Optimising Translation Tools for Post-Editing: Results of a User Survey

Marie Escribe[1,2] [0009-0002-7835-6755], Miguel Ángel Candel-Mora[1][0000-0001-8754-6046]

[1] Universitat Politècnica de València, Spain
[2] LanguageWire, Spain
mcescrib@doctor.upv.es, mcandel@upv.es

Abstract. The advances in Machine Translation led to the application of this technology in the translation workflow, thus resulting in Post-Editing being performed in Computer-Assisted Translation tools. While these tools have experienced a rather steady evolution since the commercialisation of the first systems, a plethora of enhancements can now be envisaged due to the latest technological developments in Artificial Intelligence. In this context, the present study seeks to elucidate Post-Editing needs and to identify potential features which could cater to those. To that end, the outcomes of a user survey were examined. This analysis allowed for determining the most complex types of Machine Translation errors, editing actions and decisions during Post-Editing. A feature wish list was also proposed, and the results allowed for identifying the most popular functionalities, which are aligned with Post-Editing needs.

Keywords: Post-Editing, Computer-Assisted Translation, user needs.

1 Introduction

The landscape of Computer-Assisted Translation (CAT) tools has undergone a significant evolution in recent years, transitioning from the early days of Translation Memories (TMs) and Term Bases (TBs) to the contemporary era of hybrid approaches integrating TMs with Post-Editing (PE) of Machine Translation (MT) output. PE has become a common practice in the translation workflow, and this evolution has entailed changing needs and expectations for language professionals. As a result, translators engaging in PE tasks had to navigate this change, adapt to new ways of working and new guidelines, and learn to find a balance between efficiency and quality.

While CAT technology did not seem to adjust to PE initially, the industry is witnessing a progressive incorporation of state-of-the-art features, such as Quality Estimation (QE) and Artificial Intelligence (AI) assistants based on Large Language Models (LLMs). In today's rapidly evolving technological landscape, collecting user opinions is paramount in software development. The primary objective of the present study hence lies in gathering and analysing post-editors' feedback with a view to formulating recommendations to optimise translation tools specifically for PE.

Proceedings of New Trends in Translation and Technology, pages 63–75
3–6 July, 2024, Varna, Bulgaria.
https://doi.org/10.26615/issn.2815-4711.2024_006

2 Related Work

In 1980, Martin Kay described the Translator's Amanuensis [1], a system designed to help translators by centralising various functionalities and automating certain tasks. In the following decade, the first commercial CAT tools were developed, and today a wide range of tools is available to translators. Following the great improvements in quality, MT has been introduced in the translation workflow, and PE can now occur in CAT tools. Despite this change, the main components of CAT remain TMs and TBs. PE is however a different task, and requires a specific skill set, in particular when it comes to error handling, decision making and guidelines application [2].

A few attempts to adapt tools for PE have been introduced. For example, interactive MT consists of generating MT outputs dynamically based on translators' input [3]. QE models can provide word and segment-level quality information, as implemented in IntelliCAT [4], which also comes with alternative translations. More diverse interaction modalities, such as voice and touch commands, have been explored in the Multi-Modal Post-Editing (MMPE) interface [5]. Improved contextual visualisation of the resulting document has also been suggested [6] as a way to overcome decontextualisation issues due to text segmentation [7]. An analysis of behavioural patterns can also serve to predict user actions, and thus build adaptive software, as done in the case of Escriba [8]. More recently, the integration of AI assistants has allowed for satisfying the need for in-context definitions (e.g. Matecat's AI assistant[1]) and could potentially be extended by identifying keywords and producing images to ensure a good understanding of source content, as suggested by Alonso and Nunes Vieira [6]. Rewriting assistance (e.g. changing the style or degree of formality) is also available today via SmartCat's AI actions[2] or the OpenAI Translator plug-in for Trados[3].

In the light of prospective developments, it is essential to find out which of the features discussed above would cater to real user needs and thus enhance the PE experience. In that regard, Moorkens and O'Brien [9] investigated user attitudes towards PE interfaces and identified several suggestions for improvement, including confidence scores and dynamic MT adaptation. The authors also concluded that translation tools required to be improved before even considering PE, with higher customisability for example. Ad-hoc studies revolving around the evaluation of specific features are also relevant to gain a deeper understanding of the impact of certain features on PE. For example, a study by Béchara et al. [10] found that confidence scores contributed to improving efficiency and reducing cognitive load.

However, focusing on the broader picture is at least equally important. With the recent advances in technology, new possibilities which could not have been envisaged a few years ago – such as integrating AI assistants in CAT tools – now become conceivable. But instead of developing new functionalities only because it is technically feasible, one should question their value for end users. It therefore appears essential to gather

[1] https://guides.matecat.com/ai-assistant (accessed: 28/04/2024)
[2] https://www.smartcat.com/news/gpt-4-release/ (accessed: 28/04/2024)
[3] https://community.rws.com/product-groups/trados-portfolio/rws-appstore/w/wiki/6651/openai-translator (accessed: 28/04/2024)

post-editors' opinions. The present study aims to address this gap by analysing the results of a PE user survey.

3 Methodology

This section describes the methodology used to design the survey. It consists of three main parts, covering demographical data, PE experience and a feature wish list.

The first section, Demographics, consists of nine elements to clarify the background of respondents, including age, technical literacy level, employment type, experience in translation and in PE, number of PE jobs usually completed per month, domain (with answers based on LanguageWire's list of industries[4]), usual language combination and most frequently used tool(s). All questions are close-ended, except the language combination, where respondents are advised to enter their answer in the format "Source>Target".

The second section focuses on the User Experience (UX) of post-editors. It is divided into four subsections: overall satisfaction, PE actions, PE environment characteristics, and UX elements (extracted from various usability models and UX questionnaires, such as [11], [12] and [13]). This section comprises 37 questions in total (33 of which are ratings, three are multi-choice, and one is open-ended). As the present work focuses on the optimisation of translation tools for PE, the analysis presented herein is limited to the overall satisfaction and PE actions, while the PE environment characteristics and UX elements will be the object of a follow-up study. The overall satisfaction subsection includes two questions about the general satisfaction with tools for PE and with MT quality. The PE actions subsection was composed of three multiple-choice questions:

- *Which type of MT error(s) do you consider the most challenging to correct?* For this question, a list of 25 MT errors inspired by Costa et al. [14] is provided, and respondents are offered an *Other* option to indicate errors which do not appear in this list.
- *Which editing action(s) do you find the most challenging?* The possible answers include the four editing actions (deleting, inserting, moving and replacing) [15], as well as *It depends on the context* and *None*.
- *Which decision(s) do you find the most challenging?* Four decisions are provided (deciding whether to edit a segment, which correction to apply to an MT error, order in which corrections should be addressed, when and how often to refer to external resources) together with an *Other* option.

Finally, the last section consists of a feature wish list, where respondents are asked to indicate the relevance of potential PE-centred features on a 5-point Likert scale. The list contains 11 features derived from previous work of Escribe and Candel-Mora [16]. For each feature, a short definition is also provided as follows:

- ***Interactive post-editing****: the translation is generated and updated on the fly, as you are typing.*

[4] https://www.languagewire.com/en/industries (accessed: 28/04/2024)

- ***Segment-based alternative suggestions****: offers alternative translations for each segment.*
- ***Word-based alternative suggestions****: offers alternative translations for each target word.*
- ***Confidence scores (quality estimation)****: for each MT segment, a score indicates the confidence estimation of MT quality.*
- ***Knowledge feature****: provides definitions, examples or images upon request to help you understand difficult concepts in the source text.*
- ***Rewriting assistance****: offers rewriting suggestions for pre-defined situations, such as changing the style or shortening a translation.*
- ***Multimodal interactions****: using different interaction modalities, including touch and speech input, as well as text to speech.*
- ***Document view****: switch from segment to document view, where the text can be revised and modified in context.*
- ***Effort prediction****: an expected time to completion for the current project is provided based on past performance.*
- ***Attention monitoring****: an analysis of behavioural patterns detects drops in attention and displays a warning or offers assistance.*
- ***Post-task feedback****: after project completion, provides immediate feedback based on the proportion of accepted, amended and rejected segments.*

In addition, two open-ended questions were included to collect feedback on the items listed above and to gather new feature ideas.

4 Survey Results

The survey remained open for approximately one month (from 12/02/2024 to 10/03/2024). It was distributed to professional translators via direct emails or messages and was shared as a public post via LinkedIn. In addition, a second LinkedIn post was published in a private group for freelance language experts working for LanguageWire. A total of 126 answers were collected.

4.1 Demographics

Most survey respondents are between 25 and 44 years old (Fig. 1), 50% (63) of them consider to be proficient with technology (Fig. 2), and the majority are freelancers (Fig. 3). The seven respondents who answered “other” include one student, one Language Service Provider (LSP) owner, two freelancers with another full-time job, and three scholars. The distribution of the PE workload is rather unbalanced, with 35% (43) completing less than five PE jobs per month and 25% (32) completing more than 20 (Fig. 4). Nevertheless, these figures only provide an overall idea, as the word count per PE job is not specified here. 42% (53) respondents have over 10 years of experience in translation (Fig. 5). For PE however, most respondents have between 1 and 6 years of experience (Fig. 6), with more specifically 1-3 years for 33% (42), and 4-6 years for 38% (48).

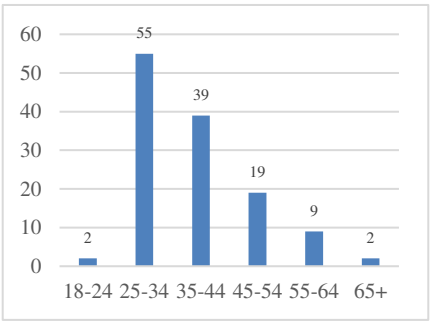


Fig. 1. Age distribution.

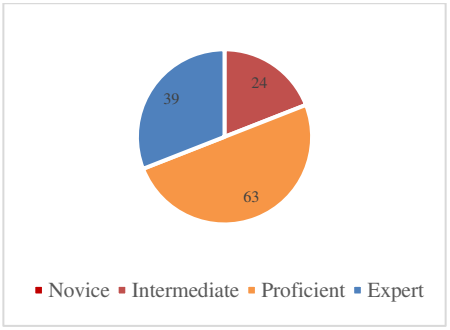


Fig. 2. Technical literacy level.

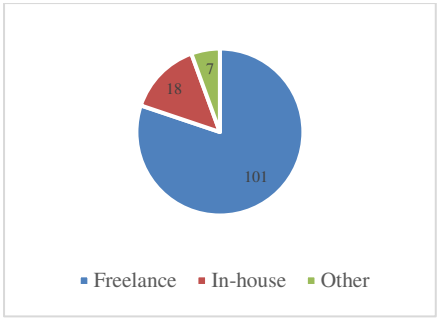


Fig. 3. Employment type.

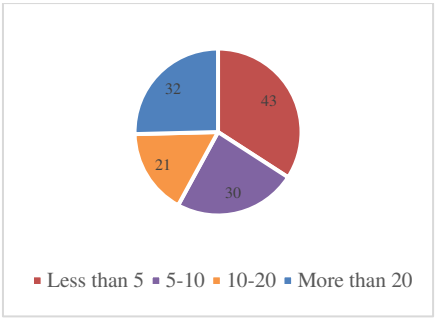


Fig. 4. Monthly workload (PE jobs).

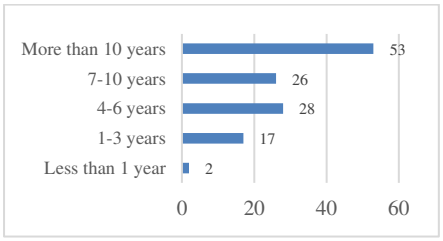


Fig. 5. Experience in translation.

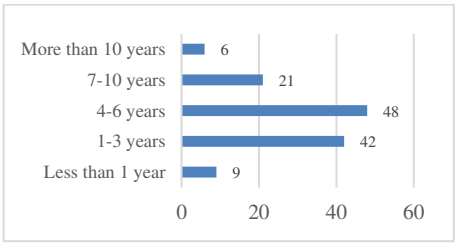


Fig. 6. Experience in PE.

The most recurring target languages (Fig. 7) are Italian (32), Spanish (27) and English (22). Respondents work in a wide variety of domains, with the top three including marketing and advertising, technology, and life sciences and healthcare (Fig. 8). Apart from the domains listed in the close-ended question, some respondents mentioned videogames (5), literature (2), education (2) and human resources (2).

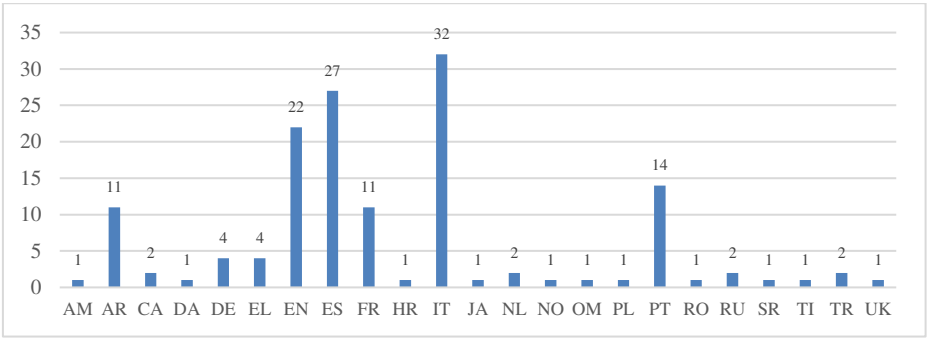


Fig. 7. Target languages (ISO 639 two-letter codes).

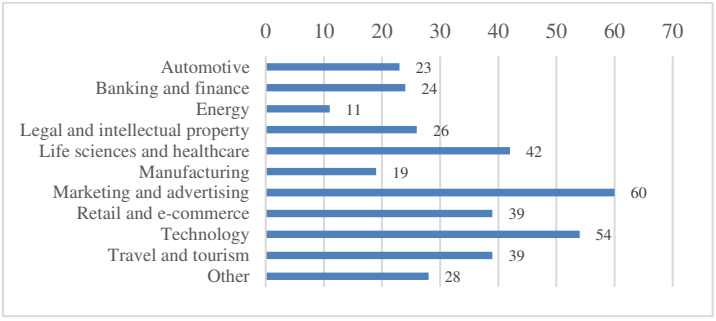


Fig. 8. Domains.

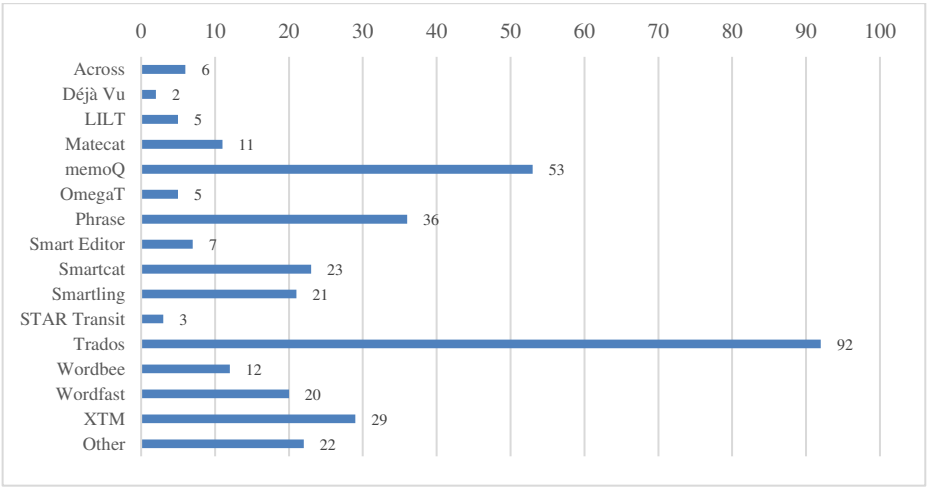
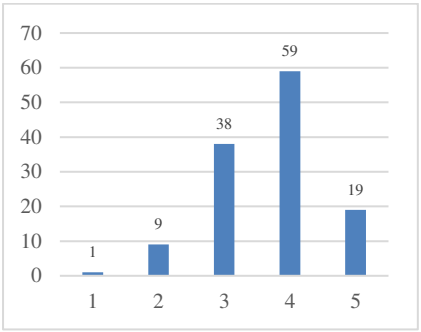


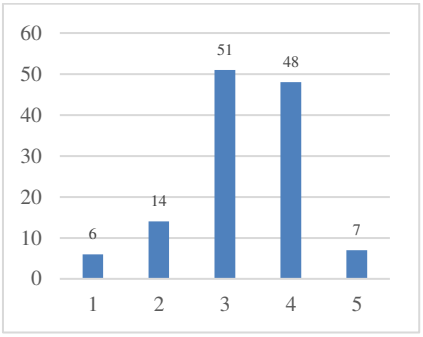
Fig. 9. Most frequently used tool(s).

As far as tool usage is concerned, Trados with 92 users and memoQ with 53 users lead the list, followed by Phrase with 36 users (Fig. 9). Several respondents also mentioned GlobalLink (4) and Polyglot (3).

4.2 Satisfaction and PE needs

The overall satisfaction with translation tools for PE received an average score of 3.68 (on a 5-point Likert scale), with 46% (59) of respondents giving a score of 4, while 30% (38) were more neutral and gave a score of 3 (Fig. 10). The outcome is more nuanced in the case of the overall satisfaction with MT quality. While the average satisfaction is 3.29, 40% (51) respondents were neutral, and 38% (48) were satisfied (Fig. 11).





1: Not satisfied at all – 5: Extremely satisfied.

Fig. 10. Satisfaction with tools for PE.

Fig. 11. Satisfaction with MT quality.

Regarding MT error handling, cultural adaptation appears to be the most challenging (76), followed by mistranslation (60) and source language interference (60). Frequently mentioned errors also comprise adherence to guidelines and reference material, document-level errors and idiomaticity (Fig. 12). It should be mentioned here that out of the seven free-text answers collected for this question, four respondents mentioned issues related to tag handling. As illustrated in Fig. 13, the difficulty of editing actions appears to be correlated with the context (81), with the most challenging one being replacing parts of the MT output (40). When it comes to decision making (Fig. 14), the most challenging decisions seem to be deciding whether to edit a segment or not (64), followed by deciding when and how often to refer to external resources (48). In addition, six comments were provided in the *Other* answer option. Two focus on consistency, including consistency between the MT output and TM content, as well as consistency in style and terminology, which can be challenging for long projects. Two other comments emphasised the difficulty of estimating the severity of an error and the extent to which a segment should be modified.

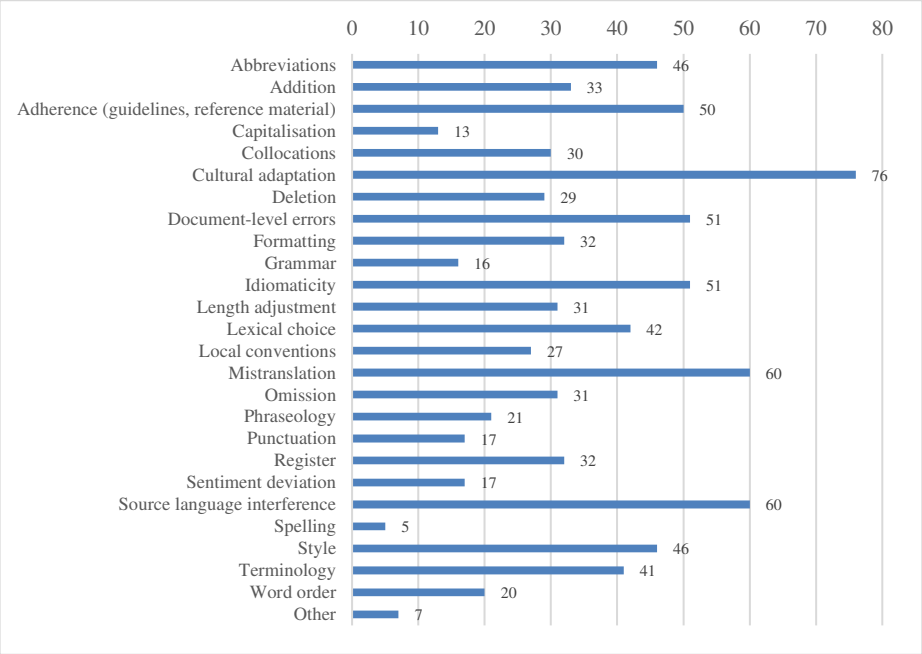


Fig. 12. MT error handling difficulty.

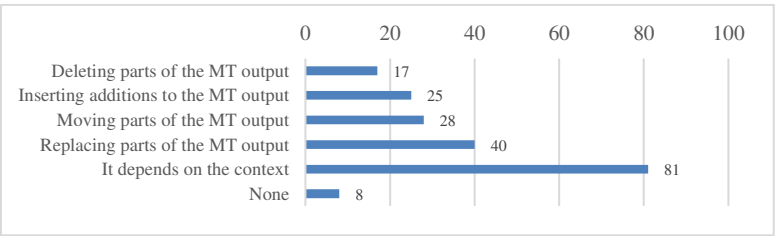


Fig. 13. Editing action difficulty.

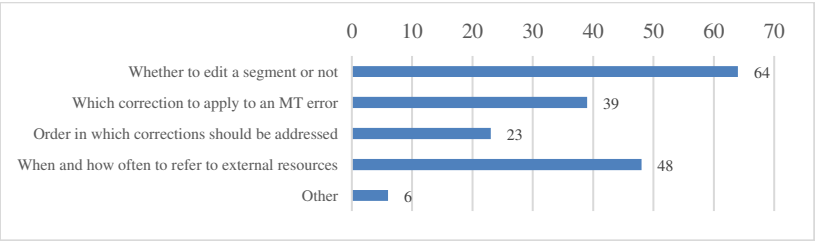


Fig. 14. Decision difficulty.

4.3 Feature wish list

Most features in the wish list appeared to be relevant since none received an average score below 3 (Fig. 15). The feature with the highest score is "document view" with an average of 4.23, followed by the "knowledge feature" (4.06) and "rewriting assistance" (3.96). These features also received the highest number of votes in the most positive category, with 63 for "document view", 61 for the "knowledge feature", and 49 for "rewriting assistance" (Fig. 16). Conversely, the least popular items, "attention monitoring" (3.03) and "multimodal interactions" (3.10) received the highest number of negative votes (21 in both cases).

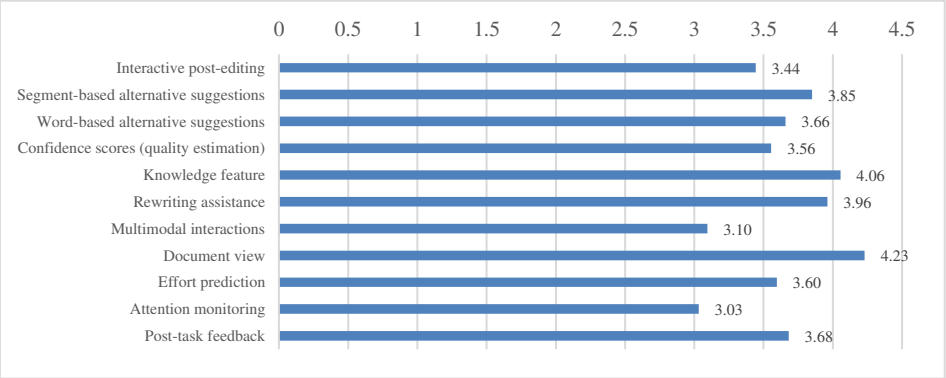


Fig. 15. Average feature scoring.

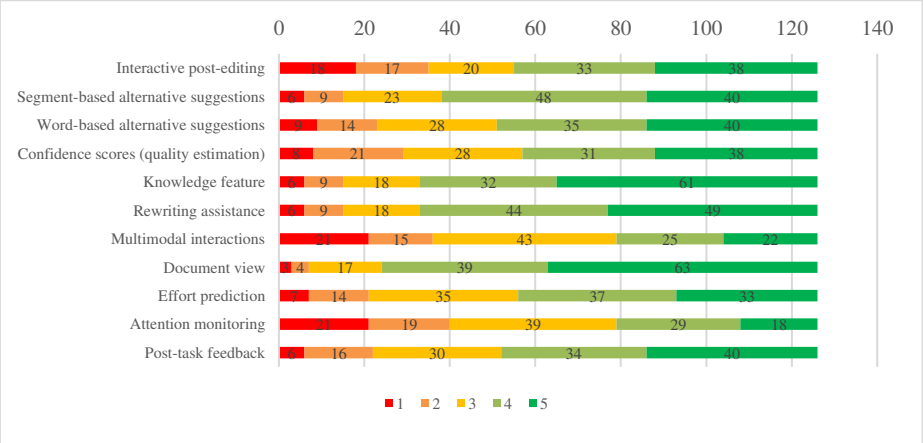


Fig. 16. Score distribution (1: Not relevant at all – 5: Very relevant.).

Respondents had the possibility to leave additional feedback on the features of the wish list. QE attracted the attention of four linguists, who seemed to be in favour of this functionality, but also advised that it could also "be a double-edged sword" because it

can be "very useful if the estimation is extremely accurate" but "distracting otherwise". One respondent mentioned that confidence scores could help in the decision-making process since "the tricky part of MT is to spot when it's actually not that good even though it seems so". Nonetheless, pricing appears to be a concern in this case, since LSPs can use QE to filter out segments above a predefined confidence score, and one person reported a negative experience in that regard. Similarly, another respondent was worried that an attention monitoring feature could result in "micro-managing and hyper-analysis of productivity". Alternative translations, in contrast, generated enthusiasm. One respondent suggested to provide alternatives based on past corrections implemented during PE, and another one proposed to combine this feature with the knowledge feature, and thus providing alternative translations together with "definitions or more context, to better understand which option to choose". Another respondent highlighted the importance of designing features that let translators stay in control and focus on their task without interfering heavily in their thinking process: "I do not enjoy features that interfere too much with my input, but I value features that support the translation work (instead of feeling like I support the machine, not the other way around) and enhance my productivity by cutting corners (e.g., features such as Confidence scores and Knowledge feature […]) and allowing me to focus fully in the translated text".

Finally, respondents were given the chance to describe other feature ideas. In total, 12 people shared their suggestions. Some answers mentioned features such as tag insertion, search and replace, access to TBs, history tracking and quality assessment. However, these functionalities are already available in various CAT tools today, and they are not specific to PE. Therefore, we focus here on more original suggestions. Below is the list of features extracted from these answers:

- **Unpopulated segments**: leaving target segments empty when the MT quality is not satisfactory in order to avoid the "anchoring effect"
- **Highlighting unedited MT output**: clearly differentiating segments which have been post-edited from those which still require to be checked ("with MT target language already present in the target segment, it's often hard to discern at a glance what you have and haven't edited and makes it harder to 'trust' that you're 100% ready to sign off on a segment, which slows down decision-making")
- **AI assistance** (i.e. integrating GPT technology), including
    - Idiomaticity support: assistance to provide support for idiomatic expressions
    - Inclusive suggestions: rewriting target segments in a more inclusive language
- **Correction propagation**: automatically applying a lexical or terminological change to the rest of the text
- **Time tracker**: integrated time tracking functionality which could provide productivity statistics for each project
- **MT engine comparison**: analysing the differences between outputs from different MT systems in order to identify which one is the most similar to the

final text after PE ("so that, if one engine seems to 'think' more in the way I do, I can realise that and start using that engine for that type of text")

Arguably, differentiating unedited MT output is to some extent already covered in several tools. For example, in Trados and Smart Editor, segments which were checked are indicated with a specific icon (a green checkmark in both cases). However, this suggestion may reveal a need to distinguish more clearly unedited machine-generated content.

AI assistance does not come as a surprise in the era of GenAI. Some integrations, like the OpenAI Translator plug-in for Trados, already offers the possibility to reformulate translations based on predefined prompts.

The proposal for correction propagation is reminiscent of online adaptation of automatic PE systems, including PEPr [17] and the incremental adaptation introduced by Escribe and Mitkov [18]. While having an online model running always in production can be expensive, this suggestion shows that adaptive systems capable of leveraging human input on the fly would be highly beneficial.

4.4 Discussion

The average satisfaction with translation tools for PE (3.68) indicates that there is substantial room for improvement. The difficulties encountered in MT error handling seem to be rather well aligned with the feature wish list. Cultural adaptation for example, could be supported by AI assistance for rewriting thanks to tailored prompts. In the case of other errors considered challenging (mistranslation, source language interference, and idiomaticity), it can be more difficult to provide support since these errors require a thorough understanding of the source and target languages. However, QE may come in handy to spot such errors. Adherence to guidelines and reference material also appeared to be challenging, which is in line with the decisions considered difficult, in particular deciding when and how often to refer to external resources. In that regard, the knowledge feature, which was among top voted items, could be helpful, in particular if it can be customised to extract contextual information from reference documentation. As suggested by one respondent, such a feature could also use this contextual knowledge to provide alternative translations. Document-level errors, which also appear among the most challenging MT errors, could be easier to handle with a propagation mechanism and potentially with enhanced visualisation of the document, which was the highest scoring feature.

The other feature ideas, in particular highlighting unedited MT and leaving unpopulated segments, emphasise the need for clarity and for making enough space for the post-editors' thinking process. As pointed out by one respondent, the PE environment should not interfere with the main task at hand, but rather provide support for decision making and other activities related to the main task, such as researching unfamiliar concepts or checking reference material. This is reminiscent of Kay's proposal [1], whereby technology would gradually take charge of certain tasks, but without

supplanting the human translator, who would remain in the driver's seat. In this sense, the knowledge feature appears particularly pertinent.

Furthermore, some concerns were raised regarding the use of QE and attention monitoring. The consequences of using such features should be investigated further, notably when it comes to pricing models.

5 Conclusions and Future Work

This study provided insights into post-editors' needs by elucidating which MT errors, editing actions and decisions are the most challenging during PE, and by identifying possible features to cater to these needs. The results show that cultural adaptation, mistranslation and source language interference are the most laborious when it comes to error handling, followed by adherence to guidelines and reference material, document-level errors and idiomaticity; the difficulty of editing actions depend on the context; and the most challenging decisions were whether to edit a segment or not and when and how often to refer to external resources. Overall, the difficulties identified align with the preferred features, namely the document view, the knowledge feature and rewriting assistance. In addition, further suggestions were proposed by respondents, including, for example, highlighting unedited MT and automatic correction propagation. These findings should help in guiding software developers to design useful improvements to translation tools that cater to PE needs.

It should be acknowledged here that while the sample size is considered sufficient to formulate practical recommendations, it can be limited in comparison to other studies (e.g. [9]). The clear prevalence of Romance languages may also have affected the results. It would thus be recommended to extend this survey by disseminating it to more post-editors working with different target languages. Furthermore, it should be noted that the number open-ended questions was kept to a minimum in order to reduce the time required to complete the survey given the total number of questions. However, open-ended questions are essential to capture certain details and nuances, therefore it would appear relevant to combine the outcomes obtained here with interviews or focus groups to gain a deeper understanding of user needs.

Finally, it should be mentioned that this research is part of a wider project, and the portion of the survey focusing on UX dimensions will be analysed in a follow-up study with a view to designing an evaluation framework for the PE environment.

Acknowledgments. This research is conducted as part of an industrial Ph.D. agreement between the Universitat Politècnica de València and LanguageWire.

References

1. Kay, M.: The proper place of men and machines in language translation. Xerox PARC CSL-80-11 (1980)

2. Ginovart Cid, C.: The need for practice in the acquisition of the post-editing skill-set: lessons learned from the industry. Ph.D. Thesis, Universitat Pompeu Fabra (2021)
3. Peris, Á., Casacuberta, F.: Online learning for effort reduction in interactive neural machine translation. Computer Speech & Language 58(1), pp. 98–126 (2019)
4. Lee, D., Ahn, J., Park, H., Jo, J.: IntelliCAT: Intelligent machine translation post-editing with quality estimation and translation suggestion. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pp. 11–19 (2021)
5. Herbig, N., Düwel, T., Pal, S., Meladaki, K., Monshizadeh, M., Krüger, A., van Genabith, J.: MMPE: A multi-modal interface for post-editing machine translation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp 1691–1702 (2020)
6. Alonso, E., Nunes Vieira, L.: The translator's amanuensis 2020. The Journal of Specialised Translation, pp. 345–361 (2017)
7. Candel-Mora, M. Á.: Comparable corpus approach to explore the influence of computer-assisted translation systems on textuality. Procedia-Social and Behavioral Sciences, 198, pp. 67–73 (2015)
8. Porto Veloso, P.: Escriba, an adaptive web CAT tool. MA Dissertation, University of Dublin, Trinity College (2013)
9. Moorkens, J., O'Brien, S.: Assessing user interface needs of post-editors of machine translation. In Kenny, D. (ed.). Human Issues in Translation Technology. Routledge, London and New York, pp. 127–148 (2017)
10. Béchara, H., Orăsan, C., Parra Escartín, C., Zampieri, M., Lowe, W.: The role of machine translation quality estimation in the post-editing workflow. Informatics 8(3) (2021)
11. Nielsen J.: Usability Engineering. Morgan Kaufmann, San Francisco, CA (1994)
12. Brooke, J.: SUS: A 'quick and dirty' usability scale. In Jordan, P. W., Thomas, B., Weerdmeester, B. A., McClelland, I. L. (eds.). Usability Evaluation in Industry. Taylor & Francis, London, pp. 189–194 (1996)
13. Hassenzahl, M., Burmester, M., Koller, F.: AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. Mensch & Computer 2003: Interaktion in Bewegung, pp. 187–196 (2003)
14. Costa, Â., Ling, W., Luís, T., Correia, R., Coheur, L.: A linguistically motivated taxonomy for machine translation error analysis. Machine Translation 29, pp. 127–161 (2015)
15. do Carmo, F. E. M: Post-editing: a theoretical and practical challenge for translation studies and machine learning. Ph.D. Thesis, Universidade do Porto (2017)
16. Escribe, M., Candel-Mora, M. Á.: Envisioning the post-editor's workstation: a backward glance and a glimpse into the future. In: Proceedings of Translating and the Computer 45 (2023)
17. Simard, M., Foster, G.: PEPr: Post-edit propagation using phrase-based statistical machine translation'. In: Proceedings of the XIV Machine Translation Summit, pp. 191–198 (2013)
18. Escribe, M., Mitkov, R.: Applying incremental learning to post-editing systems: towards online adaptation for automatic post-editing models. In: In Pan, J., Laviosa, S. (eds). Corpora and Translation Education: New Frontiers in Translation Studies. Springer, Singapore, pp. 35–62 (2023)