	43	57	29	100	82
Intrans. - future I simple	24	67	75	75	71	50	67	58	71	96	96	96	96	71	96	92	96	79	96	54	96	25	100	100	88	100	100	62	67	67	58	21	62	54	62	96	96	58	50	67	50	67	46	67	73
Intrans. - future II progr.	4	25	50	25	25	50	50	25	50	25	75	100	75	25	75	50	75	25	75	25	100	25	25	25	25	75	75	25	25	25	75	25	75	75	75	75	25	100	50	100	25	25	25	50	50
Intrans. - future II simple	7	71	100	86	100	100	100	57	100	100	43	86	100	71	100	43	86	100	71	14	86	100	100	86	100	71	71	29	100	86	100	14	86	71	57	86	43	43	100	71	43	100	29	100	77
Intrans. - past perf. progr.	16	56	50	38	62	69	62	81	69	44	44	25	62	56	62	75	44	56	38	56	38	62	38	50	56	50	50	75	69	62	81	6	31	75	44	38	44	50	38	56	50	62	44	62	53
Intrans. - past perf. simple	18	78	72	89	83	61	78	83	61	72	61	50	89	17	89	56	28	33	28	39	61	11	100	100	78	100	100	33	83	78	78	22	100	50	89	78	44	72	61	89	78	78	50	78	67
Intrans. - past progr.	28	43	57	71	68	54	57	68	54	64	46	64	64	68	64	61	71	89	39	25	61	18	46	50	50	68	86	68	71	54	50	39	64	54	64	61	61	64	57	46	50	71	50	57	58
Intrans. - pres. perf. simple	2	100	50	100	100	100	100	100	50	50	100	100	100	100	100	100	100	100	100	0	100	50	50	50	50	100	100	100	100	100	100	50	100	100	50	100	100	100	100	100	100	100	50	100	86
Intrans. - pres. progr.	5	80	100	80	80	80	100	100	80	60	80	100	60	100	60	80	80	80	80	20	100	80	80	80	80	80	100	80	80	100	100	80	80	80	80	100	80	80	60	100	100	80	40	100	82
Intrans. - simple past	24	58	38	62	67	58	46	79	38	67	88	71	75	88	75	71	100	100	100	58	71	38	67	67	62	71	96	100	67	71	67	42	71	92	75	75	75	75	62	54	50	67	54	46	69
Intrans. - simple pres.	10	40	30	50	50	50	40	80	40	70	60	50	90	90	90	70	80	80	70	50	90	70	100	90	70	40	80	80	50	30	40	60	60	70	50	80	80	50	40	20	40	50	70	40	61
Modal	20	70	60	40	50	45	55	50	55	35	60	45	35	0	35	100	10	5	5	45	30	75	90	85	95	15	5	55	50	75	70	15	30	30	30	40	30	10	50	55	35	40	70	60	45
Modal negated	20	35	65	70	75	60	60	95	70	85	95	85	85	50	80	60	90	100	100	20	85	35	55	50	75	85	70	85	80	70	75	45	80	85	75	85	85	85	55	45	45	75	60	60	70
Reflex. - cond. I progr.	65	66	52	48	45	45	46	34	71	20	40	69	37	95	37	46	58	95	77	42	82	29	34	40	49	35	75	12	45	54	66	43	77	82	89	66	55	82	45	71	46	46	51	51	55
Reflex. - cond. I simple	112	76	70	48	59	58	70	32	72	48	71	60	49	84	49	43	92	94	97	41	70	78	49	60	70	37	88	5	59	66	82	72	89	71	93	69	74	80	68	96	78	62	69	69	67
Reflex. - cond. II progr.	97	71	72	66	66	61	69	49	71	63	60	66	75	92	75	28	81	96	89	51	89	36	47	51	74	65	82	49	66	67	81	65	80	87	81	81	64	79	72	87	58	58	72	72	69
Reflex. - cond. II simple	109	61	68	52	59	54	61	52	59	74	56	61	81	81	81	16	86	85	86	46	90	52	28	38	76	56	87	55	57	47	68	67	87	78	89	91	72	77	53	92	41	64	63	63	65
Reflex. - future I progr.	70	67	67	70	57	84	79	41	66	56	69	71	56	79	56	69	59	97	81	51	81	60	51	47	66	74	79	29	60	66	74	46	77	77	69	77	63	76	61	83	53	71	77	77	67
Reflex. - future I simple	83	70	67	73	58	82	86	58	78	52	49	34	57	70	57	78	47	86	59	31	60	57	69	58	77	58	60	55	59	76	80	82	72	67	77	64	77	61	63	87	51	76	87	87	66

Accuracy of the metrics(%) with regards to the linguistically-motivated phenomena for English-German. Continued on next page

ling. phenomenon	#	baselines								QE as a metric																			ref. based metrics																avg
		BERTscore	BLEU	BLEURT-20	COMET	YiSi-1	chrF	prismRef	spBLEU	Calibri-COMET22-QE	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	KG-BERT	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	cometoid22-wmt23	mtsamp-bleurt0p2p1-qe	mtsamp-bleurtxv1p-qe	prismSrc2	Calibri-COMET22	MEE4	MEE4_stsb_xlm	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	eBLEU	embed_llama	mt0regressor	partokengram_F	tokengram_F	
Reflex. - future II progr.	81	65	56	64	64	75	**80**	68	57	68	75	58	85	77	85	52	85	**89**	86	51	83	37	44	49	72	64	85	41	67	72	79	57	84	73	80	**86**	81	81	68	79	38	68	77	77	69
Reflex. - future II simple	56	71	66	77	54	**88**	**88**	71	64	77	95	89	84	73	84	39	98	98	**100**	50	**100**	70	66	64	68	48	54	38	52	75	82	61	89	54	89	**98**	68	64	64	88	43	64	89	89	73
Reflex. - past perf. progr.	98	60	50	67	64	**71**	66	49	51	81	79	70	**88**	85	**88**	52	74	73	83	39	77	48	48	61	83	67	77	43	65	67	69	41	66	71	70	**81**	70	73	61	72	49	73	65	65	66
Reflex. - past perf. simple	53	62	47	68	55	**74**	55	49	57	70	89	85	77	64	77	17	81	79	79	38	**98**	60	45	51	68	70	83	49	58	62	58	49	81	70	91	**92**	83	72	57	77	47	83	53	53	66
Reflex. - past progr.	5	**100**	**100**	40	**100**	**100**	**100**	80	**100**	**0**	60	60	40	40	40	**100**	**100**	**100**	**100**	40	40	40	40	60	40	40	60	40	**100**	**100**	**100**	60	80	60	80	60	80	60	**100**	**100**	**100**	40	**100**	**100**	72
Reflex. - pres. perf. progr.	33	76	48	**88**	67	76	76	45	48	61	**100**	88	88	85	88	48	**100**	**100**	**100**	42	**100**	27	61	79	82	67	**100**	24	67	79	85	52	97	79	**100**	**100**	97	**100**	67	**100**	33	73	79	79	76
Reflex. - pres. perf. simple	39	59	46	67	59	69	**72**	46	44	44	**97**	72	69	64	69	31	74	85	85	38	82	62	38	67	59	56	74	31	59	69	79	28	79	69	85	74	77	82	67	**90**	54	69	69	69	65
Reflex. - pres. progr.	99	62	51	54	56	**67**	56	37	62	51	75	66	46	**91**	46	39	58	76	70	41	79	51	58	56	52	57	44	45	58	57	64	51	71	69	74	73	**76**	70	59	72	42	61	58	58	59
Reflex. - simple past	119	71	70	73	73	73	**77**	64	71	63	73	81	82	82	82	46	95	98	**100**	34	86	55	55	60	75	61	54	63	72	71	77	57	73	**84**	71	**85**	82	77	74	79	56	82	78	78	72
Reflex. - simple pres.	138	65	65	67	67	**88**	63	64	67	55	70	75	58	87	58	47	78	**97**	73	49	71	31	56	57	72	78	54	64	67	69	80	43	70	67	70	77	72	67	77	**83**	54	67	64	64	67
trans. - future II progr.	11	73	82	73	**91**	64	82	82	82	**100**	73	**100**	**100**	91	**100**	91	**100**	**100**	**100**	36	91	73	91	82	91	**100**	**100**	64	91	73	91	36	**100**	**100**	**100**	91	82	91	73	82	73	73	82	82	84
trans. - cond. I progr.	11	55	**91**	45	64	36	82	55	**91**	36	91	**100**	55	36	55	82	91	45	**100**	36	91	73	73	73	64	**100**	**100**	55	55	55	55	55	82	73	73	**100**	**100**	82	55	45	55	18	82	82	68
trans. - cond. I simple	9	67	**100**	89	89	56	**100**	22	**100**	67	89	**100**	56	56	56	89	**100**	67	89	33	89	89	44	33	89	**100**	**100**	**0**	67	67	56	56	**100**	67	**100**	**100**	**100**	78	67	67	67	56	**100**	**100**	75
trans. - cond. II progr.	20	70	55	75	65	80	55	**85**	60	70	**95**	**95**	70	75	70	40	75	85	90	40	85	40	65	65	60	85	**95**	60	65	75	70	50	70	**90**	70	**90**	85	85	65	55	55	75	55	55	70
trans. - cond. II simple	2	50	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	50	**100**	**100**	**100**	**100**	**100**	**100**	**100**	50	50	50	**100**	**100**	**100**	50	**100**	**100**	**100**	50	**100**	**100**	**100**	**100**	**100**	**100**	**100**	50	50	**100**	**100**	**100**	90
trans. - future I progr.	12	42	**83**	75	**83**	25	50	75	75	75	75	75	75	58	75	58	**92**	67	**92**	58	75	67	75	75	75	**92**	**92**	58	75	50	50	58	67	**92**	67	**92**	**92**	**92**	50	42	42	42	67	67	69
trans. - future I simple	22	68	**95**	64	68	68	77	50	**95**	73	86	36	77	5	77	86	95	73	**100**	50	59	73	77	73	68	64	91	59	64	68	68	73	86	50	86	77	**91**	73	45	68	59	41	86	86	71
trans. - future II simple	39	62	**92**	59	77	67	85	62	90	85	72	79	87	54	87	69	74	69	72	38	**97**	72	49	44	87	67	62	33	77	82	79	33	67	62	67	79	69	69	74	69	59	74	**87**	**87**	70
trans. - past perf. progr.	16	50	69	50	**81**	**81**	**81**	56	69	**94**	**94**	38	**94**	31	**94**	75	81	50	81	38	50	62	75	69	62	44	75	62	**81**	56	56	56	69	50	75	75	**81**	38	75	62	50	44	75	75	66
trans. - past perf. simple	9	44	78	**89**	**89**	33	**89**	22	78	**100**	89	**100**	**100**	44	**100**	56	**100**	78	89	22	**100**	78	89	67	67	**100**	**100**	**0**	**89**	44	44	56	78	**89**	78	**89**	**89**	78	33	33	22	22	78	78	70
trans. - pres. perf. progr.	5	20	**80**	**80**	**80**	20	**80**	20	**80**	80	**100**	**100**	**100**	**0**	**100**	**100**	**100**	20	80	60	**100**	20	**100**	**100**	**100**	**100**	**100**	40	80	20	**0**	80	60	**100**	60	**100**	80	80	20	20	20	**0**	80	80	66
trans. - pres. perf. simple	9	33	67	**78**	**78**	44	**78**	44	67	**100**	89	**100**	**100**	22	**100**	**100**	**100**	44	**100**	33	89	67	78	78	78	**100**	**100**	44	78	33	22	56	78	**100**	78	89	89	89	22	22	33	33	78	78	70
trans. - pres. progr.	10	30	**70**	20	40	30	40	30	50	30	40	50	30	20	30	40	**90**	60	**90**	50	60	50	50	50	30	40	40	40	50	40	30	20	40	60	40	**70**	**70**	60	40	40	20	40	50	50	45
trans. - simple past	23	61	43	**96**	70	35	57	87	52	91	78	91	**96**	61	**96**	91	**96**	74	83	48	**96**	39	74	74	87	**96**	**96**	78	70	65	70	52	96	**100**	91	96	91	**100**	57	48	43	70	57	57	75
trans. - simple pres.	16	31	62	38	**75**	69	62	56	56	94	50	81	**100**	62	**100**	**100**	94	94	75	50	81	50	**100**	94	**100**	81	81	69	62	38	50	**81**	75	75	56	**81**	69	75	56	56	44	75	62	62	70
Verb valency																																													
Case government	57	82	67	75	82	**86**	70	**86**	75	75	74	81	75	75	75	72	86	**91**	82	40	82	40	81	79	84	84	89	65	75	77	75	42	**82**	**82**	79	81	75	79	81	65	79	75	72	72	76
Catenative verb	177	71	62	**87**	75	68	63	68	63	73	91	**94**	73	34	73	66	88	63	90	44	92	60	77	81	67	92	89	58	72	65	68	19	81	71	84	**90**	85	77	80	64	64	63	64	64	72
Middle voice	29	83	62	**97**	93	72	76	93	72	93	79	97	93	86	93	62	**100**	**100**	**100**	55	**100**	21	**100**	93	**100**	97	**100**	31	90	90	83	59	**100**	**100**	**100**	**100**	**100**	90	79	59	83	83	76	76	84
Passive voice	70	66	54	69	70	67	74	**81**	61	74	71	63	79	94	79	69	96	**99**	90	40	79	30	67	63	70	74	89	87	70	64	69	33	**89**	81	87	71	59	51	80	57	69	60	71	71	71
Resultative	147	73	69	84	83	81	75	**86**	76	79	76	84	78	71	78	80	78	84	77	37	83	65	82	84	84	**88**	84	73	83	73	69	51	84	**88**	86	82	80	80	75	75	67	61	75	75	77
macro avg.	8945	67	65	74	**75**	70	71	73	67	73	76	77	78	67	78	68	**84**	80	**83**	43	82	52	71	72	77	78	**82**	62	74	70	72	42	79	80	77	**84**	79	71	70	67	58	69	59	71	71
micro avg.	8945	70	65	**75**	**76**	72	69	74	68	73	74	75	74	68	74	67	79	**81**	80	44	78	47	71	71	78	77	**81**	62	74	71	73	36	79	**81**	78	**81**	77	70	73	69	59	69	56	70	71

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

			baselines								QE as a metric																ref. based metrics											
ling. category	ling. phenomenon	#	BERTscore	BLEU	BLEURT-20	COMET	YiSi-1	chrF	prismRef	spBLEU	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	KG-BERT	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	cometoid22-wmt23	prismSrc2	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	eBLEU	embed_llama	partokengram_F	tokengram_F	avg
Ambiguity	Lexical ambiguity	66	80	67	**100**	**97**	88	67	83	73	**100**	94	**100**	**100**	**100**	95	**100**	**100**	**100**	39	**100**	30	**100**	**100**	**100**	56	**98**	**100**	**100**	**100**	**100**	**100**	**100**	88	77	44	59	70	86
Coordination & ellipsis	Gapping	44	**80**	75	68	**77**	73	66	**77**	73	93	86	66	30	66	93	43	**98**	41	52	84	34	73	68	80	68	14	45	75	39	73	70	57	73	**86**	50	55	75	66
	Pseudogapping	45	87	78	**89**	**89**	**91**	87	**91**	80	**87**	80	76	40	76	69	69	69	56	47	71	53	78	67	78	53	13	64	73	49	73	60	44	**89**	84	78	62	80	70
	Sluicing	11	64	55	**82**	45	55	64	55	55	18	55	36	64	36	64	**82**	**82**	**82**	45	64	73	55	55	27	27	0	55	27	36	64	55	0	**91**	**91**	64	45	64	54
	Stripping	27	78	63	59	48	63	56	**81**	52	41	26	26	15	26	52	67	22	22	37	22	48	52	56	48	**74**	4	30	37	37	41	67	41	67	67	**85**	22	52	47
	VP-ellipsis	76	**91**	76	57	74	76	59	70	68	72	57	86	47	86	82	63	**93**	**95**	43	62	32	74	74	71	46	39	**79**	**82**	**79**	68	62	61	66	76	45	70	61	68
False friends	False friends	6	**100**	17	**100**	**100**	**100**	17	83	50	0	0	0	**100**	0	0	0	**100**	17	17	0	17	33	33	**100**	0	17	**100**	**100**	**100**	33	33	**100**	83	**100**	50	83	33	50
Function word	Focus particle	50	**56**	44	36	30	52	36	**56**	46	42	24	42	20	42	18	48	44	**62**	44	56	34	48	26	10	46	16	66	**70**	64	48	36	20	42	42	52	56	38	42
	Question tag	66	62	88	**94**	82	53	39	42	**91**	62	79	56	74	56	50	80	64	38	44	85	42	**92**	**92**	88	14	3	61	47	58	70	**79**	24	55	**76**	52	52	38	61
MWE	Collocation	79	71	68	**81**	**82**	76	67	73	72	86	**87**	**90**	65	**90**	**90**	82	86	85	39	85	34	85	85	85	34	46	82	**86**	84	**89**	84	81	70	72	70	65	66	75
	Idiom	31	61	45	65	68	58	42	**84**	42	26	13	39	39	39	68	55	**90**	45	48	10	16	52	32	52	0	90	**100**	**100**	90	13	35	32	81	45	39	45	45	50
	Verbal MWE	12	83	83	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	92	50	92	17	83	92	83	25	**100**	92	75	67	92	8	42	83	**100**	83	**100**	**100**	83	**100**	83	42	58	**100**	81
Named entity & terminology	Date	45	60	47	**87**	69	58	49	80	49	80	78	53	40	53	64	69	56	**87**	42	76	38	82	82	80	58	9	84	67	84	82	**100**	53	73	64	67	53	49	65
	Domainspecific term	150	82	71	**97**	**99**	85	79	93	81	93	57	53	65	53	92	61	87	77	51	61	44	81	83	**98**	8	68	**97**	**100**	**97**	95	**99**	94	85	81	65	51	79	77
	Measuring unit	45	87	84	**100**	84	82	89	62	84	53	60	82	33	82	2	69	42	42	47	60	**100**	**100**	**98**	**100**	80	33	64	60	47	82	53	11	**91**	62	62	87	87	68
	Proper name	3	**100**	67	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	0	**100**	0	0	**100**	**100**	67	**100**	33	**100**	**100**	**100**	33	67	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	85
Negation	Negation	34	74	74	79	68	74	**85**	68	74	68	85	85	91	85	41	71	**100**	9	59	91	9	62	29	59	91	0	**94**	71	**94**	79	71	74	65	76	76	85	85	69
Non-verbal agreement	Coreference	57	68	54	**86**	77	67	67	60	56	84	**100**	56	77	56	75	75	**100**	67	42	75	26	**100**	**100**	84	63	46	**93**	**93**	**93**	84	82	33	61	60	63	68	68	71
	Genitive	4	**75**	**75**	25	**75**	**75**	**75**	**75**	**75**	75	**100**	25	0	25	50	25	75	50	25	**100**	25	25	**100**	**100**	50	0	75	75	75	**100**	25	0	**100**	75	**100**	75	75	60
Punctuation	Direct Speech	121	42	46	53	68	**86**	41	73	55	66	29	65	27	65	**87**	55	66	43	41	61	44	66	67	36	56	1	60	71	52	66	57	12	63	66	**84**	45	45	54
Subordination	Adverbial clause	132	43	52	32	47	**55**	46	48	47	49	**74**	41	32	41	70	55	55	22	49	58	16	48	16	39	22	5	60	**92**	44	64	43	80	39	30	39	45	45	46
	Cleft sentence	24	71	54	**92**	**92**	**92**	75	79	54	96	79	75	88	75	79	92	96	**100**	54	**100**	46	88	88	88	67	79	88	92	92	**100**	92	92	75	54	75	62	62	80
	Contact clause	10	60	30	90	**100**	70	60	**100**	30	60	**100**	**100**	0	**100**	**100**	40	**100**	**100**	70	**100**	**100**	**100**	**100**	**100**	**100**	0	**100**	**100**	**100**	**100**	60	0	80	40	60	60	60	74
	Infinitive clause	24	71	75	67	**96**	79	**96**	67	67	**96**	54	83	62	83	83	71	88	75	38	83	38	75	75	79	79	33	92	**100**	92	96	96	50	71	71	58	96	96	76
	Object clause	70	66	80	73	**83**	67	**84**	79	**83**	**100**	93	**100**	23	**100**	84	21	19	21	34	**100**	37	**100**	**100**	**100**	50	0	26	80	26	**100**	84	17	71	71	57	89	89	67
	Participle clause	33	70	82	67	79	**94**	76	42	82	42	70	42	9	42	9	**100**	52	94	45	58	12	9	9	30	9	0	**82**	70	**79**	52	33	55	76	70	61	76	76	54
	Pseudo-cleft sentence	51	**84**	78	**86**	**86**	78	65	73	73	65	63	65	69	65	61	75	**100**	88	39	75	55	94	84	88	43	35	75	78	76	76	78	63	**92**	71	65	63	63	72
	Relative clause	117	66	56	66	**71**	63	**74**	66	67	81	41	64	35	64	**91**	85	78	86	47	83	36	62	74	66	25	32	54	**85**	68	**84**	**85**	32	67	58	60	71	71	64
	Subject clause	38	55	53	53	55	**61**	45	**61**	47	84	58	**97**	82	**97**	84	61	76	84	47	84	24	84	82	53	63	5	50	68	37	**84**	**87**	47	53	45	79	50	50	62
Verb tense/aspect/mood	Conditional	6	**100**	67	**100**	**100**	**100**	83	**100**	83	0	17	**83**	17	**83**	0	67	17	33	17	17	67	0	0	**83**	50	33	**100**	**100**	**100**	67	67	50	**100**	50	**100**	83	83	62
	Ditransitive	13	85	69	77	**100**	69	92	**100**	69	69	**92**	**92**	62	**92**	69	69	62	**92**	54	69	46	**92**	77	**92**	31	85	**100**	**100**	**100**	92	92	85	62	69	62	85	85	79
	Gerund	34	82	74	**100**	**100**	88	82	91	85	91	**97**	91	88	91	85	91	91	94	38	**100**	26	**100**	**100**	**100**	18	74	91	**97**	91	91	82	79	79	71	68	82	82	83
	Imperative	12	**75**	67	**75**	**75**	58	**75**	**75**	**75**	**100**	**100**	83	83	83	75	**100**	**100**	92	33	**100**	75	75	50	83	33	17	**100**	**100**	**100**	**100**	**100**	**100**	92	75	92	75	75	80
	Intransitive	3	67	33	67	67	67	**100**	67	67	**100**	**100**	67	**100**	67	**100**	**100**	**100**	**100**	0	**100**	**100**	67	67	**100**	0	0	**100**	**100**	**100**	**100**	**100**	**100**	67	**100**	33	67	67	76
	Reflexive	24	71	54	62	62	58	**88**	71	75	**100**	92	71	33	71	83	71	79	71	50	67	88	88	58	83	50	17	**100**	92	**100**	96	**100**	75	83	67	83	83	83	74
	Transitive	43	77	56	63	**84**	79	**81**	72	65	65	53	49	51	49	37	35	33	28	35	58	67	37	33	**72**	51	53	77	88	77	70	**95**	60	74	70	67	84	84	62
Verb valency	Case government	3	**100**	33	0	33	**100**	33	33	33	**67**	**67**	**67**	0	**67**	0	0	0	33	0	**67**	33	0	0	0	33	33	0	**67**	33	**67**	**67**	0	33	**67**	33	33	33	35
	Catenative verb	50	78	80	74	**90**	86	70	66	80	80	**94**	**96**	56	**96**	90	74	90	74	36	78	42	72	66	76	42	58	76	82	74	74	70	66	90	**92**	**94**	74	74	75
	Impersonal Subject	10	40	50	**80**	40	70	50	20	40	40	**100**	40	90	40	40	**100**	**100**	**100**	40	40	0	40	40	40	10	60	**100**	**100**	**100**	40	40	40	**100**	40	40	50	50	56
	Resultative	58	67	60	**69**	**71**	67	**69**	55	64	69	**83**	71	74	71	64	78	**83**	76	36	**81**	17	62	64	64	12	31	**81**	78	**79**	76	76	72	**79**	69	45	69	69	65
macro avg.		1727	73	62	74	**77**	**75**	68	72	67	70	71	68	51	68	60	65	**75**	67	40	**72**	44	68	65	**73**	41	31	77	**83**	76	77	73	55	76	69	64	66	68	66
micro avg.		1727	69	64	72	**76**	73	65	70	67	**74**	67	68	51	68	**71**	67	**74**	64	44	**72**	39	**73**	68	**71**	40	32	73	**82**	71	77	74	55	71	67	62	63	66	66

Table 6: Accuracy of the metrics(%) with regards to the linguistically-motivated phenomena for English-Russian