Machine Translation Metrics are better in evaluating Linguistic Errors on LLMs than on Encoder-Decoder Systems

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

Abstract

This year's MT metrics challenge set submission by DFKI expands previous years' linguistically motivated challenge set. It includes 137,000 items extracted from 100 MT systems for the two language directions (en→de, en→ru), covering more than 100 linguistically motivated phenomena organized in 14 linguistic categories. The metrics with the statistically significant best performance with regard to our linguistically motivated analysis are METRICX-24-HYBRID and METRICX-24 for en→de and METRICX-24 for en→ru, whereas METAMETRICS and XCOMET are in the next ranking positions in both language pairs. Metrics are more accurate in detecting linguistic errors among LLM translations than in translations based on the encoder-decoder NMT architecture. Some of the most difficult phenomena for the metrics to score are the transitive past progressive, the multiple connectors, and the ditransitive simple future I for en→de and the pseudogapping, the contact clause and the cleft sentences for en→ru. Despite its overall low performance, the LLM-based metric GEMBA performs best in scoring German negation errors.

1 Introduction

For almost two decades, the development and evaluation of machine translation (MT) have relied on automatic metrics. MT metrics aim to digest and automate various aspects of human judgment of MT output into numerical scores. Over the years, these metrics have undergone several technological changes (from measuring overlap to grammatical features and neural models). Still, at the same time, they have had to follow the technological evolution of MT systems, moving from phrase-based statistical systems to NMT encoder-decoder models and, more recently, to large language models (LLMs). As we witness the first efforts to use and evaluate LLMs in the task of MT, it is of great interest to see to what extent pre-existing MT methodologies can adapt to the needs of the new technologies. An obvious question is to what extent MT metrics developed and tested for NMT can be applied to evaluating LLMs.

This year's Metrics Task (WMT24; Freitag et al., 2024) provides a very good opportunity to evaluate the metrics under these particular circumstances, as the evaluated MT outputs have for the first time been produced by numerous LLMs (Kocmi et al., 2024). Meanwhile, the ability of LLMs to act as judges for translations is being explored through the participation of an LLM-based metric.

Given this perspective, this paper extends previous work on linguistically motivated challenge sets for MT metrics to investigate whether LLMs can influence MT evaluation. As part of this year's submission to the challenge set subtask of the WMT24 Metrics Task, we repeat the methodology of previous years to evaluate the metrics on a controlled test set that can rank them with regard to their ability to detect linguistic errors by providing fine-grained statistics for each linguistic phenomenon. We then analyze whether the metrics perform differently on MT output from LLMs as opposed to output from encoder-decoder systems. In addition, we see in which linguistic aspects the LLM-based metric performs better or worse than the specialized metrics.

The rest of the paper is structured as following: Section 2 describes briefly the generation of the challenge set. Section 3 presents and discusses the results, whereas the conclusion is given in section 4

2 Method

This year's linguistically-motivated challenge set is an extension of the challenge sets that were submitted the previous years (Avramidis and Macketanz, 2022; Avramidis et al., 2023).

The source sentences s originate from an MT evaluation test suite (Macketanz et al., 2022a). Each sentence has been carefully constructed to test one particular phenomenon. Every phenomenon is

Proceedings of the Ninth Conference on Machine Translation, pages 517–528
November 15-16, 2024 ©2024 Association for Computational Linguistics

tested by more sentences (with a minimum of 20 sentences), whereas the phenomena are aggregated in a few categories. At the moment, there are more than 100 phenomena and 14 categories.

As part of the WMT shared tasks of the previous years, these source sentences have been given to a large amount of MT systems, and their output has been evaluated by combining regular expressions and annotations by linguists, labeling every output as correct ($t \in T$) or incorrect ($\hat{t} \in T'$).

In order to use this test set to evaluate the MT metrics, we create examples in the form of $(s, \hat{t}, t, r) \in S$, where each example contains one source sentence s, one incorrect translation hypothesis $\hat{t}$, one correct translation hypothesis t and one reference translation r. The correct translation hypotheses t and the reference translations r are sampled with permutations from the same set of correct translations T. Then, we decompose the set of examples S into a blind test set S', where each example includes either an incorrect translation $(s, \hat{t}, r)$ or a correct translation (s, t, r) along with the source and the reference. The separated contrastive examples are shuffled, and we set aside a file that contains the golden truth, indicating which samples are correct or incorrect.

As part of the Metrics Task, every shuffled translation t and $hatt$ is scored by every M, given the reference r in the given blind test set S', without knowing if it is correct or incorrect. A contrastive pair scoring is considered correct if the metric delivers a score for the incorrect translation hypothesis, which is lower than the one of the correct translation hypothesis $M(s, \hat{t}, r) < M(s, t, r)$. Finally, for every phenomenon and category and for every metric, the respective accuracy is calculated by dividing the number of correctly scored contrastive pairs by the total amount of examples.

$$\text{acc}_M = \frac{|M(s, \hat{t}, r) < M(s, t, r)|}{|(s, \hat{t}, t, r)|}$$

$$(s, \hat{t}, r) \cup (s, t, r) \in S' \quad (s, \hat{t}, t, r) \in S$$

Lastly, we provide three types of score averaging:

i) **Micro-average:** This approach treats 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.

The current version of the challenge set contains MT outputs from the WMT Shared Tasks of the years 2019-2024 (Avramidis et al., 2019, 2020; Macketanz et al., 2021, 2022b; Manakhimova et al., 2023, 2024). The English to German version contains 39,463 contrastive pairs, while the English to Russian version contains 30,108 pairs.

3 Results

3.1 English-German

The comparison of the metrics based on the accuracies per category for English-German can be seen in table 2, whereas the detailed phenomena in table 4. One can see that the metrics which have the highest accuracy with statistical significance are METRICX24-HYBRID and METRICX24 (Juraska et al., 2024), with more than 80.7 % macro-average. Both metrics are very good at multi-word expressions (mostly verbal MWEs). The former is the best of all metrics at coordination/ellipsis and non-verbal agreement (genitive and personal pronoun coreference). In contrast, the latter performs best at verb valency (resultative and passive voice). The metrics “METAMETRICS” (Anugraha et al., 2024) and XCOMET (Guerreiro et al., 2023) follow in the ranking, with more than 80% macro-averaged accuracy.

The LLM-based metric GEMBA (Kocmi and Federmann, 2023) performs relatively low, with an average accuracy of 69.7%, even below the baseline non-tuned metric CHRF (Popović, 2015). It is nevertheless remarkable that this metric has the best score on negation, among all metrics (97.4%, 4.5% higher than the best system). The fact that most of the metrics will miss 10% of the negations is rather noteworthy, given the implications of such a mistake on the meaning of the sentence. It is also remarkable that a reference-less metric, METRICX24-HYBRID-QE, achieves the highest accuracy on long-distance dependencies and interrogatives, mainly on the phenomenon of negative inversion.

Some of the most difficult phenomena for the

	METRICX24	METRICX24-HYB	METAMETRICS	XCOMET
encdec vs. encdec	73.2	72.3	70.8	69.7
LLM vs. encdec	77.3	76.9	79.9	77.6
LLM vs. LLM	79.9	78.1	80.0	79.1

Table 1: Accuracy of the metrics when they evaluate contrastive pairs containing (a) MT output only by encoder/decoder systems, (b) one encoder/decoder output and one LLM output, (c) only LLM output

metrics to score are transitive past progressive, multiple connectors, and ditransitive simple future I.

3.2 English-Russian

The comparison of the metrics based on the accuracies per category for English-Russian can be seen in table 3, whereas the detailed phenomena in table 5. MetricX-24 is the clear winner in this language direction, achieving a macro-averaged accuracy of 82.5% MetricX-24 excels in ambiguity, false friends, non-verbal agreement (coreference & genitive), verb semantics, and verb valency. The ranking of the metrics is similar to the one for English-German, with METAMETRICS, METRICX24-HYBRID and XCOMET having the next position, with more than 79.6% accuracy in macro-average.

If one focuses again on the phenomenon of negation, they would notice that in English-Russian, the highest accuracy is achieved by the baseline metric CHRF, whereas most metrics perform here very low (61% on average) Some of the most difficult phenomena for this language direction are the pseudogapping, the contract clause, and the cleft sentences for en→ru.

3.3 Comparing performance of metrics over LLM vs. encoder-decoder systems

Table 1 presents the accuracies of the 4 best performing metrics on three subsets of the challenge sets. Here every subset contains contrastive pairs which consist of

(a) two MT outputs, both by encoder/decoder NMT systems
(b) one encoder/decoder and one LLM output
(c) two LLM outputs

One can see that all four metrics exhibit higher accuracy when scoring contrastive translations originating from LLMs. This indicates that despite the fact that LLM translations achieve very good performance (Kocmi et al., 2024), their fewer errors are easier to be distinguished by the automatic metrics. Whether there is a systematic reason for this phenomenon remains to be investigated.

4 Conclusion

We presented the MT metrics challenge set of DFKI for two language directions (en-de, en-ru). This year, we have expanded the set to include outputs from encoder-decoder NMT systems and LLMs. The number of test items (total of 137,000) allows for producing fine-grained scores for every linguistic phenomenon and statistically significant comparisons among the MT metrics. We also identified the best-performing metric, METRICX-24, for both language directions.

Acknowledgements

This research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft; DFG) through the project TextQ. 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 to the creation of the test suite.

References

David Anugraha, Garry Kuwanto, Lucky Susanto, Derry Tanti Wijaya, and Genta Indra Winata. 2024. Metametrics-MT: Tuning machine translation metametrics via human preference calibration. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

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, 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.

Eleftherios Avramidis, Shushen Manakhimova, Vivien Macketanz, and Sebastian Möller. 2023. Challenging the state-of-the-art machine translation metrics from a linguistic perspective. In *Proceedings of the Eighth Conference on Machine Translation*, pages 713–729, Singapore. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are llms breaking mt metrics? results of the wmt24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. Metricx-24: The google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: the LLM era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. 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.

Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can ChatGPT outperform NMT? In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245, Singapore. Association for Computational Linguistics.

Shushen Manakhimova, Vivien Macketanz, Eleftherios Avramidis, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2024. Investigating the linguistic performance of large language models in machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

A Accuracies per category

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

		metric																										
ling. category	#	MetricX-24-Hybrid	MetricX-24	metametrics	XCOMET	BLEURT-20	COMET-22	CometKiwi-XXL	MetricX-24-QE	MetricX-24-Hybrid-QE	XCOMET-QE	YiSi-1	sentinel-cand-mqm	CometKiwi	MEE4	chrfS	BERTScore	chrF	gemba	spBLEU	damonmonli	monmonli	BLEU	XLsimMqm	XLsimDA	PrismRefSmall	PrismRefMedium	avg
Ambiguity	4614.0	85.1	85.9	**89.9**	80.0	**89.7**	**89.5**	60.8	74.6	70.6	61.9	88.6	77.7	48.2	82.1	83.8	78.2	85.2	70.0	80.0	83.8	83.0	64.1	55.2	55.2	68.1	60.6	75.1
Coordination & ellipsis	4373.0	**81.3**	74.2	74.4	77.4	76.5	76.7	80.2	78.2	78.8	74.4	69.2	76.7	71.1	63.8	62.9	67.3	62.2	66.5	61.8	61.0	62.9	60.6	49.5	49.5	51.1	49.1	67.6
False friends	1389.0	79.9	78.2	78.3	73.9	72.7	**85.9**	85.2	69.8	74.3	71.1	73.1	77.0	72.1	80.4	77.1	69.2	74.9	81.9	65.9	48.6	38.2	64.1	78.3	78.3	74.3	58.8	72.4
Function word	1900.0	78.1	80.6	82.2	86.0	81.9	**87.3**	83.0	81.7	78.6	82.2	72.9	85.8	86.7	76.9	77.2	**86.9**	74.4	70.9	74.2	64.9	60.1	78.6	55.7	55.7	52.2	53.8	74.9
LDD & interrogatives	1002.0	83.4	80.1	80.8	80.6	80.3	74.5	78.7	81.8	**84.7**	81.7	59.1	68.6	78.4	64.2	59.3	64.1	57.5	58.6	61.5	66.7	64.5	60.5	62.0	62.0	49.6	47.7	68.9
MWE	5816.0	**87.0**	**87.3**	85.9	86.2	84.1	82.9	80.0	82.5	81.2	80.5	80.3	84.0	76.4	75.1	75.9	76.1	73.3	82.0	70.7	77.3	76.6	71.5	67.0	67.0	59.4	55.6	77.1
Named entity & terminology	22891.0	71.5	74.2	74.2	68.8	71.7	73.6	58.0	55.3	60.9	56.9	74.7	52.1	50.2	72.2	70.5	67.1	68.9	48.1	70.0	**75.4**	73.1	62.0	48.5	48.5	49.8	50.1	63.3
Negation	506.0	92.9	89.5	88.5	91.1	92.7	92.9	93.3	93.9	91.3	90.9	87.9	74.5	95.3	90.7	82.8	86.0	76.7	**97.4**	73.9	86.6	88.3	73.7	58.3	58.3	58.5	58.1	83.2
Non-verbal agreement	15497.0	**83.6**	80.6	77.4	80.9	78.2	73.3	80.2	82.3	82.4	79.2	65.7	76.2	72.9	65.6	66.1	63.7	65.9	72.7	64.3	59.6	59.5	62.2	57.8	57.8	51.0	49.0	69.5
Punctuation	2435.0	62.2	64.4	64.9	63.2	71.9	72.4	70.1	70.4	65.9	64.9	71.6	**80.1**	71.3	69.9	72.1	66.0	68.5	44.3	67.3	66.9	50.7	68.7	50.3	50.3	50.6	50.8	64.2
Subordination	4698.0	89.1	87.5	86.3	**89.3**	84.1	83.9	89.2	**89.5**	**89.4**	86.9	78.9	80.8	**89.8**	76.6	76.1	76.4	74.1	72.6	72.3	66.1	70.9	73.9	44.4	44.4	57.5	54.3	76.3
Verb tense/aspect/mood	10120.0	78.6	81.8	79.2	**83.4**	73.0	68.2	80.6	77.4	77.3	81.8	67.5	52.2	72.1	67.8	67.8	65.9	66.7	73.3	63.0	63.6	66.0	62.6	51.8	51.8	59.1	52.7	68.7
Verb valency	3486.0	80.8	**84.6**	**84.5**	81.7	81.7	77.0	83.8	82.9	80.3	80.9	73.7	75.5	73.2	67.4	71.1	67.9	71.6	67.4	67.2	70.5	71.3	62.3	61.2	61.2	55.0	53.9	72.6
macro avg.	78727.0	**81.0**	**80.7**	80.5	80.2	79.9	79.9	78.7	78.5	78.1	76.4	74.1	73.9	73.7	73.3	72.5	71.9	70.8	69.7	68.6	68.5	66.5	66.5	56.9	56.9	56.6	53.4	71.8
micro avg.	78727.0	**79.1**	**79.4**	78.6	77.9	77.1	76.0	73.3	73.1	74.2	72.0	72.8	67.3	66.4	70.9	70.7	68.8	69.5	64.8	68.0	68.8	67.9	64.3	53.9	53.9	54.4	51.9	69.0

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

		metric																								
ling. category	#	MetricX-24	metametrics	MetricX-24-Hybrid	XCOMET	BLEURT-20	COMET-22	CometKiwi-XXL	MetricX-24-QE	XCOMET-QE	MetricX-24-Hybrid-QE	YiSi-1	CometKiwi	BERTScore	sentinel-cand-mqm	chrfS	chrF	spBLEU	BLEU	damonmonli	monmonli	gemba	XLsimMqm	XLsimDA	PrismRefSmall	PrismRefMedium	avg
Ambiguity	3788.0	**96.9**	96.4	95.1	93.2	89.8	87.4	80.9	96.3	83.8	91.4	82.6	77.2	75.3	87.1	74.7	73.1	70.6	68.9	80.5	78.4	89.4	43.9	43.9	48.1	45.3	78.0
Coordination & ellipsis	2273.0	80.6	79.3	**81.5**	80.4	74.9	76.6	81.0	**81.4**	77.2	**81.8**	68.6	78.6	68.1	75.5	63.5	61.5	62.7	62.7	65.4	66.3	60.1	52.5	52.5	47.7	48.7	69.2
False friends	2414.0	**87.8**	83.7	86.3	76.3	82.4	83.0	69.1	69.2	68.6	68.4	**87.7**	52.4	76.3	58.0	84.9	83.2	80.7	75.8	80.8	62.2	34.0	43.2	43.2	53.1	42.0	69.3
Function word	2433.0	82.5	78.0	73.4	84.1	79.7	81.4	83.0	85.7	**86.3**	71.8	65.7	79.3	69.3	82.7	64.8	60.3	65.6	73.2	56.4	57.0	56.3	73.7	73.7	50.3	49.0	71.3
LDD & interrogatives	1939.0	85.4	86.0	**87.8**	84.8	81.8	82.6	84.9	87.3	83.4	**87.6**	65.5	77.6	66.2	**87.8**	62.5	59.9	62.0	61.5	55.1	58.3	68.1	54.2	54.2	51.6	46.4	71.3
MWE	9602.0	82.9	82.9	82.9	81.2	80.5	81.4	82.2	81.6	77.7	**83.9**	77.0	75.6	74.6	72.3	75.5	73.1	72.8	70.9	67.9	65.1	69.5	53.5	53.5	51.7	51.0	72.8
Named entity & terminology	16284.0	82.8	**84.9**	81.6	80.6	**84.9**	84.3	71.6	72.2	69.5	71.6	83.0	71.3	78.8	70.6	80.3	78.7	78.3	72.5	72.9	67.7	64.1	47.6	47.6	53.6	52.1	72.1
Negation	346.0	65.3	59.8	58.7	49.4	72.3	67.3	58.7	57.8	45.4	44.5	79.5	49.4	80.3	41.3	82.9	**83.5**	74.3	72.3	70.5	72.5	42.5	49.7	49.7	49.4	45.7	60.9
Non-verbal agreement	6755.0	**86.4**	81.5	84.4	82.3	79.6	77.4	78.7	83.0	80.9	81.5	72.1	77.6	68.7	73.1	69.4	68.2	67.4	64.6	59.6	60.9	68.4	68.4	68.4	51.2	47.5	72.1
Punctuation	363.0	73.0	71.1	72.7	71.3	68.6	**76.0**	**75.8**	63.6	70.8	67.2	**75.8**	72.2	73.3	70.5	62.0	58.4	64.7	60.9	51.0	64.5	60.9	57.3	57.3	46.3	45.5	65.2
Subordination	6625.0	74.7	74.5	71.4	75.0	72.7	77.1	**78.4**	72.5	75.2	71.7	69.3	68.6	66.9	73.5	63.8	62.4	64.3	64.0	56.4	63.3	53.6	50.5	50.5	51.0	48.1	66.0
Verb semantics	275.0	**88.0**	82.2	**88.0**	85.5	86.5	74.2	79.6	75.3	80.0	76.0	53.1	69.8	55.6	55.3	60.7	65.1	53.8	48.7	68.4	66.5	72.0	33.5	33.5	65.5	67.3	67.4
Verb tense/aspect/mood	2994.0	85.0	**86.0**	82.6	**86.2**	75.5	79.7	**85.8**	82.8	80.7	79.3	69.7	72.6	68.7	70.4	68.1	66.7	68.8	63.1	60.1	55.9	61.6	47.5	47.5	50.6	51.3	69.9
Verb valency	3022.0	**83.3**	82.3	80.4	**83.5**	76.8	76.3	80.9	82.0	81.6	81.5	69.6	73.8	72.8	72.0	72.2	71.8	69.0	67.7	66.6	64.2	66.9	60.7	60.7	51.6	46.7	71.8
macro avg.	59113.0	**82.5**	80.6	80.5	79.6	79.0	78.9	77.9	77.9	75.8	75.6	72.8	71.1	71.1	70.7	70.4	69.0	68.2	66.2	65.1	64.5	62.0	52.6	52.6	51.6	49.0	69.8
micro avg.	59113.0	**83.4**	82.8	81.7	81.4	80.7	81.0	77.8	78.7	76.2	77.5	75.8	72.9	72.9	73.0	73.1	71.4	71.2	68.5	66.8	64.8	64.4	52.8	52.8	51.7	49.3	71.3

B Accuracies per phenomenon

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

			metric																									
ling. category	ling. phenomenon	\#	XCOMET	MetricX-24	MetricX-24-Hybrid	metametrics	MetricX-24-QE	MetricX-24-Hybrid-QE	XCOMET-QE	CometKiwi-XXL	BLEURT-20	CometKiwi	COMET-22	chrfS	MEE4	chrF	gemba	BERTScore	YiSi-1	spBLEU	BLEU	monmonli	damonmonli	sentinel-cand-mqm	PrismRefSmall	XLsimDA	XLsimMqm	PrismRefMedium	avg
Ambiguity	Lexical ambiguity	4614	80	86	85	**90**	75	71	62	61	**90**	48	**89**	84	82	85	70	78	89	80	64	83	84	78	68	55	55	61	75
Coordination \& ellipsis	Gapping	605	86	87	**91**	84	87	90	77	86	87	85	84	70	69	74	73	75	78	70	64	47	47	87	62	47	47	59	74
	Pseudogapping	1565	87	85	**94**	85	93	91	82	93	88	78	87	72	72	71	82	76	73	70	70	63	64	87	53	39	39	50	75
	Right node raising	647	74	67	76	66	76	68	75	76	66	64	75	59	61	55	**81**	64	66	54	55	67	55	76	55	53	53	52	65
	Sluicing	472	64	57	61	59	62	64	57	62	57	50	61	57	62	56	34	58	63	58	57	**66**	61	59	46	53	53	46	57
	Stripping	545	59	60	66	66	54	64	59	59	69	69	66	53	56	52	40	62	**71**	54	56	71	**71**	59	46	57	57	44	59
	VP-ellipsis	539	76	66	71	64	68	71	79	**80**	68	65	65	48	45	47	51	50	55	48	45	64	65	71	36	69	69	40	61
False friends	False friends	1389	74	78	80	78	70	74	71	85	73	72	**86**	77	80	75	82	69	73	66	64	38	49	77	74	78	78	59	72
Function word	Focus particle	333	62	61	65	58	52	47	61	55	57	69	71	**77**	70	74	31	72	62	74	59	35	40	31	53	47	47	56	57
	Question tag	1567	91	85	81	87	88	85	87	89	87	90	91	77	78	74	79	90	75	74	83	65	70	**97**	52	58	58	53	79
LDD \& interrogatives	Extraposition	85	79	84	**86**	76	80	76	75	78	67	76	64	71	71	72	49	66	65	66	61	68	62	74	65	58	58	66	70
	Inversion	117	81	84	87	79	85	84	**92**	82	78	85	79	69	74	67	68	79	75	68	62	75	81	79	52	85	85	49	76
	Multiple connectors	25	44	44	56	40	44	56	32	44	44	40	36	68	64	68	0	72	72	64	68	64	28	44	36	**80**	**80**	40	51
	Negative inversion	358	76	72	73	74	73	**83**	82	70	78	67	61	46	55	44	55	58	42	53	55	57	60	47	39	65	65	35	61
	Pied-piping	49	98	**100**	94	94	**100**	**100**	98	88	94	90	92	71	80	71	69	84	67	88	88	71	76	63	59	71	71	51	82
	Polar question	13	92	77	77	77	92	69	69	92	92	77	**100**	77	85	77	**100**	85	85	77	77	38	54	**100**	62	23	23	54	74
	Preposition stranding	9	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	78	**100**	78	**100**	**100**	**100**	67	56	78	89	**100**	**100**	89	89	**100**	93
	Split infinitive	88	91	98	**100**	94	98	**100**	92	94	**100**	86	89	56	62	51	60	61	67	55	50	82	75	90	39	67	67	41	76
	Topicalization	192	86	83	90	95	90	85	81	89	82	**96**	91	65	68	64	64	64	69	68	68	52	61	92	59	45	45	60	73
	Wh-movement	66	74	77	91	73	80	85	67	74	83	73	77	70	61	67	55	47	55	61	50	91	**95**	65	62	42	42	58	68
MWE	Collocation	506	90	90	88	88	79	78	86	81	84	**91**	84	80	82	77	76	78	83	74	75	77	67	80	59	62	62	52	78
	Compound	257	94	92	95	95	94	**98**	97	94	75	88	84	70	72	72	80	73	91	61	48	68	74	96	48	67	67	46	78
	Idiom	2746	91	93	93	93	90	89	87	88	93	81	95	84	85	77	**97**	85	89	76	81	86	89	90	60	74	74	57	85
	Nominal MWE	1522	75	77	76	72	69	67	64	62	71	62	64	64	58	68	59	63	65	60	59	64	70	**78**	62	53	53	56	65
	Prepositional MWE	353	**92**	79	80	89	76	72	86	82	82	84	**91**	73	74	69	69	75	79	78	68	67	51	79	52	59	59	50	74
	Verbal MWE	432	84	**88**	**88**	83	87	81	78	79	80	66	65	72	68	72	84	69	69	68	65	71	66	70	58	86	86	57	75
Named entity \& terminology	Date	2010	69	84	75	**89**	69	66	56	63	82	64	78	76	73	76	64	77	76	71	80	65	66	75	58	64	64	56	71
	Domainspecific Term	7405	79	80	78	**91**	66	73	66	72	78	63	83	72	76	71	63	68	86	76	65	64	71	55	52	46	46	46	69
	Location	2731	70	**92**	85	88	32	34	13	15	91	18	**93**	88	89	83	31	80	87	85	55	91	90	69	65	48	48	54	65
	Measuring unit	8539	58	61	60	53	45	55	59	54	58	44	59	64	65	62	32	61	62	61	58	**80**	**80**	34	41	43	43	51	55
	Proper name	2206	**76**	74	**76**	72	75	73	75	73	69	61	70	63	64	64	70	65	72	66	59	63	64	71	49	63	63	48	67
Negation	Negation	506	91	90	93	89	94	91	91	93	93	95	93	83	91	77	**97**	86	88	74	74	88	87	75	58	58	58	58	83
Non-verbal agreement	Coreference	3340	**85**	80	78	82	81	79	80	**85**	82	65	78	77	75	78	76	68	67	76	79	47	55	79	57	58	58	56	72
	Genitive	483	82	93	90	**94**	86	87	92	75	67	78	89	68	70	68	87	75	81	71	70	67	59	74	58	64	64	49	75
	Lexical Morphology/Functional shift	2330	91	94	**97**	95	90	94	92	92	90	80	91	79	81	77	76	76	85	70	65	84	77	96	59	56	56	50	81
	Lexical Morphology/Noun formation (er)	2067	72	**77**	70	76	66	64	68	63	77	65	68	63	60	65	55	61	64	61	54	64	65	65	45	74	74	44	65

Continued on next page

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

			metric																										
ling. category	ling. phenomenon	\#	XCOMET	MetricX-24	MetricX-24-Hybrid	metametrics	MetricX-24-QE	MetricX-24-Hybrid-QE	XCOMET-QE	CometKiwi-XXL	BLEURT-20	CometKiwi	COMET-22	chrfS	MEE4	chrF	gemba	BERTScore	YiSi-1	spBLEU	BLEU	monmonli	damonmonli	sentinel-cand-mqm	PrismRefSmall	XLsimDA	XLsimMqm	PrismRefMedium	avg
	Personal Pronoun Coreference	4632	83	80	93	67	91	**94**	83	83	76	79	65	53	53	52	84	54	56	53	50	54	52	75	49	47	47	48	66
	Possession	555	89	88	88	86	89	88	88	**90**	79	87	82	74	76	70	83	72	73	68	67	81	76	66	49	49	49	50	75
	Substitution	2090	65	67	67	68	**71**	67	60	69	67	68	64	65	62	64	49	62	59	65	64	53	56	69	41	69	69	45	62
Punctuation	Quotation marks	2435	63	64	62	65	70	66	65	70	72	71	72	72	70	69	44	66	72	67	69	51	67	**80**	51	50	50	51	64
Subordination	Adverbial clause	583	92	90	92	89	90	93	93	94	83	**94**	87	69	72	68	81	74	73	70	75	62	66	88	47	59	59	47	77
	Cleft sentence	578	78	75	77	76	81	82	76	81	72	**84**	77	68	68	64	64	67	73	64	66	70	53	67	57	54	54	52	69
	Contact clause	788	98	91	94	96	97	97	98	97	96	**99**	95	78	79	74	88	80	90	72	73	77	78	97	62	35	35	59	82
	Indirect speech	113	88	79	79	78	65	70	**89**	78	65	85	75	58	66	53	56	65	62	58	58	46	59	61	50	38	38	48	64
	Infinitive clause	454	89	87	91	87	**94**	**93**	90	87	81	91	87	78	72	80	68	67	80	73	67	85	84	61	50	59	59	44	77
	Object clause	111	95	85	95	74	**97**	95	92	86	61	58	45	59	54	64	73	50	50	47	49	65	56	55	42	39	39	38	64
	Pseudo-cleft sentence	578	76	78	76	66	70	66	69	80	72	71	66	80	**82**	78	66	78	74	77	**82**	66	54	66	52	42	42	53	69
	Relative clause	560	95	96	96	95	95	96	**98**	97	94	**98**	93	76	79	75	94	84	75	73	76	67	68	81	56	65	65	53	82
	Subject clause	933	91	93	94	92	**97**	95	83	88	91	93	87	85	84	83	56	84	88	81	81	74	64	95	71	20	20	66	79
Verb tense/aspect/mood	Conditional	975	87	73	65	90	78	62	89	81	86	**95**	82	66	63	61	42	60	69	49	53	75	86	72	67	68	68	56	71
	Ditransitive - conditional I progressive	83	89	73	69	61	70	66	77	**93**	59	64	55	67	73	65	73	65	64	67	63	43	35	20	58	42	42	59	62
	Ditransitive - conditional I simple	197	91	82	86	71	**94**	92	90	87	70	87	74	70	69	70	85	68	71	69	60	74	71	55	52	45	45	49	72
	Ditransitive - conditional II progressive	130	91	92	92	87	81	**95**	88	92	85	72	78	78	82	72	48	75	76	78	72	65	71	61	55	51	51	58	75
	Ditransitive - conditional II simple	108	85	88	**93**	78	78	84	77	79	66	71	70	69	74	58	59	68	62	65	61	60	67	72	40	48	48	49	68
	Ditransitive - future I progressive	244	82	70	76	68	59	78	89	**89**	75	47	58	60	58	61	30	57	60	60	62	49	51	73	52	48	48	50	62
	Ditransitive - future I simple	217	78	70	77	61	63	86	**91**	76	61	47	45	53	54	54	28	59	54	51	58	31	38	46	41	34	34	51	55
	Ditransitive - future II progressive	210	**94**	91	82	89	88	70	89	87	67	90	66	76	81	68	92	73	75	75	75	70	65	58	46	68	68	46	75
	Ditransitive - future II simple	84	79	94	94	85	**100**	**100**	65	89	79	90	69	89	89	87	96	73	87	83	82	96	71	40	71	57	57	60	80
	Ditransitive - past perfect progressive	122	65	66	61	56	79	**92**	47	57	50	75	66	59	64	55	64	51	54	47	51	52	54	55	46	62	62	48	59
	Ditransitive - past perfect simple	160	61	67	61	57	59	**86**	64	71	63	51	52	57	56	61	20	57	60	51	53	49	50	71	52	42	42	51	56
	Ditransitive - past progressive	218	**74**	71	73	71	53	61	71	71	61	50	54	53	51	50	50	55	54	47	52	51	60	55	49	49	49	54	57
	Ditransitive - present perfect progressive	107	97	85	82	81	93	62	**98**	93	74	81	68	61	66	63	56	64	73	46	48	62	66	80	53	62	62	53	70
	Ditransitive - present perfect simple	185	90	71	77	68	60	82	**99**	96	66	40	52	52	54	52	19	64	57	52	56	43	50	62	35	41	41	46	59
	Ditransitive - present progressive	114	98	87	86	86	97	89	**99**	96	80	**99**	89	71	72	68	97	61	78	54	48	89	89	83	60	71	71	62	80
	Ditransitive - simple past	199	94	82	87	73	**99**	97	91	96	74	91	74	68	67	62	93	67	65	55	54	50	53	53	43	69	69	51	72
	Ditransitive - simple present	133	92	81	73	82	86	90	**95**	94	83	83	83	67	70	61	81	69	60	54	50	73	76	79	47	71	71	50	74
	Gerund	1119	**98**	98	**98**	98	**98**	**98**	97	95	95	97	95	78	82	73	87	79	82	73	74	80	69	60	66	22	22	52	80
	Imperative	259	**90**	75	81	83	90	77	88	**90**	80	89	86	76	78	72	88	69	74	68	62	79	69	63	57	50	50	59	75
	Intransitive - conditional I progressive	23	70	91	91	74	70	70	48	39	87	52	**100**	78	78	61	96	78	78	78	61	83	83	70	39	83	83	39	72
	Intransitive - conditional I simple	15	93	87	**100**	**100**	27	40	20	20	93	47	**100**	93	**100**	67	**100**	93	87	73	7	73	73	20	67	93	93	73	71
	Intransitive - conditional II progressive	5	80	**100**	**100**	**100**	60	**100**	60	60	80	60	80	**100**	**100**	80	**100**	**100**	60	**100**	**100**	60	60	40	80	80	80	80	81
	Intransitive - conditional II simple	2	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	50	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	98
	Intransitive - future I progressive	19	**100**	**100**	**100**	84	**100**	**100**	**100**	74	89	**100**	84	84	89	89	79	89	74	84	89	68	63	**100**	47	79	79	47	84
	Intransitive - future I simple	50	78	88	88	88	52	62	72	48	**92**	56	88	**92**	86	84	82	90	72	86	60	78	82	58	60	82	82	72	76
	Intransitive - future II progressive	18	78	78	78	78	56	50	50	61	78	72	72	**89**	78	**89**	72	83	44	83	**89**	72	72	22	72	50	50	67	69
	Intransitive - future II simple	15	93	93	93	93	93	93	87	93	80	93	67	**100**	93	93	80	93	67	**100**	**100**	80	80	20	**100**	13	13	80	81

Continued on next page

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

ling. category	ling. phenomenon	\#	metric																									
			XCOMET	MetricX-24	MetricX-24-Hybrid	metametrics	MetricX-24-QE	MetricX-24-Hybrid-QE	XCOMET-QE	CometKiwi-XXL	BLEURT-20	CometKiwi	COMET-22	chrfS	MEE4	chrF	gemba	BERTScore	YiSi-1	spBLEU	BLEU	monmonli	damonmonli	sentinel-cand-mqm	PrismRefSmall	XLsimDA	XLsimMqm	PrismRefMedium	avg
	Intransitive - past perfect progressive	81	38	56	52	48	20	16	27	25	60	70	65	72	70	63	51	64	74	67	67	69	65	**75**	49	69	69	62	56
	Intransitive - past perfect simple	31	74	71	65	87	68	68	48	58	77	81	71	87	87	84	45	**94**	71	81	84	45	55	74	71	32	32	77	69
	Intransitive - past progressive	79	70	70	**78**	76	76	56	67	66	62	61	71	61	65	56	67	65	52	62	61	59	62	34	58	75	75	56	64
	Intransitive - present perfect simple	20	95	**100**	**100**	**100**	**100**	**100**	**100**	95	**100**	95	**100**	**100**	**100**	95	**100**	**100**	70	85	85	75	85	80	90	90	90	80	93
	Intransitive - present progressive	26	92	**100**	88	92	69	58	96	69	81	69	92	96	92	77	**100**	85	77	73	58	62	65	62	73	46	46	73	77
	Intransitive - simple past	53	77	83	89	60	91	**92**	81	62	58	58	53	57	62	55	75	45	57	38	38	72	68	45	74	62	62	75	65
	Intransitive - simple present	27	63	67	78	63	67	**89**	74	67	74	**89**	78	48	44	44	85	63	63	37	33	70	74	74	56	30	30	67	63
	Modal	16	81	75	62	62	38	56	62	62	81	38	75	81	81	**88**	56	81	69	62	69	62	81	25	75	69	69	**88**	67
	Modal negated	52	90	90	85	83	**92**	81	**92**	73	71	87	85	73	75	67	67	65	73	58	63	77	79	54	58	67	67	56	74
	Reflexive - conditional I progressive	150	89	81	77	85	59	68	96	95	54	49	48	59	65	59	**100**	60	56	68	60	61	58	35	78	62	62	57	67
	Reflexive - conditional I simple	141	71	78	69	77	67	71	62	68	59	38	45	60	61	61	**82**	56	50	65	62	71	62	35	65	59	59	47	62
	Reflexive - conditional II progressive	204	92	88	88	89	72	91	89	91	69	78	61	74	74	72	**95**	65	66	68	70	63	58	36	72	52	52	55	72
	Reflexive - conditional II simple	336	92	**97**	85	95	81	85	77	79	71	69	54	70	67	71	92	61	65	67	73	57	56	55	71	38	38	45	69
	Reflexive - future I progressive	212	76	72	74	82	59	66	86	79	58	46	59	65	58	70	**96**	70	57	67	58	67	65	50	59	68	68	50	66
	Reflexive - future I simple	160	73	85	75	81	68	69	57	48	73	57	74	80	76	84	**98**	84	64	78	71	86	81	50	78	85	85	57	74
	Reflexive - future II progressive	158	73	80	70	76	66	68	68	64	70	82	71	75	73	81	68	62	78	68	65	67	67	59	61	**83**	**83**	51	70
	Reflexive - future II simple	123	81	89	74	72	89	81	82	**96**	67	67	51	75	70	76	89	58	60	63	67	67	64	31	68	48	48	50	69
	Reflexive - past perfect progressive	162	84	86	79	80	87	81	80	75	73	83	69	75	70	76	**89**	69	70	67	61	76	62	51	70	50	50	70	72
	Reflexive - past perfect simple	169	75	77	73	80	74	68	79	**85**	66	76	57	56	54	59	67	53	57	51	50	66	52	52	51	58	58	49	63
	Reflexive - past progressive	843	73	73	65	70	50	62	73	77	68	55	56	58	56	60	**91**	56	57	58	57	65	61	18	58	59	59	47	61
	Reflexive - present perfect progressive	105	76	75	72	78	63	78	70	**90**	70	54	64	63	66	75	87	62	70	60	58	69	66	20	63	69	69	50	67
	Reflexive - present perfect simple	127	60	72	68	74	**81**	68	59	61	64	61	59	67	66	65	73	64	65	58	56	51	49	37	63	53	53	51	61
	Reflexive - present progressive	586	82	87	82	70	81	74	76	76	64	51	59	63	63	65	**89**	74	77	67	59	64	62	42	53	31	31	48	65
	Reflexive - simple past	256	92	**96**	90	89	95	84	95	91	81	75	69	75	75	81	93	74	84	72	68	63	61	34	50	61	61	44	75
	Reflexive - simple present	330	78	90	80	75	74	78	72	65	69	40	50	61	59	62	**97**	65	78	60	60	67	67	36	55	31	31	45	63
	Transitive - future II progressive	21	95	90	81	71	86	71	**100**	**100**	71	**100**	76	71	71	81	76	67	71	86	86	71	67	90	71	71	71	67	79
	Transitive - conditional I progressive	18	94	**100**	**100**	83	**100**	94	83	94	67	78	61	72	67	89	72	67	56	94	94	44	44	44	83	78	78	94	78
	Transitive - conditional I simple	25	**100**	**100**	**100**	92	**100**	**100**	**100**	**100**	76	68	68	84	84	92	76	80	76	92	**100**	72	80	32	80	64	64	84	83
	Transitive - conditional II progressive	51	90	98	**100**	80	98	92	80	90	76	76	75	82	88	76	75	82	80	75	75	73	67	41	63	47	47	61	76
	Transitive - conditional II simple	20	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	70	80	70	85	85	85	**100**	85	85	80	70	65	60	10	75	50	50	85	80
	Transitive - future I progressive	35	86	66	77	63	**89**	83	63	**89**	63	69	49	60	63	63	49	54	29	69	80	43	49	71	69	43	43	71	63
	Transitive - future I simple	53	81	81	75	72	**100**	77	75	53	55	75	57	79	79	87	32	74	45	85	96	32	47	58	55	81	81	57	69
	Transitive - future II simple	201	65	80	73	58	**100**	78	81	64	60	100	70	91	95	85	29	79	76	78	95	89	24	74	53	71	71	47	72
	Transitive - past perfect progressive	18	83	67	61	67	50	44	72	22	44	**100**	50	89	78	83	28	72	61	72	72	67	44	56	39	44	44	50	60
	Transitive - past perfect simple	47	55	79	64	64	62	64	57	53	51	**83**	70	72	55	72	77	49	55	57	62	32	34	66	72	21	21	81	59
	Transitive - past progressive	14	43	57	57	43	**86**	79	29	71	43	21	43	64	50	64	14	36	36	43	36	14	14	7	57	36	36	57	44
	Transitive - present perfect progressive	23	83	61	70	61	**96**	87	91	87	57	74	70	52	35	61	65	35	35	57	70	39	43	57	48	48	48	52	61
	Transitive - present perfect simple	23	87	78	78	74	**100**	91	96	**100**	43	74	70	70	43	70	30	39	35	70	70	30	35	52	70	43	43	74	64
	Transitive - present progressive	25	**88**	72	64	64	**88**	76	64	40	52	60	56	60	56	60	28	56	36	60	64	52	60	60	60	44	44	68	59
	Transitive - simple past	53	**96**	75	77	66	85	74	85	92	81	62	57	49	47	45	45	43	26	43	47	34	51	57	49	72	72	58	61

Continued on next page

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

ling. category	ling. phenomenon	\#	metric																										
			XCOMET	MetricX-24	MetricX-24-Hybrid	metametrics	MetricX-24-QE	MetricX-24-Hybrid-QE	XCOMET-QE	CometKiwi-XXL	BLEURT-20	CometKiwi	COMET-22	chrfS	MEE4	chrF	gemba	BERTScore	YiSi-1	spBLEU	BLEU	monmonli	damonmonli	sentinel-cand-mqm	PrismRefSmall	XLsimDA	XLsimMqm	PrismRefMedium	avg
	Transitive - simple present	35	80	71	71	63	74	74	80	80	43	**94**	51	37	37	46	63	29	34	43	49	49	57	83	71	66	66	63	61
Verb valency	Case government	189	81	76	74	78	87	78	**89**	78	78	77	79	69	68	62	81	70	77	67	62	80	73	70	51	51	51	52	72
	Catenative verb	885	89	88	90	81	87	79	86	**92**	81	70	74	65	64	67	76	63	68	60	50	65	56	60	56	73	73	53	72
	Mediopassive voice	183	95	95	**99**	99	94	96	91	92	97	86	98	96	95	98	87	93	95	93	89	79	80	89	82	63	63	75	89
	Passive voice	176	77	**84**	81	84	81	82	65	62	67	83	78	55	47	57	61	52	74	51	44	44	57	78	51	47	47	55	64
	Resultative	1203	83	**88**	**88**	85	83	85	82	86	82	86	80	76	68	76	76	67	74	69	67	75	80	85	49	58	58	58	76
	Semantic roles	670	65	77	55	87	76	72	71	79	**88**	49	74	74	71	73	34	76	77	73	70	80	79	81	58	55	55	41	69
Verb semantics	Verb semantics	180	**87**	73	69	82	73	71	78	68	62	64	61	53	57	55	69	58	58	55	53	56	50	54	59	69	69	54	64
macro avg.		78727	**82**	**82**	81	79	78	78	77	77	74	73	73	71	71	70	69	69	68	67	65	65	64	63	59	57	57	57	70
micro avg.		78727	78	**79**	**79**	79	73	74	72	73	77	66	76	71	71	70	65	69	73	68	64	68	69	67	54	54	54	52	69

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

ling. category	ling. phenomenon	\#	metric																								
			MetricX-24	metametrics	XCOMET	CometKiwi-XXL	MetricX-24-Hybrid	MetricX-24-QE	COMET-22	BLEURT-20	MetricX-24-Hybrid-QE	XCOMET-QE	sentinel-cand-mqm	CometKiwi	YiSi-1	BERTScore	chrfS	spBLEU	chrF	BLEU	monmonli	damonmonli	gemba	XLsimDA	XLsimMqm	PrismRefSmall	PrismRefMedium	avg
Ambiguity	Lexical ambiguity	3788	**97**	96	93	81	95	96	87	90	91	84	87	77	83	75	75	71	73	69	78	81	89	44	44	48	45	78
Coordination \& ellipsis	Gapping	698	93	92	93	87	92	91	93	88	89	90	85	**96**	86	85	81	81	77	79	81	74	78	57	57	55	56	81
	Pseudogapping	381	**71**	67	55	64	70	67	62	63	70	54	54	57	60	59	57	54	57	55	55	53	53	39	39	40	40	57
	Right node raising	183	**78**	74	74	77	77	70	75	70	70	67	**79**	75	66	68	64	60	61	58	70	63	55	45	45	66	61	67
	Sluicing	384	80	82	86	**89**	82	82	77	73	82	85	84	77	61	63	56	55	56	54	62	57	54	48	48	35	41	67
	Stripping	375	70	69	80	76	74	79	63	67	**81**	75	69	69	62	60	56	53	54	60	68	78	35	57	57	49	50	64
	VP-ellipsis	252	80	77	80	88	85	86	75	75	**90**	73	76	83	56	56	48	53	47	49	45	55	70	65	65	44	41	66
False friends	False friends	2414	**88**	84	76	69	86	69	83	82	68	69	58	52	**88**	76	85	81	83	76	62	81	34	43	43	53	42	69
Function word	Focus particle	846	**70**	62	63	68	60	67	67	57	49	66	63	55	65	63	64	66	60	67	57	50	44	**71**	**71**	45	50	61
	Question tag	1587	89	87	95	91	81	95	89	92	84	**97**	93	92	66	73	65	65	61	77	57	60	63	75	75	53	48	77
LDD \& interrogatives	Inversion	333	82	88	84	73	90	89	79	83	**92**	78	85	67	71	68	68	66	68	67	61	60	63	47	47	56	52	71
	Modifying Comparison	90	68	71	74	87	69	78	52	**100**	71	74	98	56	44	41	33	29	29	28	56	67	67	73	73	37	40	61
	Multiple connectors	400	97	92	93	95	**98**	92	88	78	96	96	95	94	64	67	62	58	61	55	60	52	86	52	52	69	52	76

Continued on next page

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

			metric																									
ling. category	ling. phenomenon	\#	MetricX-24	metametrics	XCOMET	CometKiwi-XXL	MetricX-24-Hybrid	MetricX-24-QE	COMET-22	BLEURT-20	MetricX-24-Hybrid-QE	XCOMET-QE	sentinel-cand-mqm	CometKiwi	YiSi-1	BERTScore	chrfS	spBLEU	chrF	BLEU	monmonli	damonmonli	gemba	XLsimDA	XLsimMqm	PrismRefSmall	PrismRefMedium	avg
	Pied-piping	343	80	80	78	82	81	76	**83**	72	80	79	81	75	63	66	55	61	52	61	50	50	80	68	68	35	33	68
	Preposition stranding	393	90	90	89	90	92	**93**	88	86	90	90	92	84	75	75	70	72	67	69	64	61	66	58	58	48	46	76
	Topicalization	207	71	74	77	76	77	**86**	70	74	80	75	79	58	63	61	57	61	55	64	71	57	48	42	42	50	49	65
	Wh-movement	173	92	93	86	88	87	88	93	**95**	88	76	87	85	58	62	71	62	65	64	39	41	42	42	42	56	51	70
MWE	Collocation	2167	73	77	79	**80**	74	76	78	75	75	**80**	67	70	76	76	76	76	73	73	70	68	50	59	59	56	52	71
	Compound	1393	88	91	**92**	83	87	83	89	85	**93**	86	72	90	82	78	80	76	79	69	71	77	88	62	62	52	52	79
	Idiom	1784	**100**	98	95	93	**100**	**99**	95	95	99	91	90	99	88	79	78	73	72	73	74	75	85	67	67	51	49	83
	Nominal MWE	2166	75	72	68	67	**77**	68	71	73	74	56	52	53	68	68	75	70	74	67	57	64	57	43	43	48	51	64
	Prepositional MWE	1639	88	87	80	**92**	87	90	82	81	91	83	88	81	76	75	71	73	72	74	52	58	82	46	46	54	52	75
	Verbal MWE	453	68	63	72	**79**	60	64	69	68	63	70	65	60	68	69	67	61	64	64	75	66	59	29	29	41	48	62
Named entity \& terminology	Date	3403	**87**	81	74	69	86	82	77	83	71	64	65	69	78	70	72	71	71	68	60	77	54	38	38	55	50	68
	Domainspecific Term	3471	90	**97**	89	72	89	62	95	95	69	64	84	70	88	82	86	82	83	74	71	72	70	55	55	53	50	76
	Measuring unit	3510	63	72	69	57	59	52	77	73	54	64	53	58	76	81	81	**82**	79	81	54	56	64	45	45	57	56	64
	Onomatopeia	3401	86	86	85	79	86	84	86	**87**	84	80	73	78	85	78	82	79	82	75	78	80	74	57	57	47	48	77
	Proper Name \& Location	2160	**93**	90	90	85	92	88	85	90	89	81	82	86	90	82	81	76	80	62	81	87	64	44	44	56	58	78
	Proper name	339	78	92	93	83	83	65	**96**	80	66	63	83	91	86	94	79	80	70	54	56	60	7	24	24	67	58	69
Negation	Negation	346	65	60	49	59	59	58	67	72	45	45	41	49	79	80	83	74	**84**	72	73	71	42	50	50	49	46	61
Non-verbal agreement	Coreference	526	**86**	82	83	81	84	83	74	73	83	81	80	77	63	61	57	57	57	56	66	60	58	58	58	44	35	68
	Genitive	2068	**82**	73	71	61	72	68	68	74	61	67	60	64	72	67	72	72	72	69	44	43	56	78	78	49	47	66
	Lexical Morphology/Functional shift	1134	97	95	**98**	95	96	94	95	94	95	**97**	89	92	84	84	81	77	77	74	72	78	81	85	85	68	61	86
	Lexical Morphology/Noun formation (er)	670	97	96	95	96	**98**	**97**	**98**	96	96	94	90	87	90	90	86	77	82	63	82	80	83	66	66	50	48	84
	Personal Pronoun Coreference	1290	86	79	85	84	91	**93**	72	75	**93**	89	68	86	63	54	55	56	54	54	74	61	80	62	62	46	43	70
	Possessive Pronouns	521	80	79	77	78	80	77	70	75	**85**	69	64	68	69	64	64	61	64	63	50	62	54	45	45	51	45	66
	Substitution	546	77	75	74	76	78	80	76	74	**81**	78	**82**	76	61	65	67	62	67	63	50	56	66	48	48	46	44	67
Punctuation	Quotation marks	363	73	71	71	**76**	73	64	**76**	69	67	71	71	72	**76**	73	62	65	58	61	64	51	61	57	57	46	45	65
Subordination	Adverbial clause	1458	70	66	69	**79**	64	68	74	64	67	75	61	63	67	62	63	63	62	65	46	48	52	39	39	58	53	61
	Cleft sentence	323	77	**78**	67	65	73	72	63	68	64	65	63	47	59	62	56	60	55	60	62	77	36	38	38	45	41	60
	Complex object	229	74	72	77	79	70	76	79	**90**	67	77	65	71	72	85	71	76	72	72	87	70	45	55	55	46	40	70
	Contact clause	291	65	53	57	**76**	65	65	57	71	73	58	52	46	64	66	63	58	59	54	53	52	40	38	38	62	59	58
	Indirect speech	46	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	96	85	67	85	65	78	70	26	**100**	**100**	**100**	28	30	85
	Infinitive clause	305	95	87	**99**	95	93	75	89	85	88	**99**	95	91	71	69	67	59	64	62	70	67	60	33	33	48	47	74
	Object clause	276	71	82	93	**100**	72	57	79	68	57	97	64	97	61	70	61	68	59	64	59	51	87	87	87	35	29	70
	Participle clause	1345	77	70	67	76	71	76	**81**	75	73	62	75	67	80	69	69	68	70	65	62	61	48	44	44	53	54	66
	Pseudo-cleft sentence	369	83	83	76	73	72	75	85	73	70	77	**98**	85	58	67	60	60	58	62	75	68	62	57	57	40	38	69
	Relative clause	1088	62	76	**77**	68	57	61	74	68	60	70	72	50	71	76	68	69	68	65	74	63	51	67	67	46	43	65
	Subject clause	895	87	89	87	88	94	93	82	82	**95**	94	91	92	66	55	53	56	48	63	73	39	62	56	56	55	50	72
Verb semantics	Verb semantics	275	**88**	82	85	80	**88**	75	74	87	76	80	55	70	53	56	61	54	65	49	67	68	72	33	33	65	67	67
Verb tense/aspect/mood	Conditional	343	78	72	**89**	70	70	81	69	77	70	76	69	76	59	55	63	58	61	51	61	65	80	36	36	52	56	65
	Ditransitive	299	92	90	93	97	91	93	91	91	94	96	90	**100**	63	73	61	67	54	67	56	46	77	63	63	48	47	76
	Gerund	644	84	**85**	**85**	79	**85**	71	74	68	72	81	57	81	67	66	70	70	69	63	57	60	62	41	41	51	52	68

Continued on next page

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

			metric																									
ling. category	ling. phenomenon	\#	MetricX-24	metametrics	XCOMET	CometKiwi-XXL	MetricX-24-Hybrid	MetricX-24-QE	COMET-22	BLEURT-20	MetricX-24-Hybrid-QE	XCOMET-QE	sentinel-cand-mqm	CometKiwi	YiSi-1	BERTScore	chrfS	spBLEU	chrF	BLEU	monmonli	damonmonli	gemba	XLsimDA	XLsimMqm	PrismRefSmall	PrismRefMedium	avg
	Imperative	575	88	**89**	85	**88**	84	87	83	79	84	85	80	68	74	74	66	69	64	63	69	69	50	44	44	42	44	71
	Intransitive	103	94	91	87	**98**	94	90	94	81	94	88	93	93	81	68	68	65	65	54	58	60	56	42	42	37	41	73
	Reflexive	514	88	89	84	**94**	77	90	84	77	77	83	85	67	70	69	70	70	68	65	41	55	88	58	58	51	51	72
	Transitive	516	78	**87**	85	85	85	80	77	68	80	66	47	51	77	74	73	75	78	69	50	61	28	50	50	63	59	68
Verb valency	Case government	331	76	**86**	78	71	76	74	77	85	75	71	51	70	75	82	83	75	82	71	81	79	78	74	74	43	40	73
	Catenative verb	358	72	73	69	68	**73**	71	66	70	72	70	67	63	58	63	66	61	62	59	62	63	65	49	49	50	48	64
	Impersonal Subject	217	86	77	82	95	85	**98**	75	71	90	94	74	87	65	67	61	53	60	53	76	75	59	50	50	43	41	71
	Mediopassive voice	409	77	79	89	85	69	89	80	72	77	**90**	83	82	73	75	65	70	63	67	57	63	64	58	58	38	37	70
	Passive voice	228	94	89	87	84	92	**98**	88	84	90	84	69	82	79	83	75	75	74	84	73	83	66	74	74	52	44	79
	Resultative	660	**91**	86	**91**	87	**91**	88	78	80	88	87	82	66	72	75	75	73	76	73	65	67	73	62	62	64	55	76
	Semantic roles	270	**91**	82	83	77	85	73	79	89	87	79	60	80	71	67	79	70	81	65	80	84	53	63	63	57	50	74
	Verb semantics/Verb semantics	549	81	**83**	83	80	74	73	71	70	78	78	74	75	66	70	72	68	72	66	43	46	66	58	58	53	48	68
macro avg.		59113	**82**	81	81	81	81	80	79	79	79	78	75	75	71	70	68	67	66	64	64	63	62	54	54	50	48	70
micro avg.		59113	**83**	83	81	78	82	79	81	81	77	76	73	73	76	73	73	71	71	68	65	67	64	53	53	52	49	71