Impact of translation workflows with and without MT on textual characteristics in literary translation

Joke Daems, Paola Ruffo, and Lieve Macken
LT3, Language and Translation Technology Team
Ghent University
Belgium
{firstname.lastname}@ugent.be

Abstract

The use of machine translation is increasingly being explored for the translation of literary texts, but there is still a lot of uncertainty about the optimal translation workflow in these scenarios. While overall quality is quite good, certain textual characteristics can be different in a human translated text and a text produced by means of machine translation post-editing, which has been shown to potentially have an impact on reader perceptions and experience as well. In this study, we look at textual characteristics from short story translations from B.J. Novak's *One more thing* into Dutch. Twenty-three professional literary translators translated three short stories, in three different conditions: using Word, using the classic CAT tool Trados, and using a machine translation post-editing platform specifically designed for literary translation. We look at overall text characteristics (sentence length, type-token ratio, stylistic differences) to establish whether translation workflow has an impact on these features, and whether the three workflows lead to very different final translations or not.

1 Introduction

While originally an outrageous or at least unorthodox idea, the concept of using technology and even machine translation for literary texts has gained ground in recent years. This is evidenced by the existence of works specifically dedicated to technology and creative-text translation (Hadley et al., 2022), the existence of a Literary MT Workshop (dating back to 2019) or the present workshop on Creative-text Translation and Technology at EAMT 2024.

For certain sentences, even raw MT output is seen as comparable to human translation (Toral & Way, 2018), and post-editing NMT output has been shown to be much faster than from-scratch translation for certain language combinations (Terribile, 2023), making it a potentially fruitful way of working, even for literary texts. While time gains are very high for a language combination like English-French, post-editing is actually slower for English-Swedish (Terribile, 2023).

However, additional factors that need to be taken into account are the concerns from literary translators themselves (Daems, 2022; Ruffo, 2021) and the impact of technology-mediated literary translation on a reader's experience (Guerberof-Arenas & Toral, 2020). As part of the broader DUAL-T project[1], which aims to include literary translators' voices in the development of technology-mediated literary translation, this study explores the impact of three different conditions with different degrees of technological support on textual characteristics, which are assumed to influence reader perceptions of a final text.

We continue by exploring some relevant concepts in the field of MT for literary texts and related work on textual features and reader experiences, followed by our research methodology, analysis and results, and we end with conclusions and plans for future work.

2 Related research

An important consideration when using MT for literary text translation, is that MT has been shown to lead to a decrease in lexical richness

[1] https://www.ugent.be/en/research/explorer/eu-trackrecord/heu/heu-msca/dualt.htm

Proceedings of the 1st Workshop on Creative-text Translation and Technology, pages 57–64
June 27, 2024 ©2024 European Association for Machine Translation

(Vanmassenhove et al., 2019). Even after post-editing, the effects of the MT seem to linger, with post-edited texts having lower lexical variety and density than human translations, and having more interference from the source language (Toral, 2019).

In the context of literary machine translation specifically, research has shown that MT systems (both Google Translate and DeepL) produce texts that are lexically less diverse than human translations, and that they have lower lexical and semantic cohesion (Webster et al., 2020). The authors also calculated stylistic differences using Burrows' Delta and found that the styles of Google Translate and DeepL were quite similar, whereas the distance between both MT systems and the human translations was much greater (Webster et al., 2020). Even in a post-editing context, where a professional literary translator is actively requested to keep his own typical translator style, certain words he normally would avoid are still maintained from the MT suggestions (Winters & Kenny, 2023).

Subsequent research into reading experiences found that differences between human translations, machine translations and post-edited texts led to differences in narrative engagement, with human translations generally being rated higher (Guerberof-Arenas & Toral, 2024). However, the authors also found some surprising differences between different languages, with Catalan readers preferring human translations and Dutch readers preferring to read a text in the English source, or the post-edited version, potentially precisely *because* it remains closer to the source (Guerberof-Arenas & Toral, 2024).

These studies suggest that it is crucial to explore textual features of translations produced in different conditions, in order to (in future) explore the influence on translators style and to predict the effects on reading experiences for different kinds of readers.

3 Methodology

The goal of this study was to explore the impact of translation workflow on final text characteristics. Different translation workflows were simulated by means of three possible translation conditions: Microsoft Word without specific translation technology support, Trados Studio 2022 with a relevant translation memory and termbase, and a proprietary MTPE platform. The main textual features we wanted to explore were:

- **Average sentence length and differences in sentence alignment** between source text and translations across conditions. We hypothesized that translators would stay closer to the source text structure and text length, particularly in the MTPE condition (Toral, 2019; Webster et al., 2020), and that they would feel less constrained in the Word condition (Daems, 2022).
- **Lexical diversity** for different conditions. We hypothesized that the more technology-driven workflows would lead to lexically less diverse translations across participants (Guerberof-Arenas & Toral, 2024; Vanmassenhove et al., 2019; Webster et al., 2020).
- **Stylometric differences** between conditions and between the official human translation or MT translation and the corresponding post-edited version. We hypothesized that there would be fewer differences between MTPE texts than between the texts in the Word condition, given the expected interference from the MT output (Toral, 2019) and that MTPE texts would cluster close together with the original MT output, based on earlier findings that using MT can lead to interference with a typical translator's style (Winters & Kenny, 2023).

3.1 Participants

A total of 23 professional English-Dutch translators were recruited for this study via connections established during earlier studies in this field and professional translator associations in Flanders and the Netherlands. Participants were paid 250 euros for their participation (a session lasted 4-5 hours, so participants received 50-60 euros per hour) and received reimbursement of their travel costs.

A diverse set of participants was recruited, with an average age of 48 and an average of 12 years of literary translation experience. Looking at age bands per 10 years, six participants belonged to the youngest age range (26-35) and three participants to the oldest age range (66-75). Participants had between one year and 43 years of literary translation experience.

With regards to technology use, eight participants indicated they had experience working with

CAT tools (four indicated they also used them for literary translation), and eight participants indicated they used post-editing (four indicated they also used it for literary translation).

3.2 Text selection & data preparation

Three short stories were selected from the 2014 short story collection *One More Thing* by B. J. Novak. The selection was driven by a mix of practical factors, such as the fact that the stories were short enough to be translated in one sitting while still being a self-contained piece of narrative, results of readability analyses, and the fact that the humorous and sarcastic nature of the stories could offer some challenges to the translators. The selected texts are titled *Rome* (321 words, 30 sentences, 10.7 words per sentence on average words per sentence), *The Beautiful Girl in the Bookstore* (353 words, 27 sentences, 13.07 average words per sentence), and *They Kept Driving Faster and Outran the Rain* (303 words, 30 sentences, 10.1 average words per sentence). We include a snippet from each text below (Fragments 1-3) to show some of the typical difficulties in the texts.

> He loved saying "Rome" like that. "Head into Rome," "swing by Rome." It was just the nearest place to them. How cool was that! Rome, the city of legends, of conquerors, of history, of myth—this was where he bought *batteries*! The place that people saved up to visit their whole lives: for him, this really was simply the place where he might fill up on gas one day and where the next day he'd have to know the right shop to pick up flowers for his wife to thank her for making dinner—with ingredients he had also picked up in Rome. Rome! That's all Rome was to him! Nothing special at all!

Fragment 1: Snippet from the story *Rome*, containing examples of multi-word expressions, repetition, contrast, and complex syntactic structures.

> There was a magnifying glass built out of a knotted clunk of iron with a foggy lens that magically made even the most serious face, her boyfriend's face, for example, evaporate into a vague and bloated and goofy smile that never failed to make her laugh.
> Things like that.
> "How good does this book smell," she said, pulling a paperback from a shelf. "Like dust on a bottle of vanilla."

Fragment 2: Snippet from the story *The Beautiful Girl in the Bookstore*, containing examples of compound nominals, complex syntactic structures, metaphor and original images.

> *"I love the fauna here at the hotel."*
> *"Wait, what's fauna?"*
> *"Plants, flowers, right?"*
> *"Right, but 'flora* and *fauna.' Isn't flora flowers?"*
> *"Then what's fauna?"*
> *"Don't know. Let's look it up later."*
> *"K."*
> *"K."*

Fragment 3: Snippet from the story *They Kept Driving Faster and Outran the Rain*, containing examples of dialogue and colloquial language use.

Another reason for selecting this collection was that it has been translated into Dutch (*Onverzameld Werk*, translated by Jevgenia Lodewijks, Lydia Meeder, and Maarten van der Werf), so it was possible to create a translation memory and termbase from this material. The translation memory contained the entire collection in English and Dutch, with the exception of the three short stories selected for the study. To generate the termbase, Sketch Engine was used to automatically extract key terms from the entire collection, in this case *including* the three short stories to ensure that at least some terms would be recognized in the texts during translation. As non-commercial research studies form an exception to copyright, no formal permission was sought. We did ensure that none of the material was made public in any way. All participants completed the experiment on the researcher's device, which had a local copy of the translation memory. The MTPE platform connects to an MT system via API to ensure that no data is shared with the company.

3.3 Experimental setup

Participants read an information letter and signed an informed consent form. They then read a translation brief providing some background information on the short story collection and they were instructed to provide a final translation of publishable quality (to the best of their ability in the respective conditions). Participants completed a survey about their professional background and experience with technology.

All participants translated each of the three texts and worked in each of the three translation conditions. While the Microsoft Word condition was always the first condition, the order of the other conditions (Trados and the MTPE platform) was mixed across participants to control for task order effects. Word was used as a baseline, since it is the workflow most literary translators are familiar with, and it allowed us to have less

workflow-text combinations to work with. Text order was also mixed and balanced so that each combination of text and condition appeared a similar number of times. During translation, translators were allowed to use online resources. The translation process was logged using keystroke logging (Inputlog 8.0) and screen recording (OBS Studio 29.1.3).

At the end of the session, participants received a survey where they could rank the different translation workflows and they also took part in an in-depth interview to explore their attitudes and experiences in more detail. The process measure analyses and interview data form the focus of other publications within the broader DUAL-T project (currently under review).

3.4 Data processing

All the produced translations were saved as simple text files for further processing. Texts were first split into sentences and tokenized using Stanza (Qi et al., 2020) then manually aligned for comparison across texts. A couple of different textual features were studied:

Average sentence length: A custom Python script was used to calculate and compare the average sentence length across conditions.

Alignment types: Based on the manual sentence alignment, we could determine how frequently participants diverged from the source text structures and decided to split or merge sentences.

Lexical diversity: Moving average TTR was calculated with the default window size of 50 words, using the lexical-diversity package[2] in Python.

Stylometric differences: The stylo package in R (Eder et al., 2016) was used to calculate classic Delta distances (Burrows, 2002) between texts and explore stylometric differences across texts and conditions. First, stylo generates a list of the most frequent words (MFW) in the whole corpus (the number of words is determined by the user). Then, the frequency of each of those words is checked for each text in the corpus. Burrow's Delta uses *z-scores* (normalized word frequencies) to calculate how big the difference is between the word use in a given text and the corpus as a whole. While this seems like a relatively simplistic approach, the method has proven very successful in authorship attribution, showing that texts with similar scores are generally written by the same author. We created a mini corpus for each source text, containing all translations of that text, including the reference human translation and the machine translation. Given that the words used are very different in each source text, we did not perform the analysis on the corpus as a whole (all translations would simply cluster together per source text). We used stylo to create a bootstrap consensus tree using the 100-500 most frequent words (with 100-word increments). This means that stylo performs a cluster analysis (calculating Burrow's Delta and showing how close different translations cluster together based on those values) for the 100 MFW, 200 MFW, 300 MFW, 400 MFW, and 500 MFW and then combines the results of those cluster analyses to generate the consensus tree (texts that cluster together for at least 50% of the cluster analyses will cluster together in the consensus tree).

4 Results

4.1 Average sentence length and alignment types

Table 1 shows the difference in number of words, sentences and average sentence length for the original source texts, the MT version and the human reference translation. From this, we can see that MT generally stays closer to the source text length than the reference human translator does, and that the number of sentences is exactly the same. A human translator introduces a bit more variability, although the difference in number of sentences is minimal. Average sentence length was lower in the human reference translation for text 1, but higher for text 2 and 3.

TEXT	words	sentences	avg. sentence length
ST1	321	30	10.70
MT1	320	30	10.67
REF1	292	29	10.07
ST2	353	27	13.07
MT2	350	27	12.96
REF2	390	28	13.93
ST3	303	30	10.10
MT3	305	30	10.17
REF3	323	30	10.77

Table 1: Number of words, sentences, and average sentence length for the original source text, machine translation, and reference human translation for each text.

[2] https://pypi.org/project/lexical-diversity/

When we compare this to the average sentence length and ranges for the texts produced by the participants in the different conditions (Table 2), we actually do not see that much difference between the different conditions for each text. Even when considering individual variability across participants by looking at the range between the lowest and highest possible average sentence length, the Word condition does not lead to the greatest variability (as expected), with the exception of text 3, where the range of scores is greater than that in the Trados and MTPE conditions.

CONDITION & TEXT	mean	min	max
MTPE 1	10.55	9.5	11.76
Trados 1	10.53	9.9	11.47
Word 1	10.68	10.1	11.48
MTPE 2	13.25	12.59	13.96
Trados 2	13.32	12.52	14.07
Word 2	13.12	12.48	13.52
MTPE 3	10.48	10.17	10.93
Trados 3	10.38	10.03	10.8
Word 3	10.45	9.97	11.13

Table 2: Descriptive statistics for the average sentence length for the different conditions and texts.

Like they were in the reference translation, differences in alignment are quite rare in the corpus we collected as well. For text one, there were four instances of alignment changes in the MTPE condition, one in Trados, and two in Word. For text two, there was only one instance of alignment change in the MTPE condition, and one in the Word condition. Text three elicited five alignment changes in total, three in the Trados condition and two in the Word condition. While we hypothesized that the MTPE and Trados environments would create more constraints for literary translators by forcing them to translate on a sentence by sentence level, these numbers show that condition did not obviously limit or encourage changes in sentence alignment between source and target.

4.2 Lexical diversity

The expectation based on previous research was that type-token ratio would be much higher for the Word and Trados conditions compared to the MTPE condition, as MTPE texts have been shown to be lexically less diverse. Based on the average MATTR scores across participants, however (Figure 1), this hypothesis cannot be confirmed.

For text 1 and 2, MTPE is actually the condition with the greatest lexical diversity, and for text 3, the differences between MTPE and Word (the condition with the expected greatest diversity) are minimal. Scores for raw MT output are, perhaps surprisingly, also quite high.

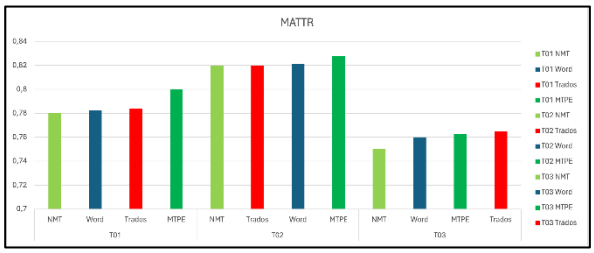

Figure 1: average MATTR for each text and condition, with the MT output as a reference score.

Looking at individual participant scores, we found there are three participants who scored lower on lexical diversity than the raw MT output for text 1 (one in the Trados and two in the Word condition), seven participants for text 2 (two in the MTPE condition, three in the Trados condition, and two in the Word condition), and one for text 3 (in the Trados condition). This shows that post-editing or even regular human translation does not automatically lead to a greater level of lexical diversity.

4.3 Stylometric differences

Figures 2-4 depict the bootstrap consensus tree for each text. In all three, we can see a clear cluster around the machine translation, which exclusively contains translations produced in the MTPE condition (with the exception of one Word translation for text 2). This means that there does seem to be some stylometric similarity between the MT output and a majority of MTPE texts. On the other hand, there are still MTPE translations that appear as part of a mixed cluster (together with Trados and Word conditions) or in isolation, indicating that MT output does not always determine the stylistic outcome of the final MTPE product. The human reference translation is stylistically closest to a translation from the Word or Trados condition, but never to a translation from the MTPE condition.

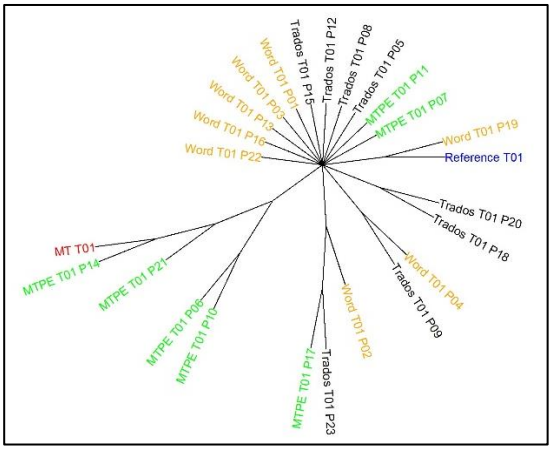

Figure 2: Bootstrap consensus tree for text 1. 100-500 most frequent words without culling, Classic Delta distance Consensus 0.5.

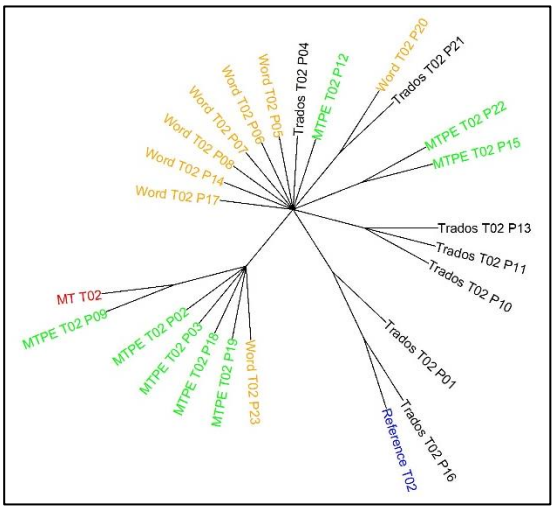

Figure 3: Bootstrap consensus tree for text 2. 100-500 most frequent words without culling, Classic Delta distance Consensus 0.5.

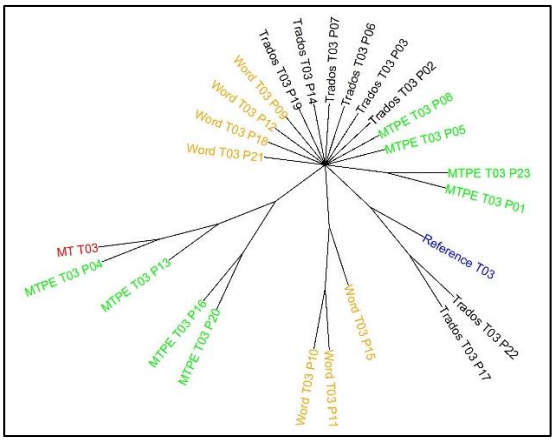

Figure 4: Bootstrap consensus tree for text 3. 100-500 most frequent words without culling, Classic Delta distance Consensus 0.5.

5 Discussion and Conclusions

The main goal of this exploratory study was to establish the impact of translation workflow on textual differences. We compared the translations produced by professional literary translators in three conditions: using Microsoft Word, using Trados, and using a proprietary MTPE tool. The hypothesis was that translations produced in Word would showcase the most individuality and divergence from the source text as the blank page does not offer specific constraints, whereas the MTPE was expected to remain closest to source and/or MT output as it offers the MT output as a starting point. Trados was expected to lead to some constraints (particularly by forcing translators to work on a segment level), but fewer than the MTPE workflow (as there was no translation to start from here).

We compared average sentence length and changes in sentence alignment, lexical diversity, and stylometric differences. Average sentence length did not seem to differ remarkably across conditions. Earlier research on sentence patterns in English and Dutch showed that, in contrast with academic texts, newspaper articles, and leaflets, sentence length for short stories can actually be similar in both languages (Tavecchio, 2010). The present study shows that the condition in which the text was produced does not change this. Changes in sentence alignment were also relatively rare, and occurred in all three conditions, contrary to expectations that they would be most frequent in the Word condition. Based on previous research, we expected MTPE to be less lexically diverse than translations in other conditions, but this could not be confirmed either (on the contrary, MTPE was the most lexically diverse for 2/3 texts). Stylometric analysis based on Burrow's Delta (2002) did show some similarities between MT output and a majority of translations produced in the MTPE condition, indicating that there is some similarity in their word use.

The analysis presented in this paper is a preliminary analysis of textual features in our dataset that contradicts some core assumptions about the 'homogenisation' of MTPE texts, and at the same time encourages additional exploration of the data. For future work, we aim to conduct more extensive analyses on this data, e.g., by exploring if translation workflow influences different metrics of syntactic equivalence (Vanroy et al., 2021). As Winters and Kenny suggest, studies like this "usually branch into richer qualitative analyses on the

basis of their initial quantitative findings" (2023, p. 70). This is precisely what we aim to do. We are currently annotating all texts on the basis of units of creative potential and creative shifts (Guerberof-Arenas & Toral, 2022), multi-word units (Colson, 2019), and translation relations (Zhai et al., 2018) in order to get a more in-depth understanding of translation choices and how they are (not) mediated by the different workflows.

Acknowledgements

This project has received funding from the European Union's Horizon Europe (HORIZON) research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101062428.

References

Burrows, J. (2002). 'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, *17*(3), 267–287. https://doi.org/10.1093/llc/17.3.267

Colson, J.-P. (2019). Multi-word Units in Machine Translation: Why the Tip of the Iceberg Remains Problematic – and a Tentative Corpus-driven Solution. *Proceedings of the Third International Conference, Europhras 2019, Computational and Corpus-Based Phraseology*, 145–156. https://doi.org/10.26615/978-2-9701095-6-3_020

Daems, J. (2022). Dutch literary translators' use and perceived usefulness of technology: The role of awareness and attitude. In Hadley, James and Taivalkoski-Shilov, Kristiina and Teixeira, Carlos and Toral, Antonio (Ed.), *Using Technologies for Creative-Text Translation* (pp. 40–65). Routledge.

Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, *8*(1), 107–121.

Guerberof-Arenas, A., & Toral, A. (2020). The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces*, *9*(2), 255–282. https://doi.org/10.1075/ts.20035.gue

Guerberof-Arenas, A., & Toral, A. (2022). Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces*, *11*(2), 184–212. https://doi.org/10.1075/ts.21025.gue

Guerberof-Arenas, A., & Toral, A. (2024). To be or not to be: A translation reception study of a literary text translated into Dutch and Catalan using machine translation. *Target. International Journal of Translation Studies*, *36*(2), 215–244. https://doi.org/10.1075/target.22134.gue

Hadley, J. L., Taivalkoski-Shilov, K., Teixeira, C. S. C., & Toral, A. (Eds.). (2022). *Using Technologies for Creative-Text Translation.* Routledge. https://doi.org/10.4324/9781003094159

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages* (arXiv:2003.07082). arXiv. https://doi.org/10.48550/arXiv.2003.07082

Ruffo, P. (2021). *In-between role and technology: Literary translators on navigating the new socio-technological paradigm.* Heriot-Watt University.

Tavecchio, L. M. (2010). *Sentence patterns in English and Dutch: A contrastive corpus analysis* [PhD-Thesis - Research and graduation internal]. LOT.

Terribile, S. (2023). Is post-editing really faster than human translation? *Translation Spaces*. https://doi.org/10.1075/ts.22044.ter

Toral, A. (2019). Post-editese: An Exacerbated Translationese. In M. Forcada, A. Way, B. Haddow, & R. Sennrich (Eds.), *Proceedings of Machine Translation Summit XVII: Research Track* (pp. 273–281). European Association for Machine Translation. https://aclanthology.org/W19-6627

Toral, A., & Way, A. (2018). What Level of Quality Can Neural Machine Translation Attain on Literary Text? In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation Quality Assessment: From Principles to Practice* (pp. 263–287). Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7_12

Vanmassenhove, E., Shterionov, D., & Way, A. (2019). Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. In M. Forcada, A. Way, B. Haddow, & R. Sennrich (Eds.), *Proceedings of Machine Translation Summit XVII: Research Track* (pp. 222–232). European Association for Machine Translation. https://aclanthology.org/W19-6622

Vanroy, B., Clercq, O. D., Tezcan, A., Daems, J., & Macken, L. (2021). Metrics of Syntactic Equivalence to Assess Translation Difficulty. In M. Carl (Ed.), *Explorations in Empirical Translation Process Research* (pp. 259–294).

Springer International Publishing. https://doi.org/10.1007/978-3-030-69777-8_10

Webster, R., Fonteyne, M., Tezcan, A., Macken, L., & Daems, J. (2020). Gutenberg goes neural: Comparing features of Dutch human translations with raw neural machine translation outputs in a corpus of English literary classics. *INFORMATICS-BASEL*, *7*(3), 21.

Winters, M., & Kenny, D. (2023). Mark My Keywords. In A. Rothwell, A. Way, & R. Youdale, *Computer-Assisted Literary Translation* (1st ed., pp. 69–88). Routledge. https://doi.org/10.4324/9781003357391-5

Zhai, Y., Max, A., & Vilnat, A. (2018). Construction of a Multilingual Corpus Annotated with Translation Relations. In P. Machonis, A. Barreiro, K. Kocijan, & M. Silberztein (Eds.), *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing* (pp. 102–111). Association for Computational Linguistics. https://aclanthology.org/W18-3814