


ARTICLE

# Evaluating optimal reference translations\*

Vilém Zouhar<sup>1,2</sup> , Věra Kloudová<sup>2</sup>, Martin Popel<sup>2</sup> and Ondřej Bojar<sup>2</sup>

<sup>1</sup>Department of Computer Science, ETH Zürich, Zürich, Switzerland and <sup>2</sup>Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czechia

**Corresponding author:** Vilém Zouhar; Email: [vilem.zouhar@gmail.com](mailto:vilem.zouhar@gmail.com)

(Received 31 January 2023; revised 20 June 2023; accepted 11 August 2023; first published online 8 May 2024)

Special Issue on ‘**The Role of Context in Neural Machine Translation Systems and its Evaluation**’, guest-edited by Rebecca Knowles and Sheila Castilho

## Abstract

The overall translation quality reached by current machine translation (MT) systems for high-resourced language pairs is remarkably good. Standard methods of evaluation are not suitable nor intended to uncover the many translation errors and quality deficiencies that still persist. Furthermore, the quality of standard reference translations is commonly questioned and comparable quality levels have been reached by MT alone in several language pairs. Navigating further research in these high-resource settings is thus difficult. In this paper, we propose a methodology for creating more reliable document-level human reference translations, called “optimal reference translations,” with the simple aim to raise the bar of what should be deemed “human translation quality.” We evaluate the obtained document-level optimal reference translations in comparison with “standard” ones, confirming a significant quality increase and also documenting the relationship between evaluation and translation editing.

**Keywords:** Language Resources; Post-Editing; Document Translation; Translation Evaluation; Translatology

## 1. Introduction

Machine translation (MT) is routinely evaluated using various segment-level similarity metrics against one or more reference translations. At the same time, reference translations acquired in the standard way are often criticized for their flaws of various types. For several high-resourced language pairs, MT quality reaches levels comparable to the quality of the reference translation (Hassan et al. 2018; Freitag et al. 2022) and sometimes MT even significantly surpasses humans in a particular evaluation setting (Popel et al. 2020). Given this, one could conclude that state-of-the-art MT has reached the point where reference-based evaluation is no longer reliable and we have to resort to other methods (such as targeted expert evaluation of particular outputs), even if they are costly, subjective, and possibly impossible to automate.

The narrow goal of the presented work is to allow for an “extension of the expiry date” for reference-based evaluation methods. In a broader perspective, we want to formulate a methodology for creating reference translations which avoid the often-observed deficiencies of “standard” or “professional” reference translations, be it multiple interfering phenomena, inappropriate expressions, ignorance of topic-focus articulation (information structure), or other abundant

\*Apart from the affiliated institutions and the funding for this project (Ministry of Education, Youth and Sports of the Czech Republic LM2018101 LINDAT/CLARIAH-CZ), the last author is additionally funded by the 19-26934X grant of the Czech Science Foundation (NEUREM3) and the third author by the Horizon Europe Innovation grant no. 101070350 (HPLT).

Article updated 21 May 2024.

shortcomings in the translation, indicating their authors' insensitivity to the topic itself, but above all to the source and target language. To this end, we introduce so-called optimal reference translations (ORT), which are intended to represent optimal (ideal or excellent) human translations (should they be the subject of a translation quality evaluation).<sup>a</sup> We focus on document-level translation and evaluation, which is in line with current trends in MT research (Maruf, Saleh, and Haffari 2019; Ma, Zhang, and Zhou 2020; Gete et al. 2022; Castilho 2022) and also this special issue of NLE. We hope that ORT will represent a new approach to the evaluation of excellent MT outputs by becoming a gold standard in the true sense of the word. Our work is concerned with the following questions:

- How to navigate future MT research for languages for which the quality level of MT is already very good?
- Is it worth creating an expensive optimal reference translation to compare with MT?
- If various groups of annotators evaluate optimal reference and standard translations, will they all recognize the difference in quality?

Subsequently, our contributions are as follows:

- definition of optimal reference translation and an in-depth analysis of evaluations and the relationship between evaluation and translation editing;
- reflection on what it means to be a high-quality translation for different types of annotators;
- publication of the *Optimal Reference Translations of English→Czech* dataset with a subset evaluated in aforementioned manner.

After discussing related work in this context (Section 2), we focus on defining ORT and describe its creation process (Section 3). Next, we describe our evaluation campaign of ORT, the data, annotation interface, and annotation instructions (Section 3.2). We then turn to a statistical perspective of our data and measure the predictability of human ratings (e.g. *Overall* rating from *Spelling, Style, Meaning*, etc.) using automated metrics (Section 4). We pay special attention to predicting document-level rating from segment level. In the penultimate Section 5, we provide a detailed qualitative analysis of human annotations and discuss this work in the greater perspective of human evaluation of translations (Section 6). Analysis code and collected data are publicly available.<sup>b</sup>

## 2. Related work

Evaluating translations (machine or human) is without doubt an extremely demanding discipline. Researchers have recently contributed several possible ways to approach the evaluation of translation quality in high-resource settings. We focus on the latest findings in this area, which—like our contribution—look for a possible new direction where future translation quality evaluation can proceed. The presented study is primarily concerned with the evaluation of human translations (“standard” vs. our optimal references) but the same evaluation methodology is applicable to machine translation.

Recently, Freitag et al. (2022) discussed metrics that were evaluated on how well they correlate with human ratings at the system and segment level. They recommended using neural-based metrics instead of overlap metrics like BLEU which correlate poorly with human ratings, and

<sup>a</sup>While we have no formal proof that the translations actually reached this optimality (i.e. nothing can be better), we are confident that the result of translato-logist collaboration comes close to this bar, especially given the evaluation results.

<sup>b</sup>[github.com/ufal/optimal-reference-translations](https://github.com/ufal/optimal-reference-translations) [huggingface.co/datasets/zouharvi/optimal-reference-translations](https://huggingface.co/datasets/zouharvi/optimal-reference-translations)

demonstrated their superiority across four different domains. Another relevant finding was that expert-based evaluation (MQM, Multidimensional Quality Metrics, Lommel *et al.* 2014) is more reliable than DA (Direct Assessment, Graham *et al.* (2013)), as already confirmed by Freitag *et al.* (2021). The MQM method relies on a fine-grained error analysis and is used for quality assurance in the translation industry. Popović (2020) proposed a novel method for manual evaluation of MT outputs based on marking issues in the translated text but not assigning any scores, nor classifying errors. The advantage of this method is that it can be used in various settings (any genre/domain and language pair, any generated text).

Other unresolved issues in the field of translation evaluation include the question of whether it is better to evaluate in a source- or reference-based fashion. As evidenced by, for example, Kocmi *et al.* (2022), reference-based human judgements are biased by unstable quality of references. For some language pairs and directions, however, it is still the main method of assessment. Licht *et al.* (2022) proposed a new scoring metric which is focused primarily on meaning and emphasises adequacy rather than fluency, for several reasons (e.g. meaning preservation is a pressing challenge for low-resource language pairs and assessing fluency is much more subjective).

Methods for automatic human translation quality estimation exist (Specia and Shah 2014; Yuan 2018), though the field focuses primarily on machine translation quality estimation. Furthermore, the definition of translation quality remains elusive and is plagued by subjectivity and low assessment agreement (House 2001; Kunilovskaya *et al.* 2015; Guerberof 2017).

### 3. Optimal reference translations

Our *optimal reference translation* (ORT) represents the ideal translation solution under the given conditions. Its creation is accompanied by the following phases and factors:

- diversity at the beginning (multiple translations are available from different translators, that is, in principle there are at least two independently-created translations available),
- discussion among experienced translation theoreticians/ linguists in search for the best possible solutions, leading to consensus,
- editing the newly created translations, reaching a point where none of the translation creators comes up with a better solution.

Another important condition is the documentation of all stages of the translation creation (archiving the initial solutions, notes on shortcomings, suggestions for other potential solutions, notes on translation strategies and procedures, record of the discussion among the authors, reasons why a solution was rejected, record of the amount of time spent on each text, etc.). The final characteristic of the creation of an optimal reference translation is the considerable amount of time spent by the creators on the analysis, discussion, and creation of new translations. In our definition of ORT, optimality therefore refers to:

- a carefully thought-out and documented translation process, and
- the quality of the resulting translation.

It however does not include the time aspect, in the sense of minimizing the time spent on the translation process. This choice is likely one more key distinction from “professional” translation. Incontestably, more than one version of ORT may be produced. The resulting ORT may vary depending on the individuality of its creators. Of course, the creators take into account the purpose and intended audience of ORT, just like in standard translations, but different collectives of ORT creators may perceive the intended purpose and audience differently or consider finer details of these aspects. Moreover, factors such as idiolect, age, experience, etc. can also play a large role, but unlike standard translations, there must always be a consensus among the creators of ORT.

### 3.1 Translation creation

The underlying dataset without the evaluation has already been described in Czech (Kloudová et al. 2023). The 130 original English texts (news articles available from the Internet, covering topics ranging from politics and economics to sports and social events) were translated from English into Czech by three human translators for the Conference on Machine Translation 2020 (WMT20). The three translators were hired by WMT organizers from a translation agency. The resulting three independent parallel Czech translations (P1, P2, P3) serve as basic reference translations, from which a final “optimal reference translation” could be synthesized. It was anticipated that our creators of ORT (two translators-cum-theoreticians—professionals who deal with translation from both a practical and a theoretical point of view)<sup>c</sup> would always choose the best translation solutions from the existing three versions, or create new solutions if necessary. However, the available translations from the first-stage translators were often of insufficient quality. Therefore, in the creation of our optimal reference translations, more emphasis was placed on the input of the creators of the final version rather than on the synthesis of existing translations.

The process of creating our ORT can be described as follows: our tandem of translators-cum-theoreticians worked as a translator and revisor pair. One of them produced a first version, which the other carefully compared with the original and critiqued if necessary. Notes on the first version of the translations were given in the form of comments on individual segments of the text. The author of the first version of the translations subsequently accepted or, with justification (and subsequent discussion), did not accept the suggestions in the comments. The crucial point in the discussion was always that the final solution should be fully in line with the beliefs of both translation authors. It is worth mentioning that the discussion between the two creators had, to a large extent, the written form of exchanging notes. ORT thus do not demand live, synchronous, attention of the creators.

The result of this process was two versions of ORT (many more versions could have evolved, though, our priority was not diversity, but above all quality—so we decided to create two versions in parallel, N1 and N2). The first version (denoted N1) is closer to the original both in terms of meaning and linguistic (especially syntactic) structure. The second version (N2) is probably more readable, idiomatic and fluent, being even closer to the Czech news style, both syntactically, for example, by emphasizing the ordering of syntactic elements typical of news reporting, and lexically, for example, by a more varied choice of synonyms. The presented work is centred around the evaluation of the various human translations. Because N2 has not been created for all segments of the translation (not all the original segments allowed an appropriate linguistic variation, that is, N1 was identical to N2), we decided not to use it. Thus, four translations were included in the evaluation—one optimal reference translation (N1) in addition to the three existing human translations.

In Figure 1, we show the sources and example translations (P1, P2, P3), together with one of the two versions of our optimal translation, N1. During the evaluation, each translation can be further edited by annotators in which case we identify the resulting segment as, for example, “P1 EDIT by annotator A4.” We will encounter examples in Section 5.

### 3.2 Annotation campaign

#### 3.2.1 Annotators

We hired 11 native Czech annotators for the evaluation of translations in three groups: (1) four professional translators,<sup>d</sup> (2) four non-experts, (3) three students of MA Study Programme

<sup>c</sup>One of them is a co-author of this article. However, the translations are later independently evaluated, and hence, to the best of our knowledge and conscience, we do not consider this to be a conflict of interest or otherwise a methodological flaw.

<sup>d</sup>Defining who is a professional translator is not easy. The factors influencing the degree of professionalism of a translator include, among others, education and experience. Professional translators in our study have at least one of the following: (1) completed an M.A. degree programme in English-Czech translation studies, (2) completed an M.A. degree programme in interpreting or philology, or (3) have at least 10 years of translation experience.

Key	Text
SRC	Professor Blair Grubb, Vice-Principal (Education) at the University, said: “To get to this stage, our students and graduates faced competition from peers attending some of the world’s top universities.”
N1	Prorektor pro oblast vzdělávání profesor Blair Grubb prohlásil: „Aby se naši studenti dostali až sem, museli čelit konkurenci svých vrstevníků, kteří studují na nejlepších světových univerzitách.“ <i>Professor Blair Grubb, Vice-Principal for Education, said: “To get to this point, our students have had to face competition from their peers studying at the world’s best universities.”</i>
P1	Profesor Blair Grubb, univerzitní zástupce ředitele pro vzdělávání, uvedl: „Abychom se dostali až do této fáze, museli naši studenti a absolventi čelit konkurenci svých vrstevníků, kteří studují na nejlepších světových univerzitách.“ <i>Professor Blair Grubb, the University’s Deputy Director of Education, said: “For us, to get to this stage, our students and graduates have had to face competition from their peers studying at the world’s best universities.”</i>
P2	Profesor Blair Grubb, zástupce děkana (vzdělávání) na univerzitě, řekl: “Aby se dostali do této fáze, čelili naši studenti a absolventi konkurenci svým vrstevníkům, kteří navštěvují některé z nejlepších světových univerzit.” <i>Professor Blair Grubb, Associate Dean (Education) at the University, said: “To get to this stage, our students and graduates have faced competition to their peers who attend some of the world’s top universities.”</i>
P3	Profesor a zástupce ředitele pro vzdělávání Blair Brubb university uvedl: “Aby se naši studenti a absolventi dostali do této fáze, museli čelit vrstevníkům z několika nejlepších universit světa.” <i>Professor and Deputy Director of Education at Blair Brubb University said: “To get to this stage, our students and graduates have had to face peers from several of the best universities in the world.”</i>

**Figure 1.** Example translations of the same source into Czech. Literal transcriptions of the translations are shown in *italics*. **N1**: translator collaboration (optimal translation), **P1**: professional translation agency (post-edited MT), **P2, P3**: professional translation agency.

Translation and/or Interpreting: Czech and English at the Institute for Translation Studies.<sup>e</sup> Their proficiency and end-campaign questionnaire responses are presented in Section 4.1.

3.2.2 Data

Out of the original data (Section 3.1), we randomly selected, with manual verification, 8 consecutive segments in 20 documents which were to be annotated. We refer to these 8 segments as documents because they contain most of the documents’ main points. Each segment corresponds approximately to one sentence, though they are longer (31 source tokens on average) than what we would find typical for the news domain. The data contain document-level phenomena (e.g. discourse), so segments cannot be translated and evaluated independently.

3.2.3 Annotation interface

We provided the annotators with online spreadsheets which showed the source text and all four translation hypotheses. This way each translation could be compared against the others while having the context available (e.g. to check for consistency). Each hypothesis column was distinguished by a colour, as shown in Figure 2, and based on annotator feedback (Section 4.1), we believe that it was manageable to perform annotations despite the amount of information shown. We showed the rest of segments in the source language for context but did not provide any translation hypotheses for the annotators to consult or rate. Each of the 20 documents was shown in a separate tab/sheet. The annotators worked on the evaluation in a span of 3 months in an uncontrolled environment.

<sup>e</sup>utrl.ff.cuni.cz

Source	Translation 1	T0 Spelling	T0 Terminology	T0 Grammar	T0 Meaning	T0 Style	T0 Pragmatics	T0 Overall	Translation 2	T1 Spelling	T1 Terminology	T1 Grammar	T1 Meaning	T1 Style	T1 Pragmatics	T1 Overall	Translation 3	T2 Spelling	T2 Terminology	T2 Grammar	T2 Meaning	T2 Style	T2 Pragmatics	T2 Overall	Translation 4	T3 Spelling	T3 Terminology	T3 Grammar	T3 Meaning	T3 Style	T3 Pragmatics	T3 Overall
Military Sees Frustrating Trend As Suicides Spike	Translation 1								Translation 2								Translation 3								Translation 4							
Military suicides surged this year to a record high among active duty troops, continuing a deadly trend that Pentagon officials say is frustrating and they are struggling to counter.	Počet sebevražd mezi vojáky v aktivní službě letos vzrostl na rekordní úroveň a pokračuje v hroznivém trendu, který frustruje úředníky Pentagonu, kteří se mu snaží čelit.	100	100	100	100	100	100	100	Sebevraždy vojáků v letošním roce prudce vzrostly k rekordnímu počtu u jednotek v aktivní službě. Pokračuje tak vražedný trend, o kterém úředníci Pentagonu říkají, že je nepřijemný, a snaží se mu čelit.	100	100	100	100	100	100	100	U vojáků v aktivní službě letos rekordně vzrostly počty sebevražd, což je děsivý trend, který frustruje úředníky Pentagonu, kteří se jej snaží odvrátit.	100	100	100	100	100	100	100	Mezi vojáky v aktivní službě letos došlo k alarmujícímu nárůstu počtu sebevražd. Pokračuje tak smrtící trend, který zástupci Pentagonu označili jako frustrující a jenž se snaží čelit.	100	100	100	100	100	100	100
The Army, Navy and Marine Corps all saw the rate of suicides go up as well as the overall numbers, with only the Air Force showing a decrease, according to data released by the Pentagon Thursday. Suicides among members of the Reserves and the National Guard also grew.	Armáda, námořnictvo a námořní pěchota registrují rostoucí počet sebevražd, stejně jako jejich celkový počet, přičemž pouze u letectva došlo k poklesu, jak vyplývá z údajů zveřejněných ve čtvrtek Pentagonem. Rovněž narostl počet sebevražd mezi příslušníky rezerv a Národní gardy.	100	100	100	100	100	100	100	Pozemní vojsko, námořnictvo i námořní pěchota zaznamenaly nárůst počtu sebevražd, stejně jako celek, kdy pouze letectvo vykazuje pokles, podle údajů zveřejněných Pentagonem ve čtvrtek. Sebevraždy mezi členy záloh a Národní gardy také vzrostly.	100	100	100	100	100	100	100	Podle dat vydaných ve čtvrtek Pentagonem pozoruje armáda, námořnictvo i námořní pěchota nárůst počtu sebevražd, jediné letectvo zaznává pokles. Sebevraždy mezi členy záloh a národní gardy také vzrostly.	100	100	100	100	100	100	100	Jak vyplývá z údajů, které ve čtvrtek zveřejnil Pentagon, zvyšující se tempo nárůstu sebevražd i jejich celkový počet byly zaznamenány jak v armádě, tak v námořnictvu a námořní pěchotě, pokles počtu sebevražd vykazuje pouze letectvo. K nárůstu počtu sebevražd došlo také mezi příslušníky rezervních složek a národní gardy.	100	100	100	100	100	100	100
The difficulties involved in identifying service members with possible problems and finding ways to prevent suicides were underscored earlier this month when the Navy reported that three crew members who served on the USS George H.W. Bush took their own lives within a week.	Na začátku tohoto měsíce vyšlo najevo, jak obtížné může být rozpoznat vojáky v aktivní službě, kteří mohou mít problémy, a také jakým způsobem zabránit sebevraždám. Tehdy totiž námořnictvo oznámilo, že tři členové posádky, kteří sloužili na lodi USS George H.W. Bush, si vzali život během jednoho týdne.	100	100	100	100	100	100	100	Potíže s identifikací členů armády s možnými problémy a nalezením způsobů, jak předjet sebevraždě, vyšly ještě více najevo v tomto měsíci, kdy námořnictvo nahlásilo, že si tři členové posádky lodi USS George H.W. Bush vzali život během jednoho týdne.	100	100	100	100	100	100	100	Problémy s rozlišením příslušníků armády, kteří by mohli mít problémy a hledání způsobů prevence sebevražd nabyly na důležitosti ještě více, když v minulém měsíci námořnictvo potvrdilo, že si tři členové posádky, kteří sloužili na lodi USS George H.W. Bush, vzali život během jednoho týdne.	100	100	100	100	100	100	100	Na potíže spojené s tím, jak rozpoznat, kteří členové armády mohou trpět problémy, a jak najít způsoby, jak by pomohly sebevraždám předcházet, bylo poukázáno dříve tento měsíc, když námořnictvo oznámilo, že si během jediného týdne vzali život hned tři členové posádky, kteří sloužili na lodi USS George H. W. Bush.	100	100	100	100	100	100	100
Asked about the deaths in the crew of the aircraft carrier, Defense Secretary Mark Esper said, "I wish I could tell you we have an answer to prevent further, future suicides in the Armed Services. We don't. We are caught up in what some call a national epidemic of suicide among our youth."	Na dotaz ohledně úmrtí členů posádky letadlové lodi odpověděl ministr obrany Mark Esper: "Přál bych si, abych vám mohl odpovědět, že existuje způsob, jak zabránit dalším sebevraždám v ozbrojených složkách. Ale nemůžeme. Nacházíme se v situaci, kterou někteří nazývají národní epidemií sebevražd mezi dnešní mládeží."	100	100	100	100	100	100	100	Ministr obrany Mark Esper na dotaz ohledně úmrtí v řadách posádky letecké lodi řekl: "Přál bych si, abych vám mohl říci, že víme, jak předejít dalším, budoucím sebevraždám v ozbrojených složkách. Ale my to nevíme. J sme v situaci, kterou někteří nazývají národní epidemií sebevražd mezi mladými."	100	100	100	100	100	100	100	Ministr Mark Esper na dotaz ohledně úmrtí členů posádky na letadlové lodi odpověděl: "Přál bych si, abych mohl říci, že víme, jak zabránit případným dalším sebevraždám v ozbrojených silách. My to nevíme. J sme lapani v situaci, kterou někteří nazývají národní epidemií sebevražd mezi mladými."	100	100	100	100	100	100	100	Na dotaz ohledně úmrtí členů posádky této letadlové lodi ministr obrany Mark Esper odpověděl: "Přál bych si, abych mohl říci, že víme, jak případným dalším sebevraždám v ozbrojených silách předcházet. Ale my to nevíme. J sme lapani v situaci, kterou někteří nazývají národní epidemií sebevražd mezi našimi mladými."	100	100	100	100	100	100	100
(hidden rows)	(hidden rows)								(hidden rows)								(hidden rows)								(hidden rows)							
Document rating		5.6	5.3	5.6	5.3	5.3	5.4	5.4	5.8	5.6	5.8	5.6	5.6	5.6	5.6	5.6		4.1	4	4.3	4	4	4	4		5.6	5.3	5.6	5.3	5.3	5.4	5.4
Document time (minutes)		75																														

Figure 2. First 5 rows of a screen for a single document with source and 4 translations in parallel. Screens were accessed by annotators in an online spreadsheet programme. Note: Scalable graphics—zoom in.



### 3.2.4 Annotation instructions

The task for annotators was three-fold, see Section 8 for the full annotation guidelines.

- Grade each segment translation on a decimal scale from 0 (least) to 6 (most) in categories *Spelling*, *Terminology*, *Grammar*, *Meaning*, *Style*, *Pragmatics* and *Overall* (e.g. 4.0 or 5.8). This scale was chosen to balance the number of attraction points for annotators (integers) and to contain a middle point (3).
- Grade each document as a whole on the same scale and categories.
- If a segment would not receive the highest grade, there would be something wrong in the translation. Therefore, the annotators should edit the hypothesis translation into a state to which they would give it the maximal scores.

## 4. Quantitative analysis

### 4.1 Annotator questionnaire

After the annotation campaign, the annotators filled a brief survey with questions about their perception of the task and their strategy.

We did not constrain the annotators in what order they should perform the annotations. As a result, they employed various approaches, most popular being *segment-category-translation*.<sup>f</sup> While we attempted to not introduce a bias, almost all annotators filled in categories one by one as they were organized in the user interface.<sup>g</sup> This could have an effect on the rating. For example, by establishing and drawing attention to the specific 6 features, the final *Overall* rating may be influenced primarily by them and it would not have been if the ordering was reversed. *Pragmatics* and *Overall* were reported as the hardest to evaluate, while *Spelling* was the easiest, especially because errors in spelling can be seen even without deeper translational analysis and there were not many of them in the translations. The annotators self-reported utilizing the preceding and following context around half the time to check for document-level consistency. While they proceeded mostly linearly, about 20% (self-reported estimate) of previously completed segments were later changed. We intentionally shuffled the ordering of translations (columns in each sheet) so that the annotators would not build a bias towards the translation source in, for example, the second column. However, the annotators reported that despite this, they were sometimes able to recognize a specific translation source based on various artefacts, such as systematically not translating or localizing foreign names.

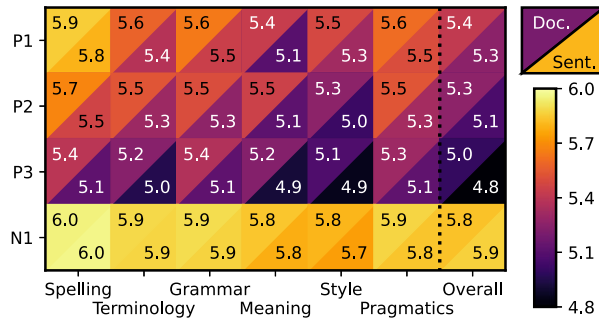
### 4.2 Collected annotations

We do not do any preprocessing or filtering of the collected data. This is justified by our all annotators working on the same set of documents and by the fact that we have established connections with each of the annotators and deem them trustworthy. Any bias of an annotator's rating would therefore be present in all documents which would not hinder even absolute comparisons. Nevertheless, we examine annotator variation later in this section. In total for 20 documents, we collected:

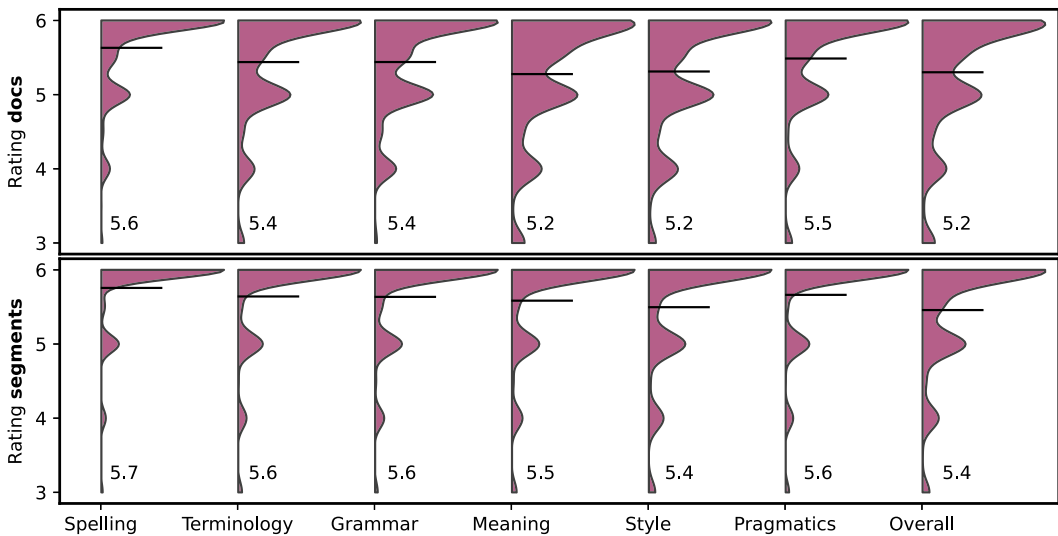
- 7k segment-level annotations (1.8k annotations of 4 translation hypotheses). Each hypothesis is edited unless it received a very high score (in 4k cases). This amounts to 49k ratings across all categories.
- 880 document-level annotations (220 annotations of 4 translation hypotheses.) This amounts to 6.2k ratings across all categories.

<sup>f</sup>That is, first finish all annotation categories in a translation, then all annotation categories in the second translation, etc., and afterwards move to the second segment.

<sup>g</sup>That is, starting from *Spelling* and ending with *Overall*.



**Figure 3.** Averages of ratings for different translation sources on document (top-left) and segment (bottom-right) level across features.



**Figure 4.** Distribution densities of ratings of each collected variable (thin tail cropped  $\geq 3$  for higher resolution of high-density values). Numbers and horizontal lines show feature means.

### 4.3 Quality of initial translations

Recall the grading scale from 0 (least) to 6 (most). The translation sources (P1, P2, and P3) were of varying quality, as shown in Figure 3. Overwhelmingly, N1 was evaluated the highest followed by P1, P2 and P3, in this order. Furthermore, there is a strong connection between the ratings on segment and document level and also across evaluation categories.

The density distribution of features in Figure 4 shows the natural tendency of annotators to use integer scores. It also shows that all features are heavily skewed towards high scores and that on average documents receive lower scores than their segments.

### 4.4 Inter-annotator agreement

To measure inter-annotator agreement, we aggregate pairwise annotator Pearson correlations on the segment level.<sup>h</sup> At first, this agreement is quite low ( $\rho = 0.33$ ). It can however be explained upon closer inspection of agreement across translations. While inter-annotator correlations for

<sup>h</sup>Even though the data are not normally distributed, the Pearson correlation reveals agreement controlled for each annotator's mean and variance.



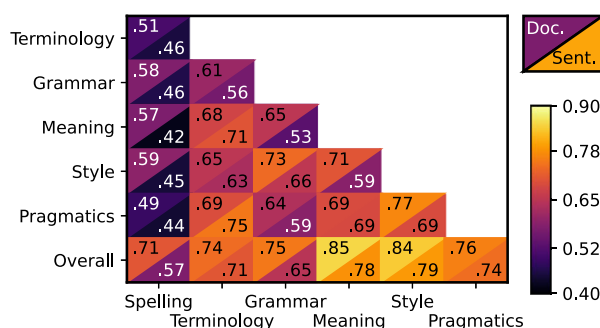


Figure 5. Pearson's correlations between individual features on document (top-left) and segment (bottom-right) level.

the worst translation P3 were  $\rho = 0.50$ , the best translation had  $\rho = 0.13$ . We hypothesize that with less variance and therefore signal for rating, the inter-annotator agreement drops. This is even more visible from the pairwise annotator correlations for the *Grammar* category, in which N1 has made almost no errors ( $\rho = 0.03$ ). In 28% of cases, the ordering of *Overall* scores for segments was the same between pairs of annotators and in 66% of cases they differed by only one transposition. In other words, the difference in the score ordering was 2 positions or more only in 8% of cases. Further individual effects of annotators are discussed in Section 4.6.

#### 4.5 Modelling overall quality from components

In this section, we attempt to model the *Overall* category based on individual categories, degree of translation editing, and individual annotators.

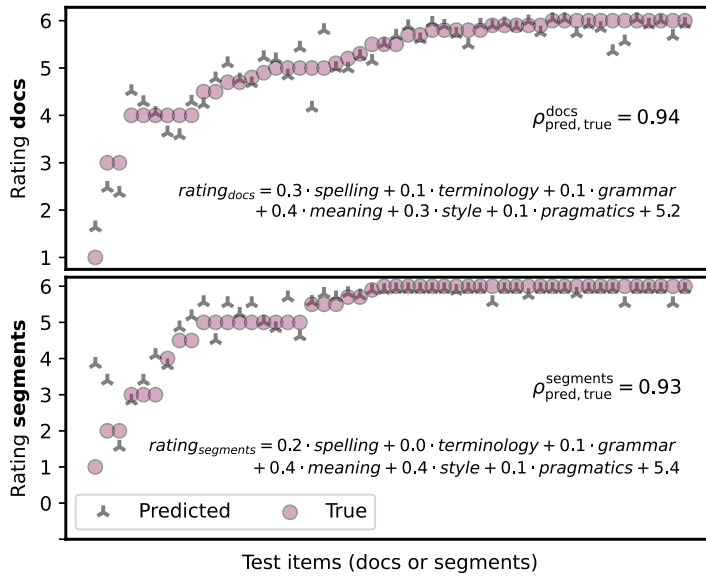
##### 4.5.1 Other categories individually

We first consider the predictability of individual categories and measure it using Pearson's correlation (0 = no relationship, 1 = perfect linear relationship). For both the document and segment level, we observe similar correlations, see Figure 5. Notably, spelling is much less predictive of other categories than the rest. A possible explanation is that this was the least common mistake and the values are therefore concentrated around the highest possible score (Figure 4). *Overall* correlates the most with *Meaning* and *Style*. This can be explained similarly because those features had the largest variances.

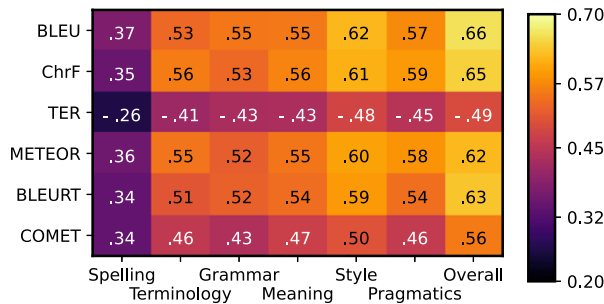
##### 4.5.2 Linear regression on other categories

We treat the prediction of *Overall* from other categories as a regression task with 6 numerical input features (*Spelling*, *Terminology*, etc) and one numerical output feature (*Overall*). We subtract the mean to preserve only the variance to be able to interpret the learned coefficients of a linear regression model. We split document- and segment-level ratings into train/test as 778/100 and 6925/100, respectively. Figure 6 shows the results of fitting two linear regression models together with the coefficients of individual variables. Because the distributions of features are similar, as documented in Figure 4, we can interpret the magnitude of the coefficient as the importance in determining the *Overall* score. For both the document and segment level, *Spelling* and *Meaning* have the highest impact while *Terminology* and *Style* have the least impact.<sup>1</sup> The linear regression model is further negatively affected by the non-linearity of the human bias towards round

<sup>1</sup>These interpretations are, however, not fully conclusive because of a possible latent co-dependent. The *Overall* variable may in reality be largely dependent on another variable *X* for which we do not have annotations. One hypothetical translation source could be very good if measured on the *X* variable and also *Overall* and unrelated to that also good in the *Spelling* level, which would yield similar results to those presented.



**Figure 6.** Predictions of linear regression models (on document and segment level) for all test set items sorted by true *Overall* score. Formulas show fitted coefficients and Pearson's correlations with the true scores. Only a random subset of points shown for visibility.



**Figure 7.** Segment-level Pearson's correlations between the collected scores and automated metrics between the original and edited versions of a segment. Colour is based on absolute value of the correlation (note TER).<sup>j</sup>

numbers, which the model is not able to take into consideration. The fitted coefficients are confirmed by annotator responses in the questionnaire in which *Meaning*, *Style*, and *Pragmatics* were most important to them when evaluating *Overall*.

#### 4.5.3 Automated metrics

As mentioned in Section 3.2, annotators were tasked to post-edit texts to a state which they would be content with. As a result, the annotators post-edited 62% of all the segments on average. We compute several automatic metric scores between the original and edited versions of segments and compare them to the collected scores, such as *Overall*. This allows us to answer the question: *Does the post-edited distance (as measured by automated metrics) correspond to the annotator score (negatively)?* The results in Figure 7 show that there is very little difference between individual

<sup>j</sup>Most metrics are scored from, for example, 0 (lower quality) to 100 (higher quality). For TER, it is the opposite (lower values mean higher quality). This explains the negative correlations.

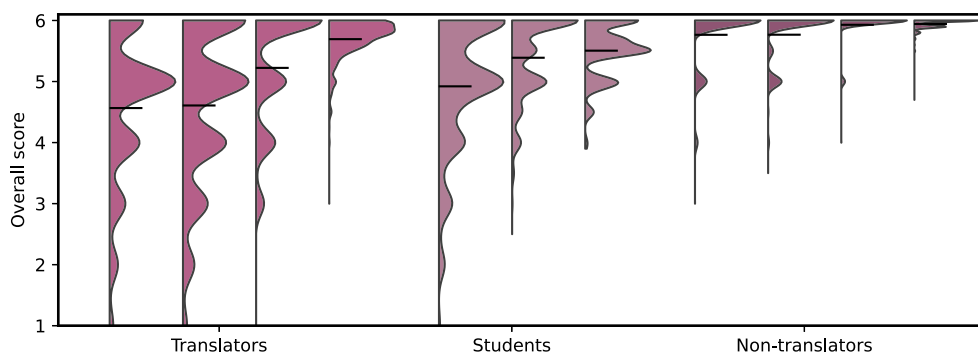


Figure 8. Distribution densities of ratings of *Overall* for individual annotators.

metrics. Most score categories are equally predictive with the exception of *Overall* (most) and *Spelling* (least). The explanation for this phenomena for *Spelling* is again (Section 4.4) much lower variance. Overall, the more the annotators changed the original text in their post-editing, the lower score they assigned to the hypothesis. Including the metrics in the prediction of *Overall* in Section 4.5.2 does not provide any additional improvement on top of other categories (final segment-level  $\rho$  is still 0.93).

#### 4.6 Annotator differences

Recall that we considered three types of annotators: professional translators, students of translation and non-translators. Despite the same annotation guidelines, their approach to the task was vastly different. For example, Figure 8 shows the distribution of segment-level ratings of *Overall*. Professional translators produced much more varying and spread-out distribution, especially compared to non-translators, who rated most segments very high. The group differences should be taken into account when modelling the annotation process statistically. When predicting segment-level *Overall* from other categories, as in Section 4.5.2, the individual annotator Pearson correlation ranges from as high as 0.98 to as low as 0.59. Similar to results of Karpinska *et al.* (2021), we find that expert annotators are important and have less noise. The average correlations with *Overall* for the translator, student and non-translator groups are 0.93, 0.91 and 0.80, respectively. The expertise feature alone yields 0.36 correlation with *Overall* and users alone 0.45. This is expected as the groups and users have different means of the variable. This information can be used in combination with other predictive features to push the segment-level correlation from 0.93 (Figure 6) to 0.95. Greater improvement is achieved when combined with the editing distance, such as pushing BLEU from 0.66 (Figure 7) to 0.76 when individual annotators are considered as an input feature (one-hot encoded).

#### 4.7 modelling document-level scores

Our annotation instructions explicitly reminded annotators to always consider the context. In other words, already our segment-level scores reflect the coherence and cohesion of the whole text, that is, how the text is organized and structured in the previous and/or subsequent segments. This is a rather important difference from automatic segment-level evaluation which discards any context. Annotators reported that in deciding document-level scores, they focused on the segments which were previously rated the lowest: that means, an individual poorly rated segment greatly influences the rating of the whole. We consider this observation essential for various future translation evaluations. We confirm this with results in Figure 9 where the *min* aggregation of segment-level ratings is a good prediction (comparable to or slightly better than *avg*) of the document-level rating. Based on segment-level ratings, we are able to predict document-level *Overall* quality with  $\rho = 0.71$ .

Category	min	max	avg	med
Spelling	0.38	0.17	0.36	0.30
Terminology	0.65	0.33	0.66	0.62
Grammar	0.64	0.36	0.66	0.60
Meaning	0.57	0.28	0.52	0.48
Style	0.70	0.26	0.69	0.63
Pragmatics	0.66	0.42	0.69	0.64
Overall	0.71	0.32	0.68	0.61

**Figure 9.** Pearson’s correlations of predictions from segment-level aggregations to document-level scores. For example, for the *Overall* category with *min* aggregation:  $\rho(\{d^{\text{Overall}} : d \in D\}, \{\min\{s^{\text{Overall}} : s \in d\} : d \in D\})$ .

It is worth noting that a similarly high correlation ( $\rho = 0.70$ ) is achieved when predicting the document-level *Style* from the corresponding segment-level ratings. This category was supposed to reflect also the coherence and cohesion of the document. Annotators saw the entire original text but only evaluated certain translated segments. However, they were assumed to have read the entire source text and to use the information for their evaluation. This was reflected in the *Style* category.

**Example 1.**  
**SOURCE:** *The All England Club, which **hosts the Wimbledon tournament**, handed the fine to Williams after she reportedly caused damage during a practice round **on the outside courts** on June 30, according to The Associated Press and CNN.*  
**ORIG:** *Klub All England Club, který **hostí turnament**, udělil dle zpravodajů The Associated Press a CNN Williams pokutu poté, co 30. června údajně způsobila škodu během cvičného kola **venku na kurtech**.*  
**EDITED:** *Klub All England Club, který **pořádá wimbledonský turnaj**, udělil dle zpravodajů The Associated Press a CNN Williamsové pokutu poté, co 30. června údajně způsobila škodu během cvičného kola na **venkovních kurtech**.*

In Example 1, the original translator (ORIG) did not consider the context of the whole document, translated only word for word and committed numerous interferences. It is completely unusual in the context of Wimbledon to use the phrase “*hostí turnament*” (hosts the tournament, both words being examples of lexical interference from English). In the context of tennis, the phrase “*venku na kurtech*” (out on the courts) is also unusual. In Czech, feminine names are typically marked using Czech morphology (e.g. *Serena Williams* → *Serena Williamsová*), which is the form predominantly found in the press. In this sentence, the name Williams follows the names The Associated Press and CNN, which is very confusing for the Czech reader. The feminine form thus makes the whole text easier to interpret and understand. The evaluator has correctly intervened in the text by using collocations such as “*pořádá wimbledonský turnaj*” (hosts the Wimbledon tournament) and writes about “*venkovní kurty*” (the outside courts) and uses the feminine form “*Williamsová*.” All these changes are highlighted in bold in Example 1 and demonstrate the evaluator’s (translation student) sense of textual continuity and their knowledge of the overall global context, which should have been the task of the original translator.

5. Qualitative analysis

If we take a closer look at the evaluation of all four translations by individual annotators, several types of qualitative comparisons can be made. We focus on the following two perspectives: characteristics of the segments (1) for which N1 scores worse than P{1,2,3} and (2) for which there

are the biggest differences in ratings. Even though we include example translations in Czech, we provide explanations in English which are self-contained and hence do not require any knowledge of Czech.

### 5.1 N1 scores worse than P{1,2,3}

N1 was evaluated with the highest scores in comparison to P1, P2, and P3, across all assessed features (see Figure 3). However, there is a small number of segments in N1 which were evaluated worse than those in P{1,2,3}. For a better overview, the frequencies at which the translations P{1,2,3} were evaluated better than N1 in the *Overall* category are P1: 6.16%, P2: 4.96%, P3: 3.99%. We selected these segments (for each category, not only for *Overall*) and analysed them. In most cases, our analysis revealed that the evaluation and the related editing of the translation was conditioned by the erroneous judgement of the annotators, who did not check the correct wording/meaning/usage in Czech and were tempted by the source text and/or the wrong parallel translations P{1,2,3}. In other words, the optimal reference translation N1 stood the test and our analysis confirmed its quality, rather than the evaluators' judgement.

Furthermore, we also encounter a reduced (imperfect) rating of some segments in N1, although no errors are apparent and no changes in the edited version have occurred in comparison to the original version. This finding is valid for all evaluated categories without exception. We list here a number of such segments with reduced (imperfect) rating for each category: *Spelling*: 1.0%, *Terminology*: 1.8%, *Grammar*: 2.0%, *Meaning*: 1.7%, *Style*: 2.8%, *Pragmatics*: 1.5%, *Overall*: 0.5%, any category: 5.4%. We perform a detailed qualitative analysis across all the seven rating categories.

#### 5.1.1 Spelling

In the spelling category, the following segment in Example 2 demonstrates an ignorance on the part of the annotator (non-translator) and failure to reflect the correct spelling and declension of the name *Narendra Modi* in Czech (correctly in nominative singular: *Naréndra Módí*) and of the Czech equivalent to the verb *harass* (correctly: *perzekvovat*, although it often appears incorrectly as *perzekuovat* in the language usage). The proposed edits are wrong.

#### Example 2.

**SOURCE:** *Sources said the action was in line with Prime Minister Narendra Modi's address to the nation [. . .] when he had said some black sheep in the tax administration may have misused their powers and harassed taxpayers [. . .]*

**N1 ORIG** (rating: 4.0): *Podle zdrojů akce souvisí s projevem premiéra Naréndry Módího k národu [. . .] v němž prohlásil, že jisté černé ovce v systému daňové správy podle všeho zneužívaly své pravomoci a perzekvovaly daňové poplatníky [. . .]*

**N1 EDIT:** *Podle zdrojů akce souvisí s projevem premiéra Nara[!]ndra Modi k národu [. . .] v němž prohlásil, že jisté černé ovce v systému daňové správy podle všeho zneužívaly své pravomoci a perzekuovaly daňové poplatníky [. . .]*

There are more segments with incorrect or unnecessary spelling edits. Unfortunately, some annotators not only erroneously "correct" what is actually right but also miscategorize the changes. We find erroneously corrected morphology in this category, etc. For example, for the source *The man pleaded guilty to seven charges involving [. . .]* the correct structure *Muž se přiznal k sedmi trestným činům (dative case) týkajícím se (dative case) [. . .]* has been edited to *Muž se přiznal k sedmi trestným činům (dative case) týkajících se (genitive case, grammatical incongruency) [. . .]*.

It is quite surprising the extent to which annotators do not verify and follow up information, leaving errors in translations that are contained in the original. This is particularly evident in the

spelling category. An example is a typo in the original: (*Pete*) *Townsend* (correctly *Townshend*; however, spelled correctly in the previous segment of the source text). P1 has *Townsend*, P2 *Townshed* [!], P3 *Townsend*. N1 uses the corrected form *Townshend*, but this form has been edited in the evaluation with the result *Townsend*.

### Example 3.

**SOURCE:** *Sony, Disney Back To Work On Third Spider-Man Film*

**N1 ORIG** (rating 3.0): *Sony a Disney opět<sub>(again)</sub> spolupracují na třetím filmu o Spider-Manovi*

**N1 EDIT:** *Sony a Disney Ø spolupracují na třetím filmu o Spider-Manovi*

#### 5.1.2 Terminology

In the terminology category, we also detected unnecessary or erroneous corrections. For example, the correction of the segment in Example 3 does not fall under terminology (and demonstrates, inter alia, the annotator's failure to verify the information; this time the annotator was actually a professional translator). The proposed edit is not correct/necessary.

#### 5.1.3 Grammar

In the grammar category, the above-mentioned segment (*Sony, Disney Back To Work On Third Spider-Man Film*) plays an interesting role, rated also 3.0 in this category, without any other changes. The proposed change (mentioned above) does not reflect grammar or spelling. As it turned out, this segment achieved the same rating from this annotator in other categories, too, namely meaning, style, and overall quality. It affects the meaning only, though, being an erroneous change.

We agree with some changes in syntax, for example, in Example 4 (rating 5.0 for this segment by a student annotator).

### Example 4.

**SOURCE:** *Homes were flooded and people waded through streets with water up to their knees in scenes normally seen only at the height of the monsoon.*

**N1 ORIG:** *Domy byly zaplavené a na ulicích se lidé brodili po kolena vodou, což bývá běžně k vidění jen v době, kdy monzunové období vrcholí.*

**N1 EDIT:** *Byly zaplaveny některé domy a na ulicích se lidé brodili po kolena ve vodě, což bývá běžně k vidění jen v době, kdy vrcholí monzunové období.*

#### 5.1.4 Meaning

In the meaning category, we observe several inconsistencies in evaluating translated segments for N1 vs P{1,2,3}. For example, reduced rating for N1 (4.0, by a non-translator) occurs in the following segment in Example 5, though, there are no changes in the edited version. Both the translations P1 and P2 score 6.0, even though there are several erroneous meaning units. The *initial therapy* is *počáteční léčba* in Czech, not *vstupní*, and the verb *require* does not mean here that the patient himself required the therapy but that his/her medical condition required it. The expression *chief medical officer* refers to the Czech equivalent *hlavní/vedoucí/vrchní lékař*, not *ředitel resortu zdravotnictví* (= *Director of the Ministry of Health*).

### Example 5.

**SOURCE:** *All but one patient had gone through initial therapy. That patient did require the recollection of stem cells, chief medical officer James Stein said.*

**N1 ORIG** (rating 4.0): *Všichni pacienti až na jednoho absolvovali počáteční léčbu. U dotyčného pacienta bylo zapotřebí provést nový odběr kmenových buněk, uvedl hlavní lékař James Stein.*

**P1 ORIG** (rating 6.0): *Kromě jednoho prošli všichni pacienti počáteční terapií. Dotyčný pacient požadoval nový odběr kmenových buněk, uvedl hlavní lékař James Stein.*



**P2 ORIG** (rating 6.0): *Všichni kromě jednoho pacienta prošli vstupní léčbou. Tento pacient vyžadoval znovuoibrání kmenových buněk, uvedl ředitel resortu zdravotnictví James Stein.*

**N1 EDIT** = N1 ORIG **P1 EDIT** = P1 ORIG

**P2 EDIT** = P2 ORIG

### 5.1.5 Style

In the style category, we agree with some edits made for N1, for example in the following segment in Example 6 rated 4.0. However, the rating of P1 is 6.0, although there have been very similar modifications to the style in the edited version of P1 as those in N1, and furthermore, the annotator (translator) uses the translation strategy proposed in N1 ORIG for his P1 EDIT version.

#### Example 6.

**SOURCE:** *Using data and artificial intelligence to try and boost revenues is part of HSBC's broader push to squeeze more out of its large physical network and client data, a key priority for interim Chief Executive Noel Quinn.*

**N1 ORIG:** *Využití dat a umělé inteligence ke zvýšení příjmů je součástí širší strategie HSBC, která tak chce vytěžit více ze své rozsáhlé fyzické sítě a klientských dat, což je klíčovou prioritou prozatímního generálního ředitele Noela Quinna.*

**N1 EDIT:** *Využití dat a umělé inteligence ke zvýšení příjmů je součástí širší strategie HSBC, která tak chce ze své rozsáhlé fyzické sítě a klientských dat vytěžit víc. Jde o jednu z hlavních priorit prozatímního generálního ředitele Noela Quinna.*

**P1 ORIG:** *Využití dat a umělé inteligence ke zvýšení výnosů je součástí širšího tlaku na HSBC, aby vytěžila více ze své rozsáhlé fyzické sítě klientů a klientských dat, což je klíčovou prioritou dočasného generálního ředitele banky Noela Quinna.*

**P1 EDIT:** *Využití dat a umělé inteligence ke zvýšení výnosů je součástí širší strategie HSBC, která chce ze své rozsáhlé fyzické sítě klientů a z klientských dat vytěžit víc. Jde o hlavní prioritu dočasného generálního ředitele banky Noela Quinna.*

### 5.1.6 Pragmatics

The evaluation in the category of pragmatics is also inconclusive and the analysis of N1 segments rated worse than P{1,2,3} segments does not provide any convincing data. For example, one of the annotators (non-translator) evaluates Example 7 N1 5.0, whereas P3 is evaluated 6.0. However, the only change we find in the edited version of N1 is the elimination of the adjective *nadšenou*, although *nadšená chvála* is a typical collocation in Czech, in contrast to the rather unusual formulation (and too literal translation) *zářná chvála* used in P3, which remained unchanged. Furthermore, *New York Magazine* is usually used in the Czech media in its original, not translated form. Nevertheless, P3 uses *New Yorský magazín* (the adjective does not even exist in Czech). Other inappropriate or non-existent word units used in P3 include: *díky* (= *thanks to*) used in a negative context, *Gettysburgského projevu* (correctly: *Gettysburského*, without g), *pro svůj historický význam* (correctly: *pro jeho historický význam*). The word order is not based on the principle of the Czech functional sentence perspective and is non-idiomatic and non-standard. We could give many more similar examples.

#### Example 7.

**SOURCE:** *Thunberg's grim pronouncements have earned her savage criticism and glowing praise. New York Magazine called her "the Joan of Arc of climate change," while The Guardian ranked her speech alongside President Lincoln's Gettysburg Address for its historical significance.*

**N1 ORIG** (rating 5.0): *Hrozné výroky vynesly Thunbergové ostrou kritiku i nadšenou chválu. New York Magazine ji nazval, „Johankou z Arku klimatických změn“, zatímco The Guardian zařadil její projev pro jeho historický význam vedle projevu prezidenta Lincolna v Gettysburgu.*



**N1 EDIT:** *Hroživé výroky vynesly Thunbergové ostrou kritiku i Ø chválu. New York Magazine ji nazval, „Johankou z Arku klimatických změn“, zatímco The Guardian zařadil její projev pro jeho historický význam vedle projevu prezidenta Lincolna v Gettysburgu.*

**P3 ORIG** (rating 6.0): *Thunbergová si díky svým hroživým výrokům vysloužila divokou vlnu kritiky a zářnou chválu. New Yorský magazín ji nazval, „Johankou z Arku klimatické změny, “zatímco The Guardian zařadil její projev vedle Gettysburgského projevu prezidenta Lincolna pro svůj historický význam.*

**P3 EDIT** = P3 ORIG

#### 5.1.7 Overall

The overall category shows similar inconsistencies as described in previous aspects. The annotators often neglect formal, meaning, and other errors, as shown above. Example 8 shows that different types of errors in P2 and P3 have been ignored. The annotator (non-translator) correctly substitutes the word *kredit* for *úvěr* in P3, but does not recognize the wrong structure *pokud jde o dobré jméno* in P2: the Czech word *kredit* (a sum of money (credit) or other value as a loan for a specified period of time for a specified consideration (e.g. interest)) can also have a colloquial meaning *trust, respectability* which is not the case here. The collocation *public accommodations* includes all services, that is not only accommodation but also catering, cultural activities (public spaces and commercial services that are available to the general public, such as restaurants, theatres, and hotels). The Czech word *služby* (= services) is correct, not *ubytování* (= accommodation). The trickiest collocation of this segment is *jury service* which is not *soudnictví* (= judiciary) in a general sense but, more specifically, *účast v soudní porotě* (= participation in jury trials). Leaving aside all the overlooked errors, the annotator evaluates P2 and P3 better than N1, although in N1 and P2 he/she made one change, in P3 two changes of a comparable nature.

#### Example 8.

**SOURCE:** *The Equality Act would extend nondiscrimination protections to LGBTQ individuals in credit, education, employment, housing, federal financial assistance, jury service and public accommodations.*

**N1 ORIG** (rating 4.0): *Zákon o rovnosti by měl rozšířit ochranu proti diskriminaci také na příslušníky sexuálních a genderových menšin, a to v oblasti úvěrů, vzdělávání, zaměstnání, bydlení, federální finanční pomoci, účasti v soudní porotě a služeb.*

**N1 EDIT:** *Zákon o rovnosti by měl rozšířit ochranu proti diskriminaci také na příslušníky sexuálních a genderových menšin, a to v oblasti úvěrů, vzdělávání, zaměstnání, bydlení, federální finanční pomoci, soudnictví a služeb.*

**P2 ORIG** (rating 5.0): *Zákon o rovnosti by měl rozšířit ochranu proti diskriminaci na LGBTQ jedince, pokud jde o dobré jméno, vzdělání, zaměstnání, bydlení, federální finanční pomoc, činnost porotců a veřejné ubytování.*

**P2 EDIT:** *Zákon o rovnosti by měl rozšířit ochranu proti diskriminaci na LGBTQ jedince, pokud jde o dobré jméno, vzdělání, zaměstnání, bydlení, federální finanční pomoc, soudnictví a veřejné ubytování.*

**P3 ORIG** (rating 5.0): *Zákon o rovnosti by zajišťoval ochranu proti diskriminaci LGBTQ osobám v oblastech kreditu, vzdělání, zaměstnání, bydlení, federální finanční asistence, výkonu poradce a veřejného ubytování.*

**P3 EDIT:** *Zákon o rovnosti by zajišťoval ochranu proti diskriminaci LGBTQ osobám v oblastech úvěrů, vzdělání, zaměstnání, bydlení, federální finanční asistence, výkonu soudnictví a veřejného ubytování.*

Category	A1	A2	A3	A4
Spelling	0	6	6	6
Terminology	2	6	6	4
Meaning	3	5	6	6
Pragmatics	3	6	6	5
Overall	2	6	6	5

Figure 10. Scores of subset of categories for a selected segment from the translation P3 by annotator A1 (translator) and annotators A{2,3,4} (non-translators).

5.2 Individual differences in ratings

This perspective detects and analyses segments with the biggest differences in ratings among annotators. In all categories in Figure 10, we find differences of at least 4.0 (Terminology and Grammar), 5.0 (Meaning, Style, Pragmatics, and Overall), or 6.0 (Spelling).

In this part, we would like to highlight selected segments with the biggest differences across relevant categories and focus on finding out the reasons for the observed discrepancies. The following segment in Example 9 shows a very low rating and multiple changes in the edited version by annotator A1, whereas annotators A2, A3, and A4 evaluate it with (almost) best scores and overlook even obvious (to the authors and translato­logists) mistakes. In the example, we use subscripts for expressions of interests.

Example 9.

**SOURCE:** *The Central Board of Indirect Taxes and Customs (CBIC)–the agency that oversees GST and import tax collections–compulsorily retired 15 senior officers under Fundamental Rule 56 (I) on corruption and other charges, official sources said.*

**P3 ORIG:** *Ústřední komise nepřímých daní a cel (UKNDC<sub>2,4</sub>)—agentura<sub>3</sub>, která dohlíží na daně ze zboží a služeb a vybrání vývozních dávek<sub>2</sub>—odvolala 15 vedoucích úředníků na základě Základního Pravidla<sub>1</sub> 56 (J)<sub>1,3</sub> o korupci a jiných obviněních, uvedly oficiální zdroje.*

**P3 EDIT by annotator A1** (translator): *Ústřední rada pro nepřímé daně a cla (CBIC<sub>2,4</sub>)—instituce<sub>3</sub>, která dohlíží na výběr daně ze zboží a služeb adovozních daní<sub>2</sub>—odvolala patnáct vysoce postavených úředníků na základě paragrafu<sub>1</sub> 56 písm. j)<sub>1,3</sub>, o korupci a jiných obviněních, uvedly oficiální zdroje.*

**P3 EDIT by annotator A2** (non-translator): *Ústřední komise nepřímých daní a cel (CBIC<sub>2,4</sub>)—agentura<sub>3</sub>, která dohlíží na GST a daně z importu<sub>2</sub>—odvolala 15 vedoucích úředníků na základě paragrafu<sub>1</sub> 56 (J)<sub>1,3</sub> o korupci a jiných obviněních, uvedly oficiální zdroje.*

**P3 EDIT by annotator A3** (non-translator): *Ústřední komise nepřímých daní a cel (UKNDC<sub>2,4</sub>)—agentura<sub>3</sub>, která dohlíží na daně ze zboží a služeb a vybrání vývozních dávek<sub>2</sub>—odvolala 15 vedoucích úředníků na základě Základního Pravidla<sub>1</sub> 56 (J)<sub>1,3</sub> o korupci a jiných obviněních, uvedly oficiální zdroje.*

**P3 EDIT by annotator A4** (non-translator): *Ústřední komise nepřímých daní a cel (CBIC<sub>2,4</sub>)—agentura<sub>3</sub>, která dohlíží na daně ze zboží a služeb a vybrání vývozních dávek<sub>2</sub>—odvolala 15 vedoucích úředníků na základě Základního Pravidla<sub>1</sub> 56 (j)<sub>1,3</sub> o korupci a jiných obviněních, uvedly oficiální zdroje.*

Annotator A1 rightly notices errors in spelling (lower and upper case letters<sub>1</sub>), terminology (name of the institution and other terms<sub>2</sub>), meaning (unclear and contradictory statements<sub>3</sub>), pragmatics (dealing with foreign realia and abbreviations<sub>4</sub>). These individual assessments are also reflected in the category Overall. On the other hand, the annotators A2, A3, A4 do not (mostly) notice the errors mentioned above or just change the wording of the abbreviation or replace the translation with the original abbreviation (A2) while maintaining the best rating. From our point

of view, the correct annotator is A1 with their relevant, thoughtful and sensitive interventions in the text.

There are many segments with similarly unbalanced ratings in our evaluation. As the analysis shows, the biggest problem is that some annotators fail to recognize most of the errors. Problematic is also lowering the rating even though no changes were made in the edited version. It is unclear whether the annotators simply did not pay enough attention to their task and whether they would have reached the same conclusions even after a more careful consideration of the whole task. Our qualitative analysis of selected segments confirms the findings presented in Figure 3: translators are the most rigorous and careful (average rating 5.00), students are slightly less attentive (5.3), and non-translators notice errors the least (5.8).

### 5.3 Document-level phenomena

In this section, we present examples that show the extent to which authors of translations P1, P2, P3, and N1, and especially, our evaluators have considered the context of the whole document. We examined the evaluated documents, including the source text and all four translations, looking for evidence of apparent respect or disregard for the document-level context.

Documents in which certain terms occur that should be consistent throughout the text and/or should correspond in a meaningful way to the thematic and pragmatic area, appear to be appropriate material to demonstrate this (spans marked with 1).<sup>k</sup> These observations are in line with MT evaluation methods focused on terminology (Zouhar, Vojtěchová, and Bojar 2020; Semenov and Bojar 2022; Agarwal et al. 2023). Another phenomenon to be observed might be a particular way of spelling words, which generally have two or more accepted spellings and the convention is just to achieve a consistent spelling throughout the document (spans marked with 2).

Finally, for the topic-focus articulation, also called functional sentence perspective (spans marked with 3), it is crucial to respect the context of the whole document (Daneš 1974; Sgall, Hajičová, and Panevová 1986; Hajičová et al. 2013). It is concerned with the distribution of information as determined by all meaningful elements, including context. In Example 12, evaluated by a non-translator, we selectively document the distribution of the degrees of communicative dynamism over sentence elements in Czech, which determines the orientation or *perspective* of the sentence (Firbas 1992).<sup>l</sup>

Our first example in this section, Example 10, illustrates the disregard for proper terminology. Translations P1,2 and N1 were not edited. The evaluator was a non-translator.

#### Example 10.

**SOURCE:** *All but one patient had gone through initial therapy. That patient did require the recollection of stem cells, chief medical officer James Stein said.*

**P1 ORIG:** *Kromě jednoho prošli všichni pacienti počáteční terapií. Dotyčný pacient požadoval<sub>1</sub>(insisted on) nový odběr kmenových buněk, uvedl hlavní lékař<sub>1</sub>(chief medical officer) James Stein.*

**P2 ORIG:** *Všichni kromě jednoho pacienta prošli vstupní léčbou. Tento pacient vyžadoval<sub>1</sub>(demanded) znovuodebrání kmenových buněk, uvedl ředitel resortu zdravotnictví<sub>1</sub>(Director of the Ministry of Health) James Stein.*

**P3 ORIG:** *Každý, až na jednoho pacienta, prošel počáteční terapií. Tento pacient potřeboval<sub>1</sub>(needed) odebrání kmenových buněk, uvedl vrchní zdravotní důstojník<sub>1</sub>(chief medical officer in command) James Stein.*

<sup>k</sup>We mark all related spans in the source as well as in all discussed translations even if they are correct, for easy comparison.

<sup>l</sup>The examples in Example 11 also contain other translation errors, such as incorrect name translation *Grubb/Brubb* and not only those related to the document-level phenomena. Unless otherwise stated, the evaluator has left these errors in the translation even after editing. Since we present associated discussions in more detail within the previous sections, we do not elaborate on them at this point.

**P3 EDIT** (non-translator): Každý, až na jedno pacienta, prošel počáteční terapií. Tento pacient vyžadoval<sub>1</sub>(demanded) odebrání kmenových buněk, uvedl hlavní lékař<sub>1</sub>,<sup>(chief medical officer)</sup> James Stein.  
**N1 ORIG**: Všichni pacienti až na jednoho absolvovali počáteční léčbu. U dotyčného pacienta bylo zapotřebí<sub>1</sub>(it was necessary) provést nový odběr kmenových buněk, uvedl hlavní lékař<sub>1</sub>(chief medical officer) James Stein.

In the next example, evaluated by the same non-translator, P1, P2, and N1 are consistent in terminology and spelling in this document.

### Example 11.

(individual evaluation segments are separated with |||)

**SOURCE**: Three University of Dundee students have been named regional winners as top graduates in Europe. ||| The University of Dundee students were named as top graduates in their respective fields in Europe in the 2019 Global Undergraduate Awards, whilst five other students from the same university were praised by the judges. ||| Professor Blair Grubb, Vice-Principal (Education) at the University, said: "To get to this stage, our students and graduates faced competition from peers attending some of the world's top universities." ||| "I would like to offer my warmest congratulations to Scott, Chester and Lola on this fantastic achievement, alongside the other Dundee representatives who were highly commended."

**P3 ORIG**: Tři studenti Univerzity v Dundee<sub>1,2</sub> byli jmenováni regionálními vítězi jakožto jedni z nejlepších absolventů v Evropě. ||| Studenti Dundeeské University<sub>1,2</sub> byly jmenováni v soutěži Globální vysokoškolské ceny 2019 jakožto jedni z nejlepších absolventů ve svých příslušných oborech, zatímco porotci ocenili dalších pět studentů ze stejné university<sub>2</sub>. ||| Profesor a zástupce ředitele pro vzdělávání<sub>1</sub> Blair Brubb university uvedl: „Aby se naši studenti a absolventi dostali do této fáze, museli čelit vrstevníkům z několika nejlepších universit<sub>2</sub> světa.“ ||| „Chtěl bych srdečně poblahopřát Scottovi, Chesterovi a Lole za jejich fantastické úspěchy a zároveň velmi pochválit i ostatní reprezentanty Dundee<sub>1</sub>.“

**P3 EDIT** (non-translator): . . . ||| . . . ||| Profesor a zástupce ředitele pro vzdělávání<sub>1</sub> Blair Brubb Ø uvedl: „Aby se naši studenti a absolventi dostali do této fáze, museli čelit vrstevníkům z několika nejlepších universit<sub>2</sub> světa.“ ||| . . .

Our third example in this section, Example 12, is represented by the article *China Says It Didn't Fight Any War Nor Invaded Foreign Land* which discusses armed conflicts between China and other countries.

In Czech, it is common to put the adverbials of time at the beginning or in the middle of a sentence (depending on the meaning and function of other sentence elements). When appearing at the end of a sentence, they become the focus of the statement, so the communicative dynamism and sentence continuity may get broken (*in 1979/ v roce/ roku 2017, in 1979/ v roce 1979*, spans 3a). Based on the information in the previous text (Example 12), the diplomatic resolution (*diplo-matically resolved, vyřešen diplomaticky/ diplomatickou cestou*) stands in contrast to the armed conflicts, so it represents the focus of the statement and should appear at the end of the sentence (after the verb) (spans 3b). Finally, the states *Vietnam, Malaysia, the Philippines, Brunei, and Taiwan* should be placed at the end of the Czech sentence (this becomes evident after reading and understanding the entire document where we are introduced to information about which countries China has had conflicts with) (spans 3c). The word *claims* (*vznášejí nároky, mají protinároky, mají opačné nároky, si činí nárok*) belongs to the topic of the statement.

### Example 12.

#### Previous article content:

China on Friday said it has not provoked a "single war or conflict" or "invaded a single square" of foreign land, skirting any reference to the 1962 war with India. "China has always been dedicated to resolving territorial and maritime delimitation disputes through negotiation and consultation," stated an official white paper released, four days ahead of the country set to celebrate its 70th

anniversary of the leadership of the ruling Communist Party of China (CPC) on October 1. “China safeguards world peace through real actions. Over the past 70 years, China has not provoked a single war or conflict, nor invaded a single square of foreign land,” the paper titled “China and the World in the New Era” said. The white paper, while highlighting the CPS’s “peaceful rise” made no reference of the bloody 1962 war with India and the vast tracts of land, especially in the Aksai Chin area, occupied by China. The Sino-India border dispute involving 3,488-km-long Line of Actual Control (LAC) remained unresolved. China also claims Arunachal Pradesh as part of South Tibet, which India contests. So far, the two countries held 21 rounds of Special Representatives talks to resolve the border dispute.

**SOURCE:** *Besides the 1962 war, India and China had a major military standoff at Doklam in 2017 when the People’s Liberation Army (PLA) tried to lay a road close to India’s narrow Chicken Neck corridor connecting with the North-Eastern states in an area also claimed by Bhutan. ||| It was finally diplomatically resolved after which both sides pulled back their troops. ||| China also had a major military conflict with Vietnam in 1979. China claims sovereignty over all of South China Sea. Vietnam, Malaysia, the Philippines, Brunei and Taiwan have counterclaims.*

**P1 ORIG:** *Kromě války v roce 1962 měly Indie a Čína velký konflikt v Doklamu roku 2017<sub>3a</sub>, kdy se čínská lidová osvobozená armáda (ČLOA) snažila postavit silnici blízko indického úzkého Kuřecího krku spojující severo-východní<sub>1</sub> země<sub>1</sub> na území, na které si dělá nárok i Bhutan. ||| Vše bylo zcela diplomaticky vyřešeno<sub>3b</sub> poté, co obě strany stáhly svá vojska. ||| Čína měla také veliký vojenský konflikt s Vietnamem<sub>3c</sub> v roce 1979<sub>3a</sub>. Čína vyhlásila svrchovanost nad všemi moři od jihu Číny. Vietnam, Malajsie, Filipíny, Brunei a Tchaj-wan<sub>3c</sub> mají protinároky.*

**P2 ORIG:** *Kromě války v roce 1962 hrozilo mezi Indií a Čínou vypuknutí většího vojenského konfliktu u Doklamské náhorní plošiny v roce 2017<sub>3a</sub>, když se Čínská lidová osvobozená armáda (PLA) pokusila vybudovat železnici poblíž úzkého indického koridoru Kuřecí krk, který spojuje západní a východní<sub>1</sub> část Indie<sub>1</sub> v oblasti nárokové také Bhútánem. ||| Nakonec bylo vše vyřešeno diplomatickou cestou<sub>3b</sub> a obě strany stáhly své vojenské jednotky. ||| V roce 1979<sub>3a</sub> došlo také k velkému vojenskému konfliktu mezi Čínou a Vietnamem<sub>3c</sub>. Čína si nárokuje svrchovanost nad celým Jihočínským mořem. Avšak Vietnam, Malajsie, Filipíny, Brunej a Tchaj-wan<sub>3c</sub> také vznášejí územní nároky na tuto oblast.*

**P3 ORIG:** *Kromě války v roce 1962 byla mezi Indií a Čínou velká vojenská patová situace v Doklamu v roce 2017<sub>3a</sub>, kdy se Lidová osvobozená armáda pokusila položit silnici v blízkosti úzkého indického koridoru Kuřecí krk, který Indii spojuje se severovýchodními<sub>1</sub> státy<sub>1</sub> v oblasti, kterou si také nárokuje Bhútán. ||| Konflikt byl nakonec diplomaticky vyřešen<sub>3b</sub> a obě strany poté stáhly svá vojska. ||| Čína měla také větší vojenský konflikt s Vietnamem<sub>3c</sub> v roce 1979<sub>3a</sub>. Čína si nárokuje suverenitu nad celým Jihočínským mořem. Vietnam, Malajsie, Filipíny, Brunej a Tchaj-wan<sub>3c</sub> mají opačné nároky.*

**N1 ORIG:** *Kromě války v roce 1962 hrozilo mezi Indií a Čínou vypuknutí většího ozbrojeného konfliktu v roce 2017<sub>3a</sub> u Doklamské náhorní plošiny, když se Čínská lidová osvobozená armáda (ČLOA) pokusila postavit železnici poblíž úzkého indického koridoru Kuřecí krk, který spojuje severní a východní<sub>1</sub> část Indie<sub>1</sub> v oblasti nárokové také Bhútánem. ||| Konflikt byl nakonec vyřešen diplomatickou cestou<sub>3b</sub>, načež obě strany svá vojska stáhly. ||| V roce 1979<sub>3a</sub> došlo k většímu vojenskému konfliktu také mezi Čínou a Vietnamem<sub>3c</sub>. Čína si nárokuje svrchovanost nad celým Jihočínským mořem, ovšem na tuto oblast si činí nárok také Vietnam, Malajsie, Filipíny, Brunej a Tchaj-wan<sub>3c</sub>.*

## 6. Discussion

Evaluating optimal reference translation(s) is in many ways a more difficult task than evaluating a “standard” (human or machine) translation. It is already a common practice in the translation



industry to have multiple workers included in a translation of a single document (e.g. initial translator and quality assurance translator). Based on our analysis of the optimal reference translation evaluation, it turns out that it is very crucial who evaluates such translations: Do the annotators have professional translation experience, or are they students of translation, or laymen in the field? It appears that laypeople are less able to notice even critical mistakes in translations. As a result for quality assurance, hiring only annotators with lots of translating experience seems to be a requirement.

However, important to determine is who the translation is for. If it is for a wide audience who do not scrutinize the translation quality, it may not be worth the extra cost to hire highly skilled translation evaluators. In turn, for the evaluation of machine translation systems that have reached this very high level of quality, highly skilled evaluators are needed.

We do note, however, that perfect translations or annotations likely do not exist, only their approximations. The cost of uncovering more translation errors is likely hyperlinear—that is, two rounds of annotations do not uncover twice as many mistakes. Each use-case should therefore make explicit what the target quality level is and adjust the annotation protocol accordingly.

## 7. Conclusion

We defined the concept of optimal reference translation (ORT), geared towards regaining informative results in reference-based machine translation evaluation. We then performed a careful manual evaluation and post-editing of ORT in comparison with three standard professional translation. The evaluation confirms that ORT deserve their name and can be regarded as a truly golden reference. In fact, the few times when ORT did not score best were examples of errors in this follow-up annotation, not examples of ORT deficiencies. Additionally, we documented that manual evaluation at these high levels of quality **cannot** be delegated to inexperienced annotators. Only people with substantial translation experience are sensitive to the subtle differences and can provide qualified judgements.

### 7.1 Time range

To process one document in all four translations takes on average 25–75 minutes. Please indicate the time spent on the annotation of each document (in minutes) in the appropriate box in each sheet. If you are systematically outside this range, send us an email. Please note that annotating the first document usually takes much more time than annotating subsequent documents.

### 7.2 Future work

While we focused on evaluating human translations, the identical setup could be used for evaluating MT models, which we plan to address in future work. This is not part of the present work which is focused on showing that the reference translations usually used are of insufficient quality and need to be reconsidered. Our next step will be to assess which of the multitude of automatic metrics of MT quality are sensitive to the subtleties captured in our ORT and can thus be used to reliably evaluate MT outputs of high quality. This will again require careful expert manual evaluation.

## 8. Annotation guidelines

The following is the main part of instructions which were distributed to the annotators.

### *Introduction.*

The goal of this study is to annotate the translation quality in seven categories. There are 20 documents in the shared Google sheet, marked as Edit1, Edit2 etc. (Orig1, . . . are described later in the

text). The first column contains the source text in English, followed by four Czech translations. However, only eight segments should be evaluated in each document. If you don't see a translation for some segments, it is not meant to be evaluated. You will evaluate the translations both at the segment level and at the level of whole documents (or at the level of the eight continuous segments). You will also indicate a better translation if you are not satisfied with the current version. Please read the source text first. The following is a possible evaluation procedure, but it is up to you how you proceed. The next steps are (for individual translations): 1. reading the translation, 2. evaluating the segments, 3. evaluating the whole document, 4. editing the segments so that you are satisfied with the translations, 5. reading the entire newly created text and possibly making minor changes. Please keep in mind that although you are also evaluating the segments separately, they are always part of a larger text, so you should pay special attention to how they relate to each other, that is also to the coherence and cohesion of the whole text. This should also be reflected in the assessment (category "style" below).

#### *Evaluation of segments.*

Rate each of the four translations in the following seven categories on a scale from 0 (worst) to 6 (best):

- spelling, punctuation, typography, typos,
- terminology (correctness, consistency, normativity),
- grammar: morphology (word forms) and syntax (sentence structure, functional sentence perspective),
- meaning accuracy (mistranslation, addition, omission, untranslated text segment etc.),
- style (appropriateness, consistency, idiomaticity, cross-sentence coherence and cohesion),
- pragmatics (culture-specific reference, locale conventions, appropriateness for the Czech reader),
- overall quality (evaluation of the translation in all the above-mentioned categories).

#### *Important notes.*

You can rate from 0 (the worst rating) to 6 (the best rating); in addition to whole numbers (0, 1, 2, 3, 4, 5, 6), decimal numbers with one decimal place (e.g. 0.1 or 4.5) are allowed. It is not necessarily the goal to use the full range of ratings for individual translations, that is if you do not see an error in a given category (even if the translation of the rated segment is very easy and does not pose a challenge for the translator), you will rate the highest possible score (6). We leave it to the discretion of each evaluator to decide how serious they consider a particular error to be and how many points to deduct for it. If an error affects more than one category (typically, e.g., both categories 3 and 4), this should result in a reduced rating in all relevant categories.

#### *Evaluation of documents.*

Rate the entire translation at the document level in the seven categories (the same as above for segment evaluation) on a scale from 0 to 6 (the same conditions as above for segment evaluation). The rating of the whole document is on the last line of each sheet.

#### *Editing of translated segments.*

If a segment translation does not receive the highest rating (6) in overall quality, please edit the translation with minimal editing (changes, corrections) to the state that you would give the highest rating (6). To clarify, if translations 1, 2, 3, and 4 get an overall quality rating of 6, 5, 3, 6, respectively (for particular segments), you must edit translations 2 and 3 independently. The resulting translations should be based on the original translations, that is most of the time they will be different from each other even after your edits. You can use dictionaries or search the internet, but



please do not use any machine translation systems. If possible, try not to copy text segments from previous translations, even if you like them. Since you probably weren't satisfied with some of the translations and didn't give them the highest possible rating, you have edited some segments. For comparison, you can look at the original translation (OrigT), which is in another sheet. For example, for document 3, the sheet is called Orig3 and is listed just after Edit3. Edit only the EditT sheet.

## References

- Agarwal, M., Agrawal, S., Anastasopoulos, A., Bentivogli, L., Bojar, O., Borg, C., Carpuat, M., Cattoni, R., Cettolo, M., Chen, M., Chen, W., Choukri, K., Chronopoulou, A., Currey, A., Declerck, T., Dong, Q., Duh, K., Estève, Y., Federico, M., Gahbiche, S., Haddow, B., Hsu, B., Mon Htut, P., Inaguma, H., Javorský, D., Judge, J., Kano, Y., Ko, T., Kumar, R., Li, P., Ma, X., Mathur, P., Matusov, E., McNamee, P., P. McCrae, J., Murray, K., Nadejde, M., Nakamura, S., Negri, M., Nguyen, H., Niehues, J., Niu, X., Kr. Ojha, A., E. Ortega, J., Pal, P., Pino, J., van der Plas, L., Polák, P., Rippeth, E., Salesky, E., Shi, J., Sperber, M., Stüker, S., Sudoh, K., Tang, Y., Thompson, B., Tran, K., Turchi, M., Waibel, A., Wang, M., Watanabe, S., and Zevallos, R. (2023). Findings of the IWSLT 2023 evaluation campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, Toronto, Canada (in-person and online). Association for Computational Linguistics, pp. 1–61.
- Castilho S. (2022). *DELA project: Document-level machine translation evaluation*. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pp. 319–320.
- Daneš F. (1974). Functional sentence perspective and the organization of the text. *Papers On Functional Sentence Perspective* 23, 106–128.
- Firbas J. (1992). Functional sentence perspective in written and spoken communication. In *Studies in English Language*. Cambridge University Press.
- Freitag M., Foster G., Grangier D., Ratnakar V., Tan Q. and Macherey W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics* 9, 1460–1474.
- Freitag M., Rei R., Mathur N., Lo C.-k., Stewart C., Avramidis E., Kocmi T., Foster G., Lavie A. and Martins A. (2022). Results of WMT22 metrics shared task: Stop using BLEU-neural metrics are better and more robust.
- Gete H., Etchegoyhen T., Ponce D., Labaka G., Aranberri N., Corral A., Saralegi X., Ellakuria I. and Martín-Valdivia M. T. (2022). *TANDO: A corpus for document-level machine translation*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 3026–3037.
- Graham Y., Baldwin T., Moffat A. and Zobel J. (2013). *Continuous measurement scales in human evaluation of machine translation*. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, Sofia, Bulgaria, pp. 33–41.
- Guerberof A. (2017). Quality is in the eyes of the reviewer.
- Hajičová E., Partee B. H. and Sgall P. (2013). *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*. Springer.
- Hassan H., Aue A., Chen C., Chowdhary V., Clark J., Federmann C., Huang X., Junczys-Dowmunt M., Lewis W., Li M., Liu S., Liu T., Luo R., Menezes A., Qin T., Seide F., Tan X., Tian F., Wu L., Wu S., Xia Y., Zhang D., Zhang Z. and Zhou M. (2018). Achieving Human Parity on Automatic Chinese to English News Translation. arXiv preprint arXiv: 1803.05567.
- House J. (2001). Translation quality assessment: Linguistic description versus social evaluation. *Meta* 46(2), 243–257.
- Karpinska M., Akoury N. and Iyyer M. (2021). *The perils of using mechanical turk to evaluate open-ended text generation*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1265–1285.
- Kloudová V., Mraček D., Bojar O. and Popel M. (2023). Možnosti a meze tvorby tzv. optimálních referenčních překladu: po stopách překladatelštiny v profesionálních překladech zpravodajských textu (possibilities and limitations of optimal reference translations: Exploring translationese in “professional” translations of newspaper articles). *Slovo a Slovesnost* 84(2), 122–156.
- Kocmi T., Bawden R., Bojar O., Dvorkovich A., Federmann C., Fishel M., Gowda T., Graham Y., Grundkiewicz R. and Haddow B. (2022). *Findings of the 2022 Conference on Machine Translation (WMT22)*.
- Kunilovskaya M. (2015). How far do we agree on the quality of translation? *English Studies at NBU* 1(1), 18–31.
- Licht D., Gao C., Lam J., Guzman F., Diab M. and Koehn P. (2022). Consistent human evaluation of machine translation across language pairs, arXiv preprint arXiv: 2205.08533.
- Lommel A., Uszkoreit H. and Burchardt A. (2014). Multidimensional quality metrics (MQM): a framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció*, 455–463.
- Ma S., Zhang D. and Zhou M. (2020). *A simple and effective unified encoder for document-level machine translation*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3505–3511.
- Maruf S., Saleh F. and Haffari G. (2019). A survey on document-level machine translation: Methods and evaluation, arXiv preprint arXiv: 1912.08494, 5.

- Popel M., Tomkova M., Tomek J., Kaiser L., Uszkoreit J., Bojar O. and Žabokrtský Z.** (2020). Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals. *Nature Communications* **11**(1), 1–15.
- Popović M.** (2020). *Informative manual evaluation of machine translation output*. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5059–5069.
- Semenov K. and Bojar O.** (2022). Automated Evaluation Metric for Terminology Consistency in MT.
- Sgall P., Hajicová E. and Panevová J.** (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Springer Science & Business Media.
- Specia L. and Shah K.** (2014). *Predicting human translation quality*. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, Vancouver, Canada, Association for Machine Translation in the Americas, pp. 288–300.
- Yuan Y.** (2018). Human translation quality estimation: feature-based and deep learning-based, PhD thesis. University of Leeds.
- Zouhar V., Vojtěchová T. and Bojar O.** (2020). *WMT20 document-level markable error exploration*. In *Proceedings of the Fifth Conference on Machine Translation*, pp. 371–380.