

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

Eleftherios Avramidis and Vivien Macketanz

German Research Center for Artificial Intelligence (DFKI),

Speech and Language Technology, Berlin, Germany

firstname.lastname@dfki.de

Abstract

We employ a linguistically motivated challenge set in order to evaluate the state-of-the-art machine translation metrics submitted to the Metrics Shared Task of the 7th Conference for Machine Translation. The challenge set includes about 20,000 items extracted from 145 MT systems for two language directions (German \leftrightarrow English), covering more than 100 linguistically-motivated phenomena organized in a dozen of categories. The best performing metrics are YiSi-1, COMET-22 and BLEURT for German-English, and XL-DA for English-German, followed by BLEURT, COMET-22 and UniTE. Metrics in both directions are performing worst when it comes to named-entities & terminology. Particularly in German-English they are weak at detecting issues at punctuation, polar questions and idiom. In English-German, they perform worst at future II progressive of intransitive verbs, focus particles and present progressive of transitive verbs.

1 Introduction

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

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

In this paper we present empirical results on the performance of MT metrics, using an exten-

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

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

2 Related work

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

Hereby we are advancing as to that preliminary analysis by (a) increasing the number of challenge items to about 9,000-10,000, including outputs from state-of-the-art systems from 2021, (b) adding a second language direction (English-German) (c)

presenting a more fine-grained analysis, not only in the category level but also on the phenomenon level. This way we can get more confident and more generalisable empirical conclusions.

3 Method

3.1 Test suite for MT systems

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

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

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

3.2 Challenge set for MT metrics

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

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

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

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

3.3 Evaluation of metrics

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

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

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

Significant tests for comparisons: the highest metric accuracy for every phenomenon is compared to all other metric accuracies of the same phenomenon. For this, a one-tailed Z-test with $\alpha = 0.95$ is calculated. The metrics whose accuracies that are not significantly worse than the highest accuracy, are considered to share the winning position for this phenomenon. The best accuracies per

category are calculated in the same way, after aggregating the challenge items from the underlying phenomena of every category.

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

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

4 Results

The results are displayed in detail in Tables 1 and 2 in the category level and in Tables 3 and 4 for the phenomenon level, for both language directions German-English and English-German respectively.

4.1 Metric performance analysis

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

German-English The best performing metrics for German-English are YiSi-1 and COMET-22, achieving the significantly highest micro- and macro-average accuracies (84-85%), whereas for the macro-average, BLEURT is also included in the first significance cluster. Two QE based metrics, REUSE and MATESE, get the lowest accuracies.

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

Interestingly enough, two QE methods are the single winners of two particular categories, outperforming reference-aware metrics: COMET-Kiwi is the best performing system for *negation*(93%) and

HWTSC-TLM is the best performing system for *punctuation*.

English-German XL-DA is the only system which prevails in both micro- and macro-average for English-German, winning 5 individual categories, whereas another 3 systems share the first position for macro-average accuracy (BLEURT, COMET-22 and UniTE). Their average accuracies are close to 80%, which raises concerns, as this indicates that 2 out of 10 challenge items in average are not scored correctly in this language direction, even for the best performing metrics. The lowest scoring metric is MATESE in both QE and reference-based versions, very close to REUSE.

Also in this direction, QE methods manage to outperform reference-based metrics in a few categories. REUSE is the best performing metrics for *false friends*, COMET-KIWI and CROSS-QE for *function words* and MS-COMET-QE for *punctuation*.

4.2 Linguistically motivated analysis

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

4.2.1 German-English

Category-level The overall average accuracy of all metrics with regards to the linguistically motivated categories is at 77% for German-English. This indicates that the metrics failed in average to predict properly the scores for about one out of four challenge items that we provided. Even for the best categories, the accuracy achieved by most metrics is considerably below the acceptable limit of 90%.

The best performing categories in average are *false friends* and *negation* with 84-85% accuracy. For the rest of the categories, the average accuracy is less than 80%. The worst performing categories in average are *named entity and terminology* and *punctuation* with only 66% accuracy, whereas *Subordination* comes next with 71%. The lowest performing score for all systems and all categories is achieved by XL-MQM, which can only score correctly almost half of the punctuation challenge items (53%).

Phenomenon-level The best accuracy for this language pair is achieved for *Transitive, future I* where the metrics get an accuracy of 95%-100%. Another 10 phenomena score more than 85%. Four of them also refer to the future tenses of the transitive, in particular future I and future II in both the

plain and their subjunctive form. Additionally, one can see good performance in *pied-piping*, *modal future I*, *intransitive present*, *false friends*, *comma* and the *negated modal for future I subjunctive II*.

The lowest accuracy of all metrics in average is given for *polar questions* (59%), followed by *idioms* (61%). An average accuracy of less than 65% is given for some more phenomena, such as the ones including *dates*, *cleft sentences*, *internal possessors*, *locations*, *relative clauses* and *quotation marks*.

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

4.2.2 English-German

Category-level The overall average accuracy of all metrics with regards to the linguistically motivated categories is at 70% for English-German. This is 7% lower than the respective average accuracy for German-English, indicating that the metrics for this MT language direction perform worse.

The category where all metrics perform best in average is *negation* (87%), whereas the one where they perform worse is *Named entity & terminology* (57%). The rest of the categories lie in rather mediocre accuracies, between 65% and 81%. The performance of metrics in English-German is worse than German-English in all categories apart from *Negation*, *punctuation* and *subordination*, although the comparisons between the language directions have to be taken with a grain of salt, due to the fact that the two directions consist of different items.

Phenomenon-level The English-German phenomena, where metrics perform best in average are the *Contact clause*, *Negation*, *Ditransitive - present progressive* and *question tags*, achieving more than 85% of accuracy. The most difficult phenomena to score are two *Intransitive - future II progressive*, *Focus particle* and *Transitive - present progressive*, which achieve less than 40% average accuracy.

Interestingly enough, in this language direction there are metrics which scored zero accuracies in

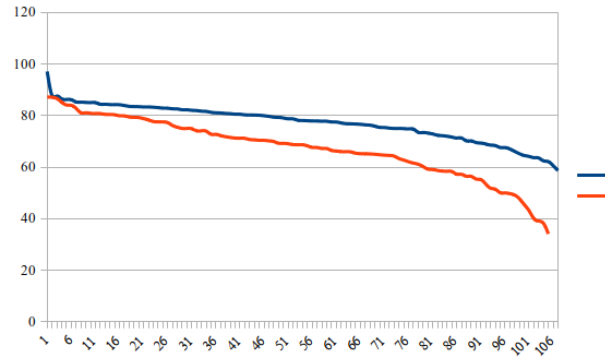


Figure 1: Plot of the accuracy of all phenomena per language direction. The accuracy percentage is shown on the vertical axis and the phenomena on the horizontal axis.

several phenomena, something that we didn't see in the opposite language direction¹. These zero accuracies are mostly relevant to rare verb-related phenomena (e.g. intransitive constructions).

A comparative plot of the accuracies for all phenomena for both language directions can be seen in Figure 1. It is very clear that English-German lacks considerably, with its lowest scored phenomena being half of the lower-scored phenomena of the opposite direction.

5 Conclusion

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

In an effort to draw conclusions after averaging accuracies, we conclude that the best performing metrics are YiSi-1, COMET-22 and BLEURT for German-English, and XL-DA for English-German, followed by BLEURT, COMET-22 and UniTE.

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

On the contrary metrics in both directions are performing worst when it comes to *named-entities & terminology*. Particularly in German-English

¹again this should take into consideration that English-German set has a participation of less systems and therefore less diversity than German-English

they are weak at detecting issues at *punctuation*, *polar questions* and *idiom*. In English-German at future II progressive of intransitive *verbs*, *focus particles* and *present progressive of transitive verbs*.

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

Acknowledgements

This research was supported by the Deutsche Forschungsgemeinschaft (DFG) through the project TextQ, and by the German Federal Ministry of Education through the project SocialWear.

References

- Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, Aljoscha Burchardt, and Sebastian Möller. 2020. Fine-grained linguistic evaluation for state-of-the-art Machine Translation. In *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics.
- Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit. 2019. [Linguistic evaluation of German-English Machine Translation using a Test Suite](#). In *Proceedings of the Fourth Conference on Machine Translation. Conference on Machine Translation (WMT-2019)*, pages 644–653, Florence, Italy. Association for Computational Linguistics.
- Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter, and Philip Williams. 2017. [A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines](#). *The Prague Bulletin of Mathematical Linguistics*, 108:159–170.
- Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. [DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352, Dublin, Ireland. Association for Computational Linguistics.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-Kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. [Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain](#).
- 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.
- Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, and Hans Uszkoreit. 2018. [Fine-grained evaluation of German-English Machine Translation based on a Test Suite](#). In *Proceedings of the Third Conference on Machine Translation (WMT18)*, pages 578–587, Brussels, Belgium. Association for Computational Linguistics.
- Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, and Hans Uszkoreit. 2022. [A linguistically motivated test suite to semi-automatically evaluate German-English machine translation output](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 936–947, Marseille, France. European Language Resources Association.
- Vivien Macketanz, Eleftherios Avramidis, Shushen Manakhimova, and Sebastian Möller. 2021. Linguistic evaluation for the 2021 state-of-the-art Machine Translation systems for German to English and English to German. In *Proceedings of the Sixth Conference on Machine Translation. (WMT21)*, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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.

category	#	baselines					qe-as-a-metric					ref-based-metrics						
		BLEU	chrF	COMET-QE	BLEURT	YiSi-1	MS-COMET-QE	HWTSC-TLM	COMET-Kiwi	KG-BERT	HWTSC-TS	CROSS-QE	XL-DA	XL-MQM	MS-COMET	COMET-22	UniTE	avg
Ambiguity	298	71	80	80	88	89	80	60	82	67	65	73	88	90	88	87	89	80
Composition	252	65	74	72	87	90	74	73	76	59	76	77	86	82	82	83	81	77
Coordination & ellipsis	316	74	77	72	79	80	73	69	82	78	72	78	79	83	71	84	81	77
False friends	90	64	78	84	93	92	73	81	88	87	91	74	90	90	93	91	93	85
Function word	586	72	73	74	83	81	70	78	81	70	81	77	84	79	83	83	83	78
LDD & interrogatives	1014	75	76	82	84	85	79	72	83	63	75	83	85	82	80	86	84	80
MWE	610	73	78	72	85	86	70	70	76	56	60	76	88	88	78	86	89	77
Named entity & terminology	861	62	67	64	68	76	63	64	65	55	61	71	67	75	62	70	64	66
Negation	76	84	88	86	88	91	88	62	93	87	74	78	89	78	82	91	92	84
Non-verbal agreement	419	74	75	74	83	76	75	66	75	62	63	72	83	81	76	84	84	75
Punctuation	293	77	69	54	70	73	67	81	55	62	73	75	61	53	59	68	59	66
Subordination	679	69	69	70	77	74	67	59	72	65	62	75	75	70	68	80	77	71
Verb tense/aspect/mood	4697	69	77	63	85	89	76	63	81	78	71	87	84	79	83	86	84	78
Verb valency	211	70	72	77	88	90	72	64	86	62	64	72	91	88	84	94	91	79
macro avg.	10402	71	75	73	83	84	73	69	78	68	70	76	82	80	78	84	82	77
micro avg.	10402	70	75	68	82	85	74	66	78	70	70	81	82	79	79	84	82	77

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

category	#	baselines					qe-as-a-metric							ref-based-metrics										
		BLEU	chrF	COMET-QE	BLEURT	YiSi-1	MS-COMET-QE	REUSE	HWTSC-TLM	COMET-Kiwi	KG-BERT	HWTSC-TS	CROSS-QE	MATESE-QE	XL-DA	XL-MQM	MEE2	MS-COMET	COMET-22	MATESE	MEE4	MEE	UniTE	avg
Ambiguity	146	71	89	72	90	87	47	15	81	55	38	38	47	37	94	89	87	81	84	61	91	73	92	69
Coordination & ellipsis	836	61	61	75	80	73	71	38	72	76	60	70	71	49	81	83	64	79	79	53	64	67	78	68
False friends	225	63	72	62	70	67	52	89	64	67	63	68	60	52	71	71	82	76	69	57	88	69	71	68
Function word	200	76	72	84	90	78	72	66	78	91	89	89	92	71	78	82	80	90	90	48	86	80	86	80
MWE	829	72	77	78	87	85	71	32	79	78	79	79	81	57	86	86	80	88	86	58	80	76	87	77
Named entity & terminology	1272	55	61	53	66	64	56	43	56	55	48	47	59	30	73	70	64	61	61	45	65	60	66	57
Negation	174	83	85	86	89	93	87	78	87	92	91	91	86	84	82	81	92	83	91	88	91	82	94	87
Non-verbal agreement	372	72	70	76	81	78	81	39	59	77	63	65	70	46	89	90	77	77	90	65	79	65	84	72
Punctuation	336	79	74	77	76	77	82	46	70	68	62	62	72	58	67	60	75	69	79	54	75	63	80	69
Subordination	994	74	74	82	80	78	80	48	81	86	82	83	82	58	85	84	74	85	84	58	76	65	84	76
Verb tense/aspect/mood	3081	62	69	57	70	69	65	54	51	70	60	59	77	49	78	74	74	67	77	53	72	75	71	66
Verb valency	480	64	71	76	84	76	77	48	65	82	70	72	74	47	78	78	71	79	82	60	69	60	78	71
macro avg.	8945	69	73	73	80	77	70	50	70	75	67	69	73	53	80	79	77	78	81	58	78	70	81	72
micro avg.	8945	65	69	67	76	73	68	48	63	73	64	65	74	49	79	77	73	73	78	55	73	69	76	69

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

category	phenomenon	#	baselines					qe-as-a-metric					ref-based-metrics					avg	
			BLEU	chrF	COMET-QE	BLEURT	YiSi-1	MS-COMET-QE	HWTSC-TLM	COMET-Kiwi	KG-BERT	HWTSC-TS	CROSS-QE	XL-DA	XL-MQM	MS-COMET	COMET-22		UniTE
Ambiguity	Lexical ambiguity	129	74	87	74	95	88	77	56	81	57	60	65	95	93	90	93	99	80
	Structural ambiguity	169	69	75	84	83	89	82	64	82	75	69	79	82	88	86	83	81	79
Composition	Compound	129	64	74	67	90	91	70	64	71	45	70	69	90	88	81	81	84	75
	Phrasal verb	123	66	74	78	85	89	79	83	81	74	82	86	81	76	82	84	79	80
Coordination & ellipsis	Gapping	51	76	76	80	82	71	92	59	100	75	75	98	80	100	82	98	88	83
	Right node raising	67	70	75	49	76	91	60	64	78	82	55	84	76	79	54	82	81	72
	Sluicing	128	75	79	73	77	78	70	77	80	76	79	66	81	76	73	80	78	76
	Stripping	70	74	76	87	84	80	79	67	77	83	71	80	77	89	74	81	80	79
False friends	False friends	90	64	78	84	93	92	73	81	88	87	91	74	90	90	93	91	93	85
Function word	Focus particle	64	75	75	69	83	88	83	70	83	88	84	70	88	67	81	84	88	80
	Modal particle	166	79	77	67	85	86	75	69	82	83	81	75	89	83	87	89	86	81
LDD & interrogatives	Question tag	356	69	71	78	82	78	65	84	81	61	80	79	81	79	82	81	81	77
	Extended adjective construction	320	80	80	89	88	88	83	79	90	61	82	93	89	86	83	90	88	84
	Extraposition	92	74	83	74	75	77	63	65	67	62	79	74	80	79	67	76	78	73
	Multiple connectors	87	79	76	69	63	76	59	67	70	64	63	68	66	57	72	68	64	68
	Pied-piping	162	78	77	97	93	93	90	73	96	70	74	90	94	90	96	95	95	88
	Polar question	51	43	45	53	63	67	67	49	69	61	53	49	61	75	55	67	65	59
	Scrambling	144	72	74	90	90	88	83	82	90	51	81	88	96	92	87	98	87	84
	Topicalization	61	85	84	79	87	87	74	66	77	77	70	69	82	70	80	82	82	78
	Wh-movement	97	62	69	58	85	77	81	56	72	66	64	75	81	73	68	81	82	72
	MWE	Collocation	190	72	79	74	91	88	80	82	84	67	65	82	91	91	81	89	92
Named entity & terminology	Idiom	133	67	69	55	76	83	42	36	44	20	17	55	89	86	65	75	89	61
	Prepositional MWE	146	79	84	71	85	86	65	82	82	72	84	84	86	86	75	85	80	80
	Verbal MWE	141	74	77	88	87	84	89	77	89	57	68	81	87	87	91	92	94	83
	Date	203	50	58	65	65	66	67	63	70	68	68	70	69	74	62	67	62	65
Non-verbal agreement	Domainspecific term	214	63	71	74	71	74	71	63	67	59	57	77	68	75	67	72	74	69
	Location	181	65	66	55	70	82	57	76	62	38	64	57	68	80	44	75	60	64
	Measuring unit	203	67	72	58	61	77	57	54	57	56	51	73	62	67	66	63	54	62
	Proper name	60	75	73	75	85	92	62	72	78	50	70	88	78	85	92	85	85	78
	Negation	76	84	88	86	88	91	88	62	93	87	74	78	89	78	82	91	92	84
Non-verbal agreement	Coreference	251	68	72	85	90	75	86	73	81	66	69	77	91	88	84	91	91	81
	External possessor	104	88	88	64	75	82	61	50	70	58	51	68	71	71	66	76	74	70
	Internal possessor	64	80	67	48	72	72	52	62	61	52	58	59	72	69	59	69	75	64

Continued on next page

category	phenomenon	#	baselines					qe-as-a-metric					ref-based-metrics						
			BLEU	chrF	COMET-QE	BLEURT	YiSi-1	MS-COMET-QE	HWTSC-TLM	COMET-Kiwi	KG-BERT	HWTSC-TS	CROSS-QE	XL-DA	XL-MQM	MS-COMET	COMET-22	UniTE	avg
Punctuation	Comma	46	91	87	78	93	89	74	83	85	87	85	91	89	80	78	89	83	85
	Quotation marks	247	75	65	49	66	70	65	81	49	57	71	72	56	48	55	64	55	62
Subordination	Adverbial clause	87	70	67	71	82	72	63	66	70	70	68	74	70	74	57	77	72	70
	Cleft sentence	109	73	66	61	67	66	51	48	66	62	64	69	66	64	66	72	69	65
	Free relative clause	70	67	71	66	77	67	76	50	60	63	56	70	74	54	66	77	79	67
	Indirect speech	119	64	70	84	81	71	66	58	80	57	58	75	83	65	65	87	85	72
	Infinitive clause	64	77	77	59	77	78	77	62	73	73	67	70	75	66	67	70	72	71
	Object clause	54	74	81	67	85	89	93	69	76	94	69	89	89	80	78	93	87	82
	Pseudo-cleft sentence	25	72	60	76	68	92	52	48	80	60	32	100	80	92	84	88	92	74
	Relative clause	71	63	66	65	65	66	58	59	63	48	59	61	62	66	68	77	70	64
	Subject clause	80	66	65	78	85	86	78	68	86	62	70	86	85	88	76	86	81	78
Verb tense/aspect/mood	Conditional	50	80	80	88	76	80	82	82	80	74	78	80	82	88	82	82	82	81
	Ditransitive - future I	121	72	71	69	92	88	89	58	93	85	68	92	92	85	83	94	94	83
	Ditransitive - future I subjunctive II	84	63	75	64	89	95	96	50	90	92	65	94	93	88	86	92	93	83
	Ditransitive - future II	97	60	71	58	82	94	64	58	98	69	69	98	85	78	87	93	87	78
	Ditransitive - future II subjunctive II	88	73	78	62	86	97	93	65	88	89	75	99	83	84	86	89	89	83
	Ditransitive - perfect	72	62	72	75	81	93	82	46	93	75	58	96	81	82	92	86	86	79
	Ditransitive - pluperfect	86	67	79	60	83	88	74	57	81	74	71	83	79	69	72	86	81	75
	Ditransitive - pluperfect subjunctive II	107	71	88	42	79	92	69	65	64	75	66	69	82	78	88	86	86	75
	Ditransitive - present	90	61	77	54	91	81	73	56	70	83	60	99	83	86	83	88	89	77
	Ditransitive - preterite	117	62	76	73	85	89	90	62	84	87	61	91	89	92	89	95	92	82
	Ditransitive - preterite subjunctive II	110	61	85	75	95	90	92	60	85	85	61	87	95	96	86	96	95	84
	Imperative	98	78	79	84	95	89	84	78	88	57	74	84	90	87	92	88	92	83
	Intransitive - future I	32	53	72	56	88	91	97	69	84	100	91	97	88	88	94	84	88	84
	Intransitive - future I subjunctive II	56	61	70	48	93	89	80	55	93	100	68	95	98	84	86	95	98	82
	Intransitive - future II	62	60	77	35	90	89	55	45	79	69	58	87	95	94	84	90	92	75
	Intransitive - future II subjunctive II	94	72	89	59	94	98	69	63	80	86	82	100	94	85	91	91	89	84
	Intransitive - perfect	61	56	59	67	84	87	54	66	62	64	66	69	75	67	77	72	72	69
	Intransitive - pluperfect	85	79	85	47	85	87	58	46	68	88	55	86	82	71	79	78	73	73
	Intransitive - pluperfect subjunctive II	79	87	90	39	97	100	52	56	78	95	76	94	96	95	99	96	97	84
	Intransitive - present	54	69	74	63	91	98	96	65	96	94	76	98	89	93	91	94	93	86
	Intransitive - preterite	46	46	63	63	89	74	78	74	93	91	83	93	85	80	74	87	74	78
	Intransitive - preterite subjunctive II	100	43	51	64	86	79	80	58	79	91	67	89	87	89	73	88	78	75
	Modal - future I	42	90	95	76	88	95	93	76	83	74	83	98	88	76	93	88	86	86

Continued on next page

category	phenomenon	#	baselines					qe-as-a-metric					ref-based-metrics							
			BLEU	chrF	COMET-QE	BLEURT	YiSi-1	MS-COMET-QE	HWTSC-TLM	COMET-Kiwi	KG-BERT	HWTSC-TS	CROSS-QE	XL-DA	XL-MQM	MS-COMET	COMET-22	UniTE	avg	
	Modal - future I subjunctive II	86	94	94	48	81	97	70	78	79	67	79	78	85	64	86	85	80	79	
	Modal - perfect	149	72	72	57	74	85	62	67	85	47	77	81	66	57	69	70	66	69	
	Modal - pluperfect	75	100	99	73	84	100	67	83	85	69	47	91	89	69	44	80	76	69	77
	Modal - pluperfect subjunctive II	61	72	80	69	79	90	80	69	85	85	79	90	80	75	89	84	82	81	
	Modal - present	30	57	73	57	93	80	80	53	90	80	80	80	73	73	80	80	77	75	
	Modal - preterite	72	61	74	65	88	88	89	54	90	93	78	92	89	81	81	89	83	81	
	Modal - preterite subjunctive II	30	80	77	53	83	83	87	43	93	87	73	87	83	70	83	80	83	78	
	Modal negated - future I	43	93	88	86	81	100	91	86	86	65	91	93	79	74	91	81	74	85	
	Modal negated - future I subjunctive II	73	92	96	79	86	97	73	79	79	77	88	92	88	75	90	86	84	85	
	Modal negated - perfect	126	50	62	63	66	72	70	60	73	63	71	88	63	61	71	63	61	66	
	Modal negated - pluperfect	126	87	99	79	90	94	63	83	74	55	93	95	88	75	87	88	83	83	
	Modal negated - pluperfect subjunctive II	81	65	74	63	73	78	69	64	59	84	79	84	79	75	81	73	70	73	
	Modal negated - present	33	79	64	58	73	70	58	48	64	64	67	88	79	76	73	67	70	68	
	Modal negated - preterite	61	66	87	67	90	89	95	38	90	95	75	82	80	80	82	85	80	80	
	Modal negated - preterite subjunctive II	77	66	83	57	91	86	92	47	91	86	75	95	83	78	84	84	84	80	
	Progressive	76	66	67	66	71	75	63	50	75	64	64	67	67	68	62	75	80	68	
	Reflexive - future I	85	76	80	74	89	82	64	84	86	75	81	85	87	81	81	89	92	82	
	Reflexive - future I subjunctive II	96	70	66	66	79	84	64	71	79	80	79	89	85	80	79	85	89	78	
	Reflexive - future II	116	83	85	43	77	97	62	40	67	72	43	73	79	74	75	87	83	71	
	Reflexive - future II subjunctive II	107	74	77	61	81	93	84	66	79	91	77	92	85	78	84	89	88	81	
	Reflexive - perfect	188	64	62	68	81	82	73	53	86	78	54	85	82	78	82	86	84	75	
	Reflexive - pluperfect	109	63	55	76	83	87	77	54	80	75	47	83	83	81	87	85	82	75	
	Reflexive - pluperfect subjunctive II	90	76	80	52	79	97	70	66	70	88	70	81	74	64	89	81	78	76	
	Reflexive - present	125	59	74	77	90	80	86	72	88	74	75	92	85	85	90	86	87	81	
	Reflexive - preterite	117	69	75	67	85	88	81	54	76	66	56	83	91	85	75	88	89	77	
	Reflexive - preterite subjunctive II	124	77	70	66	86	91	74	54	72	65	55	83	89	88	79	89	91	77	
	Transitive - future I	43	95	95	86	100	100	95	86	100	100	100	95	100	100	100	100	100	97	
	Transitive - future I subjunctive II	37	81	84	57	95	100	76	54	92	100	89	86	95	84	97	95	95	86	
	Transitive - future II	33	76	94	45	94	100	88	70	88	88	94	64	97	85	91	94	94	85	
	Transitive - future II subjunctive II	50	84	88	42	88	100	92	90	92	98	98	90	92	76	92	92	90	88	
	Transitive - perfect	99	64	80	42	81	88	91	73	79	78	86	76	76	81	88	81	79	78	
	Transitive - pluperfect	22	73	82	50	82	91	68	73	73	68	77	77	86	77	73	91	82	76	
	Transitive - pluperfect subjunctive II	39	85	97	36	64	100	33	69	49	92	67	54	74	62	97	72	72	70	
	Transitive - present	33	58	73	58	94	91	79	82	88	88	79	94	94	91	88	91	88	83	

Continued on next page

category	phenomenon	#	baselines					qe-as-a-metric						ref-based-metrics					
			BLEU	chrF	COMET-QE	BLEURT	YiSi-1	MS-COMET-QE	HWTSC-TLM	COMET-Kiwi	KG-BERT	HWTSC-TS	CROSS-QE	XL-DA	XL-MQM	MS-COMET	COMET-22	UniTE	avg
Verb valency	Transitive - preterite	57	51	63	77	86	82	67	95	93	68	91	91	100	82	100	86	82	
	Transitive - preterite subjunctive II	97	40	60	76	86	84	80	73	86	74	80	84	84	79	85	82	77	
	Case government	80	65	62	79	88	89	80	71	94	52	66	75	92	92	75	95	92	79
	Mediopassive voice	50	64	66	66	82	80	58	50	74	64	50	66	88	88	86	90	86	72
	Passive voice	33	85	82	79	91	94	73	64	94	64	61	79	91	91	88	94	94	83
	Resultative predicates	48	73	85	85	94	98	73	69	81	73	75	67	94	79	94	96	94	83
macro avg.		10402	71	76	67	83	86	74	65	79	73	71	82	83	79	80	84	83	77
micro avg.		10402	70	75	68	82	85	74	66	78	70	70	81	82	79	79	84	82	77

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

category	phenomenon	#	baselines					qe-as-a-metric							ref-based-metrics										
			BLEU	chrF	COMET-QE	BLEURT	YiSi-1	MS-COMET-QE	REUSE	HWTSC-TLM	COMET-Kiwi	KG-BERT	HWTSC-TS	CROSS-QE	MATESE-QE	XL-DA	XL-MQM	MEE2	MS-COMET	COMET-22	MATESE	MEE4	MEE	UniTE	avg
Ambiguity	Lexical ambiguity	146	71	89	72	90	87	47	15	81	55	38	38	47	37	94	89	87	81	84	61	91	73	92	69
Coordination & ellipsis	Gapping	163	58	68	77	84	74	78	58	71	80	37	67	71	67	80	88	67	83	79	68	64	71	77	71
	Pseudogapping	201	77	67	92	97	82	79	26	81	93	80	80	92	57	97	93	78	89	96	72	79	85	96	81
	Right node raising	47	64	72	94	87	83	91	9	81	87	34	91	81	81	91	96	68	85	94	9	68	66	98	74
	Sluicing	169	56	54	51	63	59	57	51	61	56	46	51	47	28	67	67	59	64	62	36	59	63	54	55
	Stripping	139	58	58	60	65	68	57	38	53	63	55	55	53	32	67	82	63	65	68	37	61	55	70	58
	VP-ellipsis	117	47	51	90	85	75	74	26	95	84	92	92	85	48	85	80	42	89	82	60	47	51	85	71
False friends	False friends	225	63	72	62	70	67	52	89	64	67	63	68	60	52	71	71	82	76	69	57	88	69	71	68
Function word	Focus particle	20	30	35	25	80	45	35	55	15	45	10	10	50	10	60	55	30	35	75	5	35	30	70	38
	Question tag	180	82	76	91	91	82	77	68	84	96	98	98	97	78	80	85	86	96	92	53	92	85	88	85
MWE	Collocation	112	61	76	78	92	88	76	46	89	88	91	91	86	57	93	96	76	86	95	74	79	56	95	80

Continued on next page

category	phenomenon	#	baselines					qe-as-a-metric							ref-based-metrics										
			BLEU	chrF	COMET-QE	BLEURT	YiSi-1	MS-COMET-QE	REUSE	HWTSC-TLM	COMET-Kiwi	KG-BERT	HWTSC-TS	CROSS-QE	MATESE-QE	XL-DA	XL-MQM	MEE2	MS-COMET	COMET-22	MATESE	MEE4	MEE	UniTE	avg
	Compound	63	51	84	89	70	87	71	3	89	98	90	90	100	68	90	97	76	87	97	33	73	68	92	78
	Idiom	266	82	75	86	95	92	83	22	93	86	73	73	92	61	97	97	86	96	98	64	85	89	97	83
	Nominal MWE	288	71	78	66	84	81	60	39	66	62	82	82	72	50	72	74	79	88	69	47	78	76	75	70
	Prepositional MWE	35	86	83	86	71	86	86	83	60	66	77	77	77	69	86	86	86	69	80	69	86	89	80	79
	Verbal MWE	65	71	74	83	89	62	48	23	69	83	65	65	58	52	85	75	74	71	86	68	77	65	82	69
Named entity & terminology	Date	234	53	60	69	74	68	81	31	80	63	83	83	93	59	67	76	60	67	65	65	60	60	71	68
	Domainspecific term	312	56	76	71	89	86	65	33	78	73	42	42	62	25	97	94	77	91	78	72	79	73	80	70
	Location	12	83	58	92	75	50	100	83	100	100	100	92	92	8	100	100	83	83	75	33	92	75	83	80
	Measuring unit	389	48	55	24	53	54	28	43	21	28	15	15	31	11	69	58	58	33	46	21	61	54	53	40
Proper name	325	61	54	58	52	53	59	62	58	64	67	64	64	39	58	55	59	58	59	36	60	55	65	57	
Negation	Negation	174	83	85	86	89	93	87	78	87	92	91	91	86	84	82	81	92	83	91	88	91	82	94	87
Non-verbal agreement	Coreference	81	86	86	96	95	75	90	41	33	77	56	60	73	60	98	100	89	85	96	93	88	73	93	79
	Genitive	206	73	68	86	73	82	73	22	83	70	62	64	63	52	85	86	75	78	84	52	76	63	79	70
	Possession	85	55	58	31	86	74	92	78	26	93	74	74	86	19	88	91	69	66	96	68	76	64	89	71
Punctuation	Quotation marks	336	79	74	77	76	77	82	46	70	68	62	62	72	58	67	60	75	69	79	54	75	63	80	69
Subordination	Adverbial clause	193	81	73	89	81	67	75	65	79	88	79	79	77	46	86	85	77	80	82	37	77	67	81	75
	Cleft sentence	179	63	57	74	60	63	61	45	72	74	80	80	59	60	73	67	68	72	71	54	68	60	71	66
	Contact clause	150	75	74	95	94	88	91	53	99	98	97	97	97	87	97	97	73	95	98	81	75	59	97	87
	Indirect speech	38	42	47	74	63	50	37	24	58	95	50	50	63	58	63	68	42	79	76	47	42	39	74	56
	Infinitive clause	85	55	80	80	86	95	96	40	78	95	71	81	99	59	89	91	61	94	93	67	69	51	93	78
	Object clause	16	38	56	62	88	62	81	31	56	81	94	94	62	0	100	100	50	81	88	19	62	56	100	66
	Pseudo-cleft sentence	73	88	89	73	66	90	86	70	81	68	88	88	85	64	85	86	85	85	75	64	88	82	71	80
	Relative clause	112	83	84	73	90	82	95	36	78	88	73	73	93	66	94	89	79	92	91	68	82	69	93	80
	Subject clause	148	90	90	84	89	91	88	33	87	89	88	88	89	37	84	82	86	86	86	57	88	78	86	81
Verb tense/aspect/mood	Conditional	106	77	70	92	94	91	98	18	86	92	92	92	87	36	92	84	84	89	87	58	89	80	92	81
	Ditransitive - conditional I progressive	72	49	61	89	93	83	61	74	94	99	74	93	99	65	93	82	71	92	92	69	67	79	92	80
	Ditransitive - conditional I simple	34	74	94	62	65	97	71	91	41	100	41	44	97	47	97	94	94	65	100	38	97	97	100	78
	Ditransitive - conditional II progressive	51	78	82	78	88	80	80	63	51	65	49	55	67	71	84	82	84	84	90	59	84	98	86	75
	Ditransitive - conditional II simple	59	64	68	66	76	66	68	59	53	69	47	49	56	53	80	78	78	69	73	51	80	78	76	66
	Ditransitive - future I progressive	61	51	62	34	62	57	79	90	84	92	49	66	51	11	79	75	56	43	66	11	59	57	69	59
	Ditransitive - future I simple	88	51	60	50	56	56	69	90	52	66	42	48	50	16	65	64	57	52	58	34	65	59	59	55
	Ditransitive - future II progressive	91	64	60	66	66	47	74	56	84	70	95	95	45	41	65	62	71	69	82	36	76	91	74	68
	Ditransitive - future II simple	49	94	94	39	86	65	86	88	76	100	59	59	92	12	86	39	90	88	92	57	90	90	96	76
	Ditransitive - past perfect progressive	91	44	58	67	60	66	66	37	51	65	73	69	75	52	78	71	60	63	71	55	62	69	68	63

Continued on next page

category	phenomenon	#	baselines					qe-as-a-metric							ref-based-metrics										
			BLEU	chrF	COMET-QE	BLEURT	YiSi-1	MS-COMET-QE	REUSE	HWTSC-TLM	COMET-Kiwi	KG-BERT	HWTSC-TS	CROSS-QE	MATESE-QE	XL-DA	XL-MQM	MEE2	MS-COMET	COMET-22	MATESE	MEE4	MEE	UniTE	avg
	Ditransitive - past perfect simple	112	62	71	40	65	72	51	43	37	56	54	38	79	30	71	57	69	74	70	41	70	71	66	58
	Ditransitive - past progressive	83	57	61	45	70	59	24	71	39	61	37	37	37	20	72	67	61	47	70	13	66	60	54	51
	Ditransitive - present perfect progressive	48	54	88	94	85	92	100	52	90	85	100	100	94	81	79	73	83	94	92	65	81	77	90	84
	Ditransitive - present perfect simple	54	37	41	35	56	30	30	33	31	33	39	30	33	28	65	59	63	41	48	26	67	65	59	43
	Ditransitive - present progressive	72	38	68	93	94	90	99	35	99	100	99	99	100	93	88	88	86	100	96	88	83	75	97	87
	Ditransitive - simple past	77	56	66	65	77	56	73	82	69	97	86	73	94	40	79	84	81	70	94	69	82	75	82	75
	Ditransitive - simple present	54	30	56	39	83	83	57	28	67	67	67	67	70	70	81	83	74	72	59	80	74	65	80	66
	Gerund	161	85	80	81	96	92	89	78	58	97	92	92	99	19	96	96	77	95	97	25	83	56	96	81
	Imperative	50	50	70	98	96	70	72	82	78	100	90	90	92	48	96	90	70	100	96	62	76	66	92	81
	Intransitive - conditional I progressive	9	89	78	100	89	100	100	100	0	100	22	22	44	67	33	56	89	89	89	44	100	78	100	72
	Intransitive - conditional I simple	3	0	33	0	67	100	100	100	33	100	100	100	100	100	67	100	67	67	100	100	100	0	67	73
	Intransitive - future I progressive	7	86	100	43	100	57	57	71	0	57	29	29	57	29	71	100	86	86	86	0	86	86	100	64
	Intransitive - future I simple	24	75	67	75	75	50	79	96	71	96	54	54	100	29	58	58	67	62	62	54	67	67	62	67
	Intransitive - future II progressive	4	50	50	0	25	50	25	25	75	75	25	25	0	0	50	75	25	25	50	0	25	50	25	34
	Intransitive - future II simple	7	100	100	86	86	100	100	100	57	100	57	57	100	43	43	43	86	100	100	14	86	57	86	77
	Intransitive - past perfect progressive	16	50	62	25	38	69	81	38	38	50	44	44	69	31	56	38	62	50	50	12	62	56	25	48
	Intransitive - past perfect simple	18	72	78	89	89	61	17	44	89	94	89	89	50	17	78	67	78	78	83	11	78	89	72	69
	Intransitive - past progressive	28	57	57	32	71	54	43	46	46	68	36	36	50	21	57	54	57	57	50	32	54	50	54	49
	Intransitive - present perfect simple	2	50	100	100	100	100	100	100	100	100	100	100	100	0	100	50	100	100	100	0	100	50	100	84
	Intransitive - present progressive	5	100	100	80	80	80	80	60	0	80	0	0	80	80	100	80	100	60	80	80	100	80	80	72
	Intransitive - simple past	24	38	46	33	62	58	58	96	96	100	100	100	100	42	62	58	58	54	71	67	71	38	71	67
	Intransitive - simple present	10	30	40	80	50	40	60	40	40	70	60	60	70	80	60	50	30	30	70	80	30	20	50	52
	Modal	20	60	55	35	40	45	95	100	50	10	45	45	15	0	35	35	70	60	25	0	75	60	45	45
	Modal negated	20	65	60	35	70	65	55	70	50	65	95	95	95	0	95	80	70	70	85	5	70	50	90	65
	Reflexive - conditional I progressive	65	52	46	35	48	45	45	28	15	38	25	25	63	65	83	85	66	35	52	63	54	77	57	50
	Reflexive - conditional I simple	112	70	70	46	48	58	57	37	9	32	32	32	100	80	86	89	78	57	64	90	68	91	60	62
	Reflexive - conditional II progressive	97	72	69	67	66	61	54	49	10	64	28	28	80	70	87	89	84	62	84	74	73	91	73	65
	Reflexive - conditional II simple	109	68	61	72	52	54	40	28	11	50	25	19	92	65	83	91	73	41	78	88	56	86	57	59
	Reflexive - future I progressive	70	67	79	61	70	84	66	64	60	59	83	83	66	47	80	76	71	61	79	53	61	74	69	69
	Reflexive - future I simple	83	67	86	39	71	76	55	63	49	61	67	67	61	41	78	72	78	55	76	53	70	80	65	65
	Reflexive - future II progressive	81	56	80	44	64	75	54	53	54	73	65	65	88	91	85	80	78	73	83	62	74	75	70	70
	Reflexive - future II simple	56	66	88	43	77	88	66	68	61	79	59	59	98	55	79	71	80	62	88	57	73	89	68	72
	Reflexive - past perfect progressive	98	50	66	56	67	71	53	51	33	66	45	45	82	60	71	72	68	68	76	44	66	72	60	61
	Reflexive - past perfect simple	53	47	55	72	68	74	47	43	25	64	23	23	98	66	79	85	66	60	87	72	64	81	68	62

Continued on next page

category	phenomenon	#	baselines					qe-as-a-metric							ref-based-metrics										
			BLEU	chrF	COMET-QE	BLEURT	YiSi-1	MS-COMET-QE	REUSE	HWTSC-TLM	COMET-Kiwi	KG-BERT	HWTSC-TS	CROSS-QE	MATESE-QE	XL-DA	XL-MQM	MEE2	MS-COMET	COMET-22	MATESE	MEE4	MEE	UniTE	avg
Verb valency	Reflexive - past progressive	5	100	100	20	40	100	20	40	20	80	20	20	60	40	80	80	100	100	80	40	100	100	80	65
	Reflexive - present perfect progressive	33	48	76	70	88	76	61	24	64	100	64	64	100	100	97	100	79	82	100	97	79	97	82	79
	Reflexive - present perfect simple	39	46	72	67	67	69	59	54	79	74	72	72	92	90	87	85	69	64	85	77	72	77	72	73
	Reflexive - present progressive	99	51	56	48	54	67	56	46	27	36	24	24	77	56	62	70	68	46	58	62	59	76	57	53
	Reflexive - simple past	119	70	77	34	73	73	61	53	37	89	73	67	83	62	82	76	75	72	91	44	69	79	72	69
	Reflexive - simple present	138	65	63	52	67	88	57	32	39	44	62	56	89	76	70	69	75	62	62	64	70	78	54	63
	Transitive - future II progressive	11	82	82	55	73	64	91	73	82	73	91	91	55	45	91	82	82	73	82	18	82	82	82	74
	Transitive - conditional I progressive	11	91	82	0	45	36	64	82	36	55	27	27	18	18	45	27	91	55	45	36	91	91	36	50
	Transitive - conditional I simple	9	100	100	11	89	56	56	67	56	100	56	56	67	33	78	44	100	78	100	56	100	100	67	71
	Transitive - conditional II progressive	20	55	55	55	75	80	60	60	35	75	50	50	40	10	70	65	65	65	85	50	75	55	55	58
	Transitive - conditional II simple	2	100	100	50	100	100	50	50	50	100	50	50	50	0	100	50	100	100	100	50	100	100	100	75
	Transitive - future I progressive	12	83	50	42	75	25	75	50	50	75	42	42	50	42	42	17	83	33	67	75	83	83	58	56
	Transitive - future I simple	22	95	77	32	64	59	82	50	41	36	36	36	50	14	59	36	91	59	55	50	86	91	59	57
	Transitive - future II simple	39	92	85	10	59	67	79	46	64	82	72	72	82	13	72	38	87	74	69	23	85	87	67	65
	Transitive - past perfect progressive	16	69	81	75	50	81	94	56	38	62	38	38	75	12	62	38	81	69	75	38	75	69	50	60
	Transitive - past perfect simple	9	78	89	44	89	33	100	44	89	100	78	78	56	0	78	78	89	44	89	67	89	78	67	71
	Transitive - present perfect progressive	5	80	80	80	80	20	100	20	100	40	60	60	40	0	60	20	80	60	60	60	80	80	40	59
	Transitive - present perfect simple	9	67	78	67	78	44	100	44	100	78	78	78	33	0	67	33	78	56	78	56	78	67	67	65
	Transitive - present progressive	10	70	40	10	20	30	60	0	40	50	40	40	40	30	50	50	60	20	30	40	50	60	30	39
	Transitive - simple past	23	43	57	52	96	35	91	61	61	87	57	57	52	43	91	70	61	65	87	70	74	43	87	65
	Transitive - simple present	16	62	62	25	38	69	50	50	31	94	31	31	44	19	50	25	62	62	81	44	62	56	44	50
	Case government	57	67	70	72	75	82	81	44	72	86	68	77	79	68	77	75	81	70	82	68	81	75	77	74
	Catenative verb	177	58	62	74	86	70	80	28	71	77	71	71	67	25	67	72	63	79	76	47	59	55	72	65
	Middle voice	29	69	83	97	93	79	72	31	90	79	86	86	76	31	97	93	86	93	83	79	90	62	93	79
	Passive voice	70	51	71	74	67	66	64	47	76	87	71	71	74	36	79	70	64	73	87	47	60	53	60	66
	Resultative	147	74	80	77	90	86	80	76	45	84	66	68	80	74	88	87	78	82	88	75	78	63	92	78
macro avg.		8945	65	70	62	74	70	70	53	60	75	62	63	72	45	77	73	74	71	79	52	74	70	74	67
micro avg.		8945	65	69	67	76	73	68	48	63	73	64	65	74	49	79	77	73	73	78	55	73	69	76	69

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