

Influence of Prompting Strategies on Lexical Variety and Syntactic Equivalence: Does Prompting Influence Students' PE Processes?

SCOURNEAU Valentin, PhD student
DE FARIA PIRES Loïc, Associate Professor
FTI-EII, UMONS

Context

- **Results of the European Language Industry Survey 2026**
 - Big players now mostly use LLMs for MT (↘ NMT)
 - MT & PE use is skyrocketing among LSPs & freelance
 - Within the last year, LLM-based MT has started to underlay a large share of PE jobs
 - ➔ PE processes & products may differ from NMT-based PE!

State of play

➤ PE process and effort

- 3 dimensions of PE effort (Krings 2001)
 - Temporal: how much time? (PE time)
 - Technical: how many operations? (keylogging, TER...)
 - Cognitive: how much cognitive load (eye-tracking, pauses...)

➤ TER

- Translation Edit Rate (Snover et al. 2006): number of edits required to go from a raw MT text to a reference translation.
- "The amount of the text edited by the post-editor as against the number of the words in the raw MT output." (Sakamoto & Bawa Mason 2024: 464), here, **at the character-level**.
- Measure of technical (and cognitive?) effort: higher TER = more edits on the raw MT version to produce the final PE product.

State of play

➤ Calques and PE effort

- NMT: reduced PE effort compared to RBMT and SMT (Koponen *et al.* 2019).
- Studies showed that “NMT reduces different types of errors and produces more fluent output specifically in morphologically rich languages.” (*ibid.*: 62)
- NMT: machine translationese / MT artifacts
 - ➔ loss of lexical diversity (Vanmassenhove *et al.*, 2019)
 - ➔ more syntactic equivalence with the source text (Hansen & Esperança-Rodier, 2022)
- LLM and prompting? Scarcely studied so far...

➤ NMT → LLM-based MT: evolution of effort?

- Few to no studies comparing NMT and LLM PE effort (recent field)
- “Unlike NMT systems that often tend towards generating translations that are either too literal or inaccurate, LLMs leverage large training data to generate context-aware translations less literally.” (Castaldo *et al.* 2025: 506).
- PE effort varies depending on the chosen LLM (*ibid.*: 510).
- Traditional gain of productivity in PE compared to HT: still observable (*ibid.* : 512).

Research questions

- 1) Can prompting help reduce known MT artifacts as measured using linguistic metrics?
- 2) Do differences in linguistic metrics in the MT underlaying the PE step lead to differences in temporal & technical PE efforts?
- 3) Do differences in linguistic metrics in the MT remain in the post-edited texts?

Methodology

1) Compilation of a corpus of 18 editorials from British newspapers

2) MT (GPT-5.0) with 2 prompts adapted from Wang et al., 2023 & from Farrell, 2023

➤ Simple prompt:

"Translate this document from English to French. Output only the translation."

➤ Detailed prompt (> < MT artifacts):

"Translate this editorial from English to French. Please keep in mind that lexical and syntactic variety is required for a human-quality final text. Calques from English must be avoided and natural-sounding French structures and expressions should be used where needed. Output only the translation."

2) Computation of linguistic metrics

➤ Lexical metrics

- Lexical density
- TTR
- MATTR-50 (Covington & McFall, 2010)

➤ Syntactic metrics (Vanroy et al., 2021)

- ASTrED
- SACr
- PoS changes
- Label changes

3) Selection of a representative editorial

4) Full PE by 2 x 6 second-year Master's students in multidisciplinary translation

5) Computations on the post-edited texts

- Linguistic metrics
- PE time
- TER

Results: Metrics

➤ Lexical metrics (18 MTs x 2)

- Detailed prompt > simple prompt
- Medium/large effect size
- $p < 0.0005$ for all metrics

➤ Syntactic metrics (18 MTs x 2)

- Detailed prompt > simple prompt
- Very large effect size
- $p < 0.0005$ for all metrics

➤ Lexical metrics (6 post-edited texts x 2)

- Detailed prompt > simple prompt (small diff.)
- Huge effect size TTR / MATTR-50: $p < 0.0005$
- Lex. density: $p < 0.001$
- Confirmed by bootstrapping

➤ Syntactic metrics (6 post-edited texts x 2)

- No difference between the prompts
- Small effect sizes
- Confirmed by bootstrapping

! small n

Examples

Source text	Raw MT - Simple prompt	Raw MT – Detailed prompt
Disagreement among member states – and the paralysing requirement for unanimity before sanctions can be imposed – has hobbled the EU's response to a genocidal war that has horrified European populations.	Les désaccords entre États membres – et l'exigence paralysante de l'unanimité avant toute imposition de sanctions – ont entravé la réponse de l'UE à une guerre génocidaire qui a horrifié les populations européennes.	Les divisions entre États membres – aggravées par l'exigence paralysante de l'unanimité avant toute sanction – ont entravé la réponse européenne face à une guerre génocidaire qui a profondément choqué les opinions publiques.
An embarrassingly one-sided trade deal struck with the US, which imposed 15% tariffs on EU exporters while removing existing tariffs on US goods, was another display of weakness.	Un accord commercial honteusement déséquilibré conclu avec les États-Unis, qui a imposé des droits de douane de 15 % aux exportateurs européens tout en supprimant les tarifs existants sur les produits américains, a constitué une autre démonstration de faiblesse.	Un accord commercial honteusement déséquilibré conclu avec les États-Unis, imposant des droits de douane de 15 % aux exportateurs européens tout en levant ceux pesant sur les produits américains, a révélé une nouvelle fois cette faiblesse.

- **Detailed prompt: more natural sentences / varied vocabulary / less calques**
- **! Neither version is exempt of meaning shifts** (not measured here)

Results: Efforts

➤ Temporal effort (PE time)

- Simple prompt: longer PE times
- Detailed prompt: shorter PE times
- Less syntactically calqued / more lexically diverse text
= quicker PE process (?)

➤ Technical effort (TER)

- Results in line with PE time (even if time does not always match TER for a same set of students)
- Simple prompt: more edits (almost double)
- Detailed prompt: less edits
- Overall, not many edits

Student	PE time (minutes) Simple prompt	PE time (minutes) Detailed prompt	
Student 1 / 7	76	69	
Student 2 / 8	78	70	
Student 3 / 9	85	70	
Student 4 / 10	85	73	
Student 5 / 11	87	81	
Student 6 / 12	89	84	
Mean	83.3 min	74.5 min	- 10.6 %
Student	TER Simple prompt	TER Detailed prompt	
Student 1 / 7	0.1304	0.0961	
Student 2 / 8	0.2366	0.0893	
Student 3 / 9	0.0857	0.1066	
Student 4 / 10	0.139	0.1096	
Student 5 / 11	0.2064	0.0577	
Student 6 / 12	0.2381	0.1228	
Mean	0.1727	0.097	- 43.8 %

Discussion: Bridging the Gap

The detailed prompt seems to have eased the PE process (lower PE time, lower TER)

Explanations:

- part of the MT artifacts already taken care of
 - Soft ceiling for syntactic metrics: needed rephrasing already done
 - Small increase in lexical metrics: rephrasings/added idioms kept by the students

Discussion

1) Can prompting help reduce known MT artifacts as measured using linguistic metrics?

➤ Yes

2) Do differences in linguistic metrics in the MT underlying the PE step lead to differences in temporal & technical PE efforts?

➤ Yes

2) Do differences in linguistic metrics in the MT remain in the post-edited texts?

➤ Somewhat for lexical metrics

➤ No for syntactic metrics

Limitations

- **Product study?**
 - Ideally, PE quality study is necessary
- **Small sample**
 - Single GenAI tool used
 - Limited number of students
- **Automated metrics**
 - Correlation with human judgement?
- **Students**
 - Short PE experience
 - Impossible to generalise
 - Typical biases when working with students
- **Single type of text**
 - Editorial (general domain)
 - Effects in specialised texts?

Future research

- **Deeper process data to further assess raw MT usefulness (esp. cognitive effort)**
 - Eye-tracking
 - Keylogging
 - Other metrics
 - **PosEdiOn** (Oliver et al. 2020)
 - Keep time stamps (order of the edits)
- **Work with professional translators to confirm results**
- **Other text types**
- **Bigger samples**
- **Related question: what about PE working conditions & job satisfaction if the arguably more satisfying part is already done via prompting?**

References

- Castaldo, Antonio, Castilho, Sheila, Moorkens, Joss, & Monti, Johanna. (2025). Extending CREAMT: Leveraging Large Language Models for Literary Translation Post-Editing. In *Proceedings of Machine Translation Summit XX: Volume 1*, 506–515, Geneva, Switzerland. European Association for Machine Translation.
- Covington, Michael A., & McFall, Joe D. (2010). Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), 94–100. <https://doi.org/10.1080/09296171003643098>
- ELIS Research. (2026). European Language Industry Survey 2026: Trends, expectations and concerns of the European language industry.
- Farrell, Michael (2023a). Current evidence of post-edits: differences between post-edited neural machine translation output and human translation revealed through human evaluation. *Proceedings of the International Conference on Human-Informed Translation and Interpreting Technology Workshop*, 52–63.
- Hansen, D., & Esperança-Rodier, E. (2022). Human-Adapted MT for Literary Texts: Reality or Fantasy? In *Proceedings of the New Trends in Translation and Technology Conference 2022*, 178–190. <https://hal.science/hal-04038025/>
- Koponen, Maarit, Salmi, Leena & Nikulin, Markku. (2019). A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. *Machine Translation* 33, 61–90 (2019). <https://doi.org/10.1007/s10590-019-09228-7>
- Krings, Hans P. (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes* (Geoffrey S. Koby, ed.), The Kent State University Press, Kent, Ohio & London, 558 p.
- Oliver, Antoni & Alvarez-Vidal, Sergi & Badia, Toni. (2020). PosEdiOn: Post-Editing Assessment in PythOn.
- Sakamoto, Akiko, & Bawa Mason, Sarah. (2024). In search of a fair MTPE pricing model: LSPs' reflections and the implications for translators. *Perspectives*, 32(3), 460–476. <https://doi.org/10.1080/0907676X.2023.2292572>
- Snover, Matthew, Dorr, Bonnie, Schwartz, Rich, Micciulla, Linnea, & Makhoul, John. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.
- Vanmassenhove, Eva, Shterionov, Dimitar, & Way, Andy (2019). Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. In *Proceedings of Machine Translation Summit XVII: Research Track* (pp. 222–232).
- European Association for Machine Translation. <https://aclanthology.org/W19-6622>
- Vanroy, Bram, De Clercq, Orphée, Tezcan, Arda, Daems, Joke, & Macken, Lieve (2021). Metrics of syntactic equivalence to assess translation difficulty. In *Explorations in empirical translation process research* (Vol. 3, 259–294). Springer. https://doi.org/10.1007/978-3-030-69777-8_10
- Wang, Longyue, Lyu, Chenyang, Ji, Tianbo, Zhang, Zhirui, Yu, Dian, Shi, Shuming, & Tu, Zhaopeng (2023). Document-Level Machine Translation with Large Language Models (arXiv:2304.02210). arXiv. <https://doi.org/10.48550/arXiv.2304.02210>