On nature and causes of observed MT errors

Maja Popović maja.popovic@adaptcentre.ie
ADAPT Centre, School of Computing, Dublin City University, Ireland

Abstract

This work describes analysis of nature and causes of MT errors observed by different evaluators under guidance of different quality criteria: adequacy, comprehension, and a not specified generic mixture of adequacy and fluency. We report results for three language pairs, two domains and eleven MT systems. Our findings indicate that, despite the fact that some of the identified phenomena depend on domain and/or language, the following set of phenomena can be considered as generally challenging for modern MT systems: rephrasing groups of words, translation of ambiguous source words, translating noun phrases, and mistranslations. Furthermore, we show that the quality criterion also has impact on error perception. Our findings indicate that comprehension and adequacy can be assessed simultaneously by different evaluators, so that comprehension, as an important quality criterion, can be included more often in human evaluations.

1 Introduction and related work

Machine translation (MT), like many other natural language generation tasks, is difficult to evaluate because there is no single correct output for a given input: for each source text, there is a large set of possible correct translations. Therefore, while costly both in time and resources, human evaluation is required to provide a reliable feedback for measuring MT quality and progress, as well as to serve as a gold standard for development of automatic evaluation metrics. While better and better automatic metrics are constantly emerging (Mathur et al., 2020; Ma et al., 2019), many of them being based on semantic word representations (embeddings), all of them represent only an approximate substitution for human assessment of translation quality. Various methods have been proposed and used for the human evaluation of MT output from the beginning of MT until now (ALPAC, 1966; White et al., 1994; Koehn and Monz, 2006; Vilar et al., 2007; Graham et al., 2013; Forcada et al., 2018; Barrault et al., 2020; Kreutzer et al., 2020; Popović, 2020a), and all of them are essentially based on some of the following three quality criteria: adequacy (how much meaning is preserved), comprehensibility (how comprehensible/readable the translation is) and fluency (grammar of the target language).

The evaluators are usually asked to assign an overall quality score for the given MT output (ALPAC, 1966; White et al., 1994; Koehn and Monz, 2006; Roturier and Bensadoun, 2011; Graham et al., 2013; Barrault et al., 2020) or to rank two or more competing outputs from best to worst (Vilar et al., 2007; Callison-Burch et al., 2008; Bojar et al., 2015). For assessing comprehension, question answering (Scarton and Specia, 2016) and filling gaps (Forcada et al., 2018) were explored, too. Recently, evaluators have been asked to highlight the observed translation errors (Kreutzer et al., 2020; Popović, 2020a).

In order to get more details about the actual errors, error classification according to a predefined error scheme is often performed. The mostly applied schemes have been the one proposed

by Vilar et al. (2006), and the MQM scheme[1] (Lommel et al., 2014) in recent years (Klubička et al., 2018; Freitag et al., 2021).

Another method to better understand particular strengths and weaknesses of MT systems is to identify nature and causes of the errors in form of linguistically motivated phenomena which, although related, often go beyond the usual error types. This type of analysis is being increasingly employed in the last years in order to better understand the ocurring errors (Popović, 2018; Arnejšek and Unk, 2020) and also to create specialised test sets ("challenge test sets" or "test suites") in order to perform more focussed evaluation procedures on identified phenomena (Isabelle et al., 2017; Šoštarić et al., 2018; Voita et al., 2019).

This work goes in this direction, but in a slightly different way: we do not try to identify the phenomena from scratch, but from translation errors already observed and highlighted by several evaluators (Kreutzer et al., 2020; Popović, 2020a). The error marking was not guided by any pre-defined error scheme, so that the evaluators had more freedom in annotating errors than in typical error classification tasks such as MQM.

We analysed the nature of these errors by tagging them with possible causes and/or plausible explanations of their origin (referred to as "phenomena"). The definition of these phenomena is based both on general linguistic knowledge as well as on phenomena related to the (machine) translation process. We did not have any pre-defined scheme for the phenomena, but we started by looking into errors and identifying the phenomena on the fly.

It is worth noting that we did not create any test suite – we do not know how many instances of each of the identified phenomena exists in the data in total, nor how many of them are correctly translated. We only analyse the observed translation errors. Nevertheless, our findings can be inspiring and useful for future work on creation of test suites.

The main goal of this work is to identify nature and causes of translation errors perceived by a set of evaluators and to get a better insight about the underlying phenomena and their impact on translation quality. In addition, we investigate the perception of major and minor errors, and also explore perception of errors for two different quality criteria: adequacy and comprehension.

We used two publicly available data sets containing English→Croatian, English→Serbian and English→German MT outputs with highligted translation errors. We first identified a set of 26 underlying phenomena around these errors and then analysed them.

2 Data sets

We worked on two publicly available data sets with highlighted MT errors: one provided by Dublin City University (*DCU*)[2] and one provided by Heidelberg University (*HU*).[3] While both data sets contain MT outputs with highlighted translation errors, there are several important differences between them.

***DCU* data set** This data set was created for purposes of MT evaluation (Popović, 2020a). The set consists of English user reviews translated into Croatian and Serbian. For each of the target languages, five different MT systems were used: three online systems (Amazon, Bing and Google) and two in-house systems based on the Sockeye[4] (Hieber et al., 2018) implementation. In total, the data set contains outputs of ten different MT systems.

Two quality criteria were used for highlighting errors: adequacy and comprehension. An important difference between the two (apart from the definition) which can lead to differences

[1] http://www.qt21.eu/mqm-definition/definition-2015-12-30.html
[2] https://github.com/m-popovic/QRev-annotations
[3] https://www.cl.uni-heidelberg.de/statnlpgroup/humanmt/
[4] https://github.com/awslabs/sockeye

in perception of errors is that seeing the source text was *required* for adequacy while seeing the source text was *forbidden* for comprehension. For both quality aspects, the evaluators were asked to concentrate on problematic parts of the text and to highlight them. They were also asked to distinguish between major and minor errors. All translations were evaluated in context – the evaluators were seeing entire reviews.

In total, 15 evaluators participated in the annotation. The largest part of the text is annotated by two evaluators, while a small part of the text (about 40 sentences) is annotated by three or four evaluators. Nothing is annotated by a single evaluator. Inter-annotator agreement in terms of Krippendorf's α is 0.61 for adequacy errors and 0.51 for comprehension errors.

***HU* data set** This data set was not created for purposes of MT evaluation, but for improving an NMT system by giving it feedback about errors (Kreutzer et al., 2020). The set consists of English TED talks translated into German by one MT system, an in-house system based on the Joey NMT[5] (Kreutzer et al., 2019) implementation.

A very important difference in comparison to the *DCU* data set is that no specific quality criterion was used: the evaluators were only asked to "highlight the errors". Usually, such "generic" criterion represents a mixture of adequacy and fluency. Also, they were not asked to distinguish between major and minor errors. Another very important fact is, since the data set is created in order to improve a system, and the used loss function did not support omissions and reordering errros, the evaluators are specifically asked not to highlight these two types of errors. As for context, translated sentences were judged in isolation, however in consecutive order as they appeared in the original documents so that a reasonable amount of context was provided.

Ten evaluators participated in this annotation, although the largest part of the text is annotated by a single evaluator. Eleven sentences are, however, annotated by all ten evaluators and the reported Krippendorf's α is 0.201.

data set	language pairs	domain	# of segments	# of MT systems	quality criterion	% of marked errors
DCU	en→sr,hr	user reviews	3334	10	adequacy	20.9
			3334	10	comprehension	24.1
HU	en→de	TED talks	302	1	not specified	13.7

Table 1: Statistics of the two analysed data sets containing MT outputs with highlighted errors.

An overview of the two data sets together with the overal percentage of highlighted words is presented in Table 1. The number of errors in the *HU* data set might be underrated due to unmarked reordering errors and omissions.

3 Identified phenomena

The errors in the described data sets were analysed in the following way: they were tagged as a particular phenomenon if 1) they were marked by at least one evaluator 2) it was possible to define a plausible cause and/or explanation for their origin. In order to motivate and facilitate future work of creating test suites and getting ideas for potential improvements of MT systems, we also tagged all corresponding English words. The analysed data sets with phenomena tags are available together with the original *DCU*[6] data set.

The identified phenomena are different by their nature: some of them are equivalent to the typical error classes (such as "mistranslation", "tense/aspect/mood") while some are going

[5] https://github.com/joeynmt/joeynmt

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

far beyond that, often bringing on several different interwining types of errors. Some of them involve single words, while others might involve a large group of words, even entire sentences. For the phenomena with larger spans, we tagged all consecutive words although not necessarily all those words are marked as errors. A typical example is negation where all words within the negation span were considered as "negation" although the evaluators might perceive only some of the words as problematic. In total, we identified 26 phenomena which will now be described and explained in alphabetical order.

ambiguity Ambiguous source words are identified as one of the most frequent causes for observed errors.
An ambiguous word is a word which can have multiple meanings, depending on the context. The translation of such word is in principle correct, but not in the given context. For example, the English verb "play" has different meanings in sentences "The children are playing in the park" and "The children are playing piano".

case Morphological form of a word (inflection) denotes incorrect case.

conjunction If a conjunction in the source language is omitted (typical for English), it can result in incorrect translation with different types of errors (lexical, morphological, order). For example, "Did you know I bought a new bike?" vs "Did you know *that* I bought a new bike?", the first sentence can provoke errors in all investigated target languages because they require a conjunction. The phenomenon involves several words around the conjunction.

determiner Incorrect or added determiner.

extra word Word(s) is/are added in the translation.

gender Morphological form of a word (inflection) denotes incorrect gender.

hallucination Translation is absolutely unrelated to the source text. For example, if the source text "Hi, how are you" is translated into "Hi, how it's going, shall we meet tomorrow?", "shall we meet tomorrow" is considered as hallucination.

"ing"-word English words with the suffix "ing" can denote present continous tense, gerund, or a noun, which might be difficult to translate properly.

mistranslation Mistranslation is one of the most frequent causes for the highlighted translation errors. It refers to an incorrect translation of the given word or phrase.

named entity A named entity generated in the target language is incorrect for some of the following reasons or a combination of them: 1) incorrectly translated 2) untranslated 3) unnecessarily translated 4) incorrectly transcribed 5) incorrect case/gender/number.
Errors related to named entities are quite frequent in user reviews, however very rare in TED talks. Also, named entities are generally easier to handle in German than in Croatian and Serbian.

negation Missing negation marker(s), added negation marker(s), or incorrectly formed negation structure involving different types of errors. The phenomenon involves all words within the negation span, possibly entire sentence.

non-existing word A word in translation does not exist either in the source or in the target language. Includes non-existing morphological variants as well as completely invented words.

noun phrase Noun phrases also belong to the most frequent causes of the highlighted translation errors. An English noun phrase consists of a head noun and additional nouns and adjectives.

Its translation can result in different types of often interwined errors (lexical, morphological, omissions, order) because formation rules for Serbian and Croatian are rather different than for English and there is often no unique solution. And even though formation rules in German are similar to the English ones, translation errors are still occurring. The examples in Table 2 represent four English noun phrases and their correct translations into Serbian, Croatian or German, together with some of the observed erroneous translations.

domain	language	noun phrase
user reviews	EN source	grill cover
	SR/HR correct	poklopac za roštilj
	MT outputs	roštilj poklopac, roštilj
	EN source	bird feeder
	SR/HR correct	hranilica za ptice
	MT outputs	hranilica ptica, ptica hranilica
TED talks	EN source	traveling salesman problem
	DE correct	Problem des Handlungsreisenden
	MT output	Reisen Verkäufer Problem
	EN source	slime mold
	DE correct	Schleimpilz
	MT output	Schlamm, Schlamm mold

Table 2: Examples of noun phrases.

number Morphological form of a word (inflection) denotes incorrect number.

omission Word(s) is/are missing in the translation: either a part of the source text is omitted, or something is not complete in the target language. This type of error cannot been found in the *HU* corpus because the evaluators were specifically instructed not to highlight it.

order Word(s) in the translation is/are at incorrect position(s). Although the evaluators of the *HU* corpus were instructed not to highlight this type of errors, a small amount of marked errors could be related to order.

passive Passive voice appears in the translation where active voice should be used, or other way round.

person (subject-verb agreement) Morphological form of a verb (inflection) denoting person does not correspond to the subject.

POS ambiguity A source word which can be interpreted as different POS tags. For example, the English word "works" can be plural of the noun "work" or third person singular of the verb "to work".

preposition Incorrect or added preposition.

pronoun Incorrect or added pronoun.

repetition Word(s) is/are unnecessarily repeated in the translation.

rephrasing Rephrasing is ranked as the most frequent cause for observed errors in all analysed data sets. It always affects more than one word, and sometimes spans over the entire sentence.
Rephrasing refers to a sequence of source words which is not translated properly for some of the following reasons or their combination: 1) the choice of each target language word looks random, both lexically and morphologically, without taking any context into account 2) rephrasing is needed in the target language but the translation follows the structure of the source language 3) rephrasing is not needed in the target language but is applied 4) rephrasing is needed in the target language but it is incorrectly applied. The phenomenon also comprises incorrect translation of multi-word expressions and collocations. It is usually manifested by several consecutive different but interwined types of errors, such as morphological (case, gender, person/tense/mood/aspect, etc.), lexical (ambiguity, mistranslation, multi-word expression), word order, etc.

Table 3 shows six groups of English source words which had to be rephrased in the given target language. Even non-speakers of the target languages can note that the correct version and the generated MT output are significantly different in several ways (order, words, endings).

domain	language	group of words to be rephrased
user reviews	EN source	tries really hard in this one
	SR/HR correct	baš se trudi u ovom
	MT output	pokušava stvarno jako teško u ovom jednom
	EN source	it does a good job of protecting
	SR/HR correct	dobro štiti
	MT output	to radi dobar posao štititi
	EN source	nowhere close
	SR/HR correct	ni približno
	MT output	nigde nije blizu
	EN source	gets his little gray cells working
	SR/HR correct	aktivira svoje male sive ćelije
	MT output	radi na svojim malim sivim ćelijama
	EN gloss	works on his little gray cells
TED talks	EN source	you name it
	DE correct	was (auch immer) Sie wollen
	MT output	Sie bennenen es
	EN source	and so am I
	DE correct	und ich auch
	MT output	und so bin ich

Table 3: Examples of rephrasing.

In all examples except the fourth one, the translation output is rather literal, namely the system failed to apply rephrasing and the output follows the structure of the source text. In the fourth example, however, the system rephrased the source text, but the applied rephrasing was incorrect and changed the meaning.

source error A word in the original text in the source language has spelling or grammar errors which resulted in incorrect translation. This type of issue has been found in user reviews but not in TED talks.

tense/aspect/mood Morpho-syntactic form of a verb (inflection, derivation, auxiliary verb) denotes incorrect tense, aspect or mood.

untranslated A word in the source language is simply copied into the translation.

4 Distribution of the observed errors over the identified phenomena

Once the phenomena were identified and tagged, for each of them the contribution was calculated as percentage of observed errors related to it. Due to the differences between the two data sets described in Section 2 as well as the two different quality criteria in the *DCU* data set, the results in Table 4 are presented separately for each of these three texts.

The numbers should be interpreted as follows: the first number in the first column means that from all highlighted adequacy errors in the *DCU* set, 17.6% are related to rephrasing, 11.2% are related to an ambiguous source word, 7.67% are related to a noun phrase, etc. The other columns are to be interpreted in the same way (second column: "from all highlighted comprehension errors in the *DCU* set", third column: "from all highlighted errors in the *HU* set"). Phenomena contributing with at least 2% of highlighted words are shown in bold.

To errors which could not be interpreted by any particular phenomenon, a tag "None" was assigned. A number of these errors is related to individual preferences of different annotators, and therefore is less frequent in the *HU* corpus which was mainly annotated by a single evaluator. Some of these words are marked due to "error propagation", when several consecutive words are marked although only one of them is actually an errors. This effect is much stronger for comprehension, because adequacy is guided by the source text.

Table 4 presents phenomena with a contribution of at least 2% of errors in at least one of the three texts. Those with at least 2% in all three texts are presented in bold. The phenomena are sorted according to their contribution to adequacy errors in the *DCU* set, but it can be noted that the contributions are very similar for comprehension errros, and also for the *HU* set.

data set:	*DCU*		*HU*
domain:	user reviews		TED talks
language pair:	en→sr, hr		en→de
quality criterion:	adequacy	comprehension	not specified
rephrasing	**17.6**	**16.6**	**21.7**
ambiguity	**11.2**	**8.98**	**13.3**
noun phrase	**7.67**	**6.65**	**7.10**
named entity	**4.63**	**4.38**	0.07
mistranslation	**4.37**	**3.10**	**13.7**
omission	**2.94**	1.38	0 (!)
gender	**2.84**	**2.41**	1.53
case	**2.45**	**2.30**	0.66
untranslated	**2.05**	1.86	**4.11**
preposition	1.02	0.90	**3.25**
extra word	0.05	0.36	**3.25**
none	27.6	38.3	21.0

Table 4: Percentages of perceived errors related to the identified phenomena: adequacy errors in *DCU* corpus (left), comprehension errors in *DCU* corpus (middle), errors in *HU* corpus (right).

Rephrasing, **ambiguous words**, **noun phrases** and **mistranslations** have very similar (high) influence on error perception in all data sets, strongly indicating that they represent challenging phenomena for modern MT systems.

Rephrasing errors seem to be partly dependent on MT system: some systems tend to stay close to the source text (generating overly literal translations) while others tend to diverge from the source (generating incorrect rephrasings). These effects should be investigated further in more details, also by creating appropriate test suites.

As for ambiguous source words, our analysis confirmed that they represent a challenge for modern NMT systems. Several test suites have already been developed (Rios Gonzales et al., 2018; Müller et al., 2018; Raganato et al., 2019), but creating more test suites covering different types of ambiguous words and various language pairs would be certainly beneficial. It should be noted that, while translation of ambiguous words can be improved by context-aware ("document-level") NMT systems, incorporating external context often could be more helpful than extending context to more sentences. For example, if a source text is a product review, it can indicate that "I will get this part" most probably means "I will buy this part of some object", while for a movie or book review "I don't get this part" probably means "I don't understand this part of a movie/book".

Mistranslations mostly consist of simply incorrect lexical choices, however a number of them looks as "false friends". Sub-word units are the most probably reason for this type of errors, but it should be investigated further in more details.

Untranslated words contribute to errors, too, although to lesser extent. The same can be observed for **omissions**, however it has to be noted that the contribution of omissions is underrated in both analysed data sets; they are not at all marked in the *HU* corpus, and even though they are marked in the *DCU* corpus by omission mark, the evaluators mostly added one single omission mark for missing phrases. Furthermore, the nature of omissions should be investigated more, for example how many of them are related to the source text and how many to the target text. Another difference between the two data sets can be seen for **named entites**: they seem to be rather problematic only in the *DCU* corpus. Therefore, errors related to named entities are probably domain and/or language dependent.

The largest difference between the two corpora can be observed for **prepositions** and **extra words**, which resulted in much more errors in the *HU* corpus. This indicates possible dependance on domain and language, but also on MT system (since only one MT system was annotated in this corpus) and on quality criterion (because it was not specified for this corpus).

Also, contribution of **gender** and especially **case** is larger in morphologically rich(er) Slavic languages than in German. It should be noted that these two phenomena include only single-word errors *exclusively* related to gender and/or case: there are more gender and case errors, but within other phenomena with larger spans: rephrasing, noun phrase, conjunction.

4.1 Major vs minor errors

As mentioned in Section 2, the evaluators of the *DCU* data set were asked to distinguish between major and minor errors. While some of the phenomena are found to be much more frequent than others, frequency of errors is not necessarily related to their importance/severity (Federico et al., 2014; Kirchhoff et al., 2014). Therefore, we further analysed *all* identified phenomena in order to determine whether they are more related to major or to minor errors. We have, however, to take into account that for the less frequent phenomena, the results of this analysis might not be sufficiently reliable.

Perceptions of each of the phenomena in the form of percentage are shown in Table 5. The numbers are to be interpreted as follows (first row, first three columns): from all words belonging to the "rephrasing" phenomenon, 32.0% are perceived as major adequacy errors, 37.6% as minor adequacy errors, and 30.3% are not perceived as errors. These correct words are often related to the phenomena with larger word spans where not all words were perceived as errors, but also to the individual preferences of different annotators.

	adequacy			comprehension		
phenomenon	major	minor	correct	major	minor	correct
rephrasing	32.0	37.6	30.3	33.6	38.0	28.4
ambiguity	**48.2**	31.5	20.3	39.2	39.2	21.6
noun phrase	35.5	34.2	30.2	33.1	35.6	31.3
named entity	27.5	**44.3**	28.2	26.6	**44.8**	28.5
mistranslation	**68.5**	18.6	13.0	**53.2**	28.0	18.8
omission	**53.7**	46.3	0	21.6	**78.1**	0.31
gender	10.6	**69.9**	19.5	13.8	**64.1**	22.1
case	15.4	**66.7**	17.9	25.2	**59.4**	15.4
untranslated	**73.2**	13.1	13.7	**64.8**	22.7	12.5
person	27.5	**57.8**	14.6	23.1	**58.5**	18.4
tense/aspect/mood	18.7	**56.9**	24.4	25.2	**50.9**	23.4
pronoun	21.1	**53.9**	24.9	21.4	**47.9**	30.6
non-existing word	**58.9**	28.7	12.3	**57.1**	33.3	9.6
source error	**68.3**	18.5	13.2	**56.6**	27.8	15.6
negation	22.1	22.9	**55.0**	25.8	28.3	**45.8**
"-ing" word	33.9	37.6	28.5	35.0	38.3	26.7
preposition	39.1	38.8	22.1	30.4	**47.8**	21.8
POS ambiguity	**46.2**	36.6	17.2	**49.1**	32.2	18.7
order	12.7	**56.9**	30.4	18.6	**54.2**	27.1
conjunction	24.8	33.1	**42.1**	**44.1**	25.8	30.1
passive	23.5	**54.9**	21.6	21.0	**58.6**	20.4
number	11.3	**72.2**	16.5	13.3	**68.1**	18.6
repetition	39.7	40.9	19.4	21.7	**69.6**	8.7
extra word	34.9	**42.9**	22.2	26.5	**55.9**	17.6
determiner	27.8	**44.4**	27.8	18.2	**45.4**	36.4
hallucination	**87.5**	0	12.5	**50.0**	0	**50.0**
none	2.00	5.60	**92.4**	4.63	7.37	**88.0**

Table 5: Percentages of words related to each of the identified phenomena perceived as major errors, minor errors or as correct.

The phenomena are again ordered according to their overall contribution to observed adequacy errors. It can be seen that ambiguity, mistranslation and untranslated words are mostly perceived as major errors, while named entitites, gender and case as minor errors. For phenomena with larger spans, namely rephrasing and noun phrase, words are equally often perceived as major errors, minor errors or as correct. Generally, for phenomena with larger spans, a number of words is perceived as correct, especially for negation and conjunction. Interestingly, perception of conjunction-related errors is rather different for comprehension: most of the words are perceived as major errors. It indicates that many of those words are hard to read although their meaning did not change.

As for omissions, they are also perceived differently for adequacy and for comprehension: mainly as major adequacy errors, but as minor comprehension errors. The main reason for this discrepance is that many omissions are not possible to perceive without access to the source text.

As for less frequent phenomena, the following tendencies can be observed: verb forms (person, tense/aspect/mood, passive), pronouns, determiners, word order, number and extra words are mainly perceived as minor errors, while non-existing words, errors in the source

text, POS ambiguity and hallucinations are mainly perceived as major errors. Repetitions and prepositions are mostly perceived as minor comprehension errors, but equally often as major and as minor adequacy errors.

The presented results indicate not only that severity of errors is perceived differently for different phenomena, but also that perception of some phenomena depends on the quality criterion. Previous work has already shown that adequacy errors are often "masked" by good fluency (Martindale and Carpuat, 2018), and also by good comprehension (Popović, 2020b). All that motivated us to investigate the differences between quality criteria for each of the identified phenomena.

4.2 Adequacy vs comprehension

Table 6 presents discrepances between the two quality criteria: *inadequate comprehensible words* are the words which changed the meaning of the source text but are perceived as comprehensible when reading the translation. On the other hand, *adequate incomprehensible words* are the words which are perceived as incomprehensible although their meaning is preserved. The results are presented only for the most prominent and most interesting phenomena.

Apart from exploring discrepancies between adequacy and comprehension errors observed by one evaluator, we also explored these discrepancies for two different evaluators. The motivation is that evaluating both criteria can be made easier if different evaluators are working on different criteria. If one single evaluator works on both criteria (as was the case with the *DCU* corpus), they have first to finish comprehension (in order not to see the source text), and then to start with adequacy. On the other hand, different evaluators could work simultaneously, thus saving time. Furthermore, while adequacy requires high proficiency in both the source and the target language, comprehension can be evaluated by fully monolingual evaluators. The results in Table 6 show that for two different evaluators all discrepances become higher (as intuitively expected), but the tendencies remain the same.

	same evaluator for A and C		different evaluators for A and C	
phenomenon	inadequate comprehensible words	adequate incomprehensible words	inadequate comprehensible words	adequate incomprehensible words
all	33.6	42.4	45.0	51.6
non-existing word	*4.31*	*9.76*	*10.0*	*15.4*
untranslated	*11.1*	*13.8*	*16.0*	*16.9*
source error	*16.2*	*14.1*	*22.9*	*19.8*
omission	**81.7**	**65.6**	**88.2**	**77.3**
hallucination	**42.8**	0	**57.1**	25.0
mistranslation	29.3	*12.4*	31.9	*16.1*
conjunction	**44.8**	**48.5**	**52.6**	**55.8**
negation	31.7	**40.5**	40.4	**48.0**
rephrasing	24.3	29.3	33.0	36.3
ambiguity	27.8	21.2	34.6	27.6
noun phrase	24.1	23.2	32.7	32.7

Table 6: Percentages of discrepances between adequacy and comprehension for the most interesting and the most prominent phenomena.

It can be seen that overall, 33% of all adequacy errors is comprehensible and more than 40% of all incomprehensible words are adequate translations. This confirms the previous findings that good comprehensibility often "masks" adequacy errors, but also shows a tendency in the opposite direction, namely "forgiving" incomprehensible errors after seeing the source text. Some of these "forgiven" errors were result of error propagation (as explained in Section 4), though, but not all of them.

For the majority of phenomena (most of them not presented in Table 6), the percentage of discrepances for the same evaluator is ranging from 20-35% (30-45% for different evaluators).

For some phenomena, however, a much lowe discrepancy can be seen in Table 6: source errors, non-existing and untranslated words result in similar perception of errors for both quality aspects.

On the other hand, there is a large number of comprehensible omissions, over 80%. This can be intuitively expected, because evaluators cannot perceive any omission related to the source text without access to it. Also, more than 65% omissions related to comprehension are "forgiven" or perceived as different error types when looking at the source text. Another phenomenon with a high discrepancy is hallucination: this type of errors is inadequate by its definition, but is often perceived as comprehensible. An opposite effect can be observed for mistranslations which are rarely observed as comprehensible.

A high discrepancy, although much smaller than for omissions, can be seen for phenomena with large spans. For missing English conjunctions and negation, there are more incomprehensible adequate words than "masked" adequacy errors. As previously mentioned, this is partly due to error propagation, but also indicates that the reader tends to "forgive" some incomprehensible parts after seeing the source text. The same tendency can be seen for the predominant phenomenon, rephrasing, although to much less extent.

5 Summary and outlook

We have carried out an extensive analysis of MT errors observed and highlighted by different evaluators according to different quality criteria. The analysis includes three language pairs, two domains and eleven NMT systems. Our main findings show that the majority of perceived errors are caused by rephrasing, ambiguous words, noun phrases and mistranslations, followed by untranslated words and omissions.

Furthermore, it is shown that perception of errors is dependent on the pre-defined quality criterion. For example, non-existing and untranslated words, as well as errors in the source text are perceived similarly for different quality aspects, but there is a large discrepance between adequacy and comprehension errors caused by negation, hallucinations and missing English conjunctions. Therefore, the ideal evaluation would include both quality criteria. However, comprehension cannot be properly assessed if the source text is seen, so that it cannot be evaluated together with adequacy, but has to be performed beforehand as a separated task. This is time and resource-consuming, so that usually a (often unspecified) combination of adequacy and fluency is used, while comprehension, although more important than fluency, is rarely included. Our findings indicate that evaluating both adequacy and comprehension can be facilitated, because it is not necessary that the same evaluators work on both quality criteria.

The findings also open several directions for future work. For some phenomena, further analysis is recommended, for example the type of rephrasing (literal translation or not), more details about the negation (span, type of negation marker(s), etc.), source vs target omissions, etc. Test suites should also be created for some of the phenomena, in order to provide more information about errors and give ideas for potential improvements of the current systems.

Acknowledgments

This project was partially funded by the European Association for Machine Translation through its 2019 sponsorship of activities programme. The ADAPT Centre for Digital Content Technology (www.adaptcentre.ie) at Dublin City University is funded by the Science Foundation Ireland Research Centres Programme (Grant13/RC/2106) and is co-funded by the European Regional Development Fund.

References

ALPAC (1966). Language and machines. Computers in translation and linguistics.

Arnejšek, M. and Unk, A. (2020). Multidimensional assessment of the eTranslation output for English–Slovene. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT 20)*, pages 383–392, Lisboa, Portugal.

Barrault, L., Biesialska, M., Bojar, O., Costa-juss, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubei, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 conference on machine translation. In *Proceedings of the Fifth Conference on Machine Translation (WMT 20)*, pages 1–55, Online.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT 15)*, pages 1–46, Lisbon, Portugal.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2008). Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation (WMT 08)*, pages 70–106, Columbus, Ohio.

Federico, M., Negri, M., Bentivogli, L., and Turchi, M. (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 14)*, pages 1643–1653, Doha, Qatar.

Forcada, M. L., Scarton, C., Specia, L., Haddow, B., and Birch, A. (2018). Exploring gap filling as a cheaper alternative to reading comprehen sion questionnaires when evaluating machine translation for gisting. In *Proceedings of the Third Conference on Machine Translation (WMT 18)*, pages 192–203, Brussels, Belgium.

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.

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*, pages 33–41, Sofia, Bulgaria.

Hieber, F., Domhan, T., Denkowski, M., Vilar, D., Sokolov, A., Clifton, A., and Post, M. (2018). The sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA 18)*, pages 200–207, Boston, MA.

Isabelle, P., Cherry, C., and Foster, G. (2017). A Challenge Set Approach to Evaluating Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 17)*, pages 2486–2496, Copenhagen, Denmark.

Kirchhoff, K., Capurro, D., and Turner, A. M. (2014). A conjoint analysis framework for evaluating user preferences in machine translation. *Machine Translation*, 28(1):117.

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

Koehn, P. and Monz, C. (2006). Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation (WMT 06)*, pages 102–121, New York City.

Kreutzer, J., Bastings, J., and Riezler, S. (2019). Joey NMT: A minimalist NMT toolkit for novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 19)*, pages 109–114, Hong Kong, China.

Kreutzer, J., Berger, N., and Riezler, S. (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 20)*.

Lommel, A., Burchardt, A., Popović, M., Harris, K., and Avramidis, Eleftherios, a. d. U. H. (2014). Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT 14)*, pages 165–172.

Ma, Q., Wei, J., Bojar, O., and Graham, Y. (2019). Results of the wmt19 metrics shared task: Segment-level and strong mt systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (WMT 19)*, pages 62–90, Florence, Italy.

Martindale, M. and Carpuat, M. (2018). Fluency over adequacy: A pilot study in measuring user trust in imperfect MT. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA 18)*, pages 13–25, Boston, MA.

Mathur, N., Wei, J., Freitag, M., Ma, Q., and Bojar, O. (2020). Results of the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation (WMT 20)*, pages 688–725, Online.

Müller, M., Rios Gonzales, A., Voita, E., and Sennr ich, R. (2018). A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation. In *Proceedings of the 3rd Conference on Machine Translation (WMT 18)*, pages 61–72, Belgium, Brussels.

Popović, M. (2018). Language-related issues for NMT and PBMT for English–German and English–Serbian. *Machine Translation*, 32(3):237–253.

Popović, M. (2020a). Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 20)*, Online.

Popović, M. (2020b). Relations between comprehensibility and adequacy errors in machine translation output. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL 20)*, Online.

Raganato, A., Scherrer, Y., and Tiedemann, J. (2019). The MuCoW Test Suite at WMT 2019: Automatically Harvested Multilingual Contrastive Word Sense Disambiguation Test Sets for Machine Translation. In *Proceedings of the 4th Conference on Machine Translation (WMT 19)*, Florence, Italy.

Rios Gonzales, A., Müller, M., and Sennrich, R. (2018). The Word Sense Disambiguation Test Suite at WMT18. In *Proceedings of the 3rd Conference on Machine Translation (WMT 18)*, pages 594–602, Belgium, Brussels.

Roturier, J. and Bensadoun, A. (2011). Evaluation of MT Systems to Translate User Generated Content. In *Proceedings of the MT Summit XIII*, Xiamen, China.

Scarton, C. and Specia, L. (2016). A reading comprehension corpus for machine translation evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 16)*, pages 3652–3658, Portorož, Slovenia. European Language Resources Association (ELRA).

Šoštarić, M., Hardmeier, C., and Stymne, S. (2018). Discourse-related language contrasts in English-Croatian human and machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT 18)*, pages 36–48, Brussels, Belgium.

Vilar, D., Leusch, G., Ney, H., and Banchs, R. E. (2007). Human evaluation of machine translation through binary system comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation (WMT 07)*, pages 96–103, Prague, Czech Republic.

Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. (2006). Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 06)*, Genoa, Italy.

Voita, E., Sennrich, R., and Titov, I. (2019). When a good translation is wrong in context: Context-aware machine translation improves on deixis, ell ipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 19)*, Florence, Italy.

White, J., O'Connell, T., and O'Mara, F. (1994). The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the 1994 Conference of Association for Machine Translation in the Americas (AMTA 94)*, pages 193–205.