Effects of different types of noise in user-generated reviews on human and machine translations including ChatGPT

Maja Popović[1], **Ekaterina Lapshinova-Koltunski**[2], **Maarit Koponen**[3]
[1] ADAPT Centre, School of Computing, Dublin City University, Ireland
maja.popovic@adaptcentre.ie
[2] Language and Information Sciences, University of Hildesheim, Germany
lapshinovakoltun@uni-hildesheim.de
[3] School of Humanities, University of Eastern Finland
maarit.koponen@uef.fi

Abstract

This paper investigates effects of noisy source texts (containing spelling and grammar errors, informal words or expressions, etc.) on human and machine translations, namely whether the noisy phenomena are kept in the translations, corrected, or caused errors. The analysed data consists of English user reviews of Amazon products translated into Croatian, Russian and Finnish by professional translators, translation students, machine translation (MT) systems, and ChatGPT language model. The results show that overall, ChatGPT and professional translators mostly correct/standardise those parts, while students are often keeping them. Furthermore, MT systems are most prone to errors while ChatGPT is more robust, but notably less robust than human translators. Finally, some of the phenomena are particularly challenging both for MT systems and for ChatGPT, especially spelling errors and informal constructions.

1 Introduction

User-generated content (UGC) plays a great role in the information society as it facilitates fast information sharing. Therefore, translation of user-generated content is extremely important as it helps to make information accessible in other languages. There is a need for machine translation of UGC, as it facilitates cross-cultural communication by fast distribution of information across languages. Therefore, understanding problems in machine translation of user-generated reviews is important as most internet users trust the recommendations posted online, which means that their correct translation is essential. However, UGC input is still challenging for MT systems as it contains a considerable amount of noise including different types of grammar and spelling errors, emoticons and other symbols, as well as informal words and expressions including abbreviations (in this work, referred to as "noisy" or "non-standard" phenomena). The MT community has become aware of the existing problem: In WMT2022[1], the 'news' task was replaced by the 'general' task in order to include other, underinvestigated, domains such as conversations, commercial product descriptions, as well as UGC (social media posts, user reviews, Kocmi et al., 2022). However, there is no clear understanding of what exactly challenges MT systems while translating UGC.

In addition, since such reviews are commonly translated automatically, we do not know how human translators would deal with such problems.

The novelty of our study is that we analyse translation of noisy phenomena in both human and machine translations. We perform our analysis on human, machine (MT) and large language model (specifically GPT3.5) translations for the three translation directions: English-Croatian, English-Finnish and English-Russian. We analyse user reviews of Amazon products which are not so noisy as social media posts, such as Reddit and Twitter data, but still contain numerous non-standard source phenomena. Our research questions include:

RQ1 Which types of noise are typical for the English user reviews at hand?

RQ2 What are the effects of those noisy phenomena onto different translations?

RQ3 Which noisy phenomena are particularly challenging for translation?

2 Related work

Although the issues of machine translation of user-generated content have been investigated in several works, many problems remain unsolved and understudied.

[1] https://www.statmt.org/wmt22/

Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024), pages 17–30
March 22, 2024 ©2024 Association for Computational Linguistics

For instance, Roturier and Bensadoun (2011) looked into the impact of the source quality in online forums onto machine-generated translations. They evaluated several systems and came to a conclusion that especially spelling errors represent a problem. Misspelled words were also addressed by Gupta et al. (2021) who analysed user-generated reviews. Further problems that the authors focused on included ungrammatical constructions and colloquial expressions.

Another approach to improve performance is to use synthesized parallel data of UGC, as shown by Marie and Fujita (2020). Berard et al. (2019) suggested a number of strategies for dealing with non-standard issues such as emoticons, emojis and others. They included placeholders for rare characters, lowcasing and error detection and generation amongst others.

Interestingly, phrase-based statistical machine translation systems seemed to outperform the analysed attention-based neuronal ones when translating UGC, as stated by Rosales Núñez et al. (2019). Another study on phrase-based statistical machine translation (van der Wees et al., 2015) attempted to describe errors occurring in UGC and their impact on the MT output. The authors reported their observations on the effects showing that various types of UGC differed in error distributions which required diverse strategies for improvement.

This confirms observations by Baldwin et al. (2013) who showed that there were both differences and similarities in English social media text types lying on a continuum of similarity ranging from microblogs to collaboratively-authored content. This variation across UGC types points to the importance of analysis on different types of texts for a better understanding of the phenomena. Besides that, most of those studies were in pre-neural and pre-generative era, which means that the current system outputs may display different effects.

Their impact of various types of artificially created noise on the quality of both statistical and neural machine translation systems was examined by Khayrallah and Koehn (2018). They showed that neural machine translation was less robust to many types of noise than statistical machine translation. The impact of various user-generated content phenomena on translation performance was also analysed by Rosales Núñez et al. (2021) who used and annotated data set of UGC. The authors also showed that traditional models (e.g. strict zero-shot ones) could not handle certain phenomena such as unknown letters.

A data set to evaluate the output of MT was presented by Fujii et al. (2020). The annotated phenomena included proper nouns, abbreviations, colloquial expressions and words deviated from their canonical forms. The evaluation results showed that such phenomena, and specifically non-canonical forms, challenge MT systems, even the widely used off-the-shelf ones. The authors also claimed that the amount of training data was not that important in handling non-standard phenomena. There is a need in special treatment against such phenomena to further improve MT systems.

Our aim is not to assess or to improve the quality of a machine translation system, but rather to analyse the nature of the problems in the user-generated reviews and to examine their impact on human translations and MT outputs including ChatGTP in three different target languages. Our work is in this way similar to approaches that present benchmark data sets or annotated data. For instance, Michel and Neubig (2018) similarly examined different types of noise in a benchmark data set consisting of noisy comments on Reddit and their professional translations.

We focus on the analysis of Amazon product reviews, which were already addressed in (Popović et al., 2021). The authors compared product reviews with movie reviews, however, in terms of overall automatic and human scores. They also reported most frequent translation errors, but without mentioning the effects of the source texts. Popovic (2021) did address the latter in identifying an error type called “source error". However a detailed analysis of this error type was missing.

While there are many studies addressing source text errors or non-standard language use and their impact on machine translated texts, analyses of these phenomena in product review translation is still insufficient.

Furthermore, a better understanding of such phenomena in not only machine but also human translation is needed. To our knowledge, there has been no work involving human translation so far.

Moreover, no further studies known to us looked into translation of UGC with the help of ChatGPT. That is why we perform an analysis of effects of non-standard phenomena in multiple human and machine translations, including translations by ChatGPT, for three translation directions.

3 Data

For our analysis, we use the publicly available corpus DiHuTra[2] (Lapshinova-Koltunski et al., 2022). The corpus contains English Amazon product reviews and their translations into three languages, Croatian, Russian and Finnish, produced by two groups of translators: several professional translators and several students. The translators were only instructed to keep the given segmentation and not to use any MT system. They did not receive any guidelines about how to treat the noise and informality in the reviews. The reason for omitting such guidelines was to collect data on different ways translator respond to such features. Therefore, the corpus is suitable for exploring the subjectivity in translating UGC.

For Croatian MT outputs, we used the two best ranked outputs by human evaluation from the WMT 2022 shared task[3] (Kocmi et al., 2022). For Russian MT outputs, we used Google Translate[4] and DeepL Translator[5]. The Finnish MT outputs were produced using OPUS-MT (Tiedemann and Thottingal, 2020) pre-trained model (opus+bt-news-2020-03-21) and Google Translate[6]. ChatGPT[7] outputs for all target languages were generated using the publicly available GPT 3.5 version. Since human translators were given only simple instructions, a similar approach was used for ChatGPT as well, namely a simple prompt "translate into Croatian/Russian/Finnish".

The data set includes 196 Amazon reviews, fourteen from each of the fourteen products/topics, consisting of 1015 segments. The number of running words and vocabulary size for the source text and for each of the translations can be seen in Table 1.

text	running words	vocabulary
en source	15,236	3,155
hr prof	13,981	4,359
hr stud	13,931	4,446
hr mt1	13,467	4,309
hr mt2	13,465	4,247
hr gpt3.5	14,170	4,265
ru prof	14,217	4,414
ru stud	14,247	4,523
ru mt1	14,472	4,348
ru mt2	14,635	4,391
ru gpt3.5	15,015	4,397
fi prof	11,709	4,612
fi stud	12,274	4,665
fi mt1	11,977	4,461
fi mt2	11,988	4,421
fi gtp3.5	12,299	4,449

Table 1: Corpus statistics.

4 RQ1: Noisy phenomena in English user reviews

Overall analysis To address the first research question, we identify different types of noisy phenomena in the source text. Without using a predefined scheme for these phenomena, we started searching for errors, informal and non-standard parts of the source and identified these phenomena on the fly. In total, at least one phenomenon was found in 597 segments (58.8%), while the remaining 418 (41.2%) were clean.

The identified phenomena, as well as their distributions in source texts can be seen in Table 2 containing absolute number of occurrences, as well as the proportion against all identified phenomena. Table reveals that non-standard capitalisation is the most frequent one, followed by incorrect combinations of punctuation and space (pun+space), non-standard punctuation marks (punctuation), and spelling errors (spelling), missing pronouns (pronoun), and informal expressions and words (informal). Less common phenomena include missing or added spaces (space), incorrect morphological forms such as number, case, tense (form), missing articles (article), incorrect/non-standard structure such as combination and order of words (structure), format conversions (format), missing verbs (verb), added/repeated content (addition), symbols such as emoticons (symbol). There are several rare phenomena, namely missing prepositions (preposition), shortened versions of words (short), lexical errors (lexical), and conjunctions.

For the overall analysis of translations in Section 5.1, we consider all the phenomena, while the detailed analysis of effects of each phenomena in Section 5.2 includes only the most frequent ones (threshold of 50 occurrences). Although this thresh-

[2] http://hdl.handle.net/21.11119/0000-000A-1BA9-A

[3] https://www.statmt.org/wmt22/translation-task.html

[4] https://translate.google.com/, accessed in February 2023

[5] https://www.deepl.com/en/translator, accessed in August 2023

[6] accessed in December 2023

[7] https://chat.openai.com/, accessed in November 2023

phenomenon	occurrences	in %
capitalisation	225	27.3
pun+space	123	14.9
punctuation	109	13.2
spelling	84	10.2
pronoun	81	9.8
informal	53	6.4
space	26	3.2
form	25	3.0
article	19	2.3
structure	17	2.1
format	16	1.9
verb	14	1.7
addition	11	1.3
symbol	9	1.1
preposition	5	0.6
shortened	5	0.6
lexical	1	0.1
conjunction	1	0.1
total	824	

Table 2: Distribution of noisy phenomena in the source text (English user reviews).

old might sound somewhat arbitrary, we believe that the results of an in-depth analysis of the less frequent and especially rarely occurring phenomena would not be reliable. For the sake of completeness, we presents the analysis of these phenomena in Appendix.

Most frequent noisy phenomena Table 3 shows examples of the predominant types of noise:

capitalisation includes example 1) with several fully capitalised words[8], example 2) with one capitalised word. Example 3) shows the English pronoun *I* which does not impact the given target languages, but was included for completeness. Examples 4) and 5) show capitalisation errors in named entities, and example 6) an incorrectly capitalised adverb.

pun+space comprises various incorrect combinations of punctuation marks and spaces: in examples 7), 8) and 9) space is missing, in 10) and 11) the space is placed before the punctuation.

punctuation includes repeated question or exclamation marks (12), missing punctuation marks (13) and punctuation errors (14).

[8]Sometimes the entire review was written in capital letters.

spelling errors result in non-existing words (15) or homophones (16 and 17).

pronouns are often omitted in the reviews (18, 19): on one hand, it does not impact the given target languages due to their pro-drop character, on the other hand, this may cause verb errors related to person and number.

informal refers to informal usage of symbols (20), spelling (21) as well as words or expressions (22).

A number of segments contains more than one non-standard phenomenon (examples 23–27). In example 23), the pronoun *this* should be in plural (*these*), and the article and the pronoun are missing (*to test first* should be *to test the first one*).

Example 24) contains several capitalisation errors (*this* at the beginning of the sentence, *i*, and *MAc* instead of *MAC*), as well as one spelling error (*isnt*).

Example 25) illustrates a named entity with incorrect capitalisation (*sherlock*) and one with both incorrect capitalisation and spelling error (*homes* instead of *Holmes*).

All words in the sentence are fully capitalised in example 26), and one of them is also incorrectly spelled (*CLAPTION* instead of *Clapton*).

A pronoun is missing at the beginning of example 27) and a comma is missing after *case*. Moreover *love* is capitalised and repeated (*LOVE LOVE LOVE*).

5 Analysis of translations

In the next step, we address the second and the third research questions. We present the results on all target languages together, because the overall tendencies are similar. The detailed results for each target language separately can be found in Appendix.

5.1 Effects of source noise on translations (RQ2)

We start with annotating translations to determine the effects caused by the phenomena identified in Section 4 (RQ2). Each target language was covered by one annotator[9], native speaker of the corresponding language with expertise and experience in both human and machine translation.

[9]An exception is the English-Russian pair, where the annotations were cross-checked by the second annotator.

phenomenon		example
capitalisation	1)	**DO NOT BUY**!
	2)	This is **NOT** a good product!
	3)	**i** just received mine
	4)	Bill **gates**
	5)	Do not order on **AMAzon**!
	6)	Very **Cheaply** made product.
pun+space	7)	This is what I needed.It was in good condition
	8)	perfect size–not too big, not too small
	9)	didn't even try to use it...just packed it up
	10)	Exactly what I need .Easy to handle.
	11)	Absolutely love the case !!
punctuation	12)	Wonderful**!!!**
	13)	I love this book[] I bought it last year[]
	14)	batteries already dead**..**
spelling	15)	Heavenly **Hiway** Hymns
	16)	It does exactly what it's supposed **too**.
	17)	the phone says **its** charging
pronoun	18)	[] Have enjoyed it for years
	19)	[] Have not even introduced markers
informal	20)	Not worth the **$$**
	21)	I was **sooo** blessed
	22)	**Yay**!
form, art, pron, pun+space	23)	I bought 2 of **this** and tried to test [] first [] **...**
cap, cap spell, cap	24)	**this** is fake MAC, **i** just received mine and super upset to find out it **isnt** real **MAc**.
cap, spell&cap	25)	**sherlock *homes***
cap, cap&spell, cap, cap	26)	**NOT *CLAPTION* MUSIC VIDEO**!
pron, pun, informal&cap	27)	[] Don't know what I would do without this case[] ***LOVE LOVE LOVE*** it.

Table 3: Examples of the most prominent noisy phenomena in English user reviews: 1)–22) represent examples of single phenomenon in a segment, 23–27) represent multiple phenomena.

The annotators were given the following instructions: for each instance of a non-standard noisy phenomenon, assign:

- "y" (yes) if the phenomenon is kept in the translation
- "n" (no) if the phenomenon is corrected in the translation, or avoided by translating in a different way
- "e" (error) if the phenomenon caused a translation error of any type (mistranslation, omission, addition, grammar error, ...)

A phenomenon that was marked as “kept" might not be replicated in the translation in the exactly same form as in the target text. Rather, a slightly modified but still informal feature might be used by the translator (see e.g. the second example in Table 6). It should be noted that an informal feature being kept in the translation does not necessarily constitute an "error". It may be an intentional choice by the translator to aim for so-called dynamic equivalence (Nida, 1964) by creating a similar effect in the translation as in the source text. In other cases, however, source text may lead to issues that are considered translation errors. A detailed analysis of the types of error found in the translated versions is outside of the scope of this paper.

Table 4 displays the distribution of effects in different translations for all target languages together. It can be noted that the noisy sources are mostly corrected by ChatGPT (about 75%), followed by professional and student translators (60-70%), while MT systems correct only about a half. Furthermore,

	n	y	e
prof	68.8	29.3	1.9
stud	62.5	**34.9**	2.6
mt	51.9	**35.2**	**12.9**
gpt	**75.7**	19.8	4.5

Table 4: Distribution of effects of all source non-standard phenomena in different translations into all languages.

student translators keep a similar amount of noise as MT systems (35%), professionals keep about 30% while ChatGPT keeps only about 20%. As for errors, almost 13% of noisy parts translated by MT systems result in errors, while ChatGPT is much more robust with only 4.5% of errors, however notably less robust than human translators with about 2-3%.

5.2 Effects of individual noisy phenomena (RQ3)

We address the most frequent phenomena as mentioned in Section 4 above. Since the overall tendencies are similar for all languages, the proportions (in %) given in Table 5 are calculated on all target languages together, while the individual results are presented in Appendix.

We observe the following tendencies:

phenomenon		n	y	e
capitalisation	prof	47.3	51.4	1.3
	stud	46.1	52.3	1.6
	mt	37.2	54.2	**8.7**
	gpt	56.4	40.0	**3.6**
pun+space	prof	75.6	23.6	0.8
	stud	64.8	**34.7**	0.5
	mt	69.9	27.2	**2.9**
	gpt	**97.5**	2.2	0.3
punctuation	prof	58.4	40.4	1.2
	stud	45.0	53.5	1.5
	mt	38.2	58.0	**3.8**
	gpt	**76.4**	**22.3**	1.2
spelling	prof	**90.9**	7.5	1.6
	stud	86.1	10.7	3.2
	mt	66.5	11.5	**22.0**
	gpt	**90.5**	2.0	**7.5**
pronoun	prof	**80.2**	18.5	1.2
	stud	76.5	21.8	1.6
	mt	75.9	10.4	**13.2**
	gpt	73.2	21.0	**5.6**
informal	prof	76.7	16.4	6.9
	stud	71.1	**20.1**	8.8
	mt	48.7	11.3	**39.9**
	gpt	74.2	13.2	**12.6**

Table 5: Effects of the most frequent source phenomena on different types of translations for all languages.

capitalisation is slightly more often kept than corrected in all types of translations with exception of ChatGPT which exhibits a reverse tendency. Furthermore, capitalisation causes rarely errors in human translations (1.3-1.6%), slightly more in ChatGPT (3.6%) and most often in MT, however less than 9%.

pun+space is almost always corrected by ChatGPT (97.5%) and frequently corrected by humans and MT. However, students keep it more often than professionals and MT systems. Less than 1% of them cause errors in human anc ChatGPT translations, and less than 3% in MT systems.

punctuation is very often corrected by ChatGTP (more than 90%) and more often corrected by professionals (58.4%) than by students (45%). Furthermore, students and MT systems keep them more often (50-60%) than professionals (40.4%) and ChatGTP (22.3%). The amount of errors in all translations is comparably slightly higher than for pun+space.

spelling is almost completely corrected by professionals and ChatGPT (over 90%) and slightly less by students (86.1%). In MT outputs, 22% of them cause errors, indicating that spelling errors are problematic for MT robustness. ChatGPT is less sensitive, but still 7.5% of them result in translation errors. Even student translators with 3.2% are notably more prone to errors than professionals.

pronoun Most of the missing pronouns do not have effect on human translations, but 13.2% of them cause errors in MT. ChatGPT is again more robust, with 5.6% of errors.

informal is often corrected by human translators and ChatGTP (about 75%). Also, students keep the informality at most (20.1%). Furthermore, almost 40% of informal constructions cause MT errors, and therefore, they should be taken into account for the MT robustness. ChatGPT is again more robust than MT systems, but still 12% of informal constructions result in translation errors.

All in all, spelling errors and informal parts represent the most prominent challenges both for MT systems and for ChatGTP, although ChatGPT is generally more robust to noise.

Other potential challenging types of noise, such as structure, space, form, verb (see Table 8 in Appendix) show the same tendencies, however they are rarely appearing in the analysed corpus so the results are not reliable and should be investigated further.

5.2.1 Examples of some specific effects

Table 6 illustrates three examples of noisy source texts and all their translations.

The **first example** contains one phenomenon only, i.e. added space (*a way* instead of *away*), which caused a mistranslation error in Croatian and Finnish MT outputs, literal translation of *give a way* in Russian MT outputs, and an omission in Russian students' translation. ChatGPT translated it correctly into all target languages.

The **second example** contains more phenomena: missing pronoun *I* at the beginning of the sentence, missing comma after *case*, and the fully capitalised informal expression *LOVE LOVE LOVE*. The missing pronoun has been kept in all translations, however, due to language properties it has an effect only on Russian translations by keeping the informal tone. The punctuation is added in some of translations, and it does not cause any errors in others. As for *LOVE LOVE LOVE*, capitalisation is kept in almost all translations except the one by Russian students. The informality is "corrected" only in the Croatian ChatGPT translation. In all other translations it is either kept (in all human translations and one Russian MT output) or caused errors (in the remaining MT outputs). The nature of errors is diverse: while in one Finnish and one Russian MT outputs this part is omitted, in the other Finnish output this part remained untranslated, and Croatian MT outputs contain incorrect disambiguation of the word *love*: an incorrect person of the verb *love* and the noun *love*. Keeping the informality is also diverse: Croatian students and Finnish professionals did not repeat the word three times, but introduced spaces/hyphens between the letters/syllables, while in the rest of the translations the three repetitions are kept. The Russian student, though, did not keep the capitalisation, and Russian ChatGPT used the word only once but added an adverb intensifying the meaning of the word. In fact, using the verb (*love*) three times should infer intensifying its meaning.

The **third example** is the most complex one, not only because of multiple phenomena but also because of ambiguity (mentioned in Section 4). Two phenomena are clear: the incorrect form of the pronoun *this* and the space before the punctuation mark ... in the end. While the incorrect form caused an error in Croatian and one of the Finnish MT outputs, the punctuation+space did not cause any, but was only kept in some of the translations.

However, the expression *to test first* is ambiguous since it can be interpreted in two ways: (a) *to test the first one*, or (b) *to test (one of) them first*. The annotator who identified the phenomena in the source language perceived the version (a) and therefore annotated the source as presented in Table 6. Further inspection revealed that different annotators as well as different translators had different interpretations. Croatian and Finnish professionals both read it as (b), and students read it as (a). Russian professionals, on the other hand, simply omitted the missing object, as did the two MT systems. In the version produced by ChatGPT, we observe the (a) reading in Croatian, the (b) reading in Russian, and the omission error in Finnish. As for annotators' interpretation, the Croatian one opted for (a) and therefore assigned an "e" to the professional translation, whereas the Finnish annotator perceived both (a) and (b) so they did not assign errors to any human translation. The Russian annotator also perceived the ambiguous reading including both (a) and (b). However, the object (*it* or *them* or *the first one*) is missing in the professional translation and in the two machine translations, so this case was tagged as an error. Although the translation by ChatGPT corresponds to the (b) reading, the annotator marked it as an error agreeing on the disambiguation as (a) suggested by the other annotators.

6 Conclusions

This work presents a detailed analysis of the effects of non-standard phenomena in source texts generated by users on both human and machine translations. While issues in machine-translated user-generated content has been already addressed and partly solved before, a better understanding of how to deal with non-standard language use in translation in general, also in human translation, is missing.

RQ1 Our results show that capitalisation, punctuation and space, spelling, missing pronouns, as well as informal usage of symbols and words belong to the most frequent noisy phenomena for Amazon product reviews written in English.

1) source	We just gave this game **a way** and kept our old one! (space)	
hr prof	Ovu smo igru proslijedili dalje i zadržali našu staru!	n
hr stud	Upravo smo vratili ovu igru i zadržali staru!!!	n
hr mt1	Upravo smo **poboljšali** ovu igru i zadržali našu staru!	e
hr mt2	Upravo smo **omogućili** ovu igru i zadržali našu staru!	e
hr gpt	Ovu novu igru smo samo poklonili i zadržali staru!	n
ru prof	Мы отдали эту игру, а себе оставили старую!	n
ru stud	В итоге мы **[]** играли в нашу старую игру!	e
ru mt1	Мы просто **дали этой игре дорогу** и сохранили старую!	e
ru mt2	Мы просто **дали этой игре дорогу** и сохранили нашу старую!	e
ru gpt	Мы просто подарили эту игру и сохранили нашу старую!	n
fi prof	Annoimme tämän pois ja pidimme vanhan versiomme!	n
fi stud	Me vain annoimme tämän pelin pois ja pidimme vanhan!	n
fi mt1	Me vain annoimme tälle pelille **keinon** ja pidimme vanhan!	e
fi mt2	Annoimme tälle pelille **tavan** ja säilytimme vanhan!	e
fi gpt	Juuri annoimme tämän pelin pois ja pidimme vanhan!	n

2) source	**[]** Don’t know what I would do without this case**[]** ***LOVE LOVE LOVE*** it. (pronoun punctuation informal capitalisation)	
hr prof	Ne znam što bih bez ove maskice. VOLIM VOLIM VOLIM je.	n n y y
hr stud	Ne znam što bih bez ove maskice – O-BO-ŽA-VAM ju.	n n y y
hr mt1	Ne znam što bih bez ove kutije **VOLI VOLI VOLI to**.	n y e y
hr mt2	Ne znam što bih napravio bez ovog slučaja **LJUBAV LJUBAV LJUBAV to**.	n y e y
hr gpt	Ne znam što bih radio bez ovog slučaja, OBOŽAVAM ga.	n n n y
ru prof	Не знаю, что бы делала без него КРУТО КРУТО КРУТО.	y n y y
ru stud	Не знаю, что бы я делал без этого чехла. Очень, очень, очень доволен.	y n y n
ru mt1	Не знаю, что бы я делал без этого чехла ЛЮБЛЮ ЛЮБЛЮ ЛЮБЛЮ.	y n y y
ru mt2	Не знаю, что бы я делал без этого чехла. **[]**	y n e e
ru gpt	Не знаю, что бы я делал без этого чехла. ОЧЕНЬ ЛЮБЛЮ его.	y n y y
fi prof	En tiedä mitä tekisin ilman tätä kuorta! R A K A S T A N.	n n y y
fi stud	En tiedä, mitä tekisin ilman tätä koteloa. RAKASTAN RAKASTAN RAKASTAN sitä.	n n y y
fi mt	En tiedä, mitä tekisin ilman tätä juttua. **[]**	n n e e
fi mt2	En tiedä mitä tekisin ilman tätä tapausta **LOVE LOVE LOVE** sitä.	n y e y
fi gpt	En tiedä, mitä tekisin ilman tätä koteloa. RAKASTAN, RAKASTAN, RAKASTAN sitä.	n n y y

3) source	I bought 2 of **this** and tried to test **[]** first **[]** **...** (form article pronoun pun+space)	
hr prof	Kupio sam 2 komada i **prvo** sam **ih** pokušao testirati ...	n e e y
hr stud	Kupio sam dva primjerka i pokušao isprobati jedan od njih...	n n n n
hr mt1	Kupio sam 2 od ovoga i **prvo** **[]** pokušao testirati ...	e e e y
hr mt2	Kupio sam 2 od ovoga i **prvo** **[]** pokušao testirati ...	e e e y
hr gpt	Kupio sam 2 ovakva proizvoda i pokušao testirati prvi...	n n n n
ru prof	Я купил 2 аккумлятора и решил проверить **[]**...	n e y n
ru stud	Я приобрел две штуки этого зарядного устройства и решил испытать первое...	n n n n
ru mt1	Я купил 2 таких и попытался сначала протестировать **[]**...	n e y n
ru mt2	Купил 2 штуки и попробовал сначала протестировать **[]**...	n e y n
ru gpt	Купил 2 штуки и решил сначала протестировать **одну из них**...	n e y n
fi prof	Ostin kaksi tällaista ja yritin ensin testata yhtä ...	n n n y
fi stud	Ostin näitä kaksi ja kokeilin ensimmäistä...	n n n n
fi mt1	Ostin **tästä** kaksi ja yritin testata **ensin** **[]**.	e e e n
fi mt2	Ostin 2 tätä ja yritin testata **ensin** **[]** ...	y e e y
fi gpt	Ostin 2 näitä ja päätin testata **ensin** **[]**...	n e e n

Table 6: Examples of effects of different non-standard phenomena on translations; example 3 could be interpreted in two ways.

RQ2 In our data, these phenomena are mostly converted into a standard form by ChatGPT, followed by professional translators, while students and MT systems are often keeping them. Furthermore, MT systems often generate a translation error, while ChatGPT is more robust to the noise in the source text.

RQ3 Our further observation is that spelling errors (especially those resulting in an existing word) and informal constructions are particularly difficult for MT systems, as well as for ChatGPT although to a less extent. The results also indicate that incorrect or non-conventional structure as well as incorrect word forms also represent a potential challenge, however further work is needed in this direction since these types of noise are not sufficiently frequent in our data.

We believe that our results are of interest for both NLP and translation studies. On the one hand, our findings can help improving robustness of MT systems. On the other hand, the work should give an idea about the guidelines for human translators if human translations are needed for user-generated texts: translator guidelines should be clear on how and if source errors should be corrected in the resulting translation. Also, the findings could be helpful for guidelines for human evaluation of translated used-generated content - what should be considered as an error and what not.

Future work should further investigate the most prominent phenomena and their sub-types. Besides that, creating challenge test sets to better understand each phenomenon could be an asset. We also plan to look into the types of translation errors in more detail. Moreover, more noisy UGC (such as social media) should be analysed as well. Furthermore, we plan to extend the analysis on outputs produced by other large language models, as well as to explore different prompts.

Limitations

We investigate only one type of user-generated content, namely user reviews. This sub-domain is relatively clear compared to other noisy types such as social media posts, as it contains less non-standard texts. Therefore, some potentially problematic phenomena do not appear at all or not sufficiently often in the analysed corpus. However, most of the analysed phenomena appear in other types of UGC, too.

Also, we investigate only English as the source language. More source languages should be explored in future work.

The annotation of each translated text was carried out by a single evaluator with an exception for Russian, where problematic cases were discussed in a team of trained linguists.

While all source sentences were translated by each of the MT systems and ChatGPT, they were not translated by each of the individual translators, but only by each group of the translators.

Using different MT systems for different target languages can be a disadvantage, but on the other hand it introduces more diversity.

Ethics Statement

The data used in this study is derived from the corpus DiHuTra which is publicly available - the corpus is hosted by Fedora Commons Repository of the Saarland University (UdS) CLARIN-D centre[10]. The DiHuTra corpus is licensed under CCBY-NC-SA4.0. The translations collected in the corpus are all anonymised and do not contain any personal information. All the authors signed a consent agreement[11]. The corpus only contains the anonymised metadata on the experience, study program, age and gender of the translators who contributed to the data collection.

Acknowledgements

ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme under Grant Agreement No. 13/RC/2106_P2. The Finnish subcorpus was supported by a Kopiosto grant awarded by the Finnish Association of Translators and Interpreters.

References

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan. Asian Federation of Natural Language Processing.

Alexandre Berard, Ioan Calapodescu, Marc Dymetman, Claude Roux, Jean-Luc Meunier, and Vassilina

[10]Persistent identifier `http://hdl.handle.net/21.11119/0000-000A-1BA9-A`

[11]The consent agreement form was made available by the corpus creators and can be viewed on the GitHub repository `https://github.com/katjakaterina/dihutra/blob/main/fortranslators/consent.pdf`

Nikoulina. 2019. Machine translation of restaurant reviews: New corpus for domain adaptation and robustness. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 168–176, Hong Kong. Association for Computational Linguistics.

Ryo Fujii, Masato Mita, Kaori Abe, Kazuaki Hanawa, Makoto Morishita, Jun Suzuki, and Kentaro Inui. 2020. PheMT: A phenomenon-wise dataset for machine translation robustness on user-generated contents. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5929–5943, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kamal Gupta, Soumya Chennabasavaraj, Nikesh Garera, and Asif Ekbal. 2021. Product review translation using phrase replacement and attention guided noise augmentation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 243–255, Virtual. Association for Machine Translation in the Americas.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski, Maja Popović, and Maarit Koponen. 2022. DiHuTra: a parallel corpus to analyse differences between human translations. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 337–338, Ghent, Belgium. European Association for Machine Translation.

Benjamin Marie and Atsushi Fujita. 2020. Synthesizing parallel data of user-generated texts with zero-shot neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:710–725.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Eugene Nida. 1964. *Toward a Science of Translating With Special Reference to Principles and Procedures Involved in Bible Translating*. Brill, Boston.

Maja Popovic. 2021. On nature and causes of observed MT errors. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 163–175, Virtual. Association for Machine Translation in the Americas.

Maja Popović, Alberto Poncelas, Marija Brkic, and Andy Way. 2021. On machine translation of user reviews. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1109–1118, Held Online. INCOMA Ltd.

José Carlos Rosales Núñez, Djamé Seddah, and Guillaume Wisniewski. 2019. Comparison between NMT and PBSMT performance for translating noisy user-generated content. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 2–14, Turku, Finland. Linköping University Electronic Press.

José Carlos Rosales Núñez, Guillaume Wisniewski, and Djamé Seddah. 2021. Noisy UGC translation at the character level: Revisiting open-vocabulary capabilities and robustness of char-based models. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 199–211, Online. Association for Computational Linguistics.

Johann Roturier and Anthony Bensadoun. 2011. Evaluation of MT systems to translate user generated content. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2015. Five shades of noise: Analyzing machine translation errors in user-generated text. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 28–37, Beijing, China. Association for Computational Linguistics.

A Appendix

A.1 Overall distribution of effects on each of the translations

(a) en-hr

	n		y		e	
prof	**604**	**73.3**	208	25.2	12	1.5
stud	553	67.1	255	31.0	16	1.9
mt1	437	53.0	309	37.5	**78**	**9.5**
mt2	435	52.8	302	36.6	**87**	**10.6**
gpt	**634**	**76.9**	**157**	**19.0**	33	4.0

(b) en-ru

	n		y		e	
prof	**538**	**65.3**	260	31.6	26	3.2
stud	506	61.4	285	34.6	33	4.0
mt1	474	57.5	288	35.0	**62**	**7.5**
mt2	511	62.0	263	31.9	**50**	**6.1**
gpt	**631**	**76.6**	**163**	**19.8**	30	3.6

(c) en-fi

	n		y		e	
prof	**558**	**67.7**	256	31.1	10	1.2
stud	486	59.0	322	39.1	16	1.9
mt1	332	40.3	274	33.2	**218**	**26.5**
mt2	376	45.6	306	37.1	**142**	**17.2**
gpt	**607**	**73.7**	**169**	**20.5**	48	5.8

Table 7: Distribution of effects of all noisy phenomena on each translation into each target language: (a) Croatian, (b) Russian, (c) Finnish.

A.2 Effects of less frequent types of noise on all target languages together

A.3 Effects of different types of noise on each of the translations

phenomenon		n	y	e
space	prof	**78.2**	18.0	3.8
(26)	stud	69.2	**26.9**	3.8
	mt	57.7	22.4	**19.9**
	gpt	73.1	21.8	**5.1**
form	prof	93.3	6.7	0
(25)	stud	**96.0**	2.7	1.3
	mt	76.0	6.0	**18.0**
	gpt	90.6	2.7	**6.7**
article	prof	94.7	0	5.3
(19)	stud	100	0	0
	mt	89.5	0.9	9.6
	gpt	94.7	0	5.3
structure	prof	**90.2**	9.8	0
(17)	stud	74.5	11.8	**13.7**
	mt	28.4	**32.4**	**39.2**
	gpt	68.6	17.7	**13.7**
format	prof	75.0	18.8	6.2
(16)	stud	37.5	47.9	14.6
	mt	41.7	45.8	12.5
	gpt	95.8	0	4.2
verb	prof	85.7	11.9	2.4
(14)	stud	78.6	21.4	0
	mt	61.9	25.0	**13.1**
	gpt	73.8	11.9	**14.3**
addition	prof	81.8	12.1	6.1
(11)	stud	78.8	18.2	3.0
	mt	77.3	9.1	13.6
	gpt	87.9	12.1	0
symbol	prof	11.1	81.5	7.4
(9)	stud	14.8	77.8	7.4
	mt	7.4	77.8	14.8
	gpt	14.8	81.5	3.7
preposition	prof	93.3	6.7	0
(5)	stud	93.3	6.7	0
	mt	76.7	6.6	16.7
	gpt	100	0	0
shortened	prof	80.0	20.0	0
(5)	stud	73.3	26.7	0
	mt	76.7	16.7	6.6
	gpt	86.7	13.3	0
lexical	prof	100	0	0
(1)	stud	100	0	0
	mt	83.3	0	16.7
	gpt	100	0	0
conjunction	prof	100	0	0
(1)	stud	66.7	33.3	0
	mt	33.3	50.0	16.7
	gpt	66.7	0	33.3

Table 8: Effects of less frequent (< 30 occurrences in source) source phenomena on different types of translations for all languages.

		en-hr			en-ru			en-fi		
phenomenon	text	n	y	e	n	y	e	n	y	e
capitalisation	prof	109	114	2	100	119	6	110	114	1
(225)	stud	115	110	0	106	110	9	90	133	2
	mt1	80	138	7	91	118	16	85	94	**46**
	mt2	80	134	11	102	113	10	64	134	**27**
	gpt	**122**	96	7	137	**82**	6	122	**92**	11
pun+space	prof	98	25	0	91	30	2	90	32	1
(123)	stud	76	47	0	81	40	2	82	41	0
	mt1	87	36	0	105	18	0	46	67	10
	mt2	87	36	0	99	15	9	92	29	2
	gpt	**120**	2	1	**119**	4	0	**121**	2	0
punctuation	prof	70	38	1	57	49	3	64	45	0
(109)	stud	55	54	0	45	59	5	47	62	0
	mt1	26	82	1	48	57	4	55	45	9
	mt2	26	82	1	59	47	3	36	46	7
	gpt	**82**	25	2	**88**	20	1	**80**	28	1
spelling	prof	**82**	2	0	70	12	2	**77**	5	2
(84)	stud	**75**	8	1	66	13	5	**76**	6	2
	mt1	57	10	**17**	62	11	**11**	37	7	**40**
	mt2	56	10	**18**	71	10	3	52	10	**22**
	gpt	**78**	1	5	**77**	2	5	**73**	2	9
pronoun	prof	80	0	1	51	28	2	64	17	0
(81)	stud	78	2	1	49	30	2	59	21	1
	mt1	64	7	**10**	35	44	2	29	19	**33**
	mt2	65	6	**10**	41	37	3	37	22	**22**
	gpt	74	3	4	44	34	3	60	14	7
informal	prof	43	8	2	41	6	6	38	12	3
(53)	stud	37	10	6	41	8	4	35	14	4
	mt1	25	4	**24**	30	8	**15**	16	8	**29**
	mt2	24	3	**26**	34	7	**12**	26	6	**21**
	gpt	36	8	9	43	7	3	39	6	8

Table 9: Effects of the most prominent source phenomena with more than 50 occurrences on each of the translations.

		en-hr			en-ru			en-fi		
phenomenon	text	n	y	e	n	y	e	n	y	e
space	prof	20	5	1	21	3	2	20	6	0
(26)	stud	21	4	1	16	8	2	17	9	0
	mt1	16	5	**5**	17	4	**5**	13	6	**7**
	mt2	17	3	**6**	16	8	2	11	9	**6**
	gpt	20	5	1	19	6	1	18	6	2
form	prof	21	4	0	25	0	0	24	1	0
(25)	stud	24	1	0	24	0	1	24	1	0
	mt1	19	3	3	25	0	0	12	1	**12**
	mt2	18	4	3	24	0	1	16	1	**8**
	gpt	23	2	0	22	0	3	23	0	2
article	prof	18	0	1	18	0	1	18	0	1
(19)	stud	19	0	0	19	0	0	19	0	0
	mt1	18	0	1	16	1	2	17	0	2
	mt2	18	0	1	16	0	3	17	0	2
	gpt	19	0	0	17	0	2	18	0	1
structure	prof	16	1	0	14	3	0	16	1	0
(17)	stud	14	0	3	12	3	2	12	3	2
	mt1	3	9	5	8	6	3	2	2	**13**
	mt2	3	9	5	10	5	2	3	2	**12**
	gpt	8	6	3	15	0	2	12	3	2
format	prof	11	2	3	16	0	0	9	7	0
(16)	stud	5	8	3	13	2	1	0	13	3
	mt1	14	1	1	5	11	0	0	11	5
	mt2	14	1	1	3	13	0	4	7	5
	gpt	16	0	0	15	0	1	15	0	1
verb	prof	14	0	0	13	0	1	9	5	0
(14)	stud	13	1	0	14	0	0	6	8	0
	mt1	9	3	2	13	1	0	5	5	4
	mt2	8	4	2	13	1	0	4	7	3
	gpt	13	0	1	12	0	2	6	5	3
addition	prof	9	2	0	10	1	0	8	1	2
(11)	stud	9	2	0	9	2	0	8	2	1
	mt1	10	1	0	9	1	1	6	1	4
	mt2	10	1	0	10	1	0	6	1	4
	gpt	10	1	0	10	1	0	9	2	0

Table 10: Effects of the source phenomena with less than 50 and more than 10 occurrences on each of the translations

phenomenon	text	en-hr n	en-hr y	en-hr e	en-ru n	en-ru y	en-ru e	en-fi n	en-fi y	en-fi e
symbol	prof	2	6	1	1	7	1	0	9	0
(9)	stud	2	6	1	2	7	0	0	8	1
	mt1	0	7	2	1	7	1	0	7	2
	mt2	0	6	3	3	6	0	0	9	0
	gpt	2	7	0	1	7	1	1	8	0
preposition	prof	5	0	0	4	1	0	5	0	0
(5)	stud	5	0	0	4	1	0	5	0	0
	mt1	4	1	0	4	0	1	4	0	1
	mt2	4	1	0	4	0	1	3	0	2
	gpt	5	0	0	5	0	0	5	0	0
shortened	prof	4	1	0	4	1	0	4	1	0
(5)	stud	3	2	0	4	1	0	4	1	0
	mt1	4	1	0	4	0	1	3	1	1
	mt2	4	1	0	5	0	0	3	2	0
	gpt	4	1	0	5	0	0	4	1	0
lexical	prof	1	0	0	1	0	0	1	0	0
(1)	stud	1	0	0	1	0	0	1	0	0
	mt1	1	0	0	1	0	0	1	0	0
	mt2	1	0	0	0	0	1	1	0	0
	gpt	1	0	0	1	0	0	1	0	0
conjunction	prof	1	0	0	1	0	0	1	0	0
(1)	stud	1	0	0	0	1	0	1	0	0
	mt1	0	1	0	0	1	0	0	0	1
	mt2	0	1	0	1	0	0	1	0	0
	gpt	1	0	0	1	0	0	0	0	1

Table 11: Effects of the source phenomena with less than 10 occurrences on each of the translations