UMONS
Université de Mons

Faculté
de Traduction
et d'Interprétation
Ecole d'Interprètes
Internationaux

NMT and LLM Post-Editing Process in Videogame Localization: How Can Students' Productions Help Us Become Better Localisation Lecturers?

Fun for All 7

Loïc DE FARIA PIRES

FTI-EII, University of Mons, 31st January 2025

Presentation outline

- Context
- Theoretical background
- PE / localisation in academic context
- Objectives
- Methodology
- Results
- Discussion
- Conclusion

Context

- Recent years: NMT quality ++ → PE as a professional practice (Cui *et al.*, 2023: 1)
- Rise of LLMs and quality of their MT outputs (Peng *et al.*, 2023: 1) → Higher relevance of generative AI on the translation market (Silva Loureiro and Ferreira, 2023 : 41)

➔ Need for translation lecturers to update their classes

➔ EMT Network requirements

➔ BUT… scarce studies on NMT/Generative AI applied to localisation teaching and practices.

Theoretical background

- Hansen and Houlmont (2022)
  - Raw MT quality can vary depending on the MT engine used ;
  - Specifically-trained engines can be a good option for VG localisation

- Rivas Ginel and Theroine (2022)
  - Gender aspects in MT applied to VG localisation
  - Some MT engines provide better results than others

- Brenner (2024: 47)
  - Ongoing research on PE methods applied to VG localisation

PE and localisation in an academic context

- Many universities : PE and localisation classes (EMT, etc…)
- Since PE is used in more and more professional settings, class sessions on NMT/LLMs applied to localisation = relevant
- But… few research papers on VG PE performed by university students… however, understanding their biases could help us teach them more relevant classes

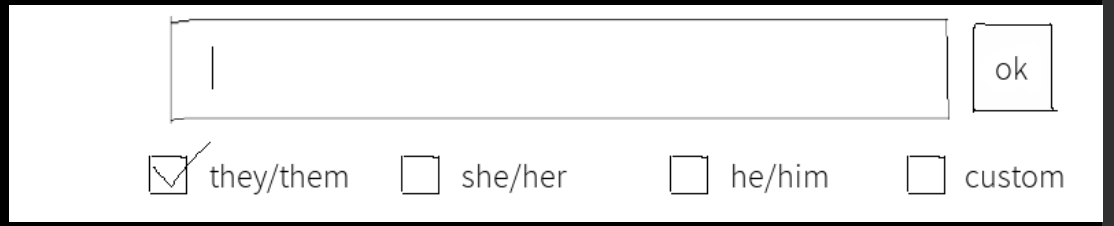
- Copet and De Faria Pires (2023)
  - Visual novel, Master's students: ChatGPT better than DeepL (human evaluation). ChatGPT ~equivalent to HT.
  - In press (2025?): Automated metrics (BLEU/HTER): ChatGPT required a few less edits than DeepL. More variety in HT compared to PE.

Objectives

- Determine the quality of Raw MT applied to VG localisation
- Identify the main problems → improve teaching practices
- Determine whether some MT engines provide for better results
  - **→ Improve the way we teach PE applied to VG localisation by focusing on frequent problems**

- Compare PE quality between 3 widely-used MT engines
- Compared PE effort between these engines
  - **→ Is MTPE more productive using one of said engines in the framework of VG localisation?**

Methodology



- Exploratory study…
- Excerpt from the Purrgatory videogame (free, Steam) – You can choose your character's gender
- Submitted to 23 students – EN-FR PE class – 3 groups (DeepL_PE n=7, Google_PE n=9, GPT_PE n=7)
  - Students: already familiar with VG localisation (followed the class during the previous academic year).
- PosEdiOn (Álvarez-Vidal & Oliver, 2023)
  - PE time (temporal effort)
  - Keystrokes (technical effort)
  - Pauses (cognitive effort)
  - Automatic metrics
- Analysis of process data + interesting phenomena

purrgatory

menu

to do

items

purgatory is spelled wrong

what if i don't sign?

(sign your name)

name .....

address .....

history

skip

the receptionist hands you a pen and a clipboard.

Results – temporal effort

DeepL	Google	GPT
00:34:26	00:52:59	00:48:31
00:31:12	00:53:31	00:45:40
00:47:36	00:51:58	01:18:08
00:47:55	00:49:27	00:54:19
00:46:15	00:41:52	00:53:26
00:31:09	00:48:18	00:56:35
00:46:56	00:52:30	00:59:36
	00:38:02	
	00:41:10	
00:40:47	**00:47:45**	**00:56:36**

PE time

-DeepL < Google < ChatGPT

-Temporal effort seems to vary depending on the engine

-ChatGPT required the longest time (!)

Possible bias because of the reduced sample

Results – technical effort

DeepL	Google	GPT
971	1348	793
92	1259	1126
460	1562	1227
1232	726	1332
928	494	2014
824	1011	514
621	1695	1860
	1236	
	763	
732,571429	**1121,55556**	**1266,57143**

Total keystrokes

-Matches temporal effort

-DeepL < Google < ChatGPT

-Variations depending on each engine

-To be confirmed with bigger sample

Results – cognitive effort

DeepL	Google	GPT
623	566	395
253	441	502
307	539	561
694	453	554
640	309	734
471	446	527
365	601	748
	465	
	437	
479	**473**	**574,428571**

Total long pauses

-Google = smallest cognitive effort (similar to DeepL)

-ChatGPT = (far) bigger cognitive effort than the other 2

-To be confirmed

Results – HTER

DeepL	Google	GPT
0,2414	0,2184	0,1552
0,0531	0,1622	0,17
0,1477	0,1649	0,212
0,2	0,1589	0,213
0,2083	0,1293	0,3321
0,1066	0,1495	0,1698
0,1615	0,2055	0,2641
	0,2396	
	0,1703	
0,1598	**0,17762222**	**0,2166**

Edit distance between Raw MT and PE

-DeepL < Google < ChatGPT

-Biggest edit distance: ChatGPT

-Corroborates the hypothetical bigger effort needed to work on ChatGPT for this text

Results – particular problems

- **Is ChatGPT that bad ?**
- 1) Lack of understanding

Source text- receptionist: hello, hello. **be a dear** and fill out this form.

DeepL - réceptionniste : bonjour, bonjour. **soyez gentil** et remplissez ce formulaire. → Gender of the main character?

Google - réceptionniste : bonjour, bonjour. **sois gentil** et remplis ce formulaire.

ChatGPT - réceptionniste : bonjour, bonjour. **Sois cher** et remplis ce formulaire. → "be expensive"

Results – particular problems

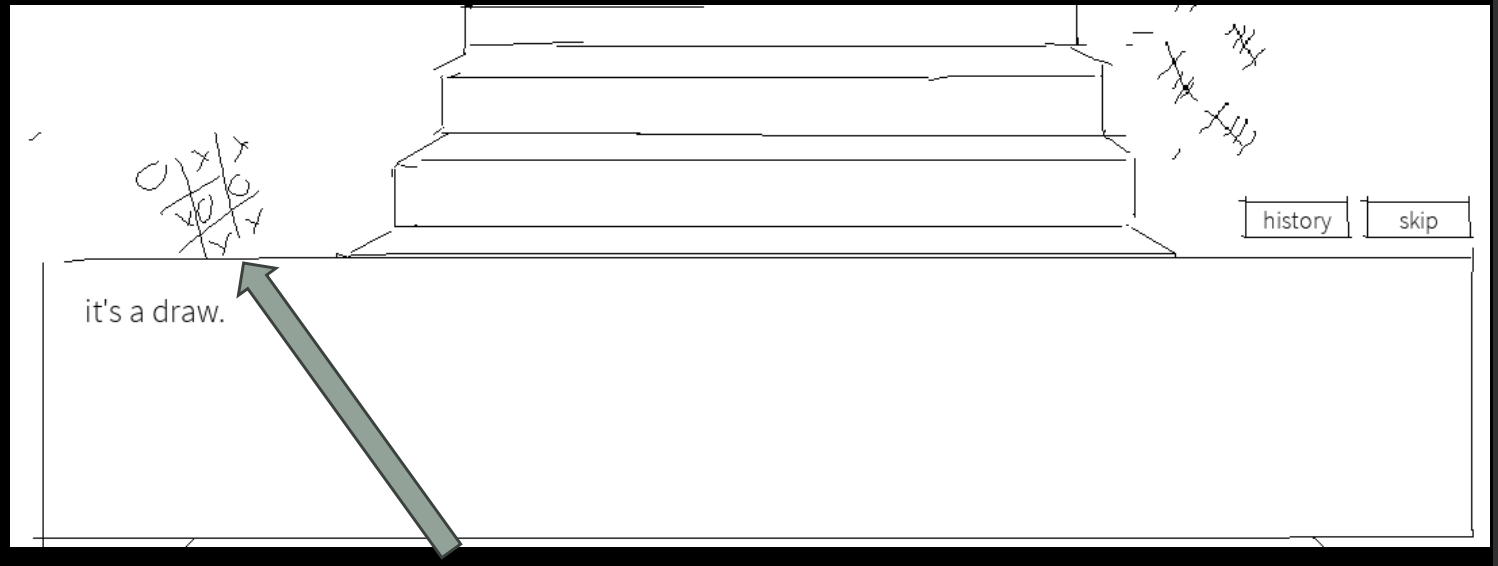
- **Is ChatGPT that bad ?**
- 2) Visual context

Source text - it's a draw

DeepL - C'est un tirage au sort

Google - c'est un match nul

ChatGPT - c'est un dessin



Problem: localisers do not often have the text before translating. If they are going to post-edit it, errors can be caused by raw MT

Discussion

- Effort seems to be bigger with ChatGPT (automatic metrics) – contradicts Copet & De Faria Pires (2023) → different for each game

- In this case, some errors were committed by ChatGPT only (lack of context)
  - But… ChatGPT's raw MT does not really contain more errors than DeepL or ChatGPT
  - Further analysis to determine why the effort was higher with it…

- In any case, the fact that efforts vary indicate that for some games, some MT engines require less effort to be post-edited. The used raw MT should therefore be carefully selected

Conclusion

- What to remember for PE applied to VG localisation in academic curricula?
  - Importance for students to be able to select the best MT engine depending on the game
    - Identify ST characteristics that influence raw MT quality for different types of games?
    - Ability for students to spot frequent MT mistakes
  - MT, if selected carefully, can be a great tool to help localisers
    - Interesting PE times (between 30 and 50 minutes for 583 source words)
    - We can teach students to identify "traditional" localisation problems when working from English into French (gender, lack of context, invented éléments…) and analyse how each MT engine deals with them.
- Product studies required as well, to determine whether these productivity levels are detrimental to PE quality…

References

- Alvarez-Vidal, S., & Oliver, A. (2023). Assessing MT with measures of PE effort. *Ampersand*, 11. https://www.sciencedirect.com/science/article/pii/S2215039023000176
- Brenner, J. (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*, 47-48, https://eamt2024.github.io/proceedings/vol2.pdf
- Copet, S., & De Faria Pires, L. (2023). Post-édition de TAN et localisation de jeux vidéo : quelles conséquences sur la qualité et l'immersion des joueurs ? *Journée d'étude Traduire le jeu vidéo: entre immersion, interactivité et interaction* [paper presentation]
- Cui, Y., Liu, X., & Cheng, Y. (2023). A Comparative Study on the Effort of Human Translation and Post-Editing in Relation to Text Types: An Eye-Tracking and Key-Logging Experiment. *SAGE Open*, *13*(1), 1-15. https://doi.org/10.1177/21582440231155849
- Ferreira, A. P., & Loureiro, A. (2023). As percepções dos alunos sobre a utilização da inteligência artificial nas aulas de práticas de tradução. *PRATICA - Revista Multimédia De Investigação Em Inovação Pedagógica E Práticas De E-Learning*, *7*(1), 33–49. https://doi.org/10.34630/pel.v7i2.5628
- Hansen, D., & Houlmont, P. Y. (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*, *2*, 257-269. https://orbi.uliege.be/bitstream/2268/294581/1/2022.amta-upg.18.pdf
- Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., & Tao, D. (2023). Towards making the most of ChatGPT for machine translation. *EMNLP 2023*, 1-12. https://doi.org/10.48550/arXiv.2303.13780
- Rivas Ginel, M. I., & Theroine S. (2022), Machine Translation and Gender biases in video game localisation: a corpus-based analysis, *Journal of Data Mining and Digital Humanities*, 1-10, https://hal.science/hal-03540605/document