sciendo

DOI: 10.2478/rjes-2024-0014

LITERAL OR LITERARY MACHINE TRANSLATION? CASE STUDY: WINNIE-THE-POOH

ALINA RĂDOI
West University of Timișoara

Abstract: *Despite its widespread commercial use, machine translation has been sparsely employed on literary works, especially on children's literature. This study focuses on reader comprehension of a chapter from Winnie-the-Pooh translated from English into Romanian using DeepL. Machine translation evaluation should focus not only on fluency and adequacy, but also on how readers perceive the target text. Twenty-five participants have annotated the passages that were difficult to understand and those that were logical but contained grammatical or stylistic errors. The survey ended with three comprehensibility and three MT user opinion questions.*
Key words*: comprehensibility, end-user expectations, informative manual evaluation, literary machine translation, Romanian, user comprehension*

1. Introduction

Literary machine translation is perceived as "the last bastion of human translation" (Toral and Way 2014:174), hence, it is a thorny issue (Toral and Way 2018). That is why it is important to see whether a machine-translated text is fit for purpose, i.e. whether it can be understood and enjoyed as much as the original.

This paper takes into consideration the end-user perspective and expectations (Castilho and O'Brien 2017; Kasperė et al. 2021) on a neural machine-translated chapter from Winnie-the-Pooh by A.A. Milne, using a survey.

When it comes to literary works, especially those aimed at children, comprehensibility and comprehension are paramount. Popović (2020:5059) defines comprehensibility as "how well a reader is able to understand the translated text without access to the original text". I would argue that comprehensibility is a characteristic of the text itself – can it be understood, is the language clear enough? Comprehension, on the other hand, is a characteristic of the reader – can they understand the text, do they perceive errors, do they even care that the text might contain errors?

Children's literature is under-represented in the world of machine translation. To my knowledge, this is the only study which concentrates on the English-Romanian language pair in the field of literary machine translation. It is also one of the very few studies which takes into account the perspective and attitudes of the machine translation end-user.

2. Related work

The literature focuses on two main areas – comparing machine translation with the human translation or assessing the level of creativity of the MT output. Some marginal work has been done on the ethical aspects of using machine translation with regards to literature and on experts' attitude towards integrating MT into their workflow.

Hansen and Esperança-Rodier (2022) focused on fine-tuning an MT system on literary data translated by the same individual translator to gauge whether the MT system can replicate their literary style. The fine-tuning process also included publicly available datasets, such as the Books corpus from the OPUS repository. Subsequently, an automatic and manual evaluation was made. The most frequent errors they identified were omissions, mistranslations, style and logical problems.

Adapting NMT systems to literary content is also what Matusov (2019) did, showing that such systems have richer vocabulary than general NMT systems. Though not focusing on literary translation per se – but nonetheless tangibly having an impact in this area –, Vanmassenhove, Shterionov and Way (2019) argue that machine translating texts leads to a loss of lexical richness.

Macken et al. (2022) focused on Dutch-English literary translation using an MT-enhanced workflow which consisted of three steps: machine translation, post-editing and revision, looking at the automatic evaluation of the (dis)similarity between the three outputs. Cespedosa Vázquez and Mitkov (2023) concluded that DeepL was the best system out of the three they used to translate prose and poetry from different time periods. According to their BLUE calculations, the more contemporary texts fared better.

Interestingly, Şahin and Gürses (2019) focused on the retranslation of literary works from English into Turkish by student translators with and without the help of machine translation. The conclusion was that using MT hindered the creativity of the participants. Still on the topic of creativity, Arenas and Toral (2022) analysed the level of creativity in three types of translation – human translation, post-editing and machine-translation. Their corpus was comprised of a short story by Kurt Vonnegut. All reviewers believed that the human translation was the best, while the MT output was considered the least suited in terms of creativity.

Looking at Asian languages, Sibuea, Budi and Jonathan (2023) talk about equivalence problems in Indonesian for the Harry Potter series.

According to Thai et al. (2022), expert literary translators prefer human-translated paragraphs rather than machine-translated ones, due to the fact that they contain discourse-disrupting errors and stylistic inconsistencies. Ruffo (2018) looks at the way in which literary translators perceive their role in a technology-driven field and their attitudes towards using technology in their work.

Finally, Taivalkoski-Shilov (2019) brings to the front the ethical issues related to machine(-assisted) literary translation, namely the relativistic view on quality, voice and noise in literary translation and text ownership.

3. Research questions

The research questions are in direct connection with the three parts of the survey used in this study, which will be detailed below.

- RQ1: How many words in the output do end-users identify as either being incomprehensible or ungrammatical / unstylistic?
- RQ2: What is the end-users' opinion on the output regarding
    - RQ2.1. its comprehensibility?
    - RQ2.2. its translator?
    - RQ2.3. its textual improvement points?
- RQ3: What is the end-users' comprehension level of the output?

Study design

a. Methodology

In order to carry out this study, I used the novel method of informative manual evaluation of machine translation output developed by Maja Popović (2020). What is interesting about this method is that the evaluators are asked to mark all problematic parts (words, phrases, sentences) of the translation. As a result, further error analyses can be carried out.

The participants in this study were asked to mark sections of the first chapter of Winnie-the-Pooh by A.A. Milne that had previously been translated using DeepL from English into Romanian. The original chapter contains 1,003 words, while the target text is made up of 1,013 words.

No personal data was formally collected on the participants, only informally, i.e. while contacting them. They were between 25 and 55 years old, both monolingual and bilingual speakers. Even though the research participants are not children themselves, it is not uncommon for adults to also read classic children's books, especially if they read them out loud to children. That is why the participants were considered machine translation end-users for the purposes of this study.

b. Survey

All participants received a Word document for carrying out the task, which contained three sections:

- Section 1 was comprised of the introduction and instructions, along with the target text that was not modified in any way, focusing on the comprehensibility of the text;
- Section 2 included three close-ended questions (multiple choice) on the participants' opinion on the target text;
- Section 3 incorporated three open-ended questions to test the participants' comprehension of the target text.

The entire document was in Romanian, namely all instructions and questions. The questions were translated into English for the purposes of this article.

In Section 1, the evaluators were informed about the nature of this study and their rights and were then given the instructions as follows:

- mark using the red highlight all parts of the text (single words, small or long phrases, or entire sentences) which are *not understandable* (it does not make sense, it is not clear what it is about, etc.);
- mark using the yellow highlight all parts of the text (words, phrases or sentences) which *might be understandable* but contain grammatical or stylistic errors.

The wording and method of marking up the errors is directly informed by Popović (2020).

Section 2 of the survey contained the following questions:

1. On the whole, the text seems…	a. easy to understand b. rather difficult to understand c. difficult to understand
2. In your opinion, the text was translated:	a. by a machine translation system (such as Google Translate) b. by a human translator
3. What would you improve to the text?	a. the style b. the grammar c. the structure

Table 1. Opinion questions

For questions 2, Google Translate was given as an example (even though the text had been translated using DeepL) because Romanian people are more used to it and can recognize what it does.

Section 3 focused on the comprehension questions below. The correct answers appear in the right-hand side column.

According to the text:	
1. What was Winnie-the-Pooh unsure of?	(What 'to live under the name of' means)
2. What should the storyteller do 'very sweetly'?	(To tell a story / To tell Winnie-the-Pooh a story)
3. What 'sort of bear' is Winnie-the-Pooh?	(One that likes to hear stories about himself)

Table 2. Comprehension questions

In both the English and the Romanian versions of these questions, the words in quotes are taken directly as they are from the original text, i.e. the ones in English are copied from the source text and the ones in Romanian (that the participants read) were taken from the DeepL output. The Results section will discuss this in more detail.

1. Results

a. Comprehensibility – Section 1

Looking at the results in Section 1, i.e. the number of words the evaluators have marked either using yellow (for grammatical or stylistic errors) or red (for no comprehension), it is evident that 16 out of 25 participants used the yellow highlight more often than the red. They identified more grammar or style errors than comprehension errors.

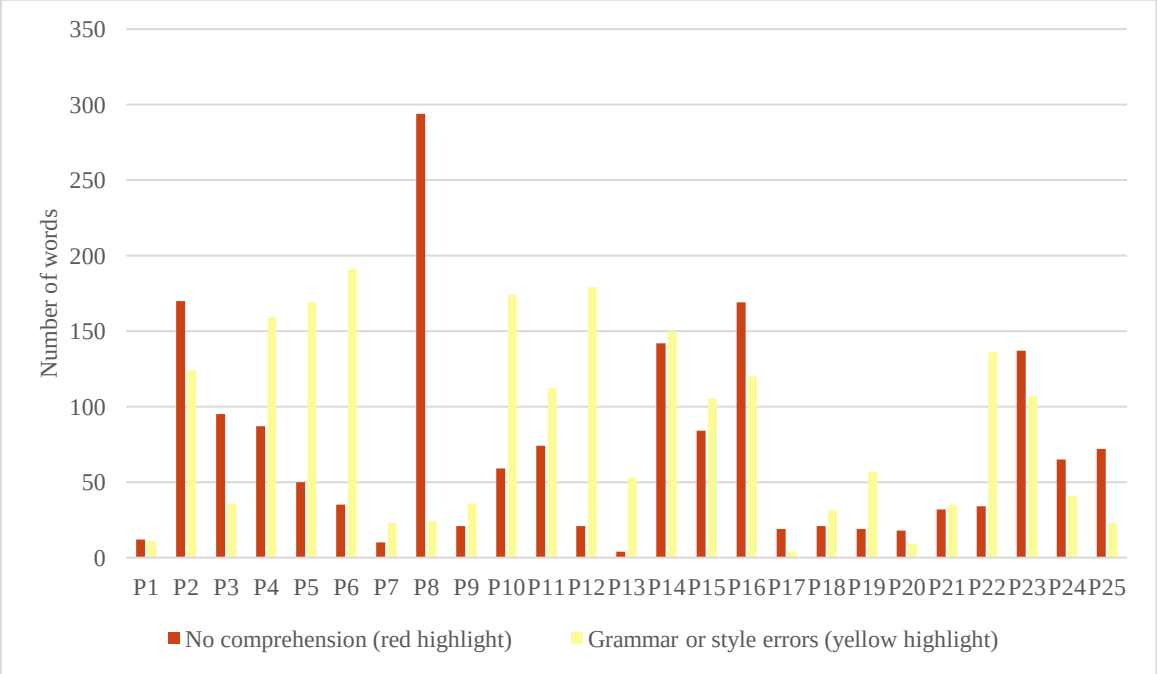

Figure 1. Number of highlighted words

Participant 8 is obviously the outlier in this study, as they believed that around 300 words out of the total 1,013 were incomprehensible. However, the yellow highlight is the most prevalent overall, which might mean that machine translation is not as fluent or is not considered as fluent by actual users as the literature suggests (Way 2018)– at least for this language combination and text genre.

The average and median number of highlighted words can be seen below.

	Red highlight (no comprehension)	**Yellow highlight (grammar or style errors)**
Average	70 words	84 words
Median	50 words	52 words

Table 3. Average and median number of highlighted words

Out of frustration and confusion, four participants wrote in-text comments or even started correcting the output, but these interventions could not be taken into consideration in any way because they did not follow the instructions. One participant highlighted one word in blue instead of red or yellow, so it was not clear what kind of error this person was pointing towards. As a result, this word was also not counted.

While evaluators have been asked to mark words, phrases or even entire paragraphs, the actual number of words is taken into account because it is the easiest and most logical way of measuring the results.

It must be noted however that grammatical and stylistic errors also include punctuation errors. Each punctuation error or rather, punctuation sign, was counted as one word in the evaluation above. This decision was taken based on the following logic: were the output post-edited, the post-editor would still need to put cognitive and time effort into correcting these punctuation errors.

The highlighted passages were not each counted as one error because it was difficult to tell where an error would start and end. For example, evaluators have highlighted wrong quotation marks as in Figure 2. In this case, the participant also identified commas and parentheses as being grammatical errors and marked them accordingly.

("Ce înseamnă "sub nume"?", a întrebat Christopher Robin.
'Înseamnă că avea numele deasupra ușii, cu litere de aur, și trăia sub el'.
Winnie-the-Pooh nu era foarte sigur', a spus Christopher Robin.
'Acum sunt', a spus o voce crescândă.
"Atunci voi merge mai departe", am spus eu).

Într-o zi, pe când se plimba, a ajuns la un loc deschis în mijlocul pădurii, iar în mijlocul acestui loc era un stejar mare și, din vârful copacului, se auzea un zumzet puternic.

Figure 2. Participant number 8's sample evaluation

It would be impossible to tell whether *"?"* would account for one, two or even three errors in the eyes of the participant. It is not this study's author's responsibility, nor wish, to judge whether or not these sample errors were identified and highlighted correctly as per Romanian grammar rules. Only the numerical evaluation is important at this point. In the case above, *"?"* was counted as three words, one for each punctuation sign, as *("* was counted as two words.

Looking at some sample errors, it becomes clear that a few of them can impede comprehension quite severely, as can be seen in the examples below.

1) and said in a deep whisper: **'*Honey!*'**
   și a spus într-o șoaptă adâncă: "**Dragă**!"
2) 'I wonder if you've got such a thing as a balloon **about** you?'
   'Mă întreb dacă ai așa ceva, un balon **în** tine?'.
3) he sang a **Complaining** Song
   a cântat un Cântec de **Plângere**

In example 1), *honey* was wrongly translated as *dragă* when it refers in fact to the sweet substance produced by bees – it is not an appellative in this case. Next, example 2) showcases the incorrect selection of a preposition – having a balloon *about* you and having a balloon *inside* yourself are two very different things. Finally, in the last example, *Cântec de Plângere* can be interpreted as being both a complaining and a crying / wailing song in Romanian.

A few participants highlighted the example below using yellow, meaning they considered it to be a grammatical or style error.

4) and the only reason **for making** a buzzing-noise that *I* [original formatting] know of is because **you're a bee**
iar singurul motiv **pentru care se face** un bâzâit, după câte știu eu, este că **ești o albină**

I can only conclude that the users tried to understand the text, but they believed that the words in bold depicted grammar or style errors. The readers were not asked to differentiate between grammar and style errors because the cognitive effort would have been too great. When piloting this survey, it became clear that participants became overwhelmed when asked to do several things at once. Contacting the participants after completing the survey revealed that some of them went through the text twice – once to catch comprehensibility errors and once to identify grammar or style errors. Others tried to do manage both activities at the same time.

b. Opinion questions – Section 2

Fifteen out of the 25 participants declared that the output was "rather difficult to understand", as can be seen in Figure 3 below. The number of people who considered that the text is difficult to understand is on par with those who deemed it easy to understand.

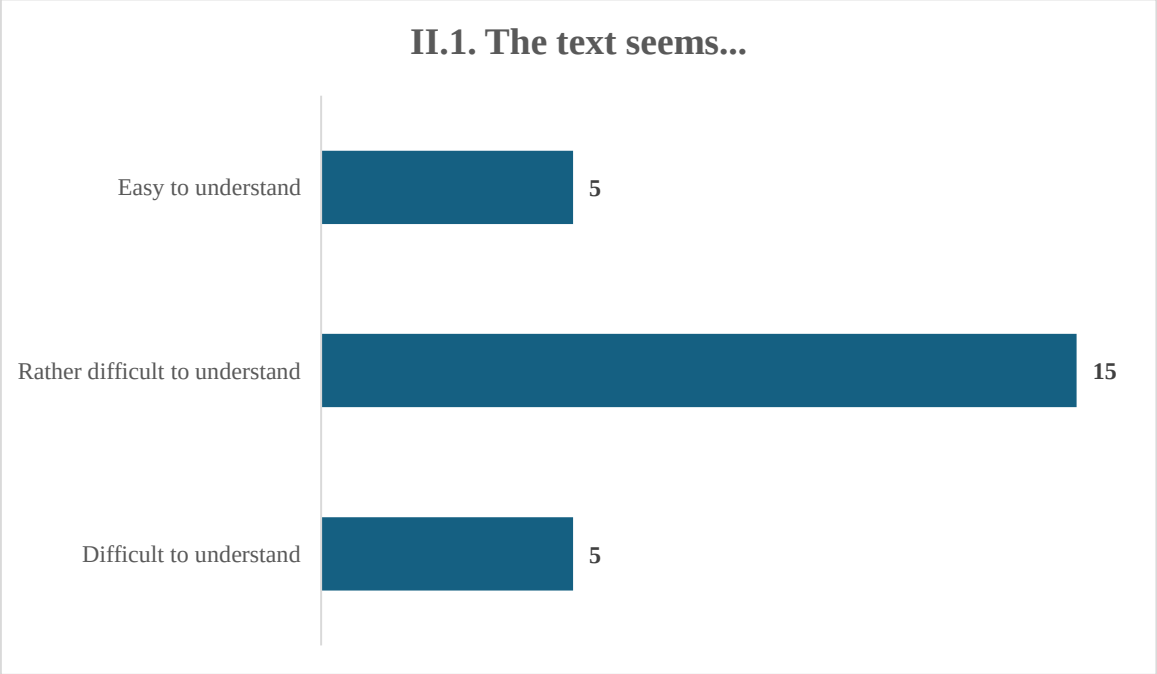

Figure 3. Answers to opinion question 1

When giving instructions to the evaluators, the author of the study mentioned that the output had been produced either by a machine translation system or by a human, so as not to disclose the real facts and to not influence the participants in any way. The overwhelming majority of participants were not hesitant as to the nature of the translator of the text – it was clear to them that the target text had been produced by a machine translation system.

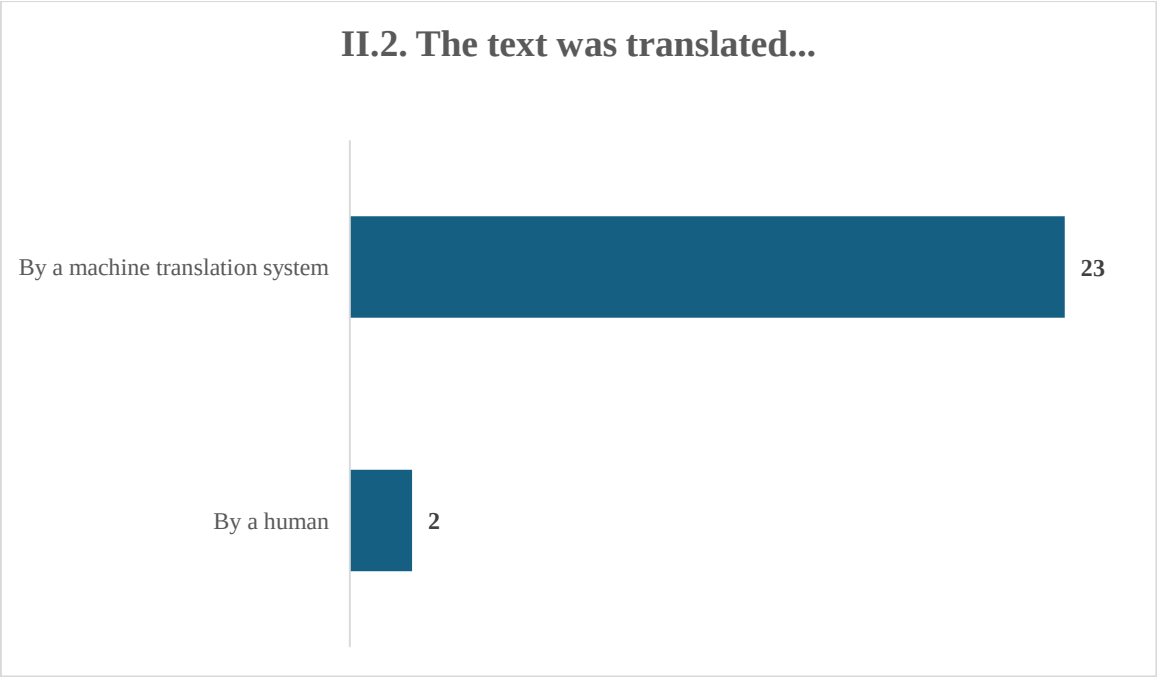

Figure 4. Answers to opinion question 2

Figure 5 showcases that many readers believe that the output needs improvement on all fronts – style, grammar and structure. The conclusion in this case is that end-users have certain standards regarding a "proper" children's story, which DeepL does not fulfil.

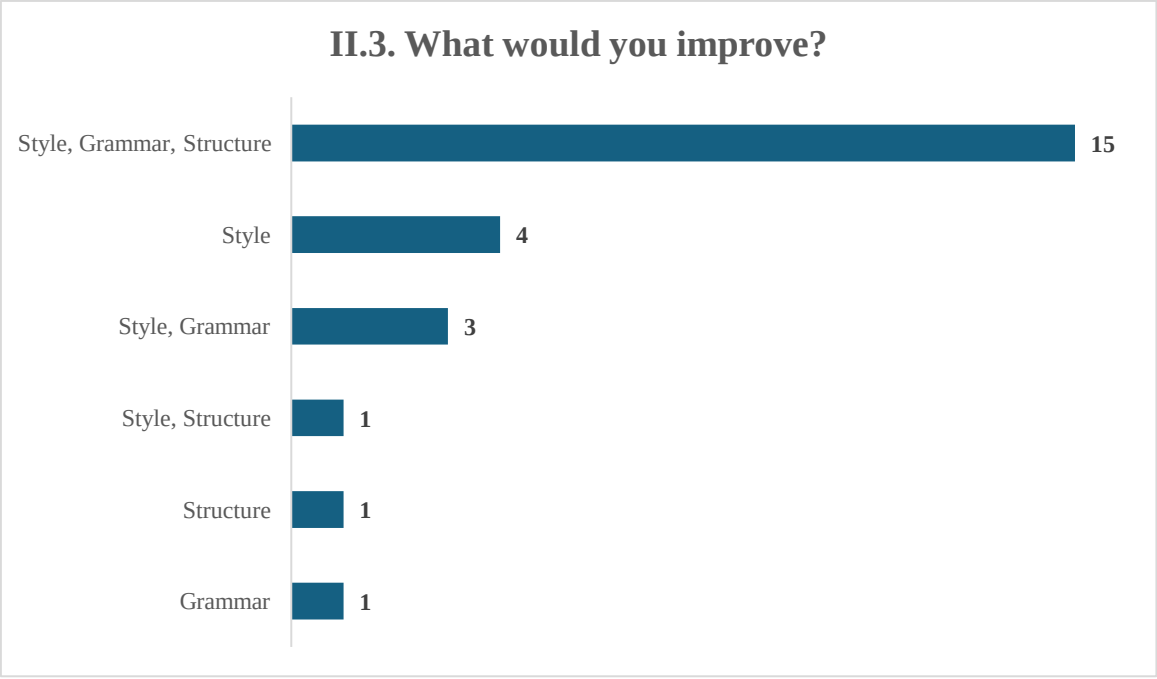

Figure 5. Answers to opinion question 3

c. Comprehension questions – Section 3

The evaluators' general level of text comprehension varies. For example, question 1 was the most difficult to answer.

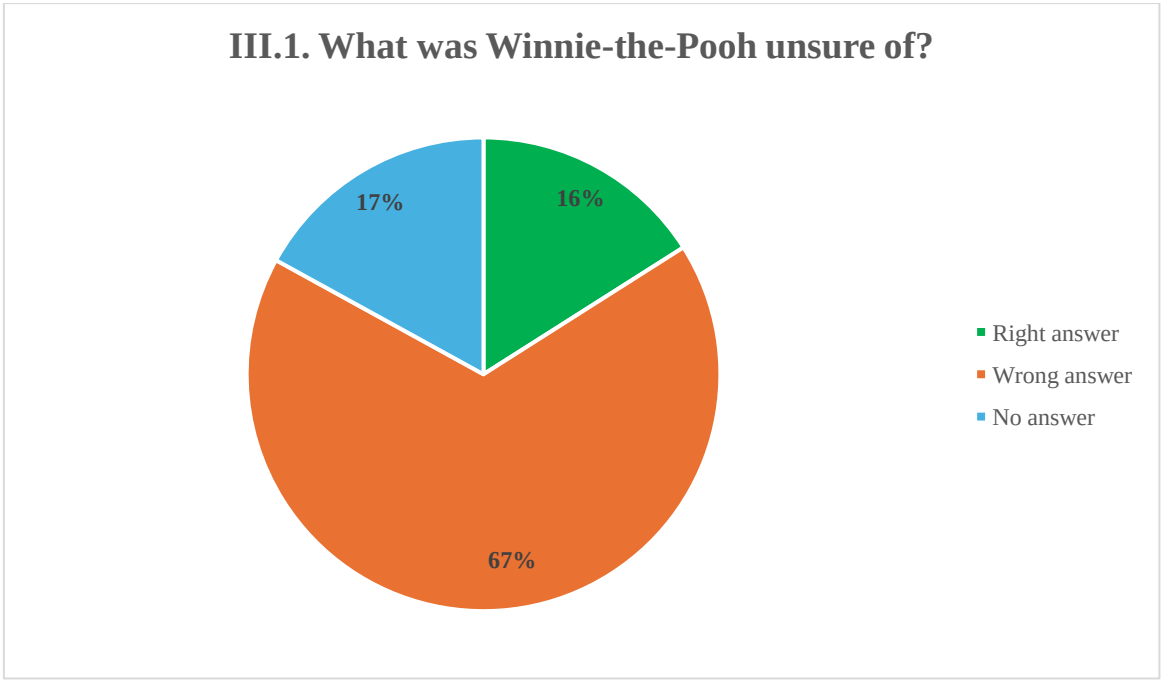

Figure 6. Answers to comprehension question 1

It is possible that the high number of wrong answers could be explained by the fact that the answer could be found in the first paragraph of the output text. Perhaps the participants did not remember or did not have the patience to go back and check what the correct answer was.

The proportion of correct answers for question 2 was much higher, as can be seen in Figure 7. Nobody gave any wrong answers, some participants only chose not to answer in any way.

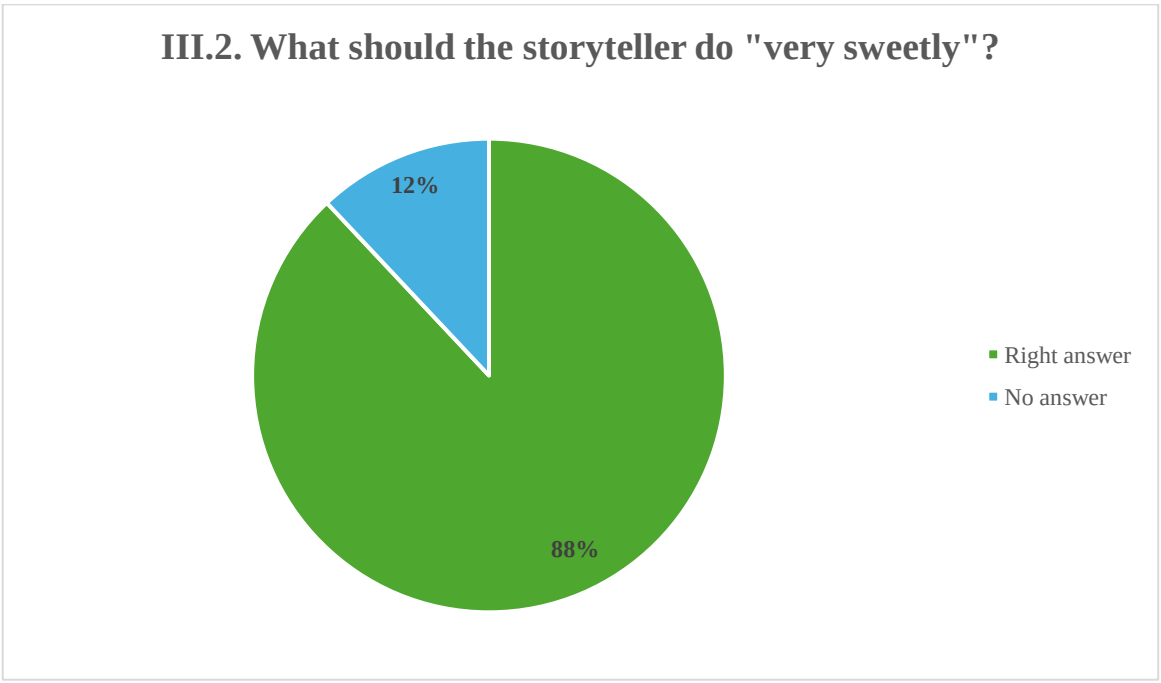

Figure 7. Answers to comprehension question 2

It is interesting to note that “very sweetly” refers to “foarte drăguț” in Romanian, which is not a collocation that a native speaker would select – *a face ceva foarte drăguț*. However, it

seems that the participants ignored this lack of fluency and were able to understand the idea behind the words.

The proportion of wrong answers for the last question is 28%. This number is however inflated because some participants jokingly pointed to the bear's gender. "Sort of bear" is the translation for "gen de urs". Some younger evaluators jokingly pointed to the gender of the bear while answering this question because "gen" means either "gender" or "type/kind of" in Romanian.

Plus, a few decided to do a short character analysis and just called Winnie-the-Pooh a narcissist. Still, the correct answer and the only one that was counted was that Winnie-the-Pooh is the sort of bear that likes to hear stories about himself.

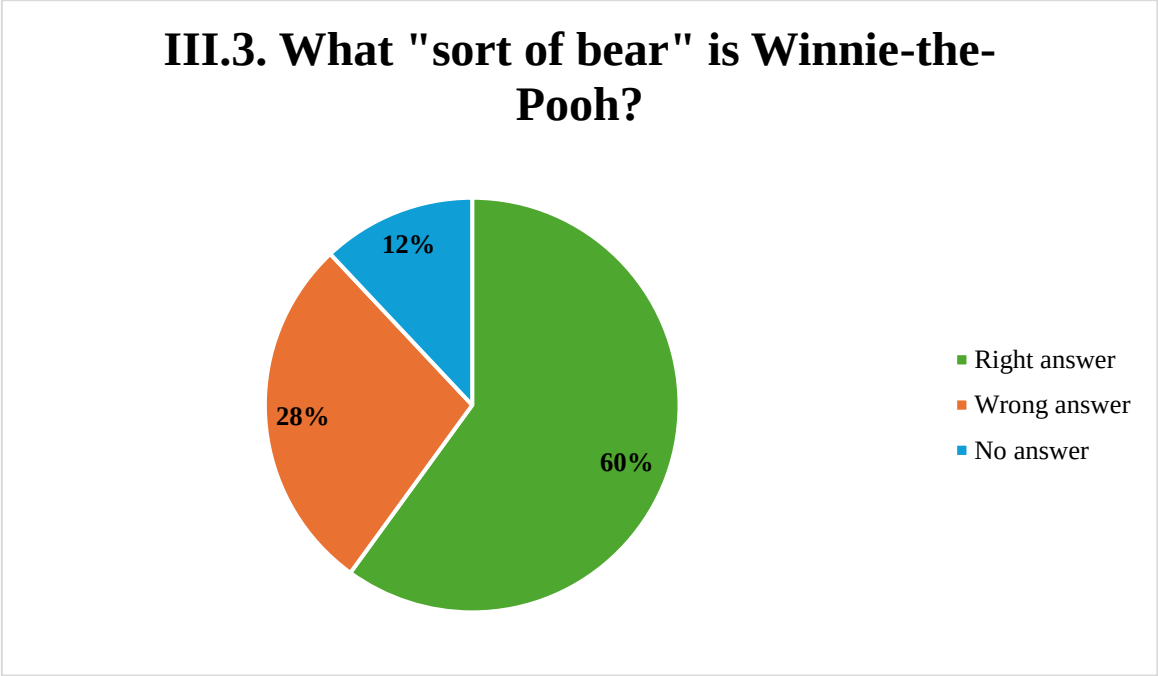

Figure 8. Answers to comprehension question 3

The overall comprehension value was 54.6%, but the individual values varied greatly. No correlation between the number of marked errors and the validity of the given comprehension answers could be drawn. For example, participant 23 marked a total of 244 words (the highest being 294) but managed to give 3 correct comprehension answers, whereas participant 9 highlighted only 57 words but still answered all three questions correctly. On the other hand, participant 7 marked 33 errors but only gave one correct answer. All participants gave at least one correct answer or did not answer these comprehension questions at all.

Conclusion

This study managed to successfully apply the informative manual evaluation method proposed by Popović (2020), focusing on machine translated children's literature. The participants were able to mark the words, sentences and phrases they deemed to be containing incomprehensible and ungrammatical or unstylistic content.

The next step would be to look into more detail at the grammar and style errors they have identified in order to better grasp what kind of mistakes there were in the target text, and what matters most to actual end-users. Maybe a system should be developed where inter-annotator agreement could be calculated, but the issue here might be that different people might assign different spans (Popović 2020) to the errors they identify.

All in all, when it comes to comprehension, since the proportion of correct answers varied so greatly, no definite conclusion could be drawn. Even though the average value was

54.6% correct answers, it is not clear what counts towards higher comprehension scores – the number and difficulty of the questions, the parts of the text that are taken into account, maybe even the age or the level of tiredness of the participants or their expectations about the text. Further investigations into this matter need to be undertaken.

In the end, this work provides a glimpse into the way end-users interact with and understand machine translated literary texts, especially for a language combination that is usually not at the forefront of research.

References:

Arenas, Ana Guerberof, and Antonio Toral. 2022. "Creativity in Translation: Machine Translation as a Constraint for Literary Texts". *Translation Spaces* 11(2):184–212. doi: 10.1075/ts.21025.gue.

Castilho, Sheila, and Sharon O'Brien. 2017. "Acceptability of Machine-Translated Content: A Multi-Language Evaluation by Translators and End-Users". *Linguistica Antverpiensia, New Series – Themes in Translation Studies* 16. doi: 10.52034/lanstts.v16i0.430.

Cespedosa Vázquez, Ana Isabel, and Ruslan Mitkov. 2023. "Machine Translation of Literary Texts: Genres, Times and Systems". *Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications*. Eds. Raquel Lázaro Gutiérrez, Antonio Pareja, and Ruslan Mitkov. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, pp. 48–53.

Hansen, Damien, and Emmanuelle Esperança-Rodier. 2022. "Human-Adapted MT for Literary Texts: Reality or Fantasy?". *NeTTT 2022*. Rhodes, Greece, pp. 178–90.

Kasperė, Ramunė, Jolita Horbačauskienė, Jurgita Motiejūnienė, Vilmantė Liubinienė, Irena Patašienė, and Martynas Patašius. 2021. "Towards Sustainable Use of Machine Translation: Usability and Perceived Quality from the End-User Perspective". *Sustainability* 13(23):13430. doi: 10.3390/su132313430.

Macken, Lieve, Bram Vanroy, Luca Desmet, and Arda Tezcan. 2022. "Literary Translation as a Three-Stage Process: Machine Translation, Post-Editing and Revision". *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, edited by Helena Moniz, Lieve Macken, Andrew Rufener, Loïc Barrault, Marta R. Costa-jussà, Christophe Declercq, Maarit Koponen, Ellie Kemp, Spyridon Pilos, Mikel L. Forcada, Carolina Scarton, Joachim Van den Bogaert, Joke Daems, Arda Tezcan, Bram Vanroy, and Margot Fonteyne. Ghent, Belgium: European Association for Machine Translation, pp. 101–10.

Matusov, Evgeny. 2019. "The Challenges of Using Neural Machine Translation for Literature". *Proceedings of the Qualities of Literary Machine Translation*. Eds. James Hadley, Maja Popović, Haithem Afli, and Andy Way. Dublin, Ireland: European Association for Machine Translation, pp. 10–19.

Popović, Maja. 2020. "Informative Manual Evaluation of Machine Translation Output". *Proceedings of the 28th International Conference on Computational Linguistics*. Eds. Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 5059–69.

Ruffo, Paola. 2018. "Human-Computer Interaction in Translation: Literary Translators on Technology and Their Roles". *Proceedings of the 40th Conference Translating and the Computer*. London, UK: AsLing, pp. 127–31.

Şahin, Mehmet, and Sabri Gürses. 2019. "Would MT Kill Creativity in Literary Retranslation?". *Proceedings of the Qualities of Literary Machine Translation*. Eds. James Hadley, Maja Popović, Haithem Afli, and Andy Way. Dublin, Ireland: European Association for Machine Translation, pp. 26–34.

Sibuea, Todo F. B., Wahyu Budi, and Kenny Jonathan. 2023. "The Equivalence Problems Produced by Machine Translation on A Literary Text: A Study on The Indonesian Translation of Harry Potter: The Order of Phoenix". *Culturalistics: Journal of Cultural, Literary, and Linguistic Studies* 7(1):1–12. doi: 10.14710/ca.v7i1.17417.

Taivalkoski-Shilov, Kristiina. 2019. "Ethical Issues Regarding Machine(-Assisted) Translation of Literary Texts". *Perspectives* 27(5):689–703. doi: 10.1080/0907676X.2018.1520907.

Toral, Antonio, and Andy Way. 2014. "Is Machine Translation Ready for Literature?". *Proceedings of Translating and the Computer 36*. London, UK: AsLing, pp. 174–76.

Toral, Antonio, and Andy Way. 2018. "What Level of Quality Can Neural Machine Translation Attain on Literary Text?". *Translation Quality Assessment: From Principles to Practice*, edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty. Cham: Springer International Publishing, pp. 263–87.

Vanmassenhove, Eva, Dimitar Shterionov, and Andy Way. 2019. "Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation". *Proceedings of Machine Translation Summit XVII: Research Track*. Eds. Mikel Forcada, Andy Way, Barry Haddow, and Rico Sennrich. Dublin, Ireland: European Association for Machine Translation, pp. 222–32.

Way, Andy. 2018. "Quality Expectations of Machine Translation". *Translation Quality Assessment: From Principles to Practice*. Eds. Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty. Cham: Springer International Publishing, pp. 159–78.

Notes on the author

Alina Rădoi is a PhD student at the West University of Timișoara, Romania. She is also a freelance translator and localizer specialized in the marketing and education field. Plus, she teaches English and German as a foreign language. Her research interests include machine translation, error analysis and end-user and expert perceptions on (machine) translated texts.