Post-édition de TAN/LLM et localisation de jeux vidéo : la traduction automatique, bénédiction pour la localisation de *visual novels* ?

NMT/LLM PE and Video game Localisation: Can Visual Novels Benefit from Machine Translation?

Simon COPET

PhD Student, FTI-EII, UMONS

Simon.COPET@alumni.umons.ac.be

Loïc DE FARIA PIRES

Associate Professor, FTI-EII, UMONS

Loic.DEFARIAPIRES@umons.ac.be

URL : https://www.unilim.fr/espaces-linguistiques/995

DOI : 10.25965/espaces-linguistiques.995

Abstract: This exploratory paper presents one of the first comparative studies of PE (post-editing) and HT (human translation) quality in the field of video game localisation. We used an excerpt from the *KillerTrait* free visual novel, selected for its particular features in terms of gender, puns and oral language. This excerpt was submitted to DeepL and ChatGPT engines, and these versions were respectively post-edited from English into French by nine and ten last-year translation students in the framework of their EN > FR post-editing class at the University of Mons. Nine other students from the group translated the text from scratch. In these productions, we analysed fidelity and fluency. We compared the results from human quality evaluation to the HTER and BLEU automatic metrics. Human evaluation shows that the texts post-edited using ChatGPT display better PE quality scores than those post-edited using DeepL in terms of fidelity. Human translation was evaluated differently by both evaluators, thus leading to an impossibility to identify any trend. In terms of fluency, no clear trend can be found. Automatic metrics show that 1) no clear trend emerges in terms of raw MT quality, 2) PE effort is slightly lower (and quality supposedly higher) in the case of the raw ChatGPT output, and 3) quality does not seem to vary when comparing the PE texts produced using each MT engine, but HT displays a higher internal variability, which could indicate that HT provides for less homogeneous productions than PE.
Keywords: NMT, LLM, Localisation, Post-Editing, TQA

Résumé : Le présent article exploratoire présente l'une des premières études comparatives de la qualité de contenus traduits humainement et post-édités en localisation de jeux vidéo. Nous nous sommes fondés sur un extrait du jeu de type *visual novel KillerTrait*, sélectionné pour ses caractéristiques particulières en termes de représentation de genre, de jeux de mots et de recours au registre oral. Cet extrait a été traduit automatiquement par DeepL et ChatGPT, dont les traductions automatiques (TA) brutes ont respectivement été post-éditées de l'anglais vers le français par neuf et dix étudiants en dernière année de Master en traduction, dans le cadre de leur cours de post-édition anglais-français à l'Université de Mons. Neuf autres étudiants inscrits au même cours ont traduit le texte humainement, sans recours à la TA. Nous avons analysé la fidélité et la fluidité de ces productions, et comparé les résultats d'une évaluation humaine aux résultats fournis par deux métriques : les scores BLEU et HTER. Les résultats de l'évaluation humaine vont en faveur d'une fidélité supérieure des contenus post-édités à partir de ChatGPT par rapport aux contenus post-édités à partir de DeepL. La fidélité des textes traduits humainement a, pour sa part, fait l'objet d'un désaccord entre les deux évaluateurs, rendant donc impossible l'identification d'une tendance. Pour la fluidité, aucune tendance claire n'a pu être dégagée. Les métriques automatiques 1) ne permettent pas ici d'identifier de tendance claire en termes de qualité de la TA brute lorsqu'elle est comparée à nos deux traductions de référence, 2) vont dans le sens d'un effort de post-édition très légèrement inférieur dans le cas de ChatGPT par rapport à DeepL, et 3) fournissent des résultats très similaires pour les textes post-édités

à partir des TA fournies par les deux moteurs de TA envisagés, mais montrent une plus grande variabilité interne pour les textes traduits humainement, témoignant de productions plus diversifiées dans ce dernier cas.
Mots clés : TAN, Grands modèles de langage (LLM), Localisation, Post-édition, Qualité

Introduction

Video game localisation is getting more and more prominence in both professional practice and academic research. In 2025, the video game market's estimated value is projected to reach €476.05bn, and this industry is expected to constantly grow in the coming years (Statita, 2025). While the vast number of commercial video games to be localised partly explains the surge in localisation demand, both the increasingly participative nature of localisation practises and the rise of independent video games contribute to "[broadening] the horizons of translation studies and [unlocking] areas of research largely underexplored to date" (Capellini, 2021, p. 1). Indeed, on the one hand, "Many developers automatically plan for the simultaneous shipment of games in more than 10 different languages, which is the case of Triple-A game companies" (Rivas Ginel, 2023, p. 16) and, on the other hand, indie game developers recur to basic localisation services to make their games playable in other parts of the world (Toftedahl *et al.,* 2018, p. 13).

Both these facts justify the importance of video game localisation on the translation professional market, though most of the market value is linked to the localisation of Triple-A games. Over the last few years, there has been a steady increase in the quantity of games released every year, which, in turn, has intensified the need for professional localisation (Rivas Ginel, 2023, p. 16), in particular into the FIGS languages (French, Italian, German and Spanish). For instance, 44% of Activision-Blizzard's revenue comes from international game sales (lowest estimate), while this number increases to 64% in the case of Ubisoft (Transphere, 2023).

However, even though video game localisation revenue is considerable, localisers regularly raise a number of concerns related to their professional practices and their working conditions, which we tried to reproduce in the experiment described in this article. Indeed, due to the very nature and confidentiality of video games, localisers do not always have access to the necessary resources. Although they are often provided with localisation kits containing elements such as glossaries, screenshots, style guides and working documents on the story, information on the characters of the game, among others (Theroine *et al.*, 2021, p. 31), such is not always the case: in her study, Rivas Ginel (2021, p. 42) found that around 25% of translators working on video game localisation did not have such files at their disposal. In the same study, Rivas Ginel (*ibid.*) also found that less than 30% of localisers had access to the actual visual environment of the game while localising. Therefore, localisers can rarely visualise their translations in context, which can lead to typical localisation problems such as translation inaccuracies, an excess of characters causing the text to truncate or overlap… This last problem has a particular incidence when localising from English into French, since "French texts [generally expand] when the source language is English because

of a more important expansion factor" (Ray, 2019, p. 94). Mangiron and O'Hagan also highlighted the lack of access to relevant resources: "The need to provide translators with contextual information is always important for all types of translation, and yet in software environments de-contextualized text fragments, which may belong to different parts of the game are routinely presented" (2013, p. 118).

In addition to this lack of access to necessary resources, working conditions can prove quite precarious for freelance localisers, who represent the majority of the localisation service providers. Both facts are underlined by Rivas Ginel: "Most video game localisers tend to be freelancers, which, in theory, should reduce their access to game content" (2022, p. 304). Deadlines are often tight for such projects (Bernal-Merino, 2007, p. 4), and very few of these professionals can claim that localising video games constitutes their only source of revenue. Therefore, tools such as NMT (Neural Machine Translation) and LLMs (Large Language Models) could prove useful for them to carry out their localisation tasks more quickly and efficiently, should these tools provide an acceptable quality in the framework of video game localisation.

1. Video game Localisation in the Classroom

Such technical and professional challenges, coupled with the increasing importance of the video game localisation industry, justify the need for a broader implementation of this discipline in translation university curricula. For instance, the European Commission's EMT group (European Master's in Translation) stated, in the 2022 version of its competence framework, that video game localisation was a part of the "translation competence": "Students know how to […] [t]ranslate and mediate in specific intracultural and intercultural contexts, for example, those involving public service translation (and interpreting), website or video game localisation and accessibility, community management, etc." (2022, p. 8). This means that the EMT network strongly encourages its member universities to implement video game localisation classes to meet the new criteria which bind them to the network.

As a result, many European universities (including the University of Mons) upgraded their curricula to include localisation classes at a Master's level. In parallel to this trend, MTPE has gained increased attention over the last few years, both in the translation industry (Girletti & Lefer, 2024) and in university curricula. To meet the growing level of productivity required by the professional market, machine translation post-editing (MTPE) seems to be a viable solution for certain text types. The goal of the present article is therefore to investigate whether such is the case for the localisation of video games characterised by a number of specific constraints such as creative elements, the necessity of gender inclusion, as well as polysemic terms.

2. NMT and Post-editing Applied to Video Game Localisation: Literature Review

Though NMT and (more recently) LLM technologies are both used for professional post-editing (Moorkens *et al.*, 2024, p. 2-3), there is little research about such practices when it comes to translating video games. It is often suggested that creative texts, such as literature, audiovisual scripts or video games, tend to resist MTPE because of some features machines struggle with, such as "sarcasm, metaphor, irony, and ambiguous elements of language that are likely to result in a word-by-word translation" (Castilho & Resende, 2022, p. 1). However, the recent advances in AI and MT paradigms have led to an improving raw MT quality for such texts and, therefore, somewhat legitimate MT use in the framework of video game localisation.

Proof of this is the existence of some recent studies in the field. For instance, Hansen & Houlmont (2022) trained a specific NMT engine with a corpus of video games (RPGs) and concluded that, if properly trained, such an engine could provide better results than a generic NMT approach (2022, p. 263) and learn particular localisation strategies, such as gender neutralisation (*ibid.*, p. 265) and therefore prove rather useful when used together with other tools such as translation memories. Similarly, Rivas Ginel and Theroine (2022) studied the use of NMT for video game localisation, and focused their research on gender bias, which is a common NMT problem (Saunders *et al.*, 2020, p. 1), especially when it comes to video games, since these are getting more and more inclusive: "The video game industry is moving with the times and increasing the number of female, gender-neutral and non-sexualised characters, leaving behind the image of a predominantly male domain" (Rivas-Ginel & Theroine, 2022, p. 2). They namely concluded that some NMT engines, such as DeepL, were better at avoiding gender bias than others, such as Google Translate and Smartcat (*ibid.*, p. 9). More recently, Brenner (2024, p. 47) presented ongoing research in which she compares several post-editing methods in a real video game localisation professional setting. To do so, she bases her analysis on process (PE speed, cognitive effort) and quality data (*ibid.*).

These few research papers seem therefore to show that academia is progressively getting interested in a trend that can be observed in the professional market: "Localization processes now often involve post-editing NMT for certain content types and projects" (Jiménez-Crespo, 2024, p. 18). Nonetheless, the way students perform PE in the field of video game localisation has, to the best of our knowledge, never been investigated, in spite of the importance of teaching good PE practises with a view to enabling students, who often fail to spot numerous MT mistakes (Volkaert *et al.*, 2022, p. 6), to be fully prepared to implement it in their future careers.

3. NMT and PE human and automatic evaluation

We intend to study the quality 1) of the raw MT output provided by the MT engines used for the present study, and 2) of the PE texts produced by Master's students.

Two evaluation types are usually used in translation and post-editing quality assessment: automatic metrics and human evaluation, which are often compared (Vilar *et al.*, 2007, p. 96) with a view to reducing biases. Indeed, both approaches have advantages and disadvantages.

On the one hand, automatic evaluation is "quick, efficient, and reproducible" (*ibid.*), since it produces numerical quality scores often based on one or several reference translations. However, the very existence of these reference translations is a problem in itself, as "the number of ways in which it is possible to express the same meaning in a language is very large" (Finch *et al.*, 2004, p. 2019). Therefore, using one or a few reference translations is restrictive and does not take into account the many ways of expressing the same idea in the target language. Many scores can be used; among the most famous are BLEU (Bilingual Evaluation Understudy), presented by Papineni *et al.* (2002), and HTER (Snover *et al.*, 2006). BLEU's main objective is to provide a numerical score used to evaluate raw MT quality (the higher the score, the better the quality), based on one or several reference translations: "The primary programming task for a BLEU implementor is to compare character n-grams of the candidate with these of the reference translation and count the number of matches" (Papineni *et al.*, 2002, p. 312). While HTER also requires reference translations (Snover *et al.*, 2009, p. 259), it is designed to measure PE effort, in that it measures the edit distance (Koponen *et al.*, 2012, p. 1) between the post-edited version and the raw MT proposal: the lower the score, the better raw MT is supposed to be (since fewer operations need to be performed to obtain a defined quality level). As stated by Álvarez-Vidal & Oliver (2023, p. 4-5): "For BLEU […], the higher the value for the automatic metric, the better is the MT quality considered. In the case of […] TER, a lower value states a higher MT quality." Such metrics, despite their well-known limitations, are used because they are "faster, easier and cheaper compared to human evaluations, which require trained bilingual evaluators" (Banerjee & Lavie, 2005, p. 65). The main problem of these metrics is that they "only capture lexical similarity and do not properly measure semantic, grammatical diversity, and sentence structure" (Lee *et al.*, 2023, p. 2).

On the other hand, human evaluation, albeit expensive, is often considered more reliable than automatic metrics, but is not exempt from problems, "such as […] subjectivity, and time-consuming nature. In addition, human evaluators may bring their own biases and preconceptions to the evaluation process, leading to inconsistent results" (*ibid.*).

Since correlation between both types of metrics can be an issue, the present study will be based upon a combination of them to see whether findings remain the same.

4. Objectives and methodological framework

In light of the above, the main objective of this paper is to investigate the quality of translation and post-editing outputs (from English into French) for a selected excerpt of a video game. This will enable us to determine if quality levels vary between translation and post-editing, and between different MT engines in the case of post-editing. Human and automatic evaluation results will be presented and compared.

The selected text is a 500-word excerpt from the *KillerTrait* free video game (Prikarin, 2025). This game, a visual novel, was chosen because it contains some translation challenges, which are typical of video games, such as familiar language, gender neutrality, a lack of context, invented names and concepts, ambiguities. Such elements usually reduce raw MT quality, as will be illustrated below. In real professional settings, translators most often do not get to play the games they need to localise, as stated by Theroine *et al.* (2021, p. 28). Therefore, the students who took part in the study did not play the game either, but were given a link to its itch.io page, which provided them with key information, such as the fact that the player can choose their pronouns before starting the game, indicating that the localised game had to remain gender-neutral in French. Moreover, the segments to be translated or post-edited were presented in the same order they were technically extracted from the game, which caused some links and transitions to disappear, as it is sometimes the case in professional settings (Chojnowski, 2016, p. 84). Twenty-eight students took part in the study: the experiment was carried out in the framework of an EN-FR post-editing class (second year of Master's) attended by all the students registered in the Master's in specialised translation programme at the University of Mons. As such, by the time the study was carried out, they had received extensive translation training and were already familiar with PE practice. Moreover, they had also followed a 30-hour EN-FR localisation class the previous year. Nine of them post-edited the DeepL raw MT (free version), ten of them worked with the same ChatGPT raw MT output (3.5, free version, using the simple prompt "translate this text from English into French"), and ten of them translated the excerpt from scratch (HT).

As previously stated, human and automatic evaluations were implemented. The human evaluation was carried out by two professional translators who were already familiar with post-editing and video game localisation. The evaluation modalities were kept simple: perceived fidelity to the source text and fluency of the target (international French) were evaluated for each segment of each production by both evaluators on a Likert scale ranging from 1 (very bad) to 5 (excellent). This

human evaluation method (measure of fidelity and fluency) is often implemented in Translation studies (for instance by Daems *et al.*, 2013), but is not perfect, as it is quite restrictive and subject to interpretation. However, both evaluators were briefed on what was to be understood by "fidelity" and "fluency", and the implementation of this approach provided for easily comparable overall quality scores. In further research, other evaluation tools such as the MQM (Multidimensional Quality Metrics) or the MTPEAS (Machine Translation Post-Editing Annotation System) tool (Lefer, Piette & Bodart, 2022) could be used to understand where variations in the evaluators' rating diverge.

As far as automatic evaluation is concerned, several metrics were used. First of all, a BLEU score was computed to compare both MT outputs to two human reference translations. Both these reference translations were provided by professional translators with experience in video game localisation and its challenges. These translators worked under the same conditions as the students who took part in the study. An HTER score was also used to measure edit distance between raw MT and the reference translation and was compared to the respective BLEU scores. Then, the HTER score was computed between the post-edited texts and the human reference translations to determine which PE texts differ most from the reference translations, while keeping in mind the restrictive nature of using a low number of reference translations. Finally, HTER scores were also used to compare each PE text to its corresponding raw MT to determine whether one of both MT engines provided a lower edit distance and required a lower PE effort. The limited number of reference translations has to be taken into account when it comes to interpreting the results, but we chose to work with professional translators to ensure that both the reference translations would display a high level of quality. As highlighted by Ehrensberger-Dow & Massey, this limitation is frequent in translation studies: "Rush jobs and unexpectedly heavy workloads may mean that translators do not have time to participate in the research process as planned […]" (2020, p. 363). This explains that, despite our efforts, it was not possible to recruit more translators to produce reference translations.

The automatic evaluation scores were obtained thanks to the PosEdiOn tool described by Oliver, Álvarez-Vidal & Badia (2020), a piece of software created to measure PE effort, which is able to compute scores such as BLEU and TER (made publicly available by the authors at https://github.com/aoliverg/PosEdiOn).

5. Results: human evaluation

5.1. Raw MT

First of all, both evaluators rated the raw MT provided by each MT engine on a Likert scale ranging from 1 to 5. They did so for each of the 41 segments of the text to evaluate fidelity to the text source and fluency of the output in French. These scores were then averaged to obtain a single fidelity and fluency score for each MT output (see Table 1).

Table 1: Raw MT—Fidelity

Fidelity	DeepL raw MT	ChatGPT raw MT
Evaluator 1	**4,0244**	**4,3171**
Evaluator 2	**4,4878**	**4,6098**

In both cases, ChatGPT raw MT was considered to display a higher fidelity to the source text than DeepL, despite evaluator 1 attributing lower scores than evaluator 2 for both MT engines.

In terms of fluency of the target language, see Table 2.

Table 2: Raw MT—Fluency

Fluency	DeepL raw MT	ChatGPT raw MT
Evaluator 1	**4,8049**	**4,8537**
Evaluator 2	**4,7561**	**4,7805**

Unlike fidelity, target text fluency does not seem to vary depending on the MT engine. Both evaluators gave slightly higher fluency scores to ChatGPT MT output, but the variation is so low that no clear trend can be highlighted here.

5.2. Post-edited Texts and Comparison with Human Translation

If we now consider the PE and HT texts provided by the students, some interesting results appear. The analysis method was the same as the one used for raw MT: for each participant, both evaluators rated the post-edited segments from 1 to 5 for fidelity and fluency. These results were then averaged (see Table 3).

Table 3: PE and HT—Fidelity

Fidelity	DeepL PE	ChatGPT PE	Human translation
Evaluator 1	**4,3089**	**4,4439**	**4,4580**
Evaluator 2	**4,4282**	**4,7049**	**4,4905**

The fidelity results confirm what was observed with raw MT fidelity scores: in all cases, PE texts produced by students who post-edited the ChatGPT output were attributed higher quality scores than PE texts produced by students who post-edited the DeepL MT output. Results vary depending on the evaluator when comparing fidelity of PE and HT. Indeed, evaluator 1 considered that the fidelity of the texts translated from scratch was equivalent to the fidelity of the texts for which the ChatGPT output was post-edited. This evaluator also considered that the fidelity of the texts for which DeepL was post-edited was lower than both the other production modalities. However, evaluator 2 considered DeepL PE fidelity to be almost equivalent to HT fidelity (which was not the case for evaluator 1), and lower than ChatGPT PE fidelity.

Table 4: PE and HT—Fluency

Fluency	DeepL PE	ChatGPT PE	Human translation
Evaluator 1	**4,7507**	**4,8098**	**4,8158**
Evaluator 2	**4,8591**	**4,7951**	**4,7751**

As was the case for raw MT, fluency scores for PE texts were really similar to each other, and this was also the case for HT, thus preventing us from highlighting any trend (see Table 4). This could be expected, since one of the main focuses of modern NMT/LLM engines is to provide an output which is fluent in the target language, even if the meaning is mistranslated (Gladkoff & Han, 2021, p. 17).

5.3. Examples

Some examples were highlighted by both human evaluators and can be of use to better understand the challenges offered by video game localisation and some typical MT/PE errors.

5.3.1. Lack of context

As previously highlighted, lack of context is a problem that localisers often have to deal with. In the excerpt used for the present study, some MT and PE errors were caused by missing context. Both the examples presented below partially explain the reasons why DeepL raw MT/PE fidelity was considered lower than ChatGPT raw MT/PE's, as these mistranslations resulting from a lack

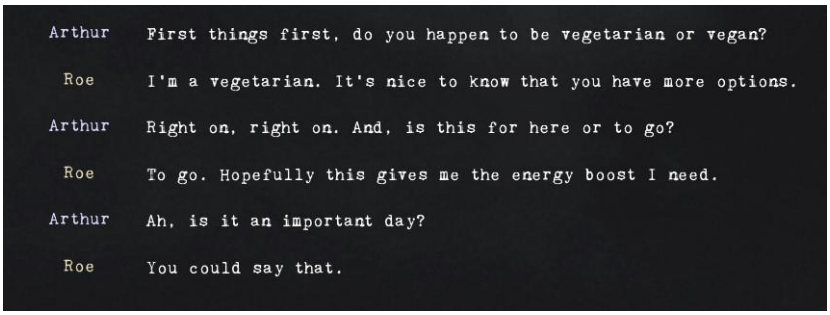
of context were only observed in the DeepL raw MT (see Table 5 – ChatGPT translated the segments correctly).

Table 5: Lack of context—Example 1

Source text	DeepL raw MT
105 Roe: I'm a vegetarian. It's nice to know that you have more options.	105 Roe : Je suis végétarien. C'est bien de savoir que vous avez plus d'options.
106 Roe: **To go**. Hopefully this gives me the energy boost I need.	106 Roe : **C'est parti**. J'espère que cela me donnera l'énergie dont j'ai besoin.

Between segments 105 and 106, the player has several options. However, the segment "Right on, right on. And, is this for here or to go?" (see Figure 1) was lost during the extraction of the text. The text was conserved as such for the students to face the same challenges as professional localisers, who often work on texts which are not presented in the same order as in the game.

Figure 1: Text Lost During Extraction



As a consequence, instead of translating "To go" by "À emporter" in French, DeepL got segment 106 wrong and translated it by "C'est parti". The participants should normally have been able to spot the mistake thanks to the overall context of the situation (character ordering a coffee in the game), but four students out of the nine who worked with DeepL failed to spot and correct the mistake.

Table 6: Lack of context—Example 2

Source text	DeepL raw MT
77 I glance down at the pictures in the **reviews** once again. They're certainly eye-catching, but they don't do it justice.	77 Je jette à nouveau un coup d'œil sur les photos de la **revue**. Elles attirent l'attention, mais elles ne rendent pas justice à la maison.

In Table 6, another mistranslation due to the segmentation of the source text and a consecutive lack of context was spotted in the DeepL raw MT. Segment 77 mentions the "reviews" that the character is reading in the games, in an app called "Hootle Maps," a reference to a well-known map app. However, DeepL failed to recognise the overall context of the game and chose the wrong

meaning, therefore translating "reviews" by "revue" into French instead of words such as "avis" or "commentaires," which correspond to what one would expect to find in a map smartphone app. Although the context was displayed in the source text (the "Hootle maps" app is clearly mentioned), six students out of the nine who post-edited the DeepL output failed to spot and correct this problem.

These examples already enable us to highlight the importance of choosing an adequate MT engine when localising video games, since some engines commit mistakes that others do not, which can lead to PE mistakes if localisers are not careful enough. Here, these PE mistakes were not committed by the students who worked on the ChatGPT output, nor were they committed in the human translations, since in both cases, students were not negatively influenced by the raw MT error, which is only found in the DeepL MT output.

5.3.2 Puns

Puns are also considered to be problematic for MT, and they represent a common challenge in video game localisation practises. For instance, in our excerpt, the lexical field of "bears" is important, since the coffee shop the main character is going to is bear-themed. The developers invented many bear-derived names for all the products sold in said coffee shop, which both MT engines struggled with (Table 7).

Table 7: Puns—Raw MT

Source text	DeepL raw MT	ChatGPT raw MT
99 A Latte Bears, Bearkaccino, Bluebeary Pie…	99 Un Latte Bears, un Bearkaccino, une Bluebeary Pie…	99 Un Latte Ours, un Bearkaccino, une Tarte aux Myrtilles…
100 Teddy Bearoll, Hambeargur, Bearnard's Cubble Tea??	100 Teddy Bearoll, Hambeargur, Bearnard's Cubble Tea	100 Teddy Bearoll, Hambeargur, Bubble Tea de Bearnard ??
101 Roe: (Someone here is REALLY fond of bears. And puns.)	101 Roe : (Quelqu'un ici aime VRAIMENT les ours. Et les jeux de mots.)	101 Roe : (Quelqu'un ici aime vraiment les ours. Et les jeux de mots.)
102 Roe: (Also, who the hell is Bearnard?)	102 Roe : (Et aussi, qui est Bearnard ?)	102 Roe : (D'ailleurs, qui diable est Bearnard ?)

As can be observed in the table, both MT engines fail to provide funny names in French, even if ChatGPT seems to achieve better results than DeepL. Here, interventions are required to find adequate solutions in French, but context is important. Some students tried to change the animal used in the puns, as displayed in Table 8 below, but this does not work, since bears appear on the screen while playing the game (see Figure 2). This information was not available for the students:

this highlights the paramount importance for localisers to be able to actually play the game they are translating, or at least have access to visual elements, even if this is pretty rare in real life conditions.

Figure 2: Necessary coherence image/text

Table 8: Puns—Students' ideas with animal change (all groups)

99 Un Bichon Latte, un toutouccino, une tarte à chien…	99 Un sachaï latte, un thé à la chamomille, des biscuits aux graines de sachia…
100 Un roulé à la saucisse, un hambeaggle, un thé-ckelà la Bernard ??	100 des rouleaux au beurre de sachahuète, un chapon aux châtaignes, un chapuccino ??
101 Roe : (Quelqu'un ici aime VRAIMENT les chiens. Et les jeux de mots.)	101 Roe : (Quelqu'un ici aime VRAIMENT les chats. Et les jeux de mots.)
102 Roe : (Et puis c'est qui ce Bernard ?)	102 Roe : (En plus, c'est qui Sacha ?)

Both these examples, based on dogs and cats, are incorrect in the framework of this game, since a change in the used lexical field betrays the visual context of the game. Some solutions could be found which match this visual information, as shown below in Table 9.

Table 9: Puns—Students' ideas without animal change (all groups)

99 Un Latte Ours, un Capourscino, Tarte au citron de Moursia…	99 Un Latte Macchiatours, un Mocaccinours, une Tarte aux Abricours…
100 Wrap au Boursin, Hamboursger, Bubble Tea de Oursula ??	100 Cinnamours Roll, Hamboursger, Bubble Tea d'Oursula ??
101 Roe : (Quelqu'un ici aime vraiment les ours. Et les jeux de mots.)	101 Roe : (Quelqu'un ici aime VRAIMENT les ours. Et les jeux de mots.)
102 Roe : (D'ailleurs, qui est Oursula ?)	102 Roe : (D'ailleurs, qui diable est Oursula ?)

5.3.3. Gender problems

Finally, gender problems are also quite common in MT use, and such bias is frequent in French because of issues in the training data: "MT systems are trained with data that contain biases present in our society and in our language" (Daems & Hackenbuchner, 2022, p. 289). As the number of video games that allow players to choose between gender increases, it becomes crucial to

implement gender-neutral strategies, which widely used MT engines cannot yet apply. Two problems are illustrated here in Table 10.

Table 10: Gender—Example 1

Source text	DeepL raw MT	ChatGPT raw MT
81 Roe: Sorry, I was a bit **distracted**.	81 Roe : Désolé, j'ai été un peu **distrait**.	81 Roe : Désolée, j'étais un peu **distraite**.

First of all, the default name of the playable character is Roe. As mentioned on the itch.io page of the video game, which students had access to during the experiment, two elements need to be taken into account: 1) "Nameable MC (default name is Roe—name will change to Fawna in a new demo)", and 2) "Ability to choose your pronouns." This means that the character can either be a woman, a man or a non-binary person… and that the text should be gender-neutral, since the game does not support several variables for a single segment, and each English segment therefore needs to be adapted using only one French segment. This was one of the reasons why this task was interesting for the students in the framework of this class. Here, DeepL raw MT chose the masculine form of the adjective, while ChatGPT chose the feminine form. For this first example, although there were some gender biases in the human translations, they were far less frequent than they were in the PE texts, which shows that gender bias in the raw MT proposals can lead translators to be negatively influenced by MT and, therefore, commit errors they would maybe not have committed should they have translated the text without MT.

The second example enables us to illustrate an interesting phenomenon (Table 11).

Table 11: Gender—Example 2

Source text	DeepL raw MT	ChatGPT raw MT
104 I'm a **vegetarian**.	104 Je suis **végétarien**.	104 Je suis **végétarien(ne)**.

Here, DeepL remained coherent with its masculine bias. However, ChatGPT decided to use the masculine form, with the feminine mark in brackets. Although this is not recommended in terms of inclusive writing, we consecutively found that more correct gender-neutral post-editing operations were carried out in the ChatGPT output than in the DeepL output, since ChatGPT's choice to display both forms led the students who worked with it to the inclusive path. Even if this segment was not necessarily correct in the texts produced by the ChatGPT group, there were less incorrect versions than in the DeepL group, which means that even if ChatGPT's version was clumsy in terms of inclusivity, it had the advantage of indicating students that both gender forms

had to be considered. Therefore, it is also important, in the framework of video game localisation, to choose an MT engine that produces as little gender bias as possible.

6. Results: automatic evaluation

6.1. BLEU and HTER scores: human translation and raw MT

The first element of the study we wanted to compare with human evaluation is its correlation with automatic metrics aiming at evaluating raw MT quality. To do so, we used the PosEdiOn tool described above. On the one hand, BLEU scores were used to assess the quality of the raw MT provided by each MT engine, on the basis of two human reference translations (since one of its main drawbacks is the restrictive nature of using a single reference translation) – see Table 12. On the other hand, we computed an HTER score to measure the quantity of edits between each of the raw MT productions and both reference translations (Table 13).

Table 12: BLEU score — Raw MT

BLEU score	DeepL	ChatGPT
MT/Reference translation 1	0.4466	**0.4854**
MT/Reference translation 2	**0.4435**	0.3999

The table 12 illustrates one of the main limitations of automatic metrics when it comes to measuring raw MT quality: the influence of each individual reference translation. In this case, the results are contradictory: ChatGPT has a higher BLEU score than DeepL for reference translation 1, but the opposite happens for reference translation 2, where DeepL performs better than ChatGPT in terms of BLEU score. Therefore, automatic scores do not enable us to detect any clear raw MT quality trend. Ideally, a higher number of reference translations should be used to determine whether one of both MT engines consistently perform better than the other one.

Table 13: HTER score—Raw MT

HTER score	DeepL	ChatGPT
MT/Reference translation 1	0.413	**0.37**
MT/Reference translation 2	**0.4272**	0.4341

The results provided by HTER in terms of raw MT quality match our BLEU score findings, in that ChatGPT performs better than DeepL for reference translation 1, and that DeepL performs better for reference translation 2. Indeed, as a lower HTER score indicates a lower edit distance between raw MT and the reference translation, and therefore a lower PE effort and a higher raw MT quality,

the fact that the HTER score is lower for ChatGPT for reference translation 1 and for DeepL for reference translation 2 suggests that fewer edits are required in the ChatGPT output compared to the DeepL output for reference translation 1, and vice versa for reference translation 2. Again, more reference translations should be used to identify any consistent trend.

6.2. HTER: PE/HT and Reference Translations

The second element we wanted to measure with automatic evaluation was the HTER score to compare the PE productions and both the reference translations, with a view to determining whether one of both MT engines enables the students to produce better PE texts than the other. We also did so with the texts humanly translated from scratch by nine of the students, in order to compute the HTER score between them and the reference translation, and then comparing these scores with those of the PE texts (Table 14).

Table 14: HTER score—HT and PE

Reference translation 1		Reference translation 2	
Production	HTER	Production	HTER
DeepL PE - 1	0.3877	DeepL PE - 1	0.4313
DeepL PE - 2	0.3819	DeepL PE - 2	0.4244
DeepL PE - 3	0.3915	DeepL PE - 3	0.4230
DeepL PE - 4	0.3626	DeepL PE - 4	0.3995
DeepL PE - 5	0.4286	DeepL PE - 5	0.4674
DeepL PE - 6	0.4148	DeepL PE - 6	0.4494
DeepL PE - 7	0.3983	DeepL PE - 7	0.4452
DeepL PE - 8	0.3887	DeepL PE - 8	0.4327
DeepL PE - 9	0.4244	DeepL PE - 9	0.4563
DeepL PE—mean	**0.3976**	**DeepL PE—mean**	**0.4366**
ChatGPT PE - 1	0.4011	ChatGPT PE - 1	0.4369
ChatGPT PE - 2	0.3846	ChatGPT PE - 2	0.4258
ChatGPT PE - 3	0.3750	ChatGPT PE - 3	0.4302
ChatGPT PE - 4	0.3750	ChatGPT PE - 4	0.4633
ChatGPT PE - 5	0.3901	ChatGPT PE - 5	0.4397
ChatGPT PE - 6	0.3654	ChatGPT PE - 6	0.4341
ChatGPT PE - 7	0.3599	ChatGPT PE - 7	0.4452
ChatGPT PE - 8	0.3626	ChatGPT PE - 8	0.4383
ChatGPT PE - 9	0.3777	ChatGPT PE - 9	0.4410
ChatGPT PE - 10	0.3664	ChatGPT PE - 10	0.4341
ChatGPT PE—mean	**0.3758**	**ChatGPT PE—mean**	**0.4389**

HT - 1	0.5509	HT - 1	0.5615
HT - 2	0.6136	HT - 2	0.6173
HT - 3	0.5083	HT - 3	0.5251
HT - 4	0.4848	HT - 4	0.5377
HT - 5	0.4605	HT - 5	0.5293
HT - 6	0.5340	HT - 6	0.5517
HT - 7	0.4660	HT - 7	0.4749
HT - 8	0.5229	HT - 8	0.5517
HT - 9	0.4633	HT - 9	0.4818
HT— mean	**0.5116**	**HT— mean**	**0.5368**

As illustrated in Table 14, it is difficult to obtain clear results for the texts post-edited using both MT engines. We can see that, for both evaluators, the variation between PE texts produced using DeepL and ChatGPT is similar (respective HTER scores of 0.3976 and 0.3758 for reference translation 1, and 0.4366 and 0.4389 for reference translation 2).

Furthermore, it can be observed that there is not much variation between the texts post-edited using the same MT engine, which could mean that PE outputs are pretty similar to each other. On the contrary, an interesting trend can be observed for HT: variation in HTER between participants for this dataset is much higher than for the PE dataset (ranging from 0.4605 to 0.6136 for reference translation 1 and from 0.4749 to 0.6173 for reference translation 2). This could mean that the students who translated the excerpt from scratch produced texts which were much more varied in terms of structure and expression compared to the PE texts, which are characterised by a low degree of variation between each other when compared with the reference translations.

6.3. HTER: PE Texts and Corresponding Raw MT

Finally, the last measure implemented in our study is an HTER metric aiming at comparing the distance between each PE text produced by our participants and its corresponding raw MT output (DeepL or ChatGPT). This provides some insight into the engine that requires less PE effort, and therefore potentially offers the better raw MT quality out of both (Table 15).

Table 15: HTER score—Raw MT/PE

HTER between raw MT and PE texts			
Production	HTER score	Production	HTER score
DeepL PE - 1	0.1580	ChatGPT PE - 1	0.2280
DeepL PE - 2	0.1608	ChatGPT PE - 2	0.1989
DeepL PE - 3	0.0540	ChatGPT PE - 3	0.1181
DeepL PE - 4	0.1417	ChatGPT PE - 4	0.1600
DeepL PE - 5	0.1619	ChatGPT PE - 5	0.1173
DeepL PE - 6	0.1157	ChatGPT PE - 6	0.0505
DeepL PE - 7	0.2124	ChatGPT PE - 7	0.1609
DeepL PE - 8	0.1275	ChatGPT PE - 8	0.1743
DeepL PE - 9	0.2131	ChatGPT PE - 9	0.1016
		ChatGPT PE - 10	0.0577
Mean HTER score	**0.1495**	**Mean HTER score**	**0.1367**

Again, the mean HTER scores are similar to both datasets, with DeepL PE texts obtaining a mean score of 0.1495 and ChatGPT PE obtaining a mean score of 0.1367. These results show that, on average, ChatGPT's raw MT output required slightly fewer edits than DeepL's raw MT output. However, a high degree of variation can be observed between participants for both MT engines, which means that some students carry out very few PE operations, while others post-edit more of the MT output. Therefore, the average score seems to indicate that ChatGPT raw MT requires slightly less PE effort than DeepL raw MT, thus supposedly offering a higher raw quality degree, but this trend should be further investigated to provide more reliable results.

7. Discussion—Comparison between human and automatic evaluation

Now that human evaluation scores and automatic metrics have been computed, this discussion section aims at providing more insight into the correlation between both evaluation methods for our video game localisation experiment.

The first evaluated element was raw MT quality. On the one hand, both human evaluators agreed that the ChatGPT raw output had a higher level of fidelity and fluency than the DeepL output. This is corroborated by the higher mean HTER score between the DeepL raw MT output and its corresponding PE texts (0.1495) than between the ChatGPT output and its corresponding PE texts (0.1367), meaning that the DeepL output required more edits and was therefore of lower quality. On the other hand, we observed that the HTER scores between the raw MT outputs and both the reference translations were contradictory when comparing both reference translations: Whereas

ChatGPT scored lower for reference translation 1, the results were the opposite for reference translation 2. This does not necessarily invalidate our results, since the restricted number of reference translations is a well-known limitation of automatic metrics, but it does indicate that, in the framework of such a video game excerpt and its particularities, a significantly higher number of reference translations would be advisable (even more for this particular video game excerpt consisting of an informal oral speech). Our findings are similar for the BLEU score: it is higher for the raw ChatGPT output in the case of reference translation 1, and for DeepL in the case of reference translation 2, thus further highlighting the need for more reference translations to identify any clear quality trend. Finally, another element should also be kept in mind: on the one hand, while it is true that a low edit distance such as the one observed here between the raw MT output and the PE texts could mean that the quality of raw MT is high, it could also mean that the students failed to detect some raw MT mistakes that needed to be post-edited, which was illustrated in section 5. As a result, using localisation professionals as participants would be a necessary step to validate our hypothesis and provide interesting results.

The other main element of the study was PE quality. Human evaluation showed that, in terms of fidelity to the source text, the texts post-edited by the students who used ChatGPT offered a higher quality level than those post-edited by the students who used DeepL. However, both evaluators disagreed on HT fidelity: while evaluator 1 rated it as higher than both PE paradigms, evaluator 2 considered HT quality to be lower than ChatGPT PE quality (but higher than DeepL PE quality). In terms of fluency, results differed between both evaluators, albeit with a low degree of variation. Automatic evaluation did not enable us to identify any clear signal towards a better PE quality for the PE texts produced using any of both MT engines, since the HTER score measured between the PE productions and both our reference translations are very similar. The HTER scores between the HT texts and both the reference translations are higher (in other words, more edits were required to get from HT texts to the reference translations) than those of PE texts, but this could be explained by the fact that students who translated the output from scratch failed to rewrite each segment number and the name of the main character, while these were present in the raw MT outputs, which were post-edited. However, the bigger variation in HTER between HT texts could be a signal that students who translated the text from scratch felt they had more freedom than the ones who post-edited, since a higher degree in variation indicates, in this case, that some outputs were significantly more different from the human reference translations than others. A similar phenomenon was observed by Castilho & Resende (2022, p. 18), who recurred to lexical variety measures: "This similarity between MT and PE texts […] was also revealed by the (h)TER scores […]. However, we can see a statistically significant difference between the HT and the PE

versions in the distribution of lexical density feature indicating that translators followed the lexical choices from the MT output, resulting in a distance from the HT lexical choices."

All in all, while automatic metrics such as BLEU score and HTER have limitations for all text types when they are based on reference translations, it is particularly true for video game excerpts that need to be localised. Indeed, the oral features of these texts and the great degree of possible expression multiply the number of reference translations needed to get satisfactory results, since every localiser would provide a significantly different version of the same source text. This highlights the need for a human evaluation of such contents.

Conclusion and Future Work

The goal of this exploratory paper was to investigate whether PE could be considered as an efficient way to carry out localisation tasks. Though our study was carried out with translation students and should be replicated with translation or localisation professionals, some interesting elements were found. In terms of fidelity to the source text, the choice of the MT engine is important: human evaluation enabled us to determine that PE texts produced using ChatGPT were, on average, better than those produced using DeepL. While HT fidelity was considered better than both PE paradigms by evaluator 1 (though ChatGPT PE came really close to HT), ChatGPT PE was considered to have a higher fidelity degree than HT by evaluator 2, thus showing that using ChatGPT could lead to an overall similar or higher degree of fidelity compared with HT. Therefore, human evaluators agree that ChatGPT could be a viable alternative to human translation for this particular excerpt, though our results cannot be generalised. For fluency, human evaluation showed that the three modes of production did provide similar results and, therefore, did not have any significant influence on this aspect of quality. However, it should be noted that the human translations, in addition to being equally fluent as the PE texts, displayed a greater variation degree than these, indicating that creativity is higher in the case of human translation. This should be studied further using lexical and syntactic variety metrics. HTER between PE texts and their respective raw MT showed that ChatGPT required a slightly lesser effort than DeepL and could be characterised by a slightly higher quality degree. Nonetheless, these results should be confirmed thanks to other methods, since we have seen that metrics such as HTER only aim at computing edit distances and do not take into account the immense number of possible translations for any text.

We also determined that raw MT could be a source of problems, namely in terms of gender, and that it fails to translate certain elements properly, especially when context is lacking. Although students should have been able to spot and correct these errors and inconsistencies, such was not

always the case, thus highlighting the need for solid training in raw MT typical mistakes for students who would like to post-edit such contents in their future careers. This study also enabled us to determine that, except for HTER scores used to measure the edit rate between a PE text and the raw MT, which was used to produce it, automatic metrics are not necessarily suited to video game localisation, or would at least require many more reference translations to produce consistent results. Future research on the matter could be carried out, and should include more reference translations, more evaluators, and be based on the latest versions of MT engines based on generative AI, since ChatGPT seems to be a promising tool for professional localisers to use.

References

Bibliography

BANERJEE Satanjeev & LAVIE Alon, 2005, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments", *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* [online], p. 65-72, available at: https://aclanthology.org/W05-0909.pdf (last accessed on September 18th, 2024).

BERNAL-MERINO Miguel Ángel, 2007, "Challenges in the translation of video games", *Tradumàtica* [online], n° 5: Localizatció de videojocs, p. 1-7, available at: https://ddd.uab.cat/pub/tradumatica/15787559n5/15787559n5a2.pdf (last accessed on September 18th, 2024).

BRENNER Judith, 2024, "The MTxGames Project: Creative Video Games and Machine Translation—Different Post-Editing Methods in the Translation Process", *Proceedings of the 25th Annual Conference of the European Association for Machine Translation* [online], vol. 2, p. 47-48, available at: https://eamt2024.github.io/proceedings/vol2.pdf (last accessed on September 18th, 2024).

CAPELLINI Thomas, 2021, *Crowdsourced translation in indie game localization*, Master's thesis [online], University of Geneva, available at: https://archive-ouverte.unige.ch/unige:156286 (last accessed on September 18th, 2024).

CASTILHO Sheila & RESENDE Natália, 2022, "Post-Editese in Literary Translations", *Information* [online], vol. 13, n° 66, p. 1-22, available at: https://www.mdpi.com/2078-2489/13/2/66 (last accessed on September 18th, 2024).

CHOJNOWSKI Ryszard, 2016, "The practical aspects of video game localization", *Styles of Communication* [online], vol. 8, n° 1, p. 69-90, available at:

https://openurl.ebsco.com/results?sid=ebsco:ocu:record&bquery=IS+2065-7943+AND+VI+8+AND+IP+1+AND+DT+2016 (last accessed on September 18th, 2024).

DAEMS Joke, MACKEN Lieve & VANDEPITTE Sonia, 2013, "Quality as the sum of its parts: A two-step approach for the identification of translation problems and translation quality assessment for HT and MT+PE", *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice* [online], p. 63-71, available at: https://aclanthology.org/2013.mtsummit-wptp.8.pdf (last accessed on September 18th, 2024).

DAEMS Joke & HACKENBUCHNER Janiça, 2022, "DeBiasByUs: Raising Awareness and Creating a Database of MT Bias", *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation* [online], p. 289-290, available at https://aclanthology.org/2022.eamt-1.pdf (last accessed on April 9th, 2025).

EHRENSBERGER-DOW Maureen & MASSEY Gary, 2020, "Translation workplace-based research", *in* Minako O'Hagan (dir.), *The Routledge Handbook of Translation and Technology*, p. 354-369, Routledge.

EMT Group, 2022, *EMT Competence Framework 2022*, Brussels, European Commission.

FINCH Andrew, AKIBA Yasuhiro & SUMITA Eiichiro, 2004, "How Does Automatic Machine Translation Evaluation Correlate With Human Scoring as the Number of Reference Translations Increases?", *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)* [online], p. 2019-2022, available at: http://www.lrec-conf.org/proceedings/lrec2004/pdf/277.pdf (last accessed on September 18th, 2024).

GIRLETTI Sabrina & LEFER Marie-Aude, 2024, "Introducing MTPE pricing in translator training: a concrete proposal for MT instructors", *The Interpreter and Translator Trainer* [online], vol. 18, n° 1, available at: https://www.tandfonline.com/doi/full/10.1080/1750399X.2023.2299914 (last accessed on September 18th, 2024).

GLADKOFF Serge & HAN Lifeng, 2022, "HOPE: A Task-Oriented and Human-Centric Evaluation Framework Using Professional Post-Editing Towards More Effective MT Evaluation", *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)* [online], p. 13-21, available at: https://aclanthology.org/2022.lrec-1.2.pdf (last accessed on September 22d, 2024).

HANSEN Damien & HOULMONT Pierre-Yves, 2022, "A Snapshot into the Possibility of Video Game Machine Translation", *Proceedings of the 15th Biennial Conference of the Association for Machine*

Translation in the Americas [online], vol. 2, p. 257-269, available at: https://aclanthology.org/2022.amta-upg.18.pdf (last accessed on September 18th, 2024).

JIMÉNEZ-CRESPO Miguel A., 2024, *Localization in Translation*, Oxon and New York, Routledge.

KOPONEN Maarit, AZIZ Wilker, RAMOS Luciana & SPECIA Lucia, 2012, "Post-editing time as a measure of cognitive effort", *Conference of the Association for Machine Translation in the Americas* [online], available at: https://aclanthology.org/2012.amta-wptp.2.pdf (last accessed on September 18th, 2024).

LEE Seungjun, LEE Jungseob, MOON Heonseok, PARK Chanjun, SEO Jaehyung, EO Sugyeong, KOO Seonmin & LIM Heuiseok, "A Survey on Evaluation Metrics for Machine Translation", *Mathematics* [online], n° 11, available at: https://www.mdpi.com/2227-7390/11/4/1006 (last accessed on September 18th, 2024).

LEFER Marie-Aude, PIETTE Justine & HANCART Romane, 2022, *MTPEAS manual*, Louvain-la-Neuve, UCLouvain, available at https://oer.uclouvain.be/jspui/bitstream/20.500.12279/829/9/MTPEAS_manual_EN_final_CC.pdf (last accessed on April 10th, 2025).

MANGIRON Carme & O'HAGAN Minako, 2013, *Game Localization—Translating for the global digital entertainment industry*, Amsterdam and Philadelphia, John Benjamins.

MOORKENS Joss, CASTILHO Sheila, GASPARI Federico, TORAL Antonio & POPOVIĆ Maja, 2024, "Proposal for a Triple Bottom Line for Translation Automation and Sustainability: An Editorial Position Paper", *The Journal of Specialised Translation* [online], n° 41, p. 2-25, available at https://www.jostrans.org/article/view/4706/4239 (last accessed on September 18th, 2024).

OLIVER Antoni, ÁLVAREZ-VIDAL Sergi & BADIA Toni, 2020, "PosEdiOn: Post-Editing Assessment in PythOn", *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* [online], p. 403-410, available at: https://aclanthology.org/2020.eamt-1.43.pdf (last accessed on April 10th, 2025).

PAPINENI Kishore, ROUKOS Salim, WARD Todd & ZHU Wei-Jing, 2004, "BLEU: a Method for Automatic Evaluation of Machine Translation", *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* [online], p. 311-318, available at https://aclanthology.org/P02-1040.pdf (last accessed on September 18th, 2024).

RAY Alice, 2019, "Playing with the Language of the Future: The Localization of Science-Fiction Terms in Video games", *in* Astrid Ensslin & Isabel Balteiro (dir.), *Approaches to Video game Discourse: Lexis, Interaction, Textuality*, p. 87-115, Bloomsbury.

RIVAS GINEL María Isabel, 2021, "Ergonomics of tools usage for video game localisation: a user survey", *Critic. Cahiers de recherches interdisciplinaires sur la traduction, l'interprétation et la communication interculturelle* [online], n° 2, p. 27-57, available at https://hal.science/hal-03465996v1 (last accessed on September 18th, 2024).

RIVAS GINEL María Isabel, 2022, "Video Game Localisation Tools: A User Survey", *in* Miguel Ibáñez Rodríguez & Carmen Cuéllar Lázaro (dir.), *De la hipótesis a la tesis: Traductología y lingüística aplicada*, p. 295-324, Albolote, Editorial Comares.

RIVAS GINEL María Isabel & THEROINE Sarah, 2022, "Machine Translation and Gender biases in video game localisation: a corpus-based analysis", *Journal of Data Mining and Digital Humanities* [online], p. 1-10, available at https://hal.science/hal-03540605/document (last accessed on September 18th, 2024).

RIVAS GINEL María Isabel, 2023, *The Ergonomics of CAT Tools for Video Game Localisation*, PhD Thesis, Université de Bourgogne.

SAUNDERS Danielle, SALLIS Rosie & BYRNE Bill, 2020, "Neural Machine Translation Doesn't Translate Gender Coreference Right Unless You Make It", *Workshop on Gender Bias in NLP (GeBNLP) 2020* [online], p. 1-10, available at https://aclanthology.org/2020.gebnlp-1.4.pdf (last accessed on September 18th, 2024).

SNOVER Matthew, DORR Bonnie, SCHWARTZ Richard, MICCIULLA Linnea & MAKHOUL John, 2007, "A Study of Translation Edit Rate With Targeted Human Annotation", *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas* [online], p. 223-231, available at https://aclanthology.org/2006.amta-papers.25.pdf (last accessed on September 18th, 2024).

SNOVER Matthew, MADNANI Nitin, DORR Bonnie & SCHWARTZ Richard, 2009, "Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric", *Proceedings of the Fourth Workshop on Statistical Machine Translation* [online], p. 259-268, available at https://aclanthology.org/W09-0441.pdf (last accessed on September 18th, 2024).

THEROINE Sarah, RIVAS GINEL María Isabel & PERRIN Aurélie, 2021, « Au cœur de la terminologie du jeu vidéo. L'absence de ressources, frein majeur pour les traducteurs », *Traduire* [online], n° 244, p. 27-40, available at https://journals.openedition.org/traduire/2285 (last accessed on September 18th, 2024).

TOFTEDAHL Marcus, BACKLUND Per & ENGSTRÖM Henrik, 2018, "Localization from an Indie Game Production Perspective: Why, When and How?", *DiGRA '18—Proceedings of the 2018*

DiGRA International Conference: The Game is the Message [online], available at https://dl.digra.org/index.php/dl/article/view/941/941 (last accessed on September 18th, 2024).

VILAR David, LEUSCH Gregor, NEY Hermann & BANCHS Rafael, 2007, "Human Evaluation of Machine Translation Through Binary System Comparisons", *Proceedings of the Second Workshop on Statistical Machine Translation* [online], pp. 96-103, available at https://aclanthology.org/W07-0713.pdf (last accessed on September 18th, 2024).

VOLKAERT Lise, GIRLETTI Sabrina, GERLACH Johanna, MUTAL Jonathan David & BOUILLON Pierrette, 2022, "Source or Target First? Comparison of Two Post-Editing Strategies with Translation Students", *Journal of Data Mining and Digital Humanities* [online], p. 1-7, available at https://hal.science/hal-03546151/document (last accessed on September 18th, 2024).

Websites

PRIKARIN, 2025, *Killer Trait*, available at https://prikarin.itch.io/killer-trait (last accessed on April 10th, 2025).

STATITA, 2025, *Games—Worldwide*, available at https://www.statista.com/outlook/amo/media/games/worldwide?currency=EUR (last accessed on April 10th, 2025).

TRANSPHERE, 2023, *Video Game Localization: The Top 2023 ROI Booster*, available at https://www.transphere.com/video-game-localization-roi/ (last accessed on April 10th, 2025).