Fine-grained linguistic evaluation for state-of-the-art Machine Translation

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel,
Aljoscha Burchardt and **Sebastian Möller**
German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
`firstname.lastname@dfki.de`

Abstract

This paper describes a test suite submission providing detailed statistics of linguistic performance for the state-of-the-art German-English systems of the Fifth Conference of Machine Translation (WMT20). The analysis covers 107 phenomena organized in 14 categories based on about 5,500 test items, including a manual annotation effort of 45 person hours. Two systems (Tohoku and VolcanTrans) appear to have significantly better test suite accuracy than the others, although the best system of WMT20 is not significantly better than the one from WMT19 in a macro-average. Additionally, we identify some linguistic phenomena where all systems suffer (such as idioms, resultative predicates and pluperfect), but we are also able to identify particular weaknesses for individual systems (such as quotation marks, lexical ambiguity and sluicing). Most of the systems of WMT19 which submitted new versions this year show improvements.

1 Introduction

Fine-grained evaluation has recently had increasing interest on several natural language processing (NLP) tasks. Focusing on particular issues gives the possibility to analyse the automatic output in ways that cannot be seen by generic metrics. This is of particular importance in the era of deep learning, which has led to high performances and differences that are relatively difficult to distinguish. Additionally, detailed evaluation can provide indications for the improvement of the systems and the data collection, or allow focusing on phenomena of the long tail that might be of particular interest for certain cases (e.g. social biases; Stanovsky et al., 2019).

The most common method for fine-grained or focused evaluation are the *test suites* (also known as *challenge sets* or *benchmarks*; Guillou and Hardmeier, 2016; Ribeiro et al., 2020). These are test sets engineered in a particular way, so that they can test the performance of NLP tasks on concrete issues (Müller et al., 2018; Bawden et al., 2018).

This paper is presenting the use of such a test suite for the evaluation of the 11 German→English Machine Translation (MT) systems that participated at the Shared Task of the Fifth Conference of Machine Translation (WMT20; Barrault et al., 2020). The evaluation applies the DFKI test suite on German-English, via 5,514 test items which cover 107 linguistically motivated phenomena organized in 14 categories. After a reference in related work (Section 2), we explain shortly the structure of the test suite (Section 3) and present the results (Section 4) and the conclusions (Section 5).

2 Related Work

The use of test suites was introduced along with the early steps of MT in the 1990's (King and Falkedal, 1990; Way, 1991; Heid and Hildenbrand, 1991). With the emergence of deep learning, recent works re-introduced test suites that focus on the evaluation of particular linguistic phenomena (e.g. pronoun translation; Guillou and Hardmeier, 2016) or more generic test suites that aim at comparing different MT technologies (Isabelle et al., 2017; Burchardt et al., 2017) and Quality Estimation methods (Avramidis et al., 2018). The test suite track of the Conference of Machine Translation has already taken place two years in a row, allowing the presentation of several test suites, focusing on various linguistic phenomena and supporting different language directions. These include work in grammatical contrasts (Cinkova and Bojar, 2018), discourse (Bojar et al., 2018), morphology (Burlot et al., 2018), pronouns (Guillou et al., 2018) and word sense disambiguation (Rios et al., 2018). When compared to the vast majority of the previous test suites, the one presented here is the only one

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

Table 1: Examples of passing and failing MT outputs

that performs a systematic evaluation of more than one hundred phenomena on the state-of-the-art systems participating in WMT20.

3 Method

The test suite is a test set that has been devised manually with the aim to allow testing the MT output for several linguistic phenomena. The entire test suite consists of subsets that test one particular phenomenon each, through several test items. Each test item of the test suite consists of a source sentence and a set of correct and/or incorrect MT outputs. At the evaluation time, the test items are given as input to the MT systems and it is tested on whether the respective MT output consists a correct translation. By observing the amount of the test items that are translated correctly, one can calculate the performance of the MT systems regarding the respective phenomenon.

The evaluation presented in this paper is based on the DFKI Test Suite for MT on German to English, which has been presented in Burchardt et al. (2017) and applied extensively in the WMT shared task of 2018 (Macketanz et al., 2018) and 2019 (Avramidis et al., 2019). The current version includes 5,560 test items in order to control 107 phenomena organised in 14 categories. Some sample test items can be seen in Table 1 whereas a more detailed list of test sentences with correct and incorrect translations can be found on GitHub[1].

3.1 Application of the test suite

The construction of the test suite has been thoroughly explained in the papers from the previous years (Avramidis et al., 2018, 2019) and depicted in Figure 1 (steps *a-c*). The test items of the test suite are

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

given as input to the MT systems (step *d*). Their MT outputs are tested using a set of rules (regular expressions or fixed strings), each rule specific for a phenomenon, that defines whether the translations are correct with respect to the tested phenomenon (step *e*). When the automatic application of the rules cannot lead to a clear decision on whether the translation is correct or not, the test item is left with a warning. The warnings are consequently resolved by human annotators with linguistic knowledge, who inspect the MT output, provide a clear judgment and also augment the set of the rules to cover similar cases in the future (step *e*).

For every system we calculate the phenomenon-specific translation accuracy as the the number of the test sentences for the phenomenon which were translated properly, divided by the number of all test sentences for this phenomenon:

$$\text{accuracy} = \frac{\text{correct translations}}{\text{sum of test items}}$$

Each phenomenon is covered by at least 20 test items[2], whereas the same test items are given to multiple systems to achieve comparisons among them. In order to achieve a fair comparison among the systems, only the test items that do not contain any warnings for any of the systems are included in the calculation.

In order to define which systems have the best performance for a particular phenomenon, all systems are compared with the system with the highest accuracy. When comparing the highest scoring system with the rest, the significance of the comparison is confirmed with a one-tailed Z-test with $\alpha = 0.95$. The systems whose difference with the best system is not significant are considered to be in the first performance cluster and indicated with boldface in the tables.

3.2 Experiment setup

In the evaluation presented in the paper, MT outputs are obtained from the 11 systems that are part of the *news translation task* of the Fifth Conference on Machine Translation (WMT20). These are 6 systems submitted by the shared task participants, one baseline system from the shared task of the biomedical domain (WMTBiomedBaseline; Bawden et al., 2020) and 4 online commercial systems whose output has been obtained by the workshop organizers and therefore have been anonymized

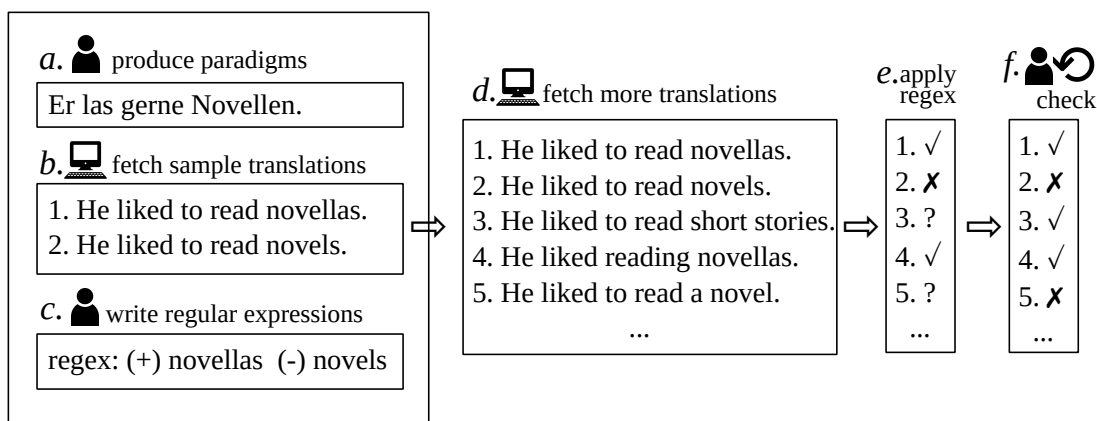[2] with the exception of 7 phenomena which have 9-19 items

347

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

(Online-A, B, G and Z). The submitted systems are OPPO (Shi et al., 2020), PROMT (Molchanov, 2020), Tohoku-AIP-NTT (Kiyono et al., 2020), UEdin (Germann, 2020), VolcTrans (Wu et al., 2020) and Zlabs, whereas a 7th system under the name "yolo"was ignored because it contained arbitrary translations. Unfortunately, contrary to previous years, very few system descriptions were provided by the time this paper was written and it is therefore not possible to associate linguistic performance with system types and settings.

As explained earlier, the application of the test suite on the output of the 11 systems left about 10% of unresolved warnings which needed to be manually edited. A human annotator with linguistic background devoted about 45 working hours in order to resolve 99% of them (resulting into 5,514 valid out of 5,560 total items).

4 Results

The accuracy of each system per linguistic category is briefly shown in Table 5 whereas the detailed statistics depicting the accuracy for every linguistic phenomenon, grouped in the respective linguistic categories are shown in Table 7. Since every category and every phenomenon have a different amount of test items, the average scores, shown in the last rows of the tables, are computed in three different ways: The first aggregates the contributions of all test items to compute the average percentages (micro-average), the second (Table 5) computes the percentages independently for each category and then takes the average (hence treating all categories equally; *category macro-average*) and the third (Table 7) computes the percentages independently for each phenomenon and then takes the average (hence treating all phenomena equally; *phenome-*

non macro-average). The significantly best systems for every category or phenomenon are bold-faced.

Very high scores do not necessarily mean that the MT of the respective grammatical phenomenon has been solved, but rather that the current test items of the test suite (which was engineered with the emergence of the first neural MT systems in 2017) are unable to expose difficulties of the systems. The artificial nature of the test suite and the variable number of test items per category and phenomenon should also be taken in consideration when doing comparisons between categories and phenomena.

4.1 Comparison between systems

Two systems are standing out for their overall performance. **Tohoku** achieves the best category macro-averaged accuracy of 88.1%, whereas it is sharing the first position with **VolcTrans** based on their micro-averaged accuracy (85.3-85.4%). The systems UEdin, Online-B, Online-G and Online-A are next. Tohoku and VolcTrans are also the best performing systems for all linguistic categories, whereas UEdin is losing in one category and Online-A is losing in two categories.

Two systems, WMTBiomedBaseline and ZLabs show very low performances and are assumed to be non state-of-the-art systems. We will therefore exclude these two systems from the discussion and conclusions for phenomena and categories. Whereas no description was available for ZLabs, the lower performance of the WMTBiomedBaseline can be attributed to the fact that it was trained with only 56% of the parallel training data used by Tohoku and no synthetic data. Among the rest of the systems, the worst performing one is Online-Z, achieving the lowest accuracy (74% on both micro- and macro-average), being on par with the best systems

Verb Valency	
Ich erinnere mich seiner.	
I remember his.	fail
I remember him.	pass
False Friends	
Er las gerne Novellen.	
He liked to read novels.	fail
He liked to read novellas.	pass

Table 2: Examples of linguistic categories with lower accuracy with passing and failing MT outputs

on only 6 categories.

BLEU scores (Papineni et al., 2002) on the official test-set are also calculated for further comparison. The order of the systems based on BLEU seems to correlate with the order given by the category macro-average with the exception of one system (OPPO). Since BLEU scores are calculated on a different test set, further investigation is needed to confirm if this correlation is of any significance. According to the official human evaluation campaign (Barrault et al., 2020, table 11), the first nine systems are tied, so it is hard to compare this system order with theirs.

4.2 Linguistic categories

The average accuracy regarding the linguistic categories ranges in relatively high numbers, between 68.9% and 97.3%. The categories with the highest accuracy in average are the **negation** (97.3%), the **composition** (85.3%), the **subordination** (85.3%) and the **named entities and terminology** (82%). The ones with the lowest accuracy are the **multiword expressions** (MWE), the **ambiguity**, the **false friends** and the **verb valency** (68.9-71.5%).

When one tries to identify weaknesses of particular systems, OPPO is suffering mostly concerning function words and long distance dependencies (LDD) / interrogatives. Online-Z and PROMT have issues with ambiguity and several systems have issues with punctuation. Some of these issues are discussed in a more fine-grained level below.

The comparison of the state-of-the-art systems with the low-resource WMTBiomedBaseline indicates that some categories, such as ambiguity and composition, are particularly sensitive to low resources, as their accuracy is proportionally lower than other categories if compared to the respective category accuracies of the state-of-the-art systems.

Table 2 contains examples from the two low accuracy categories *verb valency* and *false friends*. *Verb valency* refers to the arguments that are being controlled by the predicate. Certain verbs require a specific grammatical case. In our example, the German verb *sich erinnern* (to remember) requires a genitive object, in this case *seiner*. *Seiner*, however, can also mean *his* as in the possessive pronoun, which explains the mistranslation of *I remember his*.

False friends are words in different languages that look similar and are therefore often mistaken for being translations of one another, even though their meanings differ. The German noun *Novelle* does not translate to *novel*, but to *novella* or *short story*. While you would expect a human to make these kind of translation errors, it is surprising to see that also MT systems are prone to mistranslating false friends.

4.3 Linguistic phenomena

The accuracy regarding individual linguistic phenomena has a wide range, between very low scores (15%) and full success (100%). The phenomena which all systems had difficulty to handle were the **idioms** and the **resultative predicates**, with most systems scoring only 20% and 26% respectively. However, the overall performance on these phenomena has improved: last year only 3 systems could achieve this performance, with the majority of the systems having 5-10% less accuracy. **Modal pluperfect** is also ranging very low, scoring between 2.2% and 50.6% and similar is the case for its negated version. Other moods of the pluperfect make it particularly difficult for some systems, e.g. PROMT and Online-Z suffer in translating the **ditransitive** and **intransitive pluperfect**.

When trying to find the cases that consist a weakness for particular systems, Online-Z indicates one of the lowest scores in punctuation, which appears to derive from the fact that the system removes all **quotation marks**. A similar issue is observed with Online-G and OPPO which could correctly convey almost half of the quotation marks, whereas another two systems have some way to go. Interestingly enough, despite strongly depending on preprocessing, quotation marks are a common issue, since similar cases have been noted in previous years. In other phenomena, Online-Z has a very low accuracy for **sluicing** whereas PROMT is relatively weak concerning **lexical ambiguity**.

Table 3 contains further translation examples from linguistic phenomena with low accuracy. A *resultative predicate* is a construction that consists

Resultative Predicate	
Sie trinkt die Tasse leer.	
She drinks the cup empty.	fail
She empties the cup.	*pass*
She is drinking the whole cup.	*pass*
Intransitive Pluperfect	
Sie hatten geschlafen.	
They were sleeping.	fail
They had slept.	pass
Sluicing	
John mag die Nudeln nicht, aber er weiß nicht, warum.	
John doesn't like the noodles but he doesn't know why.	fail
John doesn't like the noodles, but he doesn't know why.	pass
Lexical Ambiguity	
Das Gericht gestern Abend war lecker.	
The court last night was delicious.	fail
The dish last night was delicious.	pass

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

WMT	categ. macro-	micro-
2018	81.0	**84.1**
2019	**87.4**	83.0
2020	**88.0**	85.1

Table 4: The accuracy (%) of the best system of each year as measured over 5,555 test items that were common over the last years. The scores that are significantly higher in each column are boldfaced

of a verb and an adjective in which the verb describes an action and the adjective describes the result of that action. In many cases, resultative predicates lead to translation errors as they do not exist in English and a literal translation leads to an ungrammatical translation, as can be seen in the example. Since none of the systems could produce a correct output for this sentence, we have provided two possible correct translations here as examples.

Intransitive verbs do not require further objects (as opposed to transitive or ditransitive verbs). *Pluperfect* is a tense which is used in German to describe completed actions that have taken place in the past. It should be translated to English in pluperfect as well. In the example, the incorrect translation contains past progressive *were sleeping* instead of the correct *had slept*.

Sluicing is a type of ellipsis that can occur in direct and indirect interrogative clauses. A wh-word precedes the part of the sentence that contains the ellipsis. In our example, all constituents following the wh-word are elided: *John mag die Nudeln nicht, aber er weiß nicht, warum ~~er die Nudeln nicht mag~~*. Sluicing exists in both German and in English: *John doesn't like the noodles, but he doesn't know why ~~he doesn't like the noodles~~*. One difference between the German and the English sluicing sentence is that in German there are two commas, while in English there is only one. Since this phenomenon concerns the complete sentence, punctuation should be correct when translating a sentence containing sluicing. In our example, the

missing comma leads to fail.

The fourth example in the Table contains the lexical ambiguity *Gericht*. *Gericht* can either mean *court* or *dish* but the context provided in the sentence (*war lecker*, English: *was delicious*) serves as disambiguation so that only a translation referring to *dish* (or to food in some way) can be a pass. Any translation referring to *court/courthouse/tribunal* or the like is a fail.

4.4 Comparison with previous years

One can notice some improvements on the overall performance of the best system, as compared with the previous two years. As seen in Table 4 from 2018 to 2019 there was a 6.4% improvement on the macro-averaged accuracy but there was no significant improvement from 2019 to 2020. The best system of 2019 was not submitted in 2020 and this is unfortunate, as it performed better than this year's best system in four categories (mostly regarding ambiguity; Table 6). When considering micro-averaged accuracy, there is significant improvement since last year (2.1%), but the best system of 2018 is competing with the one of this year, due to its high performance regarding verb tense/mood.

One can also consider the improvement of individual systems submitted to WMT from one year to another, starting from 2018. This year, only 5 of the 2019 systems submitted their new version and the yearly difference of their test suite accuracy can be seen in Table 6. The accuracy is measured over the test items that are common over all three years (or at least the last two).

All systems indicate considerable improvements on the macro-average since the previous year, ranging between 2,4% and 8,5%. Online-G had a major improvement for a second year in a row (21.6% in total), whereas it is the only system that achieved such an improvement without having an accuracy drop for any of the linguistic categories, whereas it improved 6 categories for more than 10%. PROMT

had improvement in all categories apart from verb tense/aspect/mood, UEdin deteriorated in verb tense/aspect/mood, whereas Online-B deteriorated in composition, named entity/terminology and punctuation.

The linguistic categories that improved mostly in average are the **long distance dependencies / interrogatives**, the **verb valency**, the **ambiguity** and the **punctuation**.

5 Conclusions and further work

In this paper we present the results of the application of the DFKI test suite in the output of the state-of-the-art MT systems participating in the Shared Task of the Fifth Conference of Machine Translation (WMT20). Based on about 5,500 test items, we present detailed accuracies regarding 107 phenomena organized in 14 categories. Additionally, the evolution of systems submitted also in previous years is observed.

The best system of this year is not significantly better than the one from 2019 in a macro-average, but one can see significant improvement from two years ago. The systems that seem to have the best accuracies are Tohoku and VolcanTrans. The phenomena that most systems face difficulties are again this year the idioms, the resultative predicates and some moods of the pluperfect, whereas some systems still have issues with quotation marks and lexical ambiguity.

As discussed previously, the high accuracies achieved for particular phenomena or categories raise questions on whether these phenomena are getting solved, or whether the test suite (which was originally built to challenge the systems from 2017) should raise the difficulty by including more test items.

In further work, we would like to be able to associate the performance on specific phenomena with decisions related to decisions during the development of the systems, once there is enough information about this process for all systems.

Acknowledgments

References

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

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.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 Conference on Machine Translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–54, Online. Association for Computational Linguistics.

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 2020. Findings of the WMT 2020 biomedical translation shared task: Basque, italian and russian as new additional languages. In *Proceedings of the Fifth Conference on Machine Translation*, pages 658–685, Online. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Ondřej Bojar, Jiří Mírovský, Kateřina Rysová, and Magdaléna Rysová. 2018. EvalD Reference-Less Discourse Evaluation for WMT18. In *Proceedings of the Third Conference on Machine Translation*, pages 545–549, Belgium, Brussels. Association for Computational Linguistics.

Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter, and Philip Williams. 2017. A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. *The Prague Bulletin of Mathematical Linguistics*, 108:159–170.

Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondřej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen, and François Yvon. 2018. The WMT'18 Morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English. In *Proceedings of the Third Conference on Machine Translation*, pages 550–564, Belgium, Brussels. Association for Computational Linguistics.

Silvie Cinkova and Ondřej Bojar. 2018. Testsuite on Czech–English Grammatical Contrasts. In *Proceedings of the Third Conference on Machine Translation*, pages 565–575, Belgium, Brussels. Association for Computational Linguistics.

Ulrich Germann. 2020. The University of Edinburgh's submission to the German-to-English and English-to-German Tracks in the WMT 2020 News Translation and Zero-shot Translation Robustness Tasks. In *Proceedings of the Fifth Conference on Machine Translation*, pages 196–200, Online. Association for Computational Linguistics.

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

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A Pronoun Test Suite Evaluation of the English–German MT Systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation*, pages 576–583, Belgium, Brussels. Association for Computational Linguistics.

Ulrich Heid and Elke Hildenbrand. 1991. Some practical experience with the use of test suites for the evaluation of SYSTRAN. In *the Proceedings of the Evaluators' Forum, Les Rasses*. Citeseer.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A Challenge Set Approach to Evaluating Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Shun Kiyono, Takumi Ito, Ryuto Konno, Makoto Morishita, and Jun Suzuki. 2020. Tohoku-aip-ntt at wmt 2020 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 144–154, Online. 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)*, Brussels, Belgium. Association for Computational Linguistics.

Alexander Molchanov. 2020. PROMT Systems for WMT 2020 Shared News Translation Task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 247–252, Online. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

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

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.

Annette Rios, Mathias Müller, and Rico Sennrich. 2018. The Word Sense Disambiguation Test Suite at WMT18. In *Proceedings of the Third Conference on Machine Translation*, pages 594–602, Belgium, Brussels. Association for Computational Linguistics.

Tingxun Shi, Shiyu Zhao, Xiaopu Li, Xiaoxue Wang, Qian Zhang, Di Ai, Dawei Dang, Xue Zhengshan, and JIE HAO. 2020. OPPO's machine translation systems for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 281–291, Online. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Andrew Way. 1991. Developer-Oriented Evaluation of MT Systems. In *Proceedings of the Evaluators' Forum*, pages 237–244, Les Rasses, Vaud, Switzerland. ISSCO.

Liwei Wu, Xiao Pan, Zehui Lin, Yaoming ZHU, Mingxuan Wang, and Lei Li. 2020. The Volctrans Machine Translation System for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 304–310, Online. Association for Computational Linguistics.

category	items	Tohoku	VolcTrans	UEdin	Onl-B	Onl-G	Onl-A	PROMT	OPPO	Onl-Z	ZLabs	WMTBi	avg
Ambiguity	81	**82.7**	77.8	72.8	79.0	84.0	**76.5**	64.2	**82.7**	67.9	45.7	30.9	69.5
Composition	49	**98.0**	**98.0**	93.9	93.9	95.9	93.9	**89.8**	**95.9**	85.7	49.0	44.9	85.3
Coordination & ellipsis	78	**89.7**	**91.0**	89.7	**91.0**	85.9	87.2	**87.2**	**87.2**	60.3	52.6	44.9	78.8
False friends	36	72.2	**80.6**	72.2	**80.6**	77.8	69.4	72.2	66.7	**86.1**	52.8	50.0	71.0
Function word	72	**86.1**	80.6	**86.1**	**90.3**	**90.3**	83.3	**88.9**	55.6	**88.9**	41.7	43.1	75.9
LDD & interrogatives	174	**89.1**	**86.2**	85.1	83.3	**86.8**	77.6	81.0	58.6	72.4	48.3	58.6	75.2
MWE	80	**80.0**	75.0	71.3	**77.5**	**77.5**	71.3	70.0	**78.8**	**73.8**	45.0	37.5	68.9
Named entity & terminology	89	**92.1**	84.3	87.6	82.0	82.0	**88.8**	87.6	**85.4**	68.5	70.8	73.0	82.0
Negation	20	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	95.0	**100.0**	**100.0**	95.0	80.0	**100.0**	97.3
Non-verbal agreement	61	**91.8**	**88.5**	**88.5**	86.9	**90.2**	83.6	82.0	**88.5**	**85.2**	54.1	57.4	81.5
Punctuation	60	96.7	**98.3**	**98.3**	71.7	61.7	**100.0**	**98.3**	70.0	28.3	68.3	55.0	77.0
Subordination	180	**90.6**	88.3	**91.1**	**91.1**	**92.2**	88.9	90.0	**90.6**	**87.8**	65.0	62.2	85.3
Verb tense/aspect/mood	4447	**84.6**	**85.3**	80.3	75.9	79.6	77.5	75.1	79.3	73.6	50.5	52.1	74.0
Verb valency	87	79.3	**81.6**	77.0	**81.6**	77.0	77.0	71.3	**80.5**	64.4	44.8	51.7	71.5
micro-average	5514	85.3	**85.4**	81.2	77.7	80.6	78.7	76.5	79.1	73.6	51.3	52.4	74.7
macro-average	5514	**88.1**	86.8	85.3	84.6	84.3	83.6	82.7	80.0	74.1	54.9	54.4	78.1
BLEU		43.8	43.5	42.3	41.9	41.4	40.4	39.6	43.2	35.4	31.5	32.1	35.4

Table 5: Accuracies (%) of successful translations for 11 systems and 14 categories. Boldface indicates the significantly best performing systems in each row

category	Onl-A		Onl-B		Onl-G		PROMT	UEdin		best	
	2019	2020	2019	2020	2019	2020	2020	2019	2020	2019	2020
Ambiguity	+2.6	+7.7	+1.3	+2.6	+2.6	+11.5	+16.7	+11.6	+14.1	+12.4	-9.9
Composition	+10.4	+2.1		-4.1	+12.5	+12.5	+10.4	+8.3	+8.4	+12.3	-1.3
Coordination & ellipsis				+2.2	+8.7	+19.6	+6.6		+4.4	5.1	-2.8
False friends			+2.8	+2.8		+5.6		+13.9	+5.5	+8.3	
Function word	+1.6	-1.6		+10.9	+42.2	-1.6	+6.2	+7.8	+2.7		
LDD & interrogatives	-2.2	+8.7		+6.6	+14.2	+18.4	+8.7		+15.2	+1.8	+3.4
MWE	+1.3	+5.4		+6.6	+5.3	+9.3	+14.6	+5.3	+10.7	+10.0	+1.2
Named entitiy & terminology		+3.1		-3.2	-1.5	+7.8	+15.6	+7.8	+3.1	-2.3	
Negation		-5.0		+5.0	+4		+10.0	-10.0	+10.0		+9.0
Non-verbal agreement	+6.9				+22.4	+10.4	+10.4	+10.4	+6.9	+9.8	
Punctuation	-21.9	+25.5		-3.6	-7.3	+3.7	+16.4	+21.8	+9.1	+30.0	+3.4
Subordination	-11.3	+10.3	+1.0	+7.2	+6.2	+5.1	+4.1	-2.1	+7.2	-3.8	+1.2
Verb tense/aspect/mood	+11.9	-6.3	+0.2	+2.0	+19.8	+13.9	-0.2	+5.6	-1.1	0.8	5.8
Verb valency	+1.5	+9.0		+13.5	+11.9	+6.0	+11.9	+3.0	+13.4	9.2	-1.2
micro-avg	+9.9	-4.4	+0.2	+2.4	+18.5	+13.1	+1.5	+5.6	+0.6	+2.0	+4.8
macro-avg	+0.1	+4.0	+0.4	+3.5	+12.7	+8.9	+9.3	+6.0	+7.6	+7.4	+0.6

Table 6: Difference of the test suite accuracy from one year to the next one per category, for the systems participating in the shared tasks WMT18-20, measured over the test items that are common over all these years.

phenomenon	items	Tohoku	VolcTrans	UEdin	Onl-B	Onl-G	Onl-A	PROMT	OPPO	Onl-Z	ZLabs	WMTBi	avg
Ambiguity	81	82.7	77.8	72.8	79.0	84.0	76.5	64.2	82.7	67.9	45.7	30.9	69.5
Lexical ambiguity	63	85.7	79.4	77.8	77.8	87.3	79.4	65.1	85.7	68.3	49.2	36.5	72.0
Structural ambiguity	18	72.2	72.2	55.6	83.3	72.2	66.7	61.1	72.2	66.7	33.3	11.1	60.6
Composition	49	98.0	98.0	93.9	93.9	95.9	93.9	89.8	95.9	85.7	49.0	44.9	85.3
Compound	29	96.6	96.6	89.7	93.1	93.1	89.7	86.2	96.6	82.8	51.7	48.3	84.0
Phrasal verb	20	100.0	100.0	100.0	95.0	100.0	100.0	95.0	95.0	90.0	45.0	40.0	87.3
Coordination & ellipsis	78	89.7	91.0	89.7	91.0	85.9	87.2	87.2	87.2	60.3	52.6	44.9	78.8
Gapping	20	95.0	95.0	95.0	95.0	95.0	80.0	95.0	90.0	75.0	70.0	40.0	84.1
Right node raising	20	80.0	85.0	85.0	80.0	65.0	85.0	80.0	85.0	85.0	40.0	15.0	71.4
Sluicing	18	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	11.1	72.2	100.0	89.4
Stripping	20	85.0	85.0	80.0	90.0	85.0	85.0	75.0	75.0	65.0	30.0	30.0	71.4
False friends	36	72.2	80.6	72.2	80.6	77.8	69.4	72.2	66.7	86.1	52.8	50.0	71.0
Function word	72	86.1	80.6	86.1	90.3	90.3	83.3	88.9	55.6	88.9	41.7	43.1	75.9
Focus particle	24	91.7	91.7	95.8	95.8	100.0	91.7	95.8	91.7	95.8	75.0	83.3	91.7
Modal particle	29	75.9	72.4	72.4	79.3	75.9	72.4	79.3	62.1	79.3	41.4	37.9	68.0
Question tag	19	94.7	78.9	94.7	100.0	100.0	89.5	94.7	0.0	94.7	0.0	0.0	67.9
LDD & interrogatives	174	89.1	86.2	85.1	83.3	86.8	77.6	81.0	58.6	72.4	48.3	58.6	75.2
Extended adjective construction	20	85.0	80.0	75.0	80.0	80.0	70.0	75.0	80.0	55.0	55.0	50.0	71.4
Extraposition	20	75.0	70.0	75.0	55.0	65.0	65.0	75.0	65.0	65.0	40.0	40.0	62.7
Multiple connectors	20	90.0	90.0	95.0	75.0	95.0	75.0	70.0	80.0	65.0	85.0	85.0	82.3
Pied-piping	20	90.0	85.0	90.0	90.0	85.0	80.0	90.0	95.0	65.0	30.0	55.0	77.7
Polar question	20	100.0	100.0	100.0	100.0	100.0	100.0	95.0	10.0	85.0	55.0	85.0	84.5
Scrambling	20	90.0	75.0	70.0	85.0	85.0	50.0	65.0	85.0	60.0	20.0	35.0	65.5
Topicalization	19	89.5	78.9	78.9	78.9	78.9	78.9	68.4	84.2	73.7	36.8	36.8	71.3
Wh-movement	35	91.4	100.0	91.4	94.3	97.1	91.4	97.1	8.6	94.3	57.1	71.4	81.3
MWE	80	80.0	75.0	71.3	77.5	77.5	71.3	70.0	78.8	73.8	45.0	37.5	68.9
Collocation	20	100.0	90.0	80.0	95.0	95.0	75.0	80.0	95.0	90.0	35.0	20.0	77.7
Idiom	20	25.0	20.0	15.0	20.0	20.0	15.0	5.0	20.0	20.0	0.0	5.0	15.0
Prepositional MWE	20	95.0	95.0	95.0	95.0	95.0	95.0	95.0	100.0	90.0	70.0	60.0	89.5
Verbal MWE	20	100.0	95.0	95.0	100.0	100.0	100.0	100.0	100.0	95.0	75.0	65.0	93.2
Named entity & terminology	89	92.1	84.3	87.6	82.0	82.0	88.8	87.6	85.4	68.5	70.8	73.0	82.0
Date	20	100.0	100.0	100.0	85.0	100.0	100.0	100.0	95.0	60.0	75.0	100.0	92.3
Domainspecific term	20	75.0	60.0	60.0	65.0	60.0	70.0	65.0	70.0	50.0	50.0	40.0	60.5
Location	20	95.0	95.0	95.0	95.0	90.0	95.0	90.0	95.0	90.0	85.0	90.0	92.3
Measuring unit	20	100.0	90.0	100.0	90.0	90.0	100.0	100.0	90.0	75.0	85.0	80.0	90.9
Proper name	9	88.9	66.7	77.8	66.7	55.6	66.7	77.8	66.7	66.7	44.4	33.3	64.6
Negation	20	100.0	100.0	100.0	100.0	100.0	95.0	100.0	100.0	95.0	80.0	100.0	97.3
Non-verbal agreement	61	91.8	88.5	88.5	86.9	90.2	83.6	82.0	88.5	85.2	54.1	57.4	81.5
Coreference	20	80.0	70.0	75.0	80.0	75.0	70.0	70.0	75.0	70.0	70.0	60.0	72.3
External possessor	21	100.0	95.2	90.5	85.7	95.2	81.0	81.0	90.5	90.5	14.3	42.9	78.8
Internal possessor	20	95.0	100.0	100.0	95.0	100.0	100.0	95.0	100.0	95.0	80.0	70.0	93.6

354

phenomenon	items	Tohoku	VolcTrans	UEdin	Onl-B	Onl-G	Onl-A	PROMT	OPPO	Onl-Z	ZLabs	WMTBi	avg
Punctuation	60	96.7	98.3	98.3	71.7	61.7	100.0	98.3	70.0	28.3	68.3	55.0	77.0
Comma	20	100.0	100.0	100.0	90.0	100.0	100.0	100.0	100.0	85.0	85.0	85.0	95.0
Quotation marks	40	95.0	97.5	97.5	62.5	42.5	100.0	97.5	55.0	0.0	60.0	40.0	68.0
Subordination	180	90.6	88.3	91.1	91.1	92.2	88.9	90.0	90.6	87.8	65.0	62.2	85.3
Adverbial clause	20	90.0	90.0	90.0	90.0	95.0	90.0	90.0	85.0	90.0	85.0	80.0	88.6
Cleft sentence	20	95.0	90.0	90.0	95.0	95.0	90.0	95.0	95.0	80.0	55.0	75.0	86.8
Free relative clause	20	90.0	95.0	95.0	95.0	90.0	90.0	85.0	90.0	85.0	65.0	60.0	85.5
Indirect speech	20	95.0	85.0	90.0	85.0	100.0	90.0	90.0	95.0	80.0	50.0	70.0	84.5
Infinitive clause	20	100.0	100.0	100.0	95.0	100.0	100.0	100.0	100.0	95.0	80.0	65.0	94.1
Object clause	20	95.0	95.0	95.0	100.0	100.0	95.0	95.0	100.0	95.0	75.0	70.0	92.3
Pseudo-cleft sentence	20	70.0	70.0	70.0	75.0	80.0	75.0	70.0	75.0	75.0	40.0	35.0	66.8
Relative clause	20	85.0	80.0	95.0	90.0	85.0	75.0	90.0	80.0	90.0	80.0	65.0	83.2
Subject clause	20	95.0	90.0	95.0	95.0	85.0	95.0	95.0	95.0	100.0	55.0	40.0	85.5
Verb tense/aspect/mood	4447	84.6	85.3	80.3	75.9	79.6	77.5	75.1	79.3	73.6	50.5	52.1	74.0
Conditional	19	100.0	100.0	100.0	100.0	89.5	89.5	94.7	94.7	84.2	73.7	84.2	91.9
Ditransitive - future I	36	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	91.7	75.0	66.7	93.9
Ditransitive - future I subjunctive II	36	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	75.0	55.6	93.7
Ditransitive - future II	36	100.0	100.0	100.0	100.0	86.1	100.0	97.2	100.0	100.0	69.4	91.7	94.9
Ditransitive - future II subjunctive II	36	100.0	100.0	100.0	100.0	100.0	97.2	100.0	100.0	100.0	69.4	94.4	96.5
Ditransitive - perfect	36	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.2	83.3	86.1	97.0
Ditransitive - pluperfect	36	94.4	94.4	66.7	86.1	80.6	100.0	22.2	63.9	33.3	44.4	50.0	66.9
Ditransitive - pluperfect subjunctive II	36	100.0	91.7	100.0	100.0	100.0	100.0	100.0	100.0	97.2	33.3	63.9	89.6
Ditransitive - present	36	97.2	97.2	100.0	100.0	100.0	97.2	72.2	97.2	83.3	75.0	63.9	89.4
Ditransitive - preterite	36	88.9	86.1	77.8	97.2	77.8	77.8	75.0	94.4	83.3	50.0	50.0	78.0
Ditransitive - preterite subjunctive II	36	75.0	75.0	69.4	72.2	63.9	66.7	66.7	72.2	77.8	36.1	44.4	65.4
Imperative	20	85.0	80.0	80.0	95.0	95.0	75.0	80.0	90.0	85.0	30.0	35.0	75.5
Intransitive - future I	36	97.2	97.2	97.2	97.2	100.0	97.2	100.0	97.2	97.2	86.1	66.7	93.9
Intransitive - future I subjunctive II	36	100.0	100.0	100.0	86.1	100.0	100.0	100.0	100.0	100.0	80.6	69.4	94.2
Intransitive - future II	42	88.1	90.5	95.2	95.2	88.1	92.9	100.0	97.6	92.9	61.9	35.7	85.3
Intransitive - future II subjunctive II	36	100.0	100.0	100.0	100.0	61.1	100.0	100.0	100.0	100.0	44.4	36.1	85.6
Intransitive - perfect	84	100.0	100.0	100.0	100.0	98.8	100.0	100.0	100.0	97.6	63.1	54.8	92.2
Intransitive - pluperfect	36	83.3	83.3	66.7	33.3	83.3	75.0	50.0	77.8	61.1	38.9	19.4	61.1
Intransitive - pluperfect subjunctive II	36	100.0	100.0	97.2	94.4	100.0	100.0	94.4	97.2	100.0	25.0	19.4	84.3
Intransitive - present	36	100.0	100.0	100.0	97.2	100.0	100.0	100.0	100.0	97.2	69.4	52.8	92.4
Intransitive - preterite	66	93.9	97.0	86.4	92.4	92.4	95.5	84.8	97.0	95.5	47.0	19.7	82.0
Intransitive - preterite subjunctive II	36	63.9	75.0	63.9	77.8	80.6	61.1	88.9	75.0	66.7	22.2	11.1	62.4
Modal - future I	180	79.4	90.0	73.9	78.9	79.4	73.9	72.2	75.6	75.6	62.2	59.4	74.6
Modal - future I subjunctive II	180	77.2	86.7	76.1	76.7	74.4	72.8	65.0	66.1	70.6	57.2	64.4	71.6
Modal - perfect	180	90.0	83.9	85.0	73.9	70.6	65.0	82.8	78.3	60.0	1.7	63.3	68.6
Modal - pluperfect	180	50.6	40.6	37.2	22.8	32.8	15.6	28.3	16.7	2.2	0.0	8.9	23.2
Modal - pluperfect subjunctive II	180	65.6	60.6	60.0	43.9	59.4	56.7	60.0	54.4	55.0	37.2	32.8	53.2
Modal - present	180	97.8	96.1	93.3	80.0	94.4	91.7	69.4	95.0	92.2	81.7	69.4	87.4
Modal - preterite	180	98.3	99.4	94.4	96.7	98.9	96.7	93.9	98.9	97.8	80.6	72.8	93.5

phenomenon	items	Tohoku	VolcTrans	UEdin	Onl-B	Onl-G	Onl-A	PROMT	OPPO	Onl-Z	ZLabs	WMTBi	avg
Modal - preterite subjunctive II	180	73.9	78.9	75.0	60.0	75.6	78.9	77.8	77.2	61.7	57.8	52.2	69.9
Modal negated - future I	180	80.0	92.8	75.6	78.3	79.4	76.1	75.6	79.4	77.2	66.7	67.2	77.1
Modal negated - future I subjunctive II	180	78.3	85.6	76.1	75.6	76.1	70.6	69.4	77.2	69.4	57.8	63.3	72.7
Modal negated - perfect	171	95.9	93.6	94.2	70.2	73.7	70.2	88.3	80.7	70.2	9.4	67.8	74.0
Modal negated - pluperfect	169	36.1	22.5	18.3	32.0	32.5	14.8	12.4	4.1	0.0	0.6	7.1	16.4
Modal negated - pluperfect subjunctive II	179	71.5	70.9	73.7	40.8	52.5	57.0	53.6	63.1	54.2	31.8	34.6	54.9
Modal negated - present	169	100.0	100.0	99.4	84.6	92.3	98.8	76.3	99.4	84.0	81.7	82.8	90.9
Modal negated - preterite	179	99.4	100.0	96.6	91.1	98.9	99.4	92.7	100.0	88.3	74.3	76.0	92.4
Modal negated - preterite subjunctive II	174	74.7	81.0	70.1	62.6	79.9	81.6	75.3	70.1	75.9	54.6	50.0	70.5
Progressive	19	84.2	73.7	63.2	78.9	73.7	84.2	78.9	84.2	63.2	15.8	36.8	67.0
Reflexive - future I	36	100.0	100.0	91.7	88.9	88.9	94.4	80.6	94.4	97.2	69.4	33.3	85.4
Reflexive - future I subjunctive II	36	83.3	86.1	83.3	91.7	80.6	83.3	69.4	88.9	97.2	55.6	30.6	77.3
Reflexive - future II	36	94.4	100.0	88.9	91.7	91.7	86.1	77.8	94.4	69.4	41.7	27.8	78.5
Reflexive - future II subjunctive II	36	83.3	88.9	83.3	91.7	61.1	83.3	75.0	86.1	86.1	30.6	25.0	72.2
Reflexive - perfect	36	100.0	100.0	94.4	94.4	94.4	91.7	86.1	91.7	91.7	41.7	33.3	83.6
Reflexive - pluperfect	36	91.7	100.0	83.3	86.1	91.7	80.6	80.6	97.2	91.7	30.6	19.4	77.5
Reflexive - pluperfect subjunctive II	36	86.1	83.3	80.6	91.7	88.9	80.6	75.0	80.6	88.9	36.1	22.2	74.0
Reflexive - present	36	97.2	100.0	91.7	91.7	91.7	86.1	69.4	94.4	88.9	33.3	13.9	78.0
Reflexive - preterite	36	94.4	86.1	69.4	91.7	86.1	69.4	75.0	88.9	75.0	13.9	16.7	69.7
Reflexive - preterite subjunctive II	36	83.3	75.0	61.1	77.8	80.6	61.1	69.4	75.0	66.7	22.2	19.4	62.9
Transitive - future I	42	97.6	97.6	97.6	97.6	97.6	97.6	97.6	97.6	95.2	90.5	81.0	95.2
Transitive - future I subjunctive II	36	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.2	77.8	97.7
Transitive - future II	36	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	91.7	63.9	96.0
Transitive - future II subjunctive II	36	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.2	63.9	96.5
Transitive - perfect	42	100.0	100.0	100.0	100.0	97.6	100.0	100.0	100.0	95.2	85.7	66.7	95.0
Transitive - pluperfect	36	100.0	100.0	77.8	80.6	100.0	100.0	75.0	94.4	80.6	66.7	36.1	82.8
Transitive - pluperfect subjunctive II	36	100.0	91.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	30.6	44.4	87.9
Transitive - present	48	100.0	100.0	100.0	100.0	100.0	100.0	95.8	100.0	97.9	75.0	75.0	94.9
Transitive - preterite	36	97.2	97.2	88.9	91.7	97.2	88.9	91.7	100.0	97.2	66.7	55.6	88.4
Transitive - preterite subjunctive II	36	66.7	75.0	63.9	69.4	61.1	58.3	80.6	83.3	69.4	25.0	50.0	63.9
Verb valency	87	79.3	81.6	77.0	81.6	77.0	77.0	71.3	80.5	64.4	44.8	51.7	71.5
Case government	28	92.9	96.4	85.7	92.9	82.1	89.3	78.6	92.9	82.1	42.9	57.1	81.2
Mediopassive voice	20	85.0	90.0	85.0	95.0	90.0	80.0	70.0	90.0	55.0	50.0	50.0	76.4
Passive voice	19	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	89.5	78.9	73.7	94.7
Resultative predicates	20	35.0	35.0	35.0	35.0	35.0	35.0	35.0	35.0	25.0	10.0	25.0	30.9
micro-average	5514	85.3	85.4	81.2	77.7	80.6	78.7	76.5	79.1	73.6	51.3	52.4	74.7
phenomenon macro-average	5514	89.1	88.0	85.2	85.1	85.5	83.9	82.1	83.7	78.5	53.7	51.8	78.8
category macro-average	5514	88.1	86.8	85.3	84.6	84.3	83.6	82.7	80.0	74.1	54.9	54.4	78.1
BLEU		43.8	43.5	42.3	41.9	41.4	40.4	39.6	43.2	35.4	31.5	32.1	35.4

Table 7: Accuracies (%) of successful translations for 11 systems regarding all phenomena, organized in categories. Boldface indicates the best scoring system in each row, including all systems which are not significantly inferior than the best scoring system. Grey rows average the accuracies of the phenomena per category.