WMT20 Document-Level Markable Error Exploration

Vilém Zouhar **Tereza Vojtěchová** **Ondřej Bojar**

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague, Czech Republic
{zouhar, vojtechova, bojar}@ufal.mff.cuni.cz

Abstract

Even though sentence-centric metrics are used widely in machine translation evaluation, document-level performance is at least equally important for professional usage. In this paper, we bring attention to detailed document-level evaluation focused on markables (expressions bearing most of the document meaning) and the negative impact of various markable error phenomena on the translation.

For an annotation experiment of two phases, we chose Czech and English documents translated by systems submitted to WMT20 News Translation Task. These documents are from the News, Audit and Lease domains. We show that the quality and also the kind of errors varies significantly among the domains. This systematic variance is in contrast to the automatic evaluation results.

We inspect which specific markables are problematic for MT systems and conclude with an analysis of the effect of markable error types on the MT performance measured by humans and automatic evaluation tools.

1 Introduction

This paper presents the results of our test suite for WMT20 News Translation Task.[1]

The conclusion of Vojtěchová et al. (2019), a last year's similar effort, states that expert knowledge is vital for correct and comprehensible translation of professional domains, such as Audits or Lease agreements. Furthermore, even MT systems which make fewer mistakes and score above others in both automatic and manual evaluations are prone to making fatal errors related to markable conflicts, which render the whole document translation unusable.

In this study, we aim to organize and describe a more detailed study with a higher number of annotators. We show three evaluation approaches: (1) automatic evaluation, (2) fluency and adequacy per document line and (3) detailed markable phenomena evaluation. We compare the results of this evaluation across the three domains and try to explain why all of these evaluations do not produce the same ordering of MT systems by performance.

This paper is organized accordingly: Section 1.1 defines the term "Markable", Section 1.2 describes the examined documents and Section 2 introduces the two phases of our annotation experiment and shows the annotator user interface in Section 2.3. In Section 3, we discuss the results from both phases and also automatic evaluation. The main results of this examination are shown in Section 3.5 and specific markable examples are discussed in Section 4. We conclude in Section 5.

1.1 Markable Definition

A markable in this context is an occurrence of any technical or non-technical term or expression that satisfies at least one of the following conditions:

1. The term was translated into two or more different ways *within one document*.
2. The term was translated into two or more different ways *across several translations*.
3. Two or more terms were translated to a specific expression *in one document* but have different meanings.

To be a markable, the term or expression does not have to be a named entity, but it must be vital to the understanding of the document. In the same order we show examples which satisfy the definition conditions.

1. *bytem* – It was translated within one document into *an apartment* and *a residence*.

[1] http://www.statmt.org/wmt20/translation-task.html

Document	Sentences	Direction	Markable occurrences	Description
Lease	29	cs→en en→cs	73 70	Housing lease agreement
Cars	18	cs→en	11	Brno Grand Prix competition article + highway accident report
Audit	90	cs→en en→cs	28 18	Supreme Audit Office audit report
Speech	13	en→cs	15	Greta Thunberg's U.N. speech article
Total	**269**	-	**215**	-

Table 1: Summary of examined documents with translation directions, number of lines and number of markable occurrences.

2. *rodné číslo* – It was translated in one translation to *social security number* and in another translation to *identification number*.
3. *nájemce*, *podnájemce* – They have different meanings and in one document they were both translated to tenant.

Markables were proposed first by the annotators in the first phase of annotation in Section 2.1 and then filtered manually by us.

1.2 Test Suite Composition

We selected 4 documents, 2 of which were translated in both directions totalling 6 documents. We chose 2 from the professional domain (Audit and Lease) and 2 from the News domain. The overview of their size is shown in Table 1. The number of markable occurrences is highly dependent on the document domain with the Agreement domain (Lease document) containing the most occurrences.

All of the MT systems are participants of the News Translation Task, and we test their performance even outside of this domain. Most of them were bi-directional, and we join the results from both directions when reporting their performance. The only exceptions are eTranslation (only en→cs) and PROMT_NMT (only cs→en).

1.3 Data and Tools Availability

All of the document translations and measured data are available in the project repository. Furthermore, the used online markable annotation tool written in TypeScript and Python is documented and also open-source.[2]

2 Annotation Setup

For both phases of this experiment, we used 10 native Czech annotators with English proficiency. None of them were professional audit or legal translators. Because each annotator annotated only one or two documents, the aggregated results across domains, labelled as *Total*, are of less significance than the results in individual domains.

2.1 Manual Document Evaluation

In this phase of the experiment, we wanted to measure the overall document translation quality and also to collect additional markables for use in the following experiment part. We showed the annotators the source document (in Czech) with a line highlighted and then underneath all its translation variants (in English). The current line was also highlighted. Next to every translation was a set of questions related to the just highlighted lines:

- **Adequacy**: range from 0 (worst) to 1 (best) measuring how much the translated message is content-wise correct regardless of grammatical and fluency errors.
- **Fluency**: range from 0 (worst) to 1 (best) measuring the fluency of the translation, regardless of the relation of the message to the source and the correct meaning.
- **Markables**: A text area for reporting markables for the second phase.
- **Conflicting markables**: checkbox for when there is a markable in conflict (e.g. the terminology change) with a previous occurrence in the document. This corresponds to the first condition in the markable definition in Section 1.1. The default value was *No* (no

[2] github.com/ELITR/wmt20-elitr-testsuite

conflict) because the distribution was highly imbalanced.

Bojar et al. (2016) summarize several methods for machine translation human evaluation: Fluency-Adequacy, Sentence Ranking, Sentence Comprehension, Direct Assessment, Constituent Rating and Constituent Judgement. For our purposes, we chose a method similar to Fluency-Adequacy as one of the standard sentence-centric methods. The difference to the method described is that we showed all the competing MT systems at once, together with the whole document context. Ultimately, we would like the users to rate Fluency-Adequacy of the whole documents, but we suspected that asking annotators to read the whole document and then rating it on two scales would yield unuseful biased results.

2.2 Manual Markable Evaluation

In the following phase, we focused on markables specifically. For every markable in the source, we asked the annotators to examine 11 phenomena. If the given phenomenon is present in the examined markable occurrence, a checkbox next to it should have been checked (Occurrence). Further on a scale 0–1 (not at all–most) the annotator should mark how negatively it affects the quality of the translation (Severity). We list the 11 phenomena we asked the annotators to work with:

- **Non-translated**: The markable or part of it was not translated.
- **Over-translated**: The markable was translated, but should not have been.
- **Terminology**: The translation terminology choice is terminologically misleading or erroneous.
- **Style**: An inappropriate translation style has been selected, such as too formal, colloquial, general.
- **Sense**: The meaning of the whole markable translation is different from what was intended by the source.
- **Typography**: Typographical errors in translation such as in capitalization, punctuation, special character or other typos.
- **Semantic role**: The markable has a different semantic role in translation than in the source. Without any specific linguistic theory in mind, we provided four basic roles for illustration: agent (story executor), patient (affected by the event), the addressee (recipient of the object in the event), effect (a consequence of the event).
- **Other grammar**: Other grammatical errors such as bad declension or ungrammatical form choice.
- **Inconsistency**: A different lexical translation option than the previous occurrence was used. It is enough to compare only with the previous occurrence and not with all of them.
- **Conflict**: The translation conflicts with another markable or term in the document. This and another markable translates to the same word.
- **Disappearance**: The markable does not appear in translation at all.

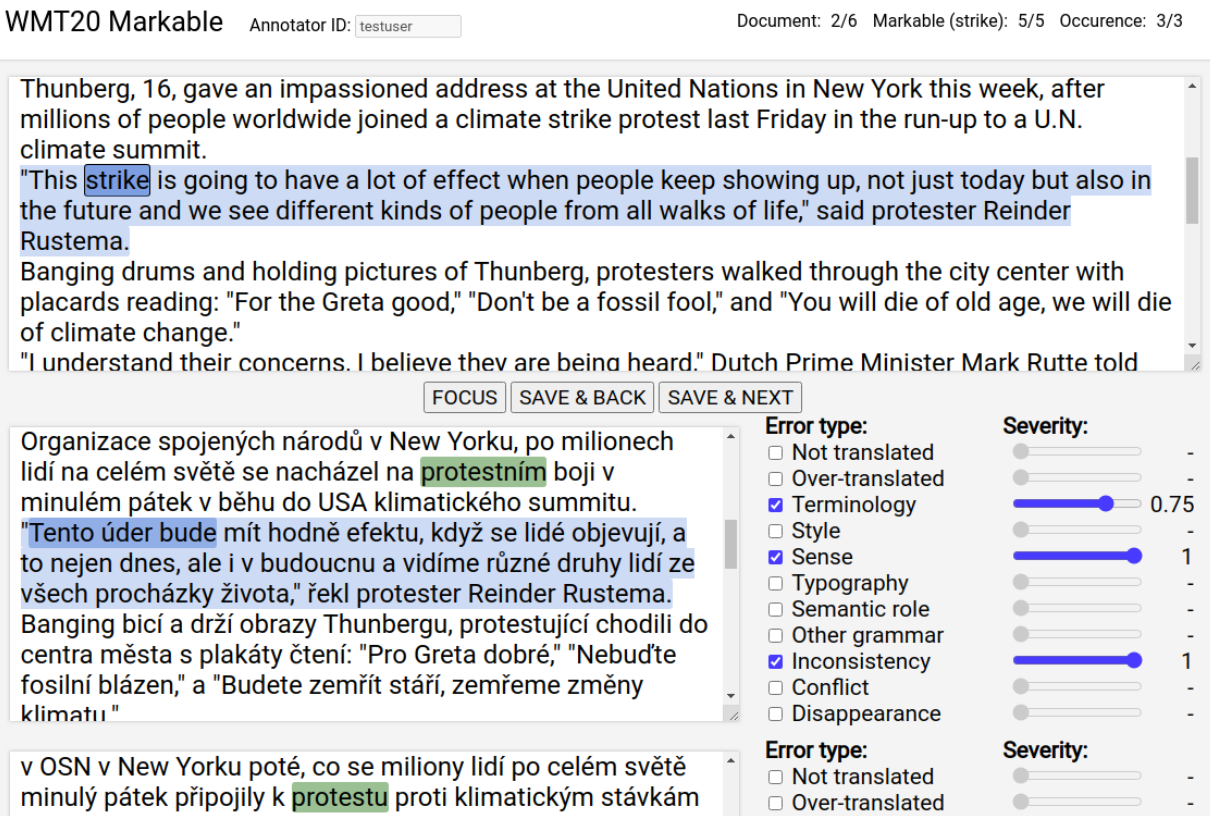
The choice to focus on markables was motivated by the aim to find a way to measure document-level performance using human annotators. A good markable translation is not a sufficient condition for document-level performance, but a necessary one. This approach is similar to Constituent Ranking/Judgement described by Bojar et al. (2016) with the difference that we chose to show all the markable occurrences in succession and in all translations in the same screen. We showed the whole translated documents context so that the annotators could refer to previous translations of the markable and the overall context.

2.3 Interface

Figure 1 shows the online interface for the second phase of this experiment. The first text area window contains the source document (e.g. in English). Below it are several translations (e.g. in Czech). Next to each translation is a set of questions. In the source, the current markable occurrence, to which the questions relate, is always displayed in dark blue. The current sentence is highlighted in the translations with light blue. The target words which probably correspond to the current markable (via automatic word alignment) are highlighted in dark blue as well. This alignment is present only for quick navigation as it is not very accurate. In translations, the remaining occurrences of a given markable are highlighted in green to simplify checking for inconsistency.

The FOCUS button is used to scroll to the current line in all text areas in case the user scrolled the view to examine the rest of the document.

In the first phase, the annotators could return to their previous answers and adjust them, but before



Thunberg, 16, gave an impassioned address at the United Nations in New York this week, after millions of people worldwide joined a climate strike protest last Friday in the run-up to a U.N. climate summit.
"This strike is going to have a lot of effect when people keep showing up, not just today but also in the future and we see different kinds of people from all walks of life," said protester Reinder Rustema.
Banging drums and holding pictures of Thunberg, protesters walked through the city center with placards reading: "For the Greta good," "Don't be a fossil fool," and "You will die of old age, we will die of climate change."
"I understand their concerns. I believe they are being heard." Dutch Prime Minister Mark Rutte told

Organizace spojených národů v New Yorku, po milionech lidí na celém světě se nacházel na protestním boji v minulém pátek v běhu do USA klimatického summitu.
"Tento úder bude mít hodně efektu, když se lidé objevují, a to nejen dnes, ale i v budoucnu a vidíme různé druhy lidí ze všech procházky života," řekl protester Reinder Rustema.
Banging bicí a drží obrazy Thunbergu, protestující chodili do centra města s plakáty čtení: "Pro Greta dobré," "Nebuďte fosilní blázen," a "Budete zemřít stáří, zemřeme změny klimatu."

Error type:	Severity:
☐ Not translated	-
☐ Over-translated	-
☑ Terminology	0.75
☐ Style	-
☑ Sense	1
☐ Typography	-
☐ Semantic role	-
☐ Other grammar	-
☑ Inconsistency	1
☐ Conflict	-
☐ Disappearance	-

v OSN v New Yorku poté, co se miliony lidí po celém světě minulý pátek připojily k protestu proti klimatickým stávkám

Error type:	Severity:
☐ Not translated	-
☐ Over-translated	-

Figure 1: Online interface for markable annotation with highlighted segments. The 12 other translations are in the rest of the page, not fully visible here.

continuing to the next line, they had to fill in the current fluency and adequacy. In the second phase, the annotators could freely return to their previous answers and adjust them. The most straightforward approach for them was to annotate a single markable occurrence across all MT systems and the switch to the next one as opposed to annotating all markable occurrences in the first translation, then all markable occurrences in the second translation, and similarly the rest.

As soon as we aggregate the statistics over multiple documents (or even translation directions), the effects of which particular annotator annotated which document can start playing a role, but we hope they cancel out on average.

3 Results

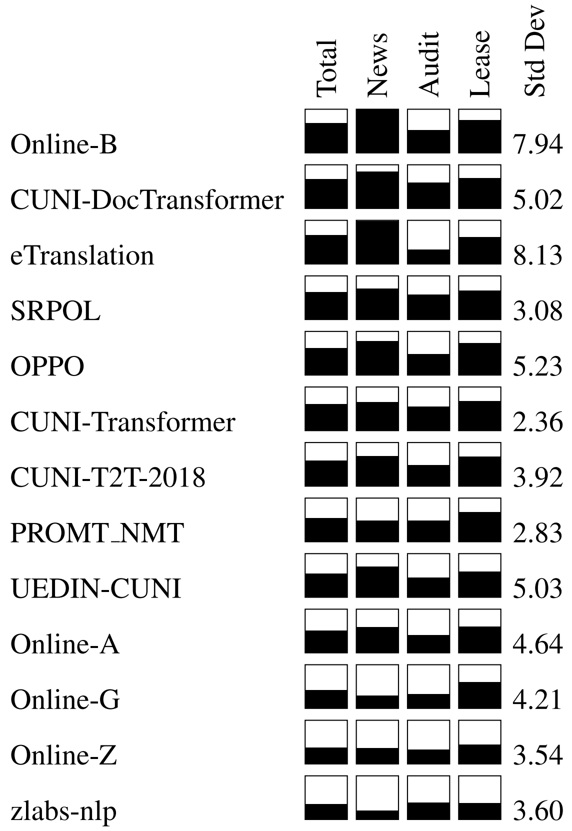
3.1 Automatic Evaluation

We measured the system quality using BLEU (Papineni et al., 2002) against a single reference. The results sorted by the score across all documents are shown in Table 2. BLEU scores across different test sets are, of course, not comparable directly. Only a very big difference, such as that of eTranslation for News and Audit (39.43% and 23.23%) suggests some statistically sound phenomena. We measured the standard deviation across MT systems within individual domains: News (6.19), Audit (2.34) and News-Lease (2.74). The Audit domain was generally the least successful for most of the submitted systems (see Table 3) and the Lease domain was more stable in terms of variance. The MT system BLEU variance over annotated lines hints that the better the system, the higher variance it has. This may be because most of the best MT systems are focused on News and fail on other domains, while the lower performant MT systems are low performant systematically across all domains.

3.2 Overall Manual Evaluation

From the first phase (Section 2.1) we collected $13 \times 328 = 4264$ line annotations. From the second phase (Section 2.2) we collected $13 \times 499 = 6487$ markable annotations. The average duration for one annotation of one translated line in the first phase was 25s, while one annotation of one system-markable occurrence in the second phase took only 8s.

Fluency and Adequacy correlate per line together



	Total	News	Audit	Lease	Std Dev
Online-B					7.94
CUNI-DocTransformer					5.02
eTranslation					8.13
SRPOL					3.08
OPPO					5.23
CUNI-Transformer					2.36
CUNI-T2T-2018					3.92
PROMT_NMT					2.83
UEDIN-CUNI					5.03
Online-A					4.64
Online-G					4.21
Online-Z					3.54
zlabs-nlp					3.60

Table 2: MT system results measured by BLEU together with standard deviation measured from all sentences. Sorted by the first column. Full black box indicates 40% BLEU, empty 15% BLEU.

strongly (0.80), and their product correlates negatively (-0.33) with the number of wrong markables. Because of this strong correlation and also the need to describe the result of the first phase by one number, we focus on Fluency×Adequacy. Table 3 shows the average Fluency×Adequacy as well as the average number of reported wrong markables per line.

Document	Mult.	Mkbs.	BLEU
Audit →cs	0.95	0.08	28.61 ±5.13
Audit →en	0.81	1.23	32.68 ±5.07
Lease →cs	0.78	0.33	33.50 ±4.96
Lease →en	0.78	0.30	35.44 ±4.94
News →en	0.74	0.65	30.68 ±5.05
News →cs	0.65	0.83	38.67 ±4.93
Average	0.79	0.73	33.57 ±4.93

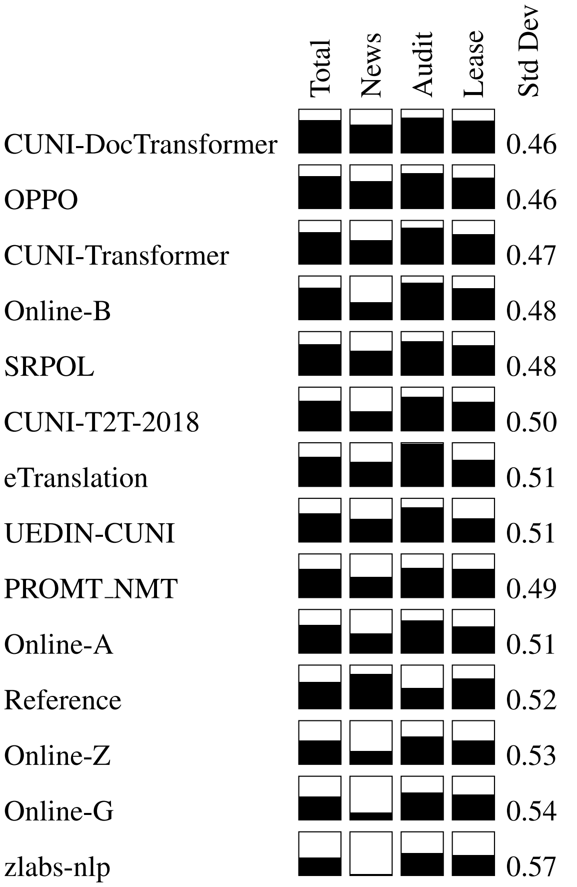
Table 3: Document average (across all systems) of Fluency×Adequacy (Mult.), number of reported wrong markables per line (Mkbs.) and BLEU.

3.3 MT System Performance

The performance per MT system and domain can be seen in Table 4. The reference translation received a comparably low rating in especially the Audit domain and fared best in the News domain. We see this as a confirmation of the last year's observation and a consequence of using non-expert annotators, who may have not annotated more complex cases thoroughly and were more content with rather general terms and language than what is correct for the specialized auditing domain.

No system has shown to be risky (high average but also with high variance). The last column in Table 4 shows, that the better the system, the more consistent it is (lower variation across documents). This did not occur with BLEU.

The ordering of systems by annotator assessment is slightly different than by automatic evaluation (Section 3.1). The automatic evaluation correlates with annotator rating (Fluency×Adequacy) with the coefficient of 0.93 (excluding Reference).



	Total	News	Audit	Lease	Std Dev
CUNI-DocTransformer					0.46
OPPO					0.46
CUNI-Transformer					0.47
Online-B					0.48
SRPOL					0.48
CUNI-T2T-2018					0.50
eTranslation					0.51
UEDIN-CUNI					0.51
PROMT_NMT					0.49
Online-A					0.51
Reference					0.52
Online-Z					0.53
Online-G					0.54
zlabs-nlp					0.57

Table 4: MT system results measured by Fluency×Adequacy together with standard deviation measured from Total. Sorted by the first column. Full black box indicates 100%, empty 40%.

Notable is the distinction in the performance of eTranslation in the Audit domain. Its BLEU in this domain (23.23%, Table 2) was below average, however it performed best of all submitted MT systems in terms of Fluency × Adequacy (98.62%, Table 4), above Reference. Closer inspection revealed that the translations were very fluent and adequate but usually used vastly different phrasing than in the Reference, leading to very low BLEU scores.

Source:
In the vast majority of cases, the obligations arising from contracts for financing were properly implemented by the beneficiaries.
Reference:
Ve většině případů byly závazky vyplývající z podmínek podpory příjemci řádně plněny.
eTranslation: (**BLEU**: 9.24%)
Ve velké většině případů příjemci řádně plnili povinnosti vyplývající ze smluv o financování.
CUNI-DocTransformer: (**BLEU**: 41.21%)
V naprosté většině případů byly závazky vyplývající ze smluv o financování příjemci řádně plněny.

Figure 2: Example translations by eTranslation and CUNI-DocTransformer together with Source and Reference. N-grams present in Reference are underlined.

The example in Figure 2 shows activization (opposite of passivization) in the translation by eTranslation (*the beneficiaries fulfilled their obligations*) instead of (*obligations were fulfilled by the beneficiaries*). This resulted in much lower n-gram precision and BLEU score in general, even though the sentence is fluent and more adequate than both the Reference and translation by CUNI-DocTransformer.

3.4 Markable Phenomena and Systems

Table 5 shows an overview of types of markable phenomena with the average number of occurrences and Severity across systems. For all systems, *Terminology* and *Conflicting markables* had the most significant impact on the translation quality. These two categories clearly differ in Severity with markable conflicts being much more severe than terminological mistakes.

Inconsistency, *Typography* and *Disappearance* phenomena also heavily impacted the translation quality, although with varying distribution of Occurrences and Severity.

Reference differs from MT systems by having higher average Occurrence, but lower average Severity (first column in Table 5). Furthermore, the Reference had a higher number of *Inconsistence* occurrences, but with lower Severity. This means that most of these *Inconsitencies* were not actual errors. This is expected, as careful word choice variation improves the style and requires having an overview of previously used terms in the document.

Over-translation occurred rarely and in those cases, mostly in names (example shown in Figure 3). *Other grammar* manifested itself most severely in gender choice when translating sentences with person names without any gender indication from English to Czech. Similarly, *Style* was marked mostly in direct speech translation. The system used informal singular form addressing instead of plural. These two phenomena are shown in Figure 4.

Source & Reference: Karolína Černá
Translation: Caroline Black

Figure 3: Example of overly-translated named entity, it is the name of one of the parties in the Lease agreement.

Source:
"How dare you?" Thunberg's U.N. speech inspires Dutch climate protesters
Reference:
"Jak se opovažujete?" projev Thunbergové v OSN inspiroval nizozemské protestující proti změnám klimatu
Translation:
"Jak se opovažuješ?" Thunbergův projev OSN inspiruje nizozemské klimatické demonstranty

Figure 4: Example of bad translation style.

Noteworthy is the correlation between phenomena across systems. The highest values were between *Sense* and *Terminology* (0.89), *Terminology* and *Inconsistency* (0.83) and *Sense* and *Other grammar* (0.82). There is no straightforward explanation of this correlation except the obvious that a good system is good across all phenomena. The correlation in the last phenomena pair suggests that the *Other grammar* category is too coarse and contains other subcategories.

3.5 Markable Phenomena and Domains

The results of markable phenomena across different domains is shown in Table 6.

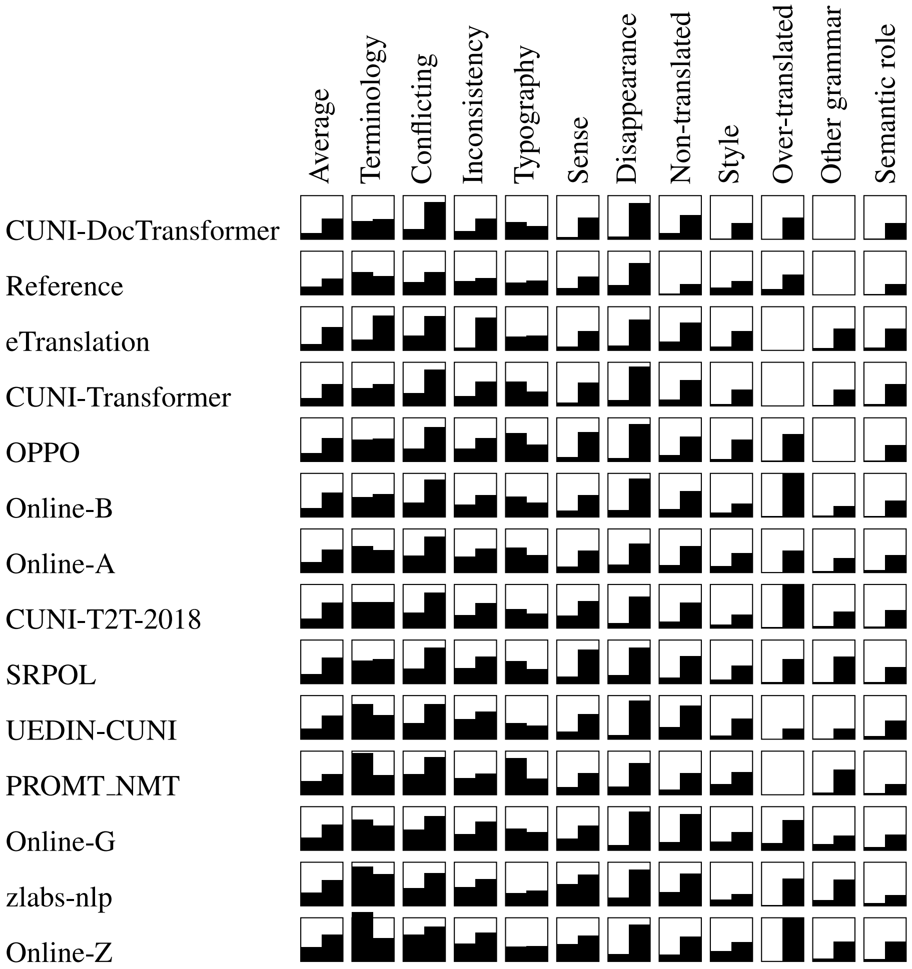


Table 5: Model results across 11 phenomena measured on markables together with their average. Each box is split into two bars: average Occurrence (left) and average Severity (right). Full left and right bars indicate occurrence in 20% of all markable instances and 100% Severity, respectively. Rows are sorted by Occurrence×Severity in the first column and columns, excluding *Average*, by the phenomena average Occurrence×Severity.

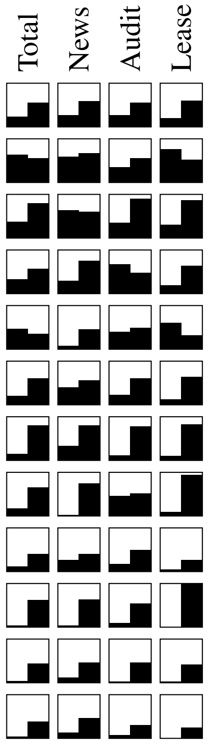
The second to last column is the correlation (across systems) between Occurrence×Severity and the BLEU score. The last column in Table 6 shows the correlation (across systems) between the two human scores: Occurrence×Severity and Fluency×Adequacy from the first phase of this experiment.

Since both BLEU and Fluency×Adequacy are positive metrics (the higher the score, the better the performance) and Occurrence×Severity is an error metric (the higher the number, the worse the performance), high negative correlations mean, that the metrics are mutually good performance predictors.

The strongest correlations are: *Conflicting* (-0.58), *Non-translated* (-0.55) and *Semantic role* (-0.41). Except for *Non-translated*, the reason is clear: BLEU is unable to check grammatical relations and never looks across sentences. We find the fact, that BLEU result was in agreement with error marking for these phenomena, to be positive.

Positive correlations (i.e. mismatches) were reached for *Disappearance* (0.28) and *Over-translated* (0.33), which is somewhat surprising because here BLEU *has* a chance to spot these errors from the technical point of view: shorter output could fire brevity penalty and missing terms where the exact wording is clear because they appear already in the source should decrease BLEU score. The overall correlation between Occurrence×Severity and Fluency×Adequacy is more significant than the correlation with BLEU. The most correlating variables are: *Sense* (-0.84), *Other grammar* (-0.84), *Terminology* (-0.81) and *Inconsistency* (-0.59).

Interesting is the markable phenomena *Disappearance* and *Sense* because of their high difference in correlations between BLEU and human score correlations.



	BLEU corr.	Mult. corr.
Average	-0.45	-0.79
Terminology	-0.38	-0.81
Conflicting	-0.58	-0.45
Inconsistency	-0.36	-0.59
Typography	-0.31	0.25
Sense	-0.29	-0.84
Disappearance	0.28	-0.46
Non-translated	-0.55	-0.50
Style	-0.07	-0.44
Over-translated	0.33	-0.37
Other grammar	-0.37	-0.84
Semantic role	-0.41	-0.24

Table 6: Document domain average (across all systems) of markable phenomena. Sorted by Occurrence×Severity in the first column. Full left and right bars indicate occurrence in 20% of all markable instances and 100% Severity, respecively. The last two columns show correlation between Occurrence×Severity and BLEU and user ratings from Phase 1, respectively.

3.6 Annotator Agreement

We would like to bring attention to inter-annotator agreement for the second annotation phase. Table 7 lists the following metrics, which are computed pairwise and then averaged:

Plain inter-annotator agreement (IAA) reports the percentage of pairs of annotations where the two annotators agree that a given phenomenon was or was not present. IAA shows high numbers in all cases but it is skewed by the heavily imbalanced class distribution: most often, a phenomenon is not present; see the left sides of squares in the leftmost column in Table 6 for distribution reference.

Cohen's Kappa (Kappa), measured also pairwise, isolates the effect of agreeing by chance and reveals that a good agreement is actually reached only in the cases of *Disappearance*, *Terminology* and *Over-translated*, which are less ambiguous to annotate. It is unclear what is the reason behind the low Kappas, but we speculate that it is due to insufficient attention of the annotators: they would perhaps agree much more often that an error occurred but they were overloaded with the complexity of the annotation task and failed to notice on their own.

Phenomenon	IAA	Kappa	Corr.	Corr.+
Disappearance	0.90	0.43	0.52	0.06
Typography	0.95	0.20	0.55	-0.13
Sense	0.91	0.17	0.73	-0.09
Style	0.94	0.24	1.00	0.19
Terminology	0.90	0.41	0.07	-0.03
Inconsistency	0.88	0.13	0.18	-0.08
Non-translated	0.94	0.20	0.64	0.30
Conflicting	0.77	0.02	1.00	0.62
Other grammar	0.96	0.10	1.00	-0.35
Semantic role	0.97	-0.01	-	0.43
Over-translated	0.98	0.37	1.00	1.00

Table 7: Annotator agreement of Occurence marking (Inter Annotator Agreement and Cohen's Kappa) and agreement in Severity (two versions of Pearson Correlation) with respect to every markable phenomenon.

Plain Pearson Correlation (Corr.) was measured on Severities in instances where both annotators marked the phenomenon as present. This, however, disregards the disagreement in cases one annotator did not mark the phenomenon. For this, we also computed Corr.+, which examines all pairs in which at least one annotator reported Severity and replaces the other with zero.

We observe a big difference in the correlations. In cases where both annotators agreed that there was an error they tend to agree on the severity of the mistake, except *Terminology* and *Inconsistency*. If the cases where only one annotator marked the error are included, then the agreement on Severity is non-existent, except *Over-translation* and *Conflicting* translation.

3.7 Translation Direction

We also examined how the language translation directions affect the results. Most notable is CUNI-DocTransformer, which performs worse when translating into Czech. With only 0.01% higher Occurence of markable phenomena, the Severity increased by 20.81%. This is not something which we observed in other systems. The translation into Czech brought on average 0.01% higher Occurrence, but the Severity on average dropped by 3.99% when switching from English→Czech to Czech→English. The explanation supported by the data is that in translation into English, CUNI-DocTransformer did not make any mistakes (or native Czech annotators did not detect them) and in translating into Czech, more issues were detected. Since the average Severity is measured across all phenomena, then the higher Severity in specific markable cases (Over-translated, Sense, Style and Disappearance) raised the overall average.

4 Annotation Examples

In the following figures (Figure 5, Figure 6 and Figure 7) we show annotated excerpts with BLEU, Fluency, Adequacy and markable phenomena severities. References are here to convey the Czech source segment meanings. They were not shown to the annotators. Examined markables are underlined.

Reference:
This Supplement No. 1 is written and signed in 2 (in words: two) copies, each of which is valid for the original.
Translation:
This Appendix 1 is drawn up and signed in two copies, each of which has the validity of the original.

BLEU: 23.59%, **Fluency:** 1, **Adequacy:** 0.9
Disappearance: 1

Figure 5: Example sentence markable (in words) annotation from Czech Lease document, translated by OPPO.

The example in Figure 5 focuses on intentional, key information duplication (for clarity and security reasons) of the number of signed copies. This duplication was however omitted in the translated output. The output is otherwise fluent and even received higher fluency than the Reference, which has an average fluency of 0.8.

Noteworthy is also another markable visible in the same figure, namely the referred section name: Appendix 1. Even though this word is different from the markable in the Reference: Supplement No. 1, it is used consistently across the whole document. Another variant of the translation is: Amendment No. 1. OPPO, together with Online-Z are the only systems which translated this markable correctly and consistently. Most of the systems (zlabs-nlp, Online-A, Online-B, Online-G, UEDIN-CUNI, CUNI-T2T-2018) switched inconsistently between the lexical choice. Other systems (SRPOL, eTranslation, CUNI-Transformer, CUNI-DocTransformer) were consistent in the main word choice, but not either in capitalization or number (e.g. Appendix No. 1 and Appendix 1).

Word variability (i.e. inconsistency) is often used to make the text more interesting, but in this context, it is vital that the term is translated consistently. Most of the systems, which outperformed even the Reference, made a severe error in this case.

Reference:
The most expensive item to be paid before the Grand Prix is the annual listing fee. This year, the fee was around 115 million Czech crowns. ”Masses of people who come to Brno to see the Grand Prix spend money here for their accommodation, food and leisure activities, which should more or less balance out the cost associated with the organization of the event, including the listing fee,” economist Petr Pelc evaluated the situation.
Translation:
The most expensive item is a breakdown fee every year before the Grand Prize. This year was about a hundred fifteen million crowns. ”Mass of people who will come to Brno at the Grand Prix will spend money on accommodation, food or entertainment, which should more or less balance the costs associated with organizing the event, including the unifying fee,” the economist Petr Pelc assessed.

BLEU: 26.59%, **Fluency:** 0.6, **Adequacy:** 0.4
Terminology: 1, **Sense:** 1, **Inconsistency:** 1

Figure 6: Example sentence markable (listing fee) annotation from Czech News document, translated by CUNI-T2T-2018.

Figure 6 shows a listing fee incorrectly translated as breakdown and unifying fee. This markable translation is interesting in the fact that systems were again very inconsistent with the markable translation choice. The wrong lexical choices were: landing, paving, parking, refill, landfill, security, zalistovacího, leasing, drop-in, back-up, reforestration, clearance, referral, padding fee and stamp duty. Good translations were: listing and registration fee.

Online-B and CUNI-DocTransformer made good and consistent lexical choices. SRPOL made good lexical choices but switched between them.

In this instance, this would not be an error, because consistency is not vital for interpreting the text.

The translation by CUNI-T2T-2018 in Figure 6 is not wrong only because of this markable translation choice, but also by poor fluency. The BLEU score, however, does not suggest, that there is anything fundamentally wrong with the translated segment despite the meaning being distorted.

Reference:
In Art. III of the Sublease agreement, entitled "Term of the Lease," the tenant and the lessee agreed that the apartment in question would be rented to the tenant for a fixed period from 13th May 2016 to 31st December 2018.
Translation:
In art. III of the apartment lease agreement, called ”sublease period”, the tenant and the tenant agreed that the apartment in question will be left to the tenant for use for a fixed period from 13. 5. 2016 to 31. 12. 2018.

BLEU: 31.95%, **Fluency:** 0.7, **Adequacy:** 0.5
Terminology: 0.5, **Sense:** 0.25, **Conflict:** 1,
Other grammar: 0.25

Figure 7: Example sentence markable (lessee) annotation from Czech News document, translated by Online-G.

The last example, in Figure 7, concerns itself with conflicting markables. In this case, two distinct markables (tenant and lessee) were merged into one translation tenant. This is a very fundamental error because, in the Lease agreement, these two markables refer to the two parties, which enter the contract.

Again, the BLEU does not suggest that anything is wrong with the translation. It could be even higher (51.06%) were it not for the localized date format in the Reference.

5 Conclusion

In this article, we compared three approaches to document translation evaluation. We saw that non-expert annotators rate most MT systems higher than Reference with Fluency and Adequacy, but Reference ranks better than most of them when inspecting markable phenomena and their Severity. Inspecting specific instances in detail, we found out that MT systems made errors in terms of markables, which no human translator would do.

Relating the current observation with the impression last year, we conclude that annotators lacking in-depth domain knowledge are not reliable for annotating on the rather broad scales of Fluency and Adequacy but they are capable of spotting term translation errors in the markable style of evaluation. This is important news because expert annotators can not be always secured. Unfortunately, the inter-annotator agreement remains generally low, possibly due to a high cognitive load with many systems annotated.

We further examined these markable phenomena and showed that especially *Sense*, *Other grammar* and *Terminology* kinds of errors negatively influence the Fluency and Adequacy the most. For BLEU the variables of highest importance were *Non-translated* and *Conflicting* errors.

In future work, we would like to examine more of the kinds of markable errors in modern MT systems and their influence on the translation quality. This description could then help researches focus on specific parts of their MT systems.

Furthermore, we would like to explore possible automated metrics, which would help in determining whether the document meaning remained intact with respect to markables.

Annotating markables appears to be easier for human annotators and more reliable for non-expert ones, and the results gave us more insight into the systems' performance than the Fluency-Adequacy method.

Acknowledgement

This study was supported in parts by the grants H2020-ICT-2018-2-825460 (ELITR) and Czech Science Foundation (grant n. 19-26934X, NEUREM3).

References

Ondřej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016. Ten years of WMT evaluation campaigns: Lessons learnt.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Tereza Vojtěchová, Michal Novák, Miloš Klouček, and Ondřej Bojar. 2019. SAO WMT19 test suite: Machine translation of audit reports. *arXiv preprint arXiv:1909.01701*.