Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 936–947
Marseille, 20-25 June 2022

A Linguistically Motivated Test Suite to Semi-Automatically Evaluate German–English Machine Translation Output

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, Hans Uszkoreit
German Research Center for Artificial Intelligence (DFKI), SLT Lab, Berlin, Germany
{vivien.macketanz, eleftherios.avramidis, aljoscha.burchardt, he.wang, shushen.manakhimova, sebastian.moeller, hans.uszkoreit}@dfki.de, renlong.ai@giance.ai, uli.strohriegel@gmail.com

Abstract

This paper presents a fine-grained test suite for the language pair German–English. The test suite is based on a number of linguistically motivated categories and phenomena and the semi-automatic evaluation is carried out with regular expressions. We describe the creation and implementation of the test suite in detail, providing a full list of all categories and phenomena. Furthermore, we present various exemplary applications of our test suite that have been implemented in the past years, like contributions to the Conference of Machine Translation, the usage of the test suite and MT outputs for quality estimation, and the expansion of the test suite to the language pair Portuguese–English. We describe how we tracked the development of the performance of various systems MT systems over the years with the help of the test suite and which categories and phenomena are prone to resulting in MT errors. For the first time, we also make a large part of our test suite publicly available to the research community.

Keywords: Test Suite, Linguistic Evaluation, Machine Translation, German, English

1. Introduction

For the longest time, the evaluation of Machine Translation (MT) has mostly concentrated on automatic metrics. However, with the rise of deep learning and Neural MT (NMT), translation outputs have become significantly better and more fluent, resulting in a need for more fine-grained evaluation techniques. Fine-grained evaluation may indicate comparative strengths and weaknesses, while aggregated scores often fail to distinguish between different well-performing systems. One method that had already been used since the beginning of MT in the 1990's are test suites, also called challenge sets (King and Falkedal, 1990; Way, 1991; Heid and Hildenbrand, 1991). While test suites passed out of mind over time, they had a recent re-emerge when the need for fine-grained evaluation arose. A test suite is a hand-designed challenge set which can be used to test the performance of NLP tasks, e.g. MT outputs, with regard to specific aspects (Müller et al., 2018; Bawden et al., 2018).

In this paper, we will present a large-scale, fine-grained test suite for German to English and English to German MT outputs, "MT-TestSuite", along with various applications examples. For the first time, a significant part of our test suite is made publicly available in GitHub[1] which can be useful for further research by the community. For the time being, we have decided to publish not the whole test set but 50% of it (i.e., 50% of test items of every phenomenon, including evaluation rules) in order to keep a number of test items a secret to be able to use them as a test set in case MT systems are trained on our test items.

The paper is structured as follows: In Section 2, we describe the related work in the field of MT evaluation. In Section 3, we provide a detailed description of our test suite, together with the evaluation process. In Section 4, we present several applications examples of our test suite. Finally, we conclude the paper in Section 5 with an outlook on future work.

2. Related Work

The most commonly used automatic metrics are BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005). While these metrics are fast, low-cost, and reproducible, they do not provide any further information about the nature of the translations errors. Even though it is common knowledge in the community that those metrics are not adequate for MT evaluation, they are still widely used for evaluatios and comparisons.

Quality estimation (QE) methods (Blatz et al., 2004; Specia et al., 2009; Avramidis et al., 2018a) are used to predict the quality of a translation without needing access to the reference translation(s).

Furthermore, some automatic metrics provide a more detailed analysis, e.g., HTER (Snover et al., 2009) which automatically measures the translation edit rate, and Hjerson (Popović, 2011) which provides an error classification. There also exist methods for manual quality analysis, e.g., MQM (Lommel et al., 2014). Further manual analysis exists in the form of human rankings which are conducted to compare the quality of a number of MT systems (Callison-Burch et al., 2007; Bojar et al., 2015). While these subjective methods are more reliable, they are also more costly and therefore limited in their use.

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

MT challenge sets and test suites have regained importance in the recent years (Guillou and Hardmeier, 2016; Isabelle et al., 2017; Burchardt et al., 2017). The main advantage of test suites is that they can provide a detailed insight into the nature of MT errors. Therefore, test suites have been part of a shared task of the yearly Conference of Machine Translation (WMT)[2] since 2018, presenting numerous test suites which focus on various linguistic aspects and language pairs (Bojar et al., 2018a; Barrault et al., 2019; Barrault et al., 2020; Akhbardeh et al., 2021). The aspects covered since then are conjunctions (Popović, 2019), grammatical contrasts (Cinkova and Bojar, 2018), discourse (Bojar et al., 2018b; Rysová et al., 2019), domain-specific translations (Vojtěchová et al., 2019), gender coreference (Kocmi et al., 2020), markables (Zouhar et al., 2020), morphology (Burlot et al., 2018), pronouns (Guillou et al., 2018), and word sense disambiguation (Rios et al., 2018; Raganato et al., 2019; Scherrer et al., 2020). While the majority of the mentioned test suites focus on a single aspect, our test suite is, to the best of our knowledge, the only test suite that performs a systematic, fine-grained evaluation of more than one hundred phenomena for two language directions.

This paper summarizes and extends experimental and development work that has been ongoing for five years and presented partially within previous reports (Macketanz et al., 2018a; Macketanz et al., 2018b; Avramidis et al., 2019; Avramidis et al., 2020; Macketanz et al., 2021).

3. The Test Suite

The test suite comprises a large test set for evaluating both German to English and English to German MT outputs[3]. It comprises around 5,000 test items per language direction. A test item always contains one sentence. The test items are categorized in 13, respectively 14 linguistic categories (depending on the language direction). The categories are in turn divided into more than 100 fine-grained phenomena. Each phenomenon is represented by at least 20 test items. For each test item, we have created a set of regular expressions to semi-automatically evaluate the correctness of MT outputs.

The classification of the test items into the categories and phenomena allows for a rather basic or more granular analysis, depending on the user's need. The classification is language-specific. For the language pair German–English, there is a large overlap in the classification between the two language directions, however, a number of categories/phenomena do only exist in either one of the language directions and few phenomena are classified in different categories.

The aim of the fine-grained phenomena is to cover as many relevant linguistic aspects as possible, meaning, aspects that might lead to translation errors. Note that we are not following any linguistic theory and define the term *linguistically-motivated phenomenon* in a broad way, covering syntactical and morphological features as well as punctuation and norms. Note also that the number of test items for the phenomena does not mirror their distribution in corpus statistics or real-world scenarios.

An overview of all categories and their corresponding phenomena per language direction can be found in Tables 3 and 4 in the Appendix.

3.1. Test Suite Creation

The creation of the test suite was a long-standing process that involved experts from different fields such as linguistics, computational linguistics, translation studies, and computer science. Figure 1 depicts the preparation process of test items.

(a) Produce paradigms: An expert with experience in German–English MT selects a number of source items that might trigger a translation error. The source items can be created from scratch, be based on known MT errors triggers, or be inspired by previous work on test suites, e.g. (Lehmann et al., 1996). It is crucial that each test item focuses on one linguistic phenomenon at a time in order to avoid noise. The items are run by at least one more expert to ensure quality control. In our example, the phenomenon under inspection is *false friends*, represented by the German word *Novellen*.

(b) Fetch sample translations: The source items are being translated by several easily available MT systems to generate various different translations.

(c) Write regular expressions: Based on the MT outputs, the expert writes regular expressions that check whether the output items are translated correctly or incorrectly with regard to the phenomenon under inspection. The regular expressions are referred to as positive or negative regular expressions. Translation errors that cannot be linked to the phenomenon are ignored. The regular expressions in our example represent a correct and an incorrect translation of the false friend *Novellen*: A correct translation would be *novellas*, whereas an incorrect translation would be *novels*.

(d) Fetch more translations: Once the test suite comprises a number of various phenomena, represented by several test items each, as well as the corresponding regular expressions, the test items can be given to more MT systems. This can be done in the scope of collaborative projects or shared tasks. That way, more varying outputs are created.

(e) Apply regex: The existing regular expressions are applied to the new translations. Based on the regular expressions, the outputs are being assigned a translation status that can be either correct, incorrect, or – if neither a positive nor a negative regular expression can be applied – to be determined. In our example, the third

[2] http://www.statmt.org/wmt22/

[3] Every following description of the test suite is a description of the complete set. The numbers of the publicly available subset are thus smaller, cf. Section 1.

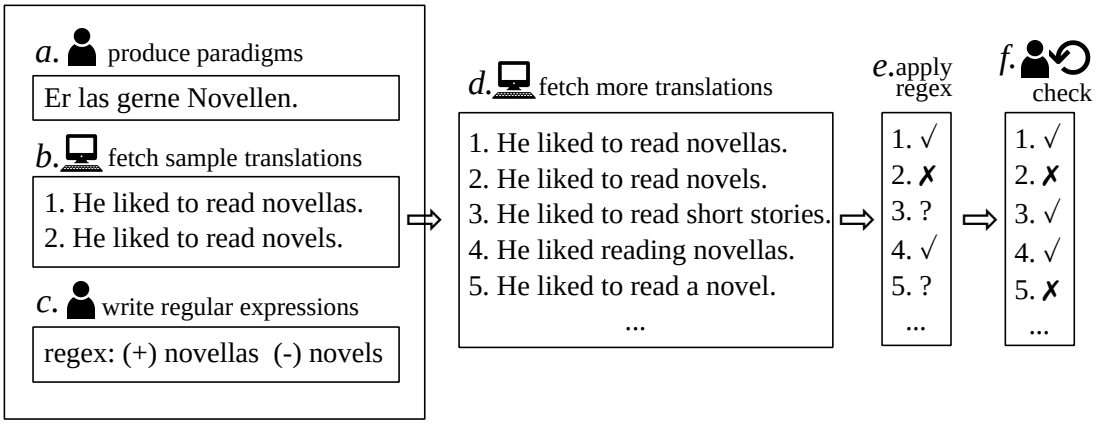

Figure 1: Example of the preparation of the test suite for one test item.

translation, containing the term *short stories*, cannot be evaluated by the regular expressions.

(f) Check: In the last step, a human expert annotator manually checks all the translations that could not be evaluated. In order to cover those outputs, the regular expressions are adapted accordingly. This way, the regular expression database grows over time, covering more and more MT outputs, and simultaneously reducing the manual effort needed to check translations that are to be determined. In our example, the positive regular expression is being adapted to additionally include the correct translation *short stories*.

Based on the evaluation conducted with the regular expressions, the phenomenon-specific translation accuracy can be computed. The accuracy is calculated by dividing the number of correctly translated test items of a phenomenon by the total number of test items of that phenomenon:

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

When comparing systems, the statistical significance of the comparison must be considered, since the test suites may contain relatively few items per category. To indicate the best system per category, one can compare the system with the highest accuracy with every other system using a one-tailed Student's t-test (Student, 1908). The systems whose difference is not significant are included in the same cluster as the first system.

3.2. Test Suite Application

Once the test suite has been developed, it can be applied in several evaluation tasks. We have created a tool for the evaluation that semi-automatically checks the MT outputs based on the regular expressions. The tool is called TQ-AutoTest (Macketanz et al., 2018a). Currently, the tool is only available for internal use, however, we are planning to release a publicly available version in the future. The evaluation process of the test suite typically comprises the following key steps:

(a) Preparation: A subset of test items or all test items of a language pair are selected, based on the categories or the phenomena. Next, the test items are mixed with distractor items in order to keep the test items under wraps.

(b) Translation: The test set, consisting of the selected test items scrambled with the distractor items, is translated by one or more MT system(s). Afterwards, the test and distractor items can be unscrambled by the tool.

(c) Evaluation: The MT outputs are semi-automatically evaluated with the help of the tool. The evaluation is based on a database of rules. The rules consist of the positive and negative regular expressions as well as positive and negative tokens. The tokens are fixed strings of fully translated test items taken from previous reports. Any unevaluated outputs are checked manually.

(d) Analysis and comparison: A statistical analysis of the MT outputs is conducted for every MT system. The translation accuracy is calculated, either on the category-level, the phenomenon-level, or on average. If multiple systems have translated the same set of test items, a system comparison can be performed. In a system comparison, the translation accuracy is calculated on the category-level, the phenomenon-level, or on average across systems, allowing to identify the best performing system(s) on the various levels. For the calculation, all systems are compared to the system with the highest accuracy. The significance of the comparison is confirmed with a one-tailed Z-test with $\alpha = 0.95$. If there are systems of which the performance does not significantly differ from the performance of the system with the highest accuracy, they are considered to be in the best performance cluster together.

3.3. Test Suite Limitations

While test suites have many advantages, like providing insights to systems' performances on different phenomena, they do also have limitations. One limitation of our test suite is that the test sentences are specifically created for the linguistic phenomena and therefore not representative of real-world MT tasks. However, there exist many various real-world MT tasks which is why

it would be impossible to cover all of them. One solution would be to create task-specific test suites in real world applications (as we have done in the past, cf. Section 4.1).
Another limitation of our test suite is that the test sentences are rather short. The longer a sentence is, the more linguistic aspects come into play which makes it more difficult to find out which aspect could be the source of a translation error. But that also means that some linguistic aspects might be more erroneous in the context of longer sentences. However, as MT is still constantly improving, we are planning to make our test sentences more complex in the future to increase their difficulty.
Lastly, we would like to point out that the translation accuracy which we calculate based on the test suite analysis can only provide indications of systems' strengths and weaknesses based on the test suite data and is not representative for real-world scenarios.

4. Application Examples of the Test Suite

Since the creation of the test suite and the implementation of the evaluation process, we have utilized the test suite in a number of different contexts. In the following sections, we will present the most relevant applications of the previous years.

4.1. Domain-specific Test Suite

Some years ago, we have used the test suite approach to compare two different types of MT from the perspective of a language service provider (LSP) (Beyer et al., 2017). We compared a state-of-the-art Neural MT (NMT) system for the general domain with a domain-trained and optimized Phrase Based MT (PBMT) system to find out whether it is already time for the LSP to invest in the – at that time – new technology of NMT.
For the system comparison, we created a small-scale domain-specific test suite to test the customer data provided by the LSP. The data came from translations of catalogues for technical tools and the test suite categories were adapted accordingly. While a simple BLEU score comparison of the two systems suggested that the PBMT system outperformed the NMT system, the test suite evaluation revealed that the NMT system performed reasonably well on various categories, despite not being trained on the specific customer data.
While the question whether LSPs should switch from PBMT to NMT is not relevant anymore nowadays, the performance of a domain-adapted MT system compared to a general MT system on domain-specific data is still of notable interest today. The insights gained from the adapted test suite approach can not be achieved by a simple BLEU comparison

4.2. WMT Shared Task: Test Suite Track

We have contributed our German – English test suite to the test suite track of the WMT in the past years. We first started doing so for the WMT18 with an initial version of the test suite and have proceeded participating every year since (Macketanz et al., 2018b; Avramidis et al., 2019; Avramidis et al., 2020; Macketanz et al., 2021). In 2021, for the first time, we provided our test suite for both language directions German to English and English to German.
For the analysis, we obtained translations from the systems that were part of the News Translation Task of WMT. In 2018 and 2019, we received outputs from 16 different systems, in 2020 from 11 systems, and in 2021 from 18 MT systems per language direction. After receiving the various outputs, a considerable amount of time had to be spent for the manual evaluation of translations that were not covered by the existing rules. For German to English, in 2018 10% to 45% of test items had to be checked manually (depending on the system); in the years following, the percentage was usually 10%, and 2021 the percentage decreased to around 6%. For English to German, the initial situation in 2021 was different, with almost no rules existing yet; therefore, around 85 % of test items had to be checked manually. Depending on the number of systems participating and the percentage of sentences having to be checked manually, the time spent ranged between 45 to 80 hours. After the manual analysis, there was usually still a small number of items left non-evaluated due to time constraints. For the comprehensive system comparison, which we provided every year, only the test items that did not contain any non-evaluated outputs for any of the MT systems were included in the calculation in order to achieve a fair comparison.

4.2.1. WMT18 Results

When we first submitted our German to English test suite to the WMT in 2018 (Macketanz et al., 2018b), the systems were performing distinctively less good than they do nowadays. The MT systems achieved 73.1% accuracy on average for all test items, and the best average was achieved by the system UCAM with 86.0%.
The categories that achieved the lowest performance by the systems were *punctuation*, *multiword expressions*, *ambiguity*, and *false friends*, with an average accuracy of less than 64%. These categories were followed by *verb tense/aspect/mood*, *non-verbal agreement*, *function words*, and *coordination & ellipsis*, with an average accuracy of around 75%. The categories with the highest accuracy were *subordination*, *negation*, and *composition*.
On the phenomenon-level, the accuracies for the phenomena *compounds* and *location* were quite high, while the phenomenon *quotation marks* exhibited a large range of accuracy, extending from 0% to almost 95%. Furthermore, the verb tenses *future II* and f*uture II subjunctive* displayed the lowest accuracies with a maximum of about 30%.

4.2.2. WMT19 Results

For the WMT19 (Avramidis et al., 2019), the MT systems achieved 75.6% accuracy on average for all Ger-

man to English test items, an improvement of 2.5% as compared to the previous year. The best average was achieved by the systems RWTH and online-A with 83.6% and 82.8%, respectively.

The systems' lowest-performance categories with around 66% accuracy were *multiword expressions* (just like the previous year) and *verb valency*. The categories with the highest performances, i.e., an accuracy of more than 80%, were *subordination*, *negation*, *composition* (as last year), as well as *function words* and *non-verbal agreement*.

On the phenomenon-level, the phenomena that reached the lowest accuracies (max. 30%) were *idiom*, *resultative predicate*, *modal pluperfect*, and *negated modal pluperfect*. However, there was also a number of phenomena with a high accuracy of more than 90%, among them the phenomena *transitive*, *intransitive* and *ditransitive verbs* in the *perfect tense* and *future tenses*, as well as *passive voice*, *polar question*, *infinitive clause*, *conditional*, *focus particle*, *location*, and *phrasal verb*.

4.2.3. WMT20 Results

For the WMT20 (Avramidis et al., 2020), the average accuracy of all systems for the German to English test items was 74.7% and the best average was achieved by the systems VolcTrans and Tohuku with 85.4% and 85.3%, respectively.

The categories with the lowest performance (around 70%, thus, higher than the previous years), were *multiword expressions*, *verb valency*, *ambiguity*, and *false friends*. The highest accuracies were achieved on the categories *negation*, *composition*, *subordination*, *named entities*, and *terminology* (around 82%-97%).

On the phenomenon-level, the most problematic phenomena were *idiom* and *resultative predicate*, with accuracies below 30%. The phenomena with the highest accuracy (above 90%) were *focus particle*, *verbal MWE*, *date*, *measuring unit*, *negation*, *internal possessor*, *comma*, *infinitive clause*, *object clause*, *conditional*, and *passive voice*, as well as many verb tenses for *intransive*, *transitive*, and *ditransitive verbs*.

4.2.4. WMT21 Results

For the WMT21 (Macketanz et al., 2021), the average accuracy of all systems for the German to English test items was 84.0%, thus improving again compared to all previous years. The best average was achieved by five systems: Online-W with 88.3%, Facebook with 88.2%, uedin with 87.4%, Online-A with 87.3%, and borderline with 87.1%.

The categories that revealed the lowest accuracies (below 86%, thus, distinctly improved) were *false friends*, *ambiguity*, *verb tense/aspect/mood*, and *multiword expressions*. The categories with the highest performance (above 90%) were *composition*, *subordination*, and *named entities & terminology*. *Negation* and *punctuation* even reached an accuracy of 100%.

On the phenomenon-level, the phenomena with the lowest accuracies remained *modal pluperfect*, *modal pluperfect negated*, *resultative predicate*, and *idioms*. Many phenomena reached an average of more than 90%, and there were some for which all systems achieved an accuracy of 100%, such as *negation*, *internal possessor*, *comma*, *ditransitive perfect*, and *intransitive future I*.

For the language direction English to German, the average accuracy of all systems was 94.8%, which is considerably higher than for the other language direction. The best average was achieved by three systems: Online-B and VolcTransGLAT with 96.9% each and FacebookAI with 97.4%.

The lowest accuracies on the category-level (below 85%) were reached for the categories *coordination & ellipsis*, *verb valency*, and *ambiguity*. The categories with the highest performance (above 95%) were *function words*, *negation*, *subordination*, and *verb tense/aspect/mood*.

On the phenomenon-level, the lowest accuracy by far was achieved by *idioms* with 14.6%, followed by *middle voice*, *pseudogapping*, and *stripping* (all below 65% accuracy). However, many phenomena reached (nearly) 100% of accuracy, like *internal possessor*, *comma*, *indirect speech*, *infinitive clause*, *object clause*, *subject clause*, *passive voice*, and *ditransitive*, *intransitive* and *transitive verbs* in many tenses.

4.2.5. WMT System Developments 2018 to 2021

Figure 2 shows the system improvements from the years 2018 to 2021 for the average performance on the German to English test suite. Only systems that participated in multiple years are included in the graph. While the Facebook AI MT system showed a small improvement from 2019 to 2021, other systems like Online-G showed a distinct improvement over the years, from an average accuracy of less than 70% in 2018 to almost 90% in 2021. In 2021, all systems are achieving an average accuracy of around 90%. The growing accuracies for the categories and phenomena raise the question whether some of the test items might have become too simple, as their original creation dates a few years back when MT was still distinctly inferior. Therefore, it might be required to increase the difficulty of the respective test items in the future.

The improvement of particular categories through the years can be seen in Table 1. The criterion for choosing the best system every year is the systems' macro-average performance on the test suite. Despite the overall improvement of the average scores, when it comes to half of the categories, some minor improvements or drops between 2019 and 2021 are not significant. A number of categories exhibit a high improvement of around 10 percentage points or more from 2018 to 2021: *false friends*, *function words*, *multiword expressions*, *non-verbal agreement*, and *verb tense/aspect/mood*. The accuracy of *punctuation* even shows an improvement of almost 20 percentage points only from 2018 to 2019, rising to more than 20 percentage points in 2021. At the same time, there are many

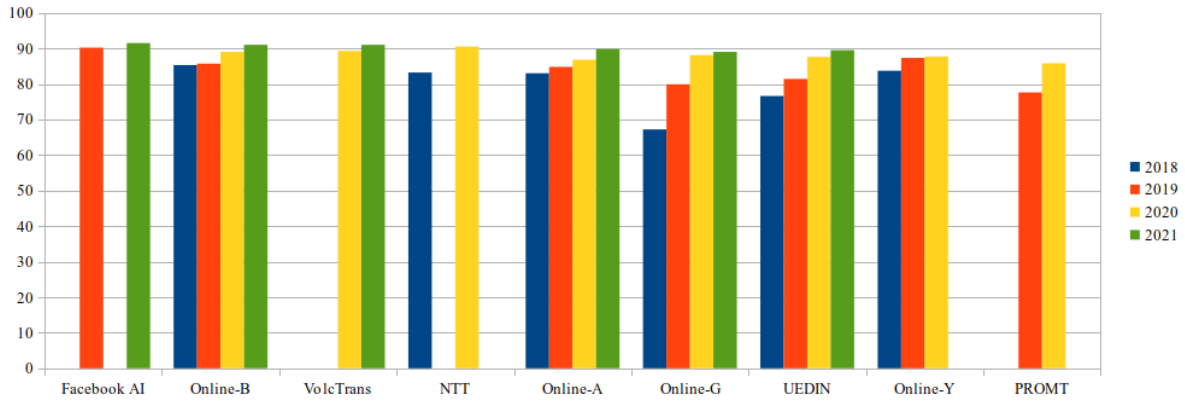

Figure 2: System improvements (accuracy macro-average) on the test suite from 2018 to 2021 for German to English, including the systems that appear at least once in the past two years.

category	count	WMT 2018	WMT 2019	WMT 2020	WMT 2021
Ambiguity	77	75.3	**92.2**	81.8	**84.4**
Composition	48	95.8	97.9	97.9	95.8
Coordination & ellipsis	74	86.5	90.5	89.2	91.9
False friends	36	75.0	75.0	72.2	86.1
Function word	65	78.5	**87.7**	**86.2**	**90.8**
LDD & interrogatives	154	81.2	**87.0**	**90.9**	**87.0**
Multiword expressions	74	74.3	**82.4**	**82.4**	**87.8**
NE & terminology	83	**89.2**	86.7	**95.2**	**95.2**
Negation	17	94.1	100.0	100.0	100.0
Non-verbal agreement	59	86.4	**91.5**	**91.5**	**96.6**
Punctuation	57	75.4	**94.7**	**98.2**	**98.2**
Subordination	160	85.0	**91.3**	**91.9**	**88.1**
Verb tense/aspect/mood	4070	74.7	80.2	**85.8**	**85.0**
Verb valency	77	75.3	81.8	81.8	80.5
micro-average	5051	76.1	81.9	**86.5**	**85.9**
macro-average	5051	81.9	88.5	88.9	**90.5**

Table 1: Comparison of the best German to English systems per year between 2018 and 2021 regarding the accuracy of every category (online-B, Facebook-AI, Tohoku, Online-W, respectively, selected based on their category macro-average). Boldface indicates the significantly best accuracies per row after a one-tailed t-test.

categories which first experience an improvement but then later a drop in their accuracy, e.g. *ambiguity*, *function words*, *long distance dependency & interrogatives*, *named entity & terminology*, *subordination*, and *verb valency*.

Table 2 presents exemplary MT outputs from the uedin systems that were submitted to the WMT shared task from 2018 to 2021 (Haddow et al., 2018; Bawden et al., 2019; Germann et al., 2020; Chen et al., 2021). We have extracted examples from the categories and phenomena that exhibited the lowest accuracies throughout the years.

The first phenomenon depicted is *collocation*. A collocation is a *multiword expresion* (MWE) that can be described as a co-occurence of two or more words which is higher than it would be expected by chance. Like all MWEs, a collocation needs to be translated as a whole, as the translation of its separate elements usually leads to a translation error. The German collocation *lieblicher Wein* can only be translated to the corresponding English collocation *sweet wine*, any other translation of *lieblich*, for example to *adorable* or *lovely*, leads to a translation error.

The second example is taken from the category *false friends*. While one might expect that false friends are only prone to lead to translation errors in human translation processes, our test suite shows that they do indeed lead to MT errors as well. The false friend example contains the German word *Novelle* which could easily mixed up with the English word *novel*. However, the German *Novelle* can only be translated as *novella* or *short story*; a translation into *novel* is incorrect as this means *Roman* in German.

The third phenomenon presented is a *negated*

	Collocation (MWE)	False friends (false friends)	Modal negated – preterite (verb tense/aspect/mood)
Test item	Simon trinkt am liebsten *lieblichen Wein*.	Dieser Autor schreibt hauptsächlich *Novellen*.	Ihr *durftet nicht lesen*.
WMT18 output	Simon prefers to drink *adorable wine*.	This author writes mainly *Novellen*.	You *are not allowed to read*.
WMT19 output	Simon loves to drink *lovely wine*.	**This author writes mainly *novellas*.**	You *are not allowed to read.*
WMT20 output	Simon prefers to drink *lovely wine*.	**This author writes mainly *novellas*.**	You *don't read.*
WMT21 output	**Simon prefers to drink *sweet wine*.**	This author mainly writes *novels*.	**You *were not allowed to read*.**

Table 2: Output examples from the uedin system from 2018 to 2021. Italic marks the part of the test item that is related to the phenomenon and boldface marks correct outputs.

modal – preterite, taken from the category *verb tense/aspect/mood*. For the translation of modal verbs it is crucial that the modal verb itself is translated correctly, as well as its tense. While the ouputs in 2018 and 2019 contain a correct translation of the modal *dürfen* as *to not be allowed to*, the tense preterite is translated as present. The output of 2020 contains an incorrect translation of the modal itself as well as the tense.

4.3. WMT Shared Task: Metrics

For the WMT21, we provided a test set, taken from our test suite, to the Metrics Shared Task (Freitag et al., 2021). The allocated test set contained a number of German to English test items with several corresponding correct and incorrect translations. The correct and incorrect outputs were taken from the existing set of test suite rules, i.e., the positive and negative tokens. This set was then used to evaluate the robustness of automatic metrics for MT outputs by turning it into a challenge set. This was done by splitting the outputs into two sets: One set with correct translations, serving as a correct reference for the metrics, and another set consisting of pairs of one correct and one incorrect translation. Ideally, the automatic metrics should give a higher score to the correct output. One of the main results was a distinct difference between metrics that did or did not use a reference.

This application of the test suite is interesting insofar as this was the first time that not only the source sentences were utilized but also the database of evaluation rules.

4.4. Quality Estimation

In Avramidis et al. (2018b) we presented an alternative method of evaluating systems for the Quality Estimation (QE) of MT, based on the linguistically-motivated test suite. Based on the 14 linguistic categories of the German to English test suite we created a test set, containing a set of translation outputs for each category, including both correct and erroneous MT outputs. Then, we measured the performance of five comparative QE systems by checking their ability to distinguish between the correct and the erroneous translations. The detailed results are much more informative about the ability of each system than the overall scores. Three of the five QE systems (logistic regression with 10 features, gradient boosting with 10 features, and gradient boosting with 139 features) have mostly complementary performance, by winning about five different categories each. Indicatively, the aforementioned logistic regression systems is the best one at *coordination & ellipsis*, *function words*, *multiword expressions*, and *nonverbal agreement*, whereas the gradient boosting system with 137 features is the best one at *long distance dependencies & interrogatives*, *named entities & terminology*, *negation*, *subordination*, and *verb valency*. The fact that different QE systems perform differently at various phenomena confirms the usefulness of the test suite.

4.5. Expansion to Portuguese to English

Until recently, the test suite only existed for the language pair German–English. However, in the future, we would like to expand it to other language pairs. The first step towards that direction was recently taken with a small-scale test suite for the language direction Portuguese to English (Avelino et al., 2022). It is the first challenge set of its kind for this language pair. The linguistically motivated categories and phenomena are inspired by our original test suite. Some of the categories/phenomena overlap with our German–English test suite, but since the categories and phenomena are always language pair-specific, many of them are differing. The Portuguese to English test suite contains 330 test items, organized in 14 categories and 66 phenomena; each phenomenon is represented by 5 test items. The test suite has been applied for the evaluation of the performance of eight MT systems. An interesting finding was that the categories with the lowest accura-

cies for Portuguese to English are similar to the ones for German–English: *Ambiguity*, *named entity & terminology*, and *verb valency* were the most problematic categories. On the phenomenon-level, all evaluated systems struggled with *direct object omission & polar questions*, a phenomenon which is very specific to this language pair. The categories with high accuracy on average were *negation*, *pronouns*, *subordination*, *verb tense/aspect/mood*, and *false friends*, some of them also being similar to the categories that displayed high accuracies for the WMT evaluations. In the future, we plan to expand the Portuguese to English test suite to the other language direction.

5. Conclusion and Future Work

We have presented our large-scale German–English test suite for the linguistically inspired evaluation of MT outputs. We have described our evaluation process in detail and further provided several application examples of our test suite: (a) The test suite track of the WMT, including an analysis of system developments in the past four years, (b) the application of a hand-tailored subset of our test suite to the evaluation of the robustness of automatic metrics, (c) MT Quality Estimation based on the test suite, and (d) an expansion of the German–English test set to the new language pair Portuguese to English.

With our test suite growing larger over time and being applied to various evaluation scenarios, we see a lot of potential for future work with the test suite:

The diagnostics from the fine-grained evaluation could in principle allow MT system developers to address issues focusing on particular phenomena, e.g., by fixing punctuation or adding targeted corpora (e.g. for language-specific idioms). Nevertheless, we are not aware of any work that has done that so far, and further research is required to determine how this could be accomplished, particularly for the more complex grammatical phenomena.

We are currently working on an automatic test item creator tool that would expand our data set further and help make test items more versatile and potentially increase their difficulty.

Furthermore, we are developing a new version of our evaluation tool, which we are planning to make publicly available. That way, external users could make use of our test suite within in the tool, which would help our rules database grow further.

With the Portuguese – English test suite, we have taken the first step to expand our test suite to other language pairs. We would like to pursue this expansion and include more language pairs over time. However, this involves a distinct amount of manual work for the research of the categories and phenomena as well as the creation of new test items, and can only be done by an expert in the source and target language. Nevertheless, we believe that the manual effort is worth the many application possibilities of a test suite.

Acknowledgements

We would particularly like to thank Silvia Hansen-Schirra and her team at the Johannes Gutenberg University Mainz for their valuable contributions to the initial creation of test suite test items.

This research was supported by the Deutsche Forschungsgemeinschaft (DFG) through the project TextQ, and by the German Federal Ministry of Education and Research (BMBF) through the project SocialWear. Earlier steps were funded by the EU-funded projects QT-Launchpad, QT21 and QT-Leap.

Appendix

Category	Phenomenon
Ambiguity	Lexical Ambiguity, Structural Ambiguity
Composition	Compound, Phrasal Verb
Coordination & Ellipsis	Gapping, Right Node Raising, Sluicing
False Friends	False Friends
Function Words	Focus Particle, Modal Particle, Question Tag
Long Distance Dependency & Interrogatives	Extended Adjective Construction, Extraposition, Multiple Connectors, Pied-piping, Polar question, Scrambling, Topicalization, Wh-Movement
Multiword Expressions	Collocation, Idiom, Prepositional MWE, Verbal MWE
Named Entity & Terminology	Date, Domain-specific Term, Location, Measuring Unit, Proper Name
Negation	Negation
Non-verbal Agreement	Coreference, External Possessor, Internal Possessor
Punctuation	Comma, Quotation marks
Subordination	Adverbial Clause, Cleft Sentence, Free Relative Clause, Indirect Speech, Infinitive Clause, Object Clause, Pseudo-cleft Sentence, Relative Clause, Subject Clause
Verb Tense/Aspect/Mood	Conditional, Ditransitive, Imperative, Intransitive, Modal, Modal Negated, Progressive, Reflexive, Transitive
Verb Valency	Case Government, Mediopassive Voice, Passive Voice, Resultative Predicate

Table 3: German to English test suite classification.

Category	Phenomenon
Ambiguity	Lexical Ambiguity
Coordination & Ellipsis	Gapping, Pseudogapping, Right Node Raising, Sluicing, Stripping, VP-ellipsis
False Friends	False Friends
Function Words	Focus Particle, Question tag
Long Distance Dependency & Interrogatives	Extraposition, Inversion, Multiple Connectors, Negative Inversion, Pied-piping, Polar Question, Preposition Stranding, Split Infinitive, Topicalization, Wh-Movement
Multiword Expressions	Collocation, Compound, Idiom, Nominal MWE, Prepositional MWE, Verbal MWE
Named Entity & Terminology	Date, Domain-specific Term, Location, Measuring Unit, Proper Name
Negation	Negation
Non-verbal Agreement	Coreference, Genitive, Possession
Punctuation	Quotation Marks
Subordination	Adverbial Clause, Cleft Sentence, Contact Clause, Indirect Speech, Infinitive Clause, Object Clause, Pseudo-cleft Sentence, Relative Clause, Subject Clause
Verb Tense/Aspect/Mood	Conditional, Ditransitive, Gerund, Imperative, Intransitive, Modal, Modal Negated, Reflexive, Transitive
Verb Valency	Case Government, Catenative Verb, Middle Voice, Passive Voice, Resultative Predicate

Table 4: English to German test suite classification.

6. Bibliographical References

Akhbardeh, F., Arkhangorodsky, A., Biesialska, M., Bojar, O., Chatterjee, R., Chaudhary, V., Costa-jussa, M. R., España-Bonet, C., Fan, A., Federmann, C., Freitag, M., Graham, Y., Grundkiewicz, R., Haddow, B., Harter, L., Heafield, K., Homan, C., Huck, M., Amponsah-Kaakyire, K., Kasai, J., Khashabi, D., Knight, K., Kocmi, T., Koehn, P., Lourie, N., Monz, C., Morishita, M., Nagata, M., Nagesh, A., Nakazawa, T., Negri, M., Pal, S., Tapo, A. A., Turchi, M., Vydrin, V., and Zampieri, M. (2021). Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online, November. Association for Computational Linguistics.

Avelino, M., Macketanz, V., Avramidis, E., and Möller, S., (2022). *A Test Suite for the Evaluation of Portuguese-English Machine Translation*, pages 15–25. 03.

Avramidis, E., Macketanz, V., Lommel, A., and Uszkoreit, H. (2018a). 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, March. Association for Machine Translation in the Americas.

Avramidis, E., Macketanz, V., Lommel, A., and Uszkoreit, H. (2018b). 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.

Avramidis, E., Macketanz, V., Strohriegel, U., and Uszkoreit, H. (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), August 1-2, Florence, Italy*. Association for Computational Linguistics, 8.

Avramidis, E., Macketanz, V., Strohriegel, U., Burchardt, A., and Möller, S. (2020). Fine-grained linguistic evaluation for state-of-the-art machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 344–354, Online, November. Association for Computational Linguistics.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019). Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August. Association for Computational Linguistics.

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 Conference on Machine Translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–54, Online, November. Association for Computational Linguistics.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (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, June. Association for Computational Linguistics.

Bawden, R., Bogoychev, N., Germann, U., Grundkiewicz, R., Kirefu, F., Miceli Barone, A. V., and Birch, A. (2019). The University of Edinburgh's Submissions to the WMT19 News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 103–115, Florence, Italy, August. Association for Computational Linguistics.

Beyer, A., Macketanz, V., Burchardt, A., and Williams, P. (2017). Can out-of-the-box nmt beat a domain-trained moses on technical data.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). Confidence Estimation for Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, page 315–es, USA. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., et al. (2015). Findings of the 2015 Workshop on Statistical Machine Translation.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., and Monz, C. (2018a). Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels, October. Association for Computational Linguistics.

Bojar, O., Mírovský, J., Rysová, K., and Rysová, M. (2018b). EvalD Reference-Less Discourse Evaluation for WMT18. In *Proceedings of the Third Conference on Machine Translation*, pages 545–549,

Belgium, Brussels, oct. Association for Computational Linguistics.

Burchardt, A., Macketanz, V., Dehdari, J., Heigold, G., Peter, J.-T., and Williams, P. (2017). A linguistic evaluation of rule-based, phrase-based, and neural MT engines. *The Prague Bulletin of Mathematical Linguistics*, 108(1):159–170.

Burlot, F., Scherrer, Y., Ravishankar, V., Bojar, O., Grönroos, S.-A., Koponen, M., Nieminen, T., and Yvon, F. (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, oct. Association for Computational Linguistics.

Callison-Burch, C., Fordyce, C. S., Koehn, P., Monz, C., and Schroeder, J. (2007). (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158.

Chen, P., Helcl, J., Germann, U., Burchell, L., Bogoychev, N., Miceli Barone, A. V., Waldendorf, J., Birch, A., and Heafield, K. (2021). The University of Edinburgh's English-German and English-Hausa submissions to the WMT21 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 104–109, Online, November. Association for Computational Linguistics.

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

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Foster, G., Lavie, A., and Bojar, O. (2021). Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online, November. Association for Computational Linguistics.

Germann, U., Grundkiewicz, R., Popel, M., Dobreva, R., Bogoychev, N., and Heafield, K. (2020). Speed-optimized, Compact Student Models that Distill Knowledge from a Larger Teacher Model: the UEDIN-CUNI Submission to the WMT 2020 News Translation Task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 190–195, Online, November. Association for Computational Linguistics.

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

Guillou, L., Hardmeier, C., Lapshinova-Koltunski, E., and Loáiciga, S. (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, oct. Association for Computational Linguistics.

Haddow, B., Bogoychev, N., Emelin, D., Germann, U., Grundkiewicz, R., Heafield, K., Miceli Barone, A. V., and Sennrich, R. (2018). The University of Edinburgh's Submissions to the WMT18 News Translation Task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 403–413, Belgium, Brussels, October. Association for Computational Linguistics.

Heid, U. and Hildenbrand, E. (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.

Isabelle, P., Cherry, C., and Foster, G. (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, September. Association for Computational Linguistics.

King, M. and Falkedal, K. (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.

Kocmi, T., Limisiewicz, T., and Stanovsky, G. (2020). Gender Coreference and Bias Evaluation at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 355–362, Online, November. Association for Computational Linguistics.

Lehmann, S., Oepen, S., Regnier-Prost, S., Netter, K., Lux, V., Klein, J., Falkedal, K., Fouvry, F., Estival, D., Dauphin, E., Compagnion, H., Baur, J., Balkan, L., and Arnold, D. (1996). TSNLP - Test Suites for Natural Language Processing. In *Proceedings of the 16th International Conference on Computational Linguistics, COLING 1996, Copenhagen, Denmark*.

Lommel, A., Uszkoreit, H., and Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Macketanz, V., Ai, R., Burchardt, A., and Uszkoreit, H. (2018a). TQ-AutoTest–An Automated Test Suite for (Machine) Translation Quality. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Macketanz, V., Avramidis, E., Burchardt, A., and Uszkoreit, H. (2018b). Fine-grained evaluation of german-english machine translation based on a test suite. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 584–593, Belgium, Brussels, October. Association for Computational Linguistics.

Macketanz, V., Avramidis, E., Manakhimova, S., and Möller, S. (2021). Linguistic evaluation for the 2021 state-of-the-art machine translation systems for

German to English and English to German. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1059–1073, Online, November. Association for Computational Linguistics.

Müller, M., Rios, A., Voita, E., and Sennrich, R. (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, October. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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, July. Association for Computational Linguistics.

Popović, M. (2011). Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 96(-1):59–68.

Popović, M. (2019). Evaluating Conjunction Disambiguation on English-to-German and French-to-German WMT 2019 Translation Hypotheses. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 464–469, Florence, Italy, August. Association for Computational Linguistics.

Raganato, A., Scherrer, Y., and Tiedemann, J. (2019). The MuCoW Test Suite at WMT 2019: Automatically Harvested Multilingual Contrastive Word Sense Disambiguation Test Sets for Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy, August. Association for Computational Linguistics.

Rios, A., Müller, M., and Sennrich, R. (2018). The Word Sense Disambiguation Test Suite at WMT18. In *Proceedings of the Third Conference on Machine Translation*, pages 594–602, Belgium, Brussels, oct. Association for Computational Linguistics.

Rysová, K., Rysová, M., Musil, T., Poláková, L., and Bojar, O. (2019). A Test Suite and Manual Evaluation of Document-Level NMT at WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 455–463, Florence, Italy, August. Association for Computational Linguistics.

Scherrer, Y., Raganato, A., and Tiedemann, J. (2020). The MUCOW word sense disambiguation test suite at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 363–368, Online, November. Association for Computational Linguistics.

Snover, M., Madnani, N., Dorr, B., and Schwartz, R. (2009). Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, Athens, Greece, March. Association for Computational Linguistics.

Specia, L., Saunders, C., Turchi, M., Wang, Z., and Shawe-Taylor, J. (2009). Improving the confidence of machine translation quality estimates. 08.

Student. (1908). The probable error of a mean. *Biometrika*, pages 1–25.

Vojtěchová, T., Novák, M., Klouček, M., and Bojar, O. (2019). SAO WMT19 Test Suite: Machine Translation of Audit Reports. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 481–493, Florence, Italy, August. Association for Computational Linguistics.

Way, A. (1991). Developer-Oriented Evaluation of MT Systems. In Kirsten Falkedal, editor, *Proceedings of the Evaluators' Forum*, pages 237–244, Les Rasses, Vaud, Switzerland, apr. ISSCO.

Zouhar, V., Vojtěchová, T., and Bojar, O. (2020). WMT20 Document-Level Markable Error Exploration. In *Proceedings of the Fifth Conference on Machine Translation*, pages 369–378, Online, November. Association for Computational Linguistics.