Reporting Scores and Agreement for Error Annotation Tasks

Maja Popović and **Anya Belz**
ADAPT Centre, School of Computing
Dublin City University, Ireland
name.surname@adaptcentre.ie

Abstract

The work reported in this paper examines different ways of aggregating scores for error annotation in MT outputs: raw error counts, error counts normalised over total number of words ('word percentage'), and error counts normalised over total number of errors ('error percentage'). We use each of these three scores to calculate inter-annotator agreement in the form of Krippendorff's α and Pearson's r and compare the obtained numbers, overall and separately for different types of errors. While each score has its advantages depending on the goal of the evaluation, we argue that the best way of estimating inter-annotator agreement using such numbers are raw counts. If the annotation process ensures that the total number of words cannot differ among the annotators (for example, due to adding omission symbols), normalising over number of words will lead to the same conclusions. In contrast, total number of errors is subjective because different annotators often perceive different numbers of errors in the same text, therefore normalising over this number can be associated with lower agreement.

1 Introduction

Manual error annotation is an increasingly investigated component of assessing the quality of automatically translated or otherwise generated text (Federico et al., 2014; Costa et al., 2015; Caseli and Inácio, 2020; Thomson and Reiter, 2020). Error annotation can be construed as a word-span labelling task, and a variety of different error labelling schemes have been proposed (Vilar et al., 2006; Lommel et al., 2014b; He et al., 2021; Belkebir and Habash, 2021; Al Sharou and Specia, 2022). If annotators are instructed to assign a predefined error type to given word-spans, the task is called error classification (Vilar et al., 2006; Costa et al., 2015; Lommel et al., 2014b). If annotators are asked only to mark the erroneous word-spans without assigning any particular error type, the task is called error marking (Kreutzer et al., 2020; Popović and Belz, 2021). Both for classification and marking, approaches differ in how they report the results from error annotation, i.e. how they convert the word-span labels obtained in the annotation process to aggregated, quantified results on the basis of which conclusions can be drawn. It is currently unclear how the choice of aggregation method affects conclusions, e.g. when comparing systems, or assessing inter-annotator agreement and reproducibility.

This paper aims to investigate three widely used ways of aggregating error scores, namely (i) raw error counts, (ii) error counts normalised by total number of words (referred to in this paper as 'word percentage'), and (iii) error counts normalised by total number of errors (referred to as 'error percentage'). We carried out our experiments on a publicly available data set consisting of annotated machine translated outputs which was also used in previous work (Popović, 2021).

2 Related Work

In an early paper reporting an approach to systematic error classification in MT outputs, Vilar et al. (2006) analyse several MT systems for two language pairs in order to obtain details about their weakest points. The results are reported as error percentages (in the above sense) for each defined error type in order to see which error types are predominant in each of the MT systems. Since in this work different annotators evaluated different texts, it was not possible to report inter-annotator agreement in any manner.

Lommel et al. (2014a) specifically address IAA for error classification in MT using the MQM[1] error scheme. IAA was calculated using both pairwise matching/accuracy between assigned labels on the word level as well as Cohen's κ coefficient on the

[1] https://themqm.org/

Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 306 - 315

word level. Although the work did not aim to evaluate MT systems, some overall results are reported: raw error counts as well as what might be called 'sentence percentage' – error counts normalised by the total number of sentences. It should be noted that the error counts did not take into account the number of words in the error span, but each span is counted as one error no matter how many words the span comprises. Lommel et al. also investigated sentence-level agreement, which led to much better agreement since evaluators often identified the same error types in a sentence although on different spans. These differences in spans are mentioned as the main reason for low κ coefficients, apart from annotators' personal preferences.

Another work which deals with IAA (Castilho, 2020) for error classification compares (a) IAA when evaluators see only isolated sentences, with (b) IAA when evaluators see larger portions of text in context (paragraphs, "documents"). Since the goal was not analysis of MT systems, a simple scheme involving four error types was used. The tool used for error annotation allowed only marking on the sentence level, so that Cohen's κ is also calculated on the sentence level. In addition, Pearson's r calculated on error counts and percentage of matched error types are reported as well. Since the focus of the work was fully on IAA, no error scores were reported.

Klubička et al. (2018) use the MQM error scheme for error classification to compare phrase-based and neural based systems for translating into Croatian. They mention that there is no standard in reporting numerical results for error classification and discuss two possibilities: (1) counting only error spans (without taking the word span into account), (2) counting number of words in error spans and normalising this word count over the sentence length (what we call word percentage). They argue that although raw error counts provide useful information, they do not enable drawing statistically meaningful conclusions about the results. They further argue that different MT systems may generate different sentence lengths which would make comparison of raw error counts unfair. Therefore, they decide to report the word percentage. As for IAA, Cohen's κ on the sentence level was calculated and compared across different error types. They obtained higher agreement coefficients than (Lommel et al., 2014c) and commented that the reason is calculating κ on sentence-level instead of word-level.

Kreutzer et al. (2020) perform error marking (annotation without classification into types) for use in a loss function in order to improve an MT system. Therefore, no scores were reported, and IAA was reported as Krippendorff's α calculated on raw error counts. The same approach, error marking and α calculated on error counts, was used by Popović and Belz (2021), together with the percentage of label matches (F-score). The error annotation results were reported in the form of word percentages.

Freitag et al. (2021) apply the MQM error scheme on large amounts of texts for two language pairs. They use raw span counts weighted with the error severity so that each minor error span is counted as 1 and each major error span as 5. A restriction of a maximum of five errors per segment is applied, and evaluators are instructed to choose the five most severe errors. As for IAA, for each rater Freitag et al. report the score and the ratio over the average score of all raters. They also report pairwise agreement, and discuss that there is no obvious best way to compute it, especially since they take into account error severity. The chosen approach groups the scores of each rater into numeric intervals, and then compares these intervals. The authors report average, minimum, and maximum pairwise annotator agreements, but do not specify how exactly these agreements were calculated.

A previous investigation (Popović, 2021) also concentrated on IAA, with the focus on pairwise word level matches for different error types, and reported overall Krippendorff's α calculated on error counts. The work reported in this paper was carried out on part of the same corpus, but with the following differences: (i) we introduce Krippendorff's α and Pearson's r calculated on three types of numerical scores, and (ii) we compare all IAA coefficients (including word overlap) and analyse the differences in depth. Also, we choose a subset of the corpus for which we have annotations by four evaluators as a result of an earlier reproduction study (Popović and Belz, 2021), in order to have more variation.

3 Annotated Data Set

Our experiments were carried out on a subset of the publicly available annotated QREV data set.[2] The full set consists of English user reviews about

[2] https://github.com/m-popovic/QRev-annotations

QRev data set	
language pair	en→hr
domain	user reviews
# of MT systems	3
# of unique segments	1217
total # of annotators	14
# of annotators per segment	4
quality criterion	Adequacy, Comprehension

Table 1: Descriptive statistics for the data set used in our analyses.

IMDb movies and Amazon products translated into Croatian and Serbian by five different MT systems for each language pair. Each text was annotated by two different evaluators. In the sub-set we used[3] (Popović and Belz, 2021), we had additional annotations obtained in a reproduction study (Popović and Belz, 2021); therefore in this subset each text was annotated by four different evaluators. The sub-set consists of Croatian MT outputs generated by three systems (Amazon Translate, Microsoft Bing and Google Translate). In total, 14 different evaluators, computational linguistics students and researchers as well as translation students fluent in the source language and native speakers of the target language, participated in the annotation of this sub-set. Some descriptive statistics for the data sub-set we used are shown in Table 1.

The annotation of MT outputs was carried out in two stages. In the first stage, a group of annotators provided error marks (*Major*, *Minor* and *None* for correct words) according to two quality criteria: Adequacy and Comprehension. For both quality aspects, the evaluators were asked to concentrate on problematic parts of the text and to highlight them. For Adequacy, they were instructed to highlight parts which entirely or partially change the meaning of the source text. For Comprehension, they were asked to mark parts which are impossible or hard to understand. They were also asked to add omission tags "XX" whenever necessary.

Since it was an error marking task, i.e. not guided by any predefined error scheme, only by the quality criteria, the evaluators had more freedom in annotating errors than in error classification tasks. Therefore, this annotation represents a descriptive type of evaluation (Rottger et al., 2022) which allows and encourages subjectivity.

The annotators were also asked to distinguish between major and minor errors, however we did not use this distinction in the experiments reported here. This could be an interesting direction for future work.

Error types were assigned in the second stage by the first author of this paper, a computational linguistics researcher with expertise in translation, who analysed the marked errors and assigned error type labels according to their cause and/or origin. The error types were not predefined by any particular error typology, but defined on the fly, while looking at the annotated text. For some words, multiple error types were identified. Some of these error types have small word spans (1–2 words) while others can involve a larger number of words, even the entire sentence. Since the evaluation protocol in the first stage allowed free annotation, evaluators often perceived the same issue but marked different words. This freedom also allowed evaluators to express their individual stylistic preferences. Moreover, it led to evaluators often marking up many consecutive words even in cases where only some of the words were actually part of an error. In some of these cases, the second-stage annotator was unable to identify an error type; to indicate this, such cases were tagged as *None*.

4 Aggregated Error Scores

We investigated the following aggregated scores:

1. **error count** $C(err\_type)$
   - number of words marked as an error type
   - ranges from 0 to total number of words in the text
2. **word percentage** $\frac{C(err\_type)}{C(all\_words)} \cdot 100$
   - number of words marked as an error type divided by the total number of words in the text
   - ranges from 0 to 100
3. **error percentage** $\frac{C(err\_type)}{C(all\_errors)} \cdot 100$
   - number of words marked as an error type divided by the total number of errors in the text
   - ranges from 0 to 100

[3] https://github.com/m-popovic/QRev-annotations/tree/master/reproduction_second-round_hr

5 Inter-Annotator Agreement

We estimated inter-annotator agreement in the following three ways:

1. **word overlap** (Popović, 2021)
    - number of words marked as errors by all annotators divided by number of all words perceived as errors by any annotator
    - ranges between 0 and 100, higher value indicating higher agreement
    - not based on aggregated error scores, only on actual words annotated as errors
2. **Krippendorff's** α (Krippendorff, 2004)
    - compares numerical scores obtained by different annotators
    - ranges between -1 and 1
    - 0 indicates no agreement, and common practice is to consider $\alpha \geq 0.667$ as acceptable and $\alpha \geq 0.8$ as strong agreement (Krippendorff, 2004)
3. **Pearson's** r
    - compares numerical scores obtained by different annotators
    - ranges between -1 and 1
    - absolute values from 0.8 to 1 are commonly considered as strong correlation, between 0.6 and 0.8 as high correlation, between 0.4 and 0.6 as moderate, between 0.2 and 0.4 as fair, and between 0 and 0.2 as weak (0 meaning no correlation at all)

Since Krippendorff's α and Pearson's r are measures computed on numerical values, we compared their values as obtained when using the three different types of aggregated scores above (raw counts, word percentage and error percentage). We additionally compared these values with word overlap values as used in previous work and calculated in a different way, independently of numerical scores.

Word overlap as well as Pearson's r are calculated on all pairs of annotators (e1-e2, e1-e3, e1-e4, e2-e3, e2-e4, e3-e4).

6 Results

As a first step, we calculated IAA measures for all error types combined, taking into account only whether a word is tagged as an error or not, independently of the error type. We compared IAA measures for all aggregation methods except calculating coefficients on error percentage (count of a particular error type normalised over the total number of errors) because without distinguishing error types it would simply be 100% and therefore does not make sense.

After observed certain tendencies in this first step, we calculated all values for each individual error type and analysed all observations in depth.

6.1 Overall results

Table 2 shows error levels as perceived by each of the four annotators in the two versions, raw count and word percentage, together with the corresponding agreement measures. As already mentioned, word overlap is not based on the quantity of errors but on actual tagged words.

First, it can be seen that agreement between annotators is higher in all cases when assessing Adequacy than Comprehension, which confirms results from previous work (Popović, 2021) where however only word overlap was reported. It should be noted that word overlap is slightly different in the present context, because it was calculated on a different subset, but the tendency is the same.

As for comparing different IAA measures, it can be noted that both coefficients, α and r, are lower for word percentage than for raw count, in all cases. Manual inspection revealed that the reason lies in variations of sentence length between annotators caused by different numbers of omission tags and/or differences in tokenisation affected by annotations.

Two examples are shown in Table 3, the sentence in the top half of the table illustrates the effect of omission tags and the sentence in the bottom half the effect of altering tokenisation through annotation. In the first sentence, all four annotators marked two errors, so the agreement on the raw error counts is perfect. However, since the fourth annotator perceived one omission error while the other three did not, this sentence became longer than others due to the added omission tag, and the word percentage became smaller. If omission tags were excluded from sentence length, word percentages for segments with omissions could become overly high and difficult to interpret.

In the second sentence in Table 3, three annotators marked two errors and one did not mark any.

		amount of errors				coefficients		
		ev1	ev2	ev3	ev4	$\alpha \uparrow$	$r \uparrow$	*overlap* $\uparrow$
Adequacy	counts	3282	3377	3910	4310	*.705*	*.714*	*59.6*
	word %	20.3	20.9	24.0	26.5	*.567*	*.579*	
Comprehension	counts	3380	3684	4336	5487	*.659*	*.687*	*57.4*
	word %	20.9	22.8	26.6	33.5	*.496*	*.523*	

Table 2: Results for all error types combined: error count and word percentage (count normalised by total number of words) for each of four annotators together with coefficients of agreement: word overlap, Krippendorff's α and Pearson's r.

original	annotated MT output	# errors	# words	word%
Usually a fan but not impressed	Obično ventilator|AMBIGUITY , ali neimpresioniran|NON_EXISTING	2	5	40.0
	Obično ventilator|AMBIGUITY , ali neimpresioniran|NON_EXISTING	2	5	40.0
	Obično ventilator|AMBIGUITY , ali neimpresioniran|NON_EXISTING	2	5	40.0
	Obično ventilator|AMBIGUITY , ali **XX|OMISSION** neimpresioniran	2	**6**	**33.3**
Wayne's World is hardly a plot driven film.	Wayneov|NE svijet|NE teško da je film pokrenut **zapletom.**	2	8	25.0
	Wayneov|NE svijet|NE teško da je film pokrenut **zapletom.**	2	8	25.0
	Wayneov svijet teško da je film pokrenut|NOUN_PHRASE **zapletom|NOUN_PHRASE .**	2	**9**	**22.2**
	Wayneov svijet teško da je film pokrenut **zapletom.**	0	8	0.0

Table 3: Examples of two sentences resulting in different lengths after annotation by inserting omission mark (above) and separating punctuation mark (below). This leads to different word percentages for same error counts.

However, the first three annotators marked different errors, one of them involved a word next to a punctuation mark. Therefore, the annotator separated the word from the punctuation mark increasing the total number of words and decreasing the word percentage.

6.2 Different error types

The next step in our experiment was to investigate the effects on different error types. Here, we want to compare the numerically based coefficients α and r when calculated on counts vs. word percentage separately for each error type. In addition, we want to investigate the error percentage (count of a particular error type normalised over total number of errors, see Section 4).

6.2.1 Aggregated error scores

The list of error types together with their quantification in the three forms (raw count, word percentage and error percentage) can be seen in Table 4, sorted by the frequency in the analysed corpus. It can be noted that the conclusions about their distribution will be exactly the same for each of the three numerical scores. For example, ambiguity is one of the predominant types according to raw count, word percentage as well as error percentage.

The question is now what happens with agreement coefficients for each of the error types when we use each of the three different numerical scores? Will word percentage be associated with lower agreement for each error type? What will happen when using error percentage?

6.2.2 Agreement measures for different error types

Tables 5 and 6 present all agreement measures, word overlap and both number-based coefficients α and r, calculated on three types of scores: count, word percentage and error percentage. The error types are ordered from lowest to highest word overlap.

It can be noted that there is very low agreement for the tag *None*, which could be expected since, as explained in Section 3, these error marks are related to evaluators' stylistic preferences as well as different perceptions of the word span.

As the results for rare error types are likely to be less reliable we also mark the frequency of each type in the first columns of Tables 5 and 6: '++'

(a) Adequacy

Adequacy issue type	count	word%	err%
REPHRASING	3197	4.93	21.49
AMBIGUITY	1841	2.84	12.37
NOUN PHRASE	1006	1.55	6.76
MISTRANSLATION	651	1.00	4.38
VERB FORM	618	0.95	4.15
NAMED ENTITY	561	0.86	3.77
CASE	529	0.82	3.56
GENDER	424	0.65	2.85
UNTRANSLATED	387	0.60	2.60
PRONOUN	338	0.52	2.27
NEGATION	333	0.51	2.24
OMISSION	288	0.44	1.94
ORDER	266	0.41	1.79
-ING	245	0.38	1.65
NON-EXISTING	216	0.33	1.45
SOURCE ERROR	207	0.32	1.39
PREPOSITION	189	0.29	1.27
POS AMBIGUITY	161	0.25	1.08
ADDITION	102	0.16	0.69
PASSIVE	101	0.16	0.68
NUMBER	84	0.13	0.56
CONJUNCTION	73	0.11	0.49
REPETITION	33	0.05	0.22
SR	18	0.03	0.12
HALLUCINATION	15	0.02	0.10
None	3755	5.79	25.24

(b) Comprehension

Comprehension issue type	count	word%	err%
REPHRASING	3368	5.18	19.94
AMBIGUITY	1670	2.57	9.89
NOUN PHRASE	992	1.53	5.87
NAMED ENTITY	594	0.91	3.52
VERB FORM	567	0.87	3.36
MISTRANSLATION	553	0.85	3.27
CASE	539	0.83	3.19
GENDER	428	0.66	2.53
UNTRANSLATED	412	0.63	2.44
NEGATION	344	0.53	2.04
PRONOUN	322	0.50	1.91
ORDER	283	0.44	1.68
OMISSION	261	0.40	1.55
-ING	244	0.38	1.44
NON-EXISTING	219	0.34	1.30
SOURCE ERROR	212	0.33	1.26
PREPOSITION	179	0.28	1.06
POS AMBIGUITY	133	0.20	0.79
ADDITION	109	0.17	0.65
PASSIVE	100	0.15	0.59
CONJUNCTION	79	0.12	0.47
NUMBER	71	0.11	0.42
REPETITION	34	0.05	0.20
SR	20	0.03	0.12
HALLUCINATION	7	0.01	0.04
None	5904	9.08	34.96

Table 4: Error levels for different Adequacy and Comprehension error types annotated by all evaluators, in the form of counts, word percentages and error percentages.

denotes error types accounting for more than 5% of all errors, '+' denotes error types accounting for 2–5%, '−' for 1–2%, and '−−' for less than 1%.

It can further be noted that for the majority of error types, the coefficients calculated on error counts indicate the same level of agreement as the word overlap. For several error types, however, the overlap is smaller (conjunction, negation, rephrasing), but this could be expected: these error types have long word spans, so annotators often perceive the same number of errors but mark different words.

As for comparing numerically based coefficients calculated on the three scores, bold values indicate that the agreement is lower than the agreement calculated on raw counts, while underlined values indicate higher agreement than counts. Overall, for a number of error types, using error percentage results in lower agreement than using word percentage or raw counts. For some types, word percentage is lower, too, as observed in the overall values in Table 2. For a few types, however, word percentage results in higher agreement. These are the Adequacy-related omission and verb form errors as well as the Comprehension-related noun-phrase and source errors. For the Comprehension-related conjunction error, the highest agreement is obtained when using error percentage.

In order to understand these observations, we further analysed all error types where conclusions about IAA differ for different numerical scores.

7 Analysis of Differing Agreement Levels

Decreased agreement when using error percentage. For a large number of error types, calculating Krippendorff's α and Pearson's r on error percentage results in much lower agreement compared to using raw counts or word percentage. In order to explain this phenomenon, further analysis was carried out on these error types, and a systematic pattern was found: the total number of errors marked by different annotators often varies notably so that identical error counts become very different error percentages.

Table 7 illustrates the phenomenon on two sentences. In the first sentence (top half of the table), two annotators perceived one mistranslation. However, for the first annotator, this was the only error in the sentence, while the other spotted a problem with the verb form. Therefore, the total number of errors for the first annotator is one and for the second one is two, resulting in double the error percentage of mistranslations for the first annotator.

A issue type	word overlap	$\alpha \uparrow$ count	$\alpha \uparrow$ word %	$\alpha \uparrow$ error %	$r \uparrow$ count	$r \uparrow$ word %	$r \uparrow$ error %
OMISSION^{-}	26.6	.236	.645	.275	.238	.650	.275
CONJUNCTION^{-}	53.0	.700	.767	.703	.706	.767	.860
ORDER^{-}	58.1	.554	.558	.500	.577	.599	.500
NEGATION^{+}	59.0	.713	.744	.778	.714	.744	.783
NAMED ENTITY^{+}	66.9	.748	.617	.647	.748	.619	.647
PREPOSITION^{-}	66.0	.689	**.649**	**.503**	.691	.649	**.583**
PRONOUN^{+}	66.5	.667	**.562**	**.454**	.667	**.562**	**.454**
REPHRASING^{++}	68.4	.772	.757	.747	.776	.762	.748
REPETITION^{--}	68.8	.879	.919	.841	.905	.933	.844
SR^{--}	70.4	.703	.698	**.580**	.703	.699	**.581**
NOUN PHRASE^{++}	70.8	.797	.757	.749	.798	.762	.749
GENDER^{+}	72.8	.758	**.523**	**.580**	.758	**.543**	**.584**
CASE^{+}	74.8	.800	**.763**	**.703**	.800	**.766**	**.707**
AMBIGUITY^{++}	75.2	.791	.744	**.596**	.794	.745	**.596**
POS AMBIGUITY^{-}	75.4	.889	**.799**	**.752**	.890	.813	**.752**
NUMBER^{--}	76.2	.771	.767	**.517**	.772	.773	**.519**
ADDITION^{--}	76.5	.742	.736	**.491**	.743	.741	**.496**
VERB FORM^{+}	76.6	.764	.825	**.650**	.764	.825	**.653**
PASSIVE^{--}	77.2	.829	**.754**	**.672**	.831	**.774**	**.690**
MISTRANSLATION^{+}	85.0	.941	.885	**.709**	.941	.885	**.718**
UNTRANSLATED^{+}	87.3	.961	.939	**.768**	.964	.939	**.771**
-ING^{-}	88.1	.899	**.751**	**.707**	.900	**.751**	**.707**
SOURCE ERROR^{-}	88.6	.884	.879	**.702**	.885	.879	**.704**
NON-EXISTING^{-}	90.7	.943	.933	**.786**	.943	.933	**.791**
HALLUCINATION^{--}	93.3	.982	.982	.997	.983	.983	.998
None	21.5	.233	.125	.131	.245	.138	.141

Table 5: Agreement coefficients for Adequacy error types ('++' denotes error types accounting for more than 5% of all errors, '+' denotes error types accounting for 2–5%, '−' for 1–2%, and '−−' for less than 1%) : word overlap together with α and r calculated on counts, word percentages and error percentages. Bold values indicate lower agreement than counts, underline values indicate higher agreement than counts.

In the second sentence (lower half of the table), all four annotators perceived one ambiguity error, but due to differences in perception of other error types (in this case presence and span of negation errors) the total number of errors, and therefore the error percentage, are different.

This finding indicates that, regardless of which score is considered best for reporting the results of the analysis, error percentage is not suitable for calculating agreement coefficients.

Increased agreement when using word percentage. For some error types, coefficients calculated on word percentage are associated with higher agreement than those calculated on raw counts and error percentage. One of these error types is Adequacy-related omission, which at first might look contradictory to the findings in the overall scores, where they contribute to decreased agreement by changing sentence length. However, increasing sentence length has another effect, namely 'smoothing' the number of omissions. An example can be in Table 8: one annotator did not perceive any omission, two perceived one omission, while one perceived two. The resulting sentence lengths are therefore different, and the increase of the sentence length is larger in the sentence with more omissions. Therefore, the difference between the amounts of omissions are smaller for word percentage than for count: while one evaluator tagged twice as many omissions than the other (2:1), the difference between word percentages is only 1.6 (40:25).

Besides omission, such tendencies can be seen in a few other error types (verb form error for Adequacy, noun phrase and source errors for Comprehension). However, the analysis revealed that in those cases, there are no reasons related to the nature of the error type itself (as is the case for omissions). The only reason is that these error types often occur in sentences where lengths are different due to tokenisation changes and/or omission annotations, as mentioned in Section 6.1.

Other differences in agreement. As previously mentioned, we carried out a more in-depth analysis of a few other kinds of differences in agreement related to certain error types.

C issue type	word overlap	$\alpha\uparrow$ count	$\alpha\uparrow$ word %	$\alpha\uparrow$ error %	$r\uparrow$ count	$r\uparrow$ word %	$r\uparrow$ error %
OMISSION^{-}	17.9	.160	.116	.092	.161	.119	.094
HALLUCINATION^{--}	28.6	.320	.320	.331	.374	.374	.362
CONJUNCTION^{--}	50.6	.712	.718	.843	.712	.718	.847
PRONOUN^{-}	58.4	.581	.404	**.320**	.590	.496	**.322**
ORDER^{-}	59.3	.552	.534	.516	.575	.553	.520
NEGATION^{+}	63.2	.758	.712	.838	.760	.713	.838
NAMED ENTITY^{+}	64.3	.684	**.600**	**.569**	.690	.604	**.570**
PREPOSITION^{-}	68.2	.694	.764	**.582**	.702	.764	**.583**
REPETITION^{--}	69.8	.879	.828	**.731**	.918	.875	**.777**
SR^{--}	70.0	.699	.719	**.634**	.702	.719	.658
GENDER^{+}	71.5	.727	**.609**	**.533**	.728	.617	**.533**
NOUN PHRASE^{++}	71.9	.791	.809	.737	.796	.812	.737
REPHRASING^{++}	72.0	.787	.748	.791	.794	.754	.791
VERB FORM^{+}	73.1	.776	.743	**.627**	.777	.745	**.627**
POS AMBIGUITY^{--}	73.7	.827	**.226**	.719	.827	**.261**	**.720**
AMBIGUITY^{++}	74.4	.783	.734	**.586**	.789	.744	**.589**
CASE^{+}	76.9	.784	.746	**.685**	.785	.748	**.690**
NUMBER^{--}	78.9	.789	.776	**.465**	.791	.777	**.466**
PASSIVE^{--}	79.3	.854	.866	.837	.858	.873	.837
SOURCE ERROR^{-}	81.4	.787	.833	.708	.793	.837	.711
MISTRANSLATION^{+}	82.2	.918	.850	**.646**	.921	.854	**.647**
ADDITION^{--}	83.2	.842	.796	**.607**	.797	.731	**.575**
-ING^{-}	85.8	.888	.824	**.700**	.893	.825	**.700**
UNTRANSLATED^{+}	87.7	.933	.870	**.775**	.934	.871	**.775**
NON-EXISTING^{-}	89.8	.920	.938	**.722**	.920	.938	**.725**
None	29.1	.317	.190	.212	.355	.218	.235

Table 6: Agreement coefficients for Comprehension error types ('++' denotes error types accounting for more than 5% of all errors, '+' denotes error types accounting for 2–5%, '−' for 1–2%, and '−−' for less than 1%): word overlap together with α and r calculated on counts, word percentages and error percentages. Bold values indicate lower agreement than counts, underline values indicate higher agreement than counts.

original	annotated MT output	# mistrans.	# errors	error%
Don't waste your money.	Nemoj|VERB trošiti novac.	0	1	0.00
	Nemoj trošiti|MISTRANSLATION novac.	1	1	100.0
	Nemoj trošiti novac.	0	0	0.00
	Nemoj|VERB trošiti|MISTRANSLATION novac.	1	2	50.0

original	annotated MT output	# ambiguity	# errors	error%
Sadly, I can't review them as they were both non-responsive.	Nažalost, ne mogu ih pregledati|AMBIGUITY jer oboje nisu reagirali	1	1	100
	Nažalost, ne mogu ih pregledati|AMBIGUITY jer oboje|NEGATION nisu|NEGATION reagirali	1	3	33.3
	Nažalost, ne mogu ih pregledati|AMBIGUITY jer oboje nisu|NEGATION reagirali|NEGATION	1	3	33.3
	Nažalost, ne mogu ih pregledati|AMBIGUITY jer oboje nisu reagirali|NEGATION	1	2	50.0

Table 7: Examples of sentences with identical error counts for a particular error type (mistranslation at the top and ambiguity below) but different total number of errors perceived by different annotators. This leads to different error percentages for same error counts.

One phenomenon is that both word percentage and error percentage decrease the agreement for certain error types including gender, case (Adequacy), named entity (Comprehension), etc. We have also observed cases where word percentage notably decreases agreement while error percentage does not, e.g. for POS ambiguity (Comprehension); in fact, conjunction error percentage even increases agreement (Comprehension).

Nevertheless, these differences in agreement are not related to error type, but rather to the circumstances in which the majority of errors occur (variations in sentence length and total number of errors). Moreover, many of the error types are relatively rare, which makes these effects even stronger. For example, conjunction errors are very rare, and al-

original	annotated MT output	# omissions	# words	word%
...from other of the genre, ...	... od ostalih žanra, ...	0	3	0.0
	... od ostalih XX|OMISSION XX|OMISSION žanra, ...	2	5	40.0
	... od ostalih XX|OMISSION žanra, ...	1	4	25.0
	... od ostalih XX|OMISSION žanra, ...	1	4	25.0

Table 8: Examples illustrating reduction in differences in omission error levels.

most all of them are marked by annotators which assign a higher total number of errors, too, so that error percentage is 'smoother' than raw error count.

It should also be noted that while the results reported here indicate that error percentage is often associated with misleadingly *low* agreement, it is theoretically possible to find misleadingly *high* agreement (for example, if one annotator marks twice as many errors in total than another, but the proportions of error types are exactly the same, the agreement on error percentages will be perfect).

8 Conclusions

This paper examined different ways of aggregating scores for error annotation in automatically generated text, more specifically MT outputs: we compared raw error counts, error counts normalised over total number of words (word percentage), and error counts normalised over total number of errors (error percentage), and the associated difference in inter-annotator agreement measures calculated on the different aggregated scores. While reporting each type of aggregated score as the evaluation result has its advantages depending on the goal of the evaluation, our experiments indicate that the overall best way to estimate inter-annotator agreement using such scores are raw counts. If the annotation process ensures that the total number of words cannot differ among the annotators (for example, due to adding omission symbols or separating punctuation marks), word percentage will lead to the same conclusions.

In contrast, error percentage can be associated with misleading agreements for a number of error types, chiefly because the total number of errors is subjective as different annotators often perceive different numbers of errors in the same text. Therefore, normalising one subjective number (error count for a particular type) by another subjective number (total number of errors) can notably influence the agreement.

Limitations

The work was carried out only on one language pair, for one translation direction. The identification of the error types in the second stage was carried out by a single annotator.

Ethics Statement

This statement follows the structure of the ARR responsible research checklist. We discuss limitations of the work presented in this paper in the previous section. No new data or computational resources were created, no computational experiments were run, no human annotation or evaluations were carried out for this paper. The work computes and analyses scores obtained from a previously annotated corpus. Results are of potential use in improving comparability and reliability in quantified error reporting.

Acknowledgements

The ADAPT SFI Centre for Digital Media Technology which is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106.

References

Khetam Al Sharou and Lucia Specia. 2022. A taxonomy and study of critical errors in machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 171–179.

Riadh Belkebir and Nizar Habash. 2021. Automatic error type annotation for arabic. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 596–606.

Helena Caseli and Marcio Inácio. 2020. Nmt and pbsmt error analyses in english to brazilian portuguese automatic translations. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3623–3629.

Sheila Castilho. 2020. On the same page? comparing inter-annotator agreement in sentence and document level human machine translation evaluation.

In *Proceedings of the Fifth Conference on Machine Translation (WMT 20)*, pages 1150–1159, Online.

Ângela Costa, Wang Ling, Tiago Luís, Rui Correia, and Luísa Coheur. 2015. A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation*, 29(2):127–161.

Marcello Federico, Matteo Negri, Luisa Bentivogli, and Marco Turchi. 2014. Assessing the impact of translation errors on machine translation quality with mixed-effects models. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1643–1653.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 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.

Jie He, Bo Peng, Yi Liao, Qun Liu, and Deyi Xiong. 2021. Tgea: An error-annotated dataset and benchmark tasks for textgeneration from pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6012–6025.

Filip Klubička, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2018. Quantitative Fine-grained Human Evaluation of Machine Translation Systems: A Case Study on English to Croatian. *Machine Translation*, 32(3):195–215.

Julia Kreutzer, Nathaniel Berger, and Stefan Riezler. 2020. Correct Me If You Can: Learning from Error Corrections and Markings. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT 2020)*.

Klaus Krippendorff. 2004. *Content Analysis, an Introduction to Its Methodology*. Sage Publications.

Arle Lommel, Maja Popović, and Aljoscha Burchardt. 2014a. Assessing inter-annotator agreement for translation error annotation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 14)*, Reykjavik, Iceland.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014b. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumatica*.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014c. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció*, (12):455–463.

Maja Popović. 2021. Agree to Disagree: Analysis of Inter-Annotator Disagreements in Human Evaluation of Machine Translation Output. In *Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL 2021)*, Punta Cana, Dominican Republic.

Maja Popović and Anya Belz. 2021. A reproduction study of an annotation-based human evaluation of MT outputs. In *Proceedings of the 14th International Conference on Natural Language Generation (INLG 2021)*, Aberdeen, Scotland, UK.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2022)*, pages 175–190, Seattle, United States.

Craig Thomson and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation (INLG 2020)*, pages 158–168, Dublin, Ireland.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).