Optimising ChatGPT for creativity in literary translation: A case study from English into Dutch, Chinese, Catalan and Spanish

Shuxiang Du[1†], **Ana Guerberof Arenas**[1†], **Antonio Toral**[2†]
Kyo Gerrits[1], **Josep Marco Borillo**[3]
[1]Centre for Language and Cognition, University of Groningen
[2]Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant
[3]Departament de Traducció i Comunicació, Universitat Jaume I
sophie321du@gmail.com

Abstract

This study examines the variability of ChatGPT's machine translation (MT) outputs across six different configurations in four languages, with a focus on creativity in a literary text. We evaluate GPT translations in different text granularity levels, temperature settings and prompting strategies with a Creativity Score formula. We found that prompting ChatGPT with a minimal instruction yields the best creative translations, with "Translate the following text into [TG] creatively" at the temperature of 1.0 outperforming other configurations and DeepL in Spanish, Dutch, and Chinese. Nonetheless, ChatGPT consistently underperforms compared to human translation (HT). All the code and data are available at https://github.com/INCREC/Optimising.

1 Introduction

The intersection of artificial intelligence (AI) and creativity in the domain of translation presents a fascinating and challenging field for research. Even if the development of machine translation (MT) technologies, especially through the advent of Large Language Models (LLMs) like ChatGPT, has reshaped the landscape of the language industries, there remains a notable gap in the creative capacities of MT outputs in comparison to that of professionals (Karpinska and Iyyer, 2023). This type of translation, often applied to literary texts, requires not just the accurate conveyance of meaning but also the preservation of style, tone, and creative nuances inherent in the source text to create an effect on the reader that is not purely information driven. Since new models offer a dialogic capacity, we explore in this paper the best set of variables to generate the most creative translations using ChatGPT.

We investigate various configurations of ChatGPT, including different text granularities, temperature settings, and prompting strategies, alongside a comparison with translations by neural machine translation (NMT) systems, and human translations as references. The primary aim is to evaluate how these configurations impact the creativity and quality of the translations produced, using both a manual creativity scoring system and automatic evaluation metrics. The experiment involves translating a short science fiction story by Kurt Vonnegut, "2BR02B", from EN (English) to ZH (Chinese), NL (Dutch), CA (Catalan), and ES (Spanish). The translations are manually annotated to assess creative shifts (CSs) and errors, providing a detailed analysis of how ChatGPT in different configurations handles the nuanced demands of literary translation. Thus, central to this research are the questions:

RQ1: What is the variability in MT outputs from ChatGPT under different settings?

RQ2: What is the optimal prompting setting for the most creative MT output using ChatGPT?

2 Related Work

LLMs like ChatGPT have demonstrated promising performances in natural language processing tasks (Kalyan, 2023). LLMs such as IOL-Research, Unbabel Tower 70B, and Claude-3.5-Sonnet are the top performing MT systems submitted to the last edition of WMT's general translation task (Kocmi et al., 2024). ChatGPT for MT has demonstrated promising applications to help users translate specific contents or entire documents, especially between high-resourced languages (Jiao et al., 2023, Hendy et al., 2023). However, whether it outperforms NMT systems or commercial MT systems is still under debate (Kalyan, 2023).

The research community has investigated the effectiveness of ChatGPT for MT in different aspects.

[†]Equal contribution

Proceedings of Machine Translation Summit XX Volume 1, pages 578–591, June 23-27, 2025

Gao et al. (2024) focused on developing advanced prompting strategies by including additional information like task, domain, and syntactic information like PoS (parts of speech) tags. The researchers tested the language pairs English ↔ Spanish, English ↔ French, and Spanish ↔ French in the domains of news, e-commerce, social, and conversational in a sentence level. They concluded that including appropriate information about the input text in the prompt, such as specifying translation task or context domain, can improve the performance of ChatGPT. ChatGPT has a higher BLEU (Papineni et al., 2002) score in four out of the six language pairs when compared to Google Translate (GT) and DeepL Translate (DeepL) with their proposed advanced prompting strategies.

In terms of text granularity levels, Wang et al. (2023) examined the performances of ChatGPT for document-level translation, covering three language pairs (Chinese ⇒ English, English ⇒ German, and English ⇒ Russian) in seven domains (news, social, fiction, Q&A from an online forum, TED, Europarl, and subtitle). The researchers reported that ChatGPT does well when the sentences in the document are combined and given at once to the model. With this prompting strategy, it exhibited better performances than commercial MT systems according to human evaluation and also outperformed most document-level NMT methods in terms of d-BLEU scores.

Temperature is a hyperparameter in LLMs that regulates the randomness in text generation by adjusting the probability distribution of potential next words (Peeperkorn et al., 2024). Decoding with higher temperatures displays greater linguistic variety, while low values tend to generate grammatically correct and more deterministic text (Ippolito et al., 2019). Peng et al. (2023) explored the impact of temperature, task, and domain information on the translation performance of ChatGPT. In the translation of English, Chinese, German, and Romanian of biomedicine, news, and e-commerce texts, the study showed that ChatGPT performance degraded with an increase in temperature in terms of both BLEU and COMET (Rei et al., 2020) scores, and hence it was recommended to use a lower temperature (recommended is 0 for their test set). Additionally, including task and domain information in the prompt enhanced the translation performance of ChatGPT consistently for both high- and low-resource languages in their research.

As MT technology advances, there is growing interest in exploring how well these systems can handle the complexities of literary translation. With respect to LLMs, Karpinska and Iyyer (2023) evaluated the performance of ChatGPT in translating literary paragraphs across 18 linguistically diverse language pairs. The authors experimented with three different prompting strategies, namely translating sentence by sentence in isolation, translating sentence by sentence in the presence of the rest of the paragraph, and translating the entire paragraph at once. According to human evaluation, when translating entire paragraphs, ChatGPT produced translations of significantly higher quality compared to other strategies and commercial systems. However, critical errors such as content omissions still occur. The findings suggest that while ChatGPT can leverage larger context units like paragraphs to enhance translation quality, this is yet not sufficient on their own for high-stakes applications like literary translation where nuanced understanding and stylistic consistency are crucial.

The challenges of literary MT lie not only in the performance of the systems but also in the evaluation of the results. Fonteyne et al. (2020) provided an in-depth evaluation of the quality of a novel translated by NMT from English ⇒ Dutch. Unlike traditional sentence-level evaluations, this study emphasized the importance of document-level analysis to better assess the coherence and cohesion of translated texts, which are crucial in literary translations. It utilized an adapted version of the SCATE error taxonomy (Tezcan et al., 2017), which considers errors at both the sentence and document levels. Again, the findings suggested that while NMT can produce a substantial portion of error-free translations, significant errors remain, particularly with complex elements like style and coherence that are vital to literary texts. Therefore, it is important to consider metrics other than error annotation when evaluating literary texts.

In the studies of ChatGPT for translation, most evaluations focus on automatic metrics like COMET and BLEU, while the specific aspect of creativity has hardly been touched upon. This could be because creativity in translation is hard to measure. Bayer-Hohenwarter (2009) proposed a framework for assessing translational creativity based on the concepts of novelty, acceptability, flexibility, and fluency. Novelty in translation is characterized by three main aspects: exceptional performance that significantly surpasses routine translation activities, uniqueness or rarity within a specific cor-

pus of translations, and non-obligatory translational shifts that indicate a high level of translator engagement and creativity. Acceptability is defined as "skopos adequacy" (Bayer-Hohenwarter, 2009, 2). This emphasizes that a creative translation must not only be innovative but also appropriate and useful within the context for which it is intended. These novelty and acceptability aspects of the framework are largely adopted in this research.

Bayer-Hohenwarter (2011) defines creative shifts as transformative operations in translation that deviate from the direct replication of the source text. These shifts are categorized into three types: abstraction, where translators generalize specific details from the source; modification, which involves alterations to better suit the cultural or contextual needs of the target text; and concretization, where translators add specific details not explicitly mentioned in the source text. Bayer-Hohenwarter proposed a systematic methodology to measure creativity in translation by identifying and analyzing these creative shifts. She defined specific "units of analysis" within the texts, identifying both "creativity units" (requiring high problem-solving capacity) and "routine units" (relatively straightforward translation tasks). The results were quantified by calculating the proportion of creative shifts versus literal reproductions. The study also examined the relationship between the frequency of creative shifts and the overall quality (acceptability) of the translations. There was a general trend suggesting that translators who produced more creative shifts also produced higher-quality translations. However, this was not a strict correlation, as some creative shifts led to errors, particularly among less experienced translators.

Guerberof-Arenas and Toral (2020, 2022) created a formula (see section 3) to quantify creativity in translations, offering a measurable way to assess and compare the creative output of different translation modalities, including MT. Their study involved the translation of literary texts from English to Catalan and Dutch. The texts were translated by professionals, post-edited by professionals, and machine translated. By applying a creativity score to the translations, they found MT outputs to be less creative than professional translations and that they limited the translator's creativity in post-editing. The quantification framework they established is used in this research.

In their Master thesis Du (2024) use this creativity index in an evaluation of ChatGPT translations of a literary text in the English ⇒ Chinese translation direction. They investigated different set-ups of ChatGPT including levels of text granularities, different temperatures, prompting strategies, and few-shot prompting. The findings indicated that the quality and creativity of ChatGPT translations vary across these configurations. The best setting in their study was a document-level translation with a temperature of 1.0 and a direct prompt to be more creative. In this paper, we replicate the experiment with more languages and a more in-depth analysis.

3 Methodology

In this section, we explain the source text (ST) used, how the target texts (TTs) were generated in the different phases of our experimentation, as well as the data annotation and analysis process.

3.1 Source Text

The study utilized a curated dataset comprising different translations of a short science fiction story by Kurt Vonnegut: *2BR02B*[1] (Vonnegut, 1999). The story is a short science fiction piece set in a future society where aging has been cured and the population is strictly controlled to remain at forty million. Individuals must volunteer for death to allow new births. It revolves around a family about to give birth to three kids and therefore in need of three volunteers to die.

This story was selected for three reasons: A) we have an existing corpus of annotations on the units of creative potential in the story (Guerberof-Arenas and Toral, 2022), B) to our knowledge it has not been translated into the target languages to date[2] and therefore we assume it has not been used in the training data of ChatGPT, and C) it requires a high level of translation creativity.

The story was processed in Python to be broken into separate paragraphs. The text overall contains 123 paragraphs, 234 segments and 2548 words. There are 185 units of creative potential (UCP) in total, annotated by two experienced translators and researchers in the previous study (Guerberof-Arenas and Toral, 2022). These are units in the ST that are expected to require translators to use problem-solving skills, as opposed to those that are regarded as routine units with little creative potential (Bayer-Hohenwarter, 2011).

[1] https://www.gutenberg.org/ebooks/21279

[2] Not found in the Unesco Translationum database https://www.unesco.org/xtrans/bsform.aspx nor on National Library of China https://www.nlc.cn/web/index.shtml

3.2 Target Text

For the target text (TT), we used the model gpt-4o-2024-08-06 with the ChatGPT API[3] to translate the text into ZH, NL, ES and CA. This version is chosen for three reasons: A) it was the latest stable model of ChatGPT when we started our experimentation, thus representing state-of-the-art performance, B) according to OpenAI[4] , this version performs better on text in non-English languages, C) in terms of data training, the cost of this version is relatively lower and the speed is faster when compared to ChatGPT-4.

Due to limited capacity and the exploratory nature of this experiment, we decided to annotate a subset of the text. We selected a series of UCPs that were previously singled out by two annotators in the ST to ensure a better representation of the creative potential of this text. In the end, 54 UCPs, present in 48 separate sentences with a total of 602 words were selected for the annotation task in the TT. To prepare the sentence-aligned files, we manually post-processed the text by extracting the 48 sentences in each translation.

Each translation of the sentences in the TT was manually annotated for a detailed comparative analysis of creativity across different translations. The annotators were four of the researchers that are experienced translators or have a language related Master degree in the selected language combinations: there was therefore one annotator per language combination.

As the baseline of the study, DeepL has been chosen to compare with ChatGPT. The reason is that in the preliminary experiment (Du, 2024) it offered a more pleasant-to-read translation than other NMT systems like Google Translate. Since DeepL is not available for CA, we used two popular NMT systems for this target language: Softcatalà's *Traductor*[5] and Google Translate.[6]

3.3 Data Collection

In this experiment, we try a range of text granularities, temperature settings, and zero-shot prompting strategies based on Du (2024) master project to generate translations with ChatGPT. The experiments and annotations were conducted between October 2024 and January 2025. Figure 1 shows an overview of the NL and ZH workflow process as an example. The workflow in each phase slightly differs for the other two languages (CA and ES).

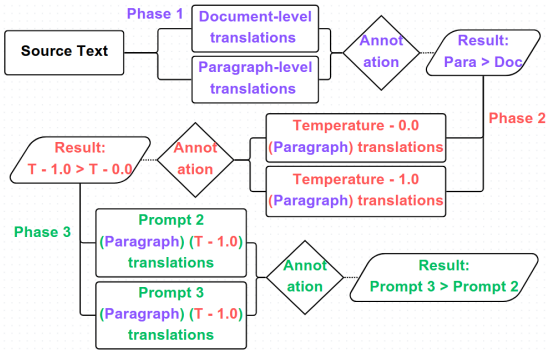

Figure 1: Workflow for ZH and NL

3.3.1 Phase 1. Text Granularity

The variable in the first phase is text granularity. We translated the text at both paragraph level (setting 1a) and document level (1b). At the paragraph level, we entered the same prompt for each paragraph in the story, with each request done separately to avoid context interference. At the document level, we entered the same prompt followed by the entire story in one request. It is worth noting that 14 of the 48 sentences involved in the evaluation process were single-sentence paragraphs.

The prompts used in phase 1 are:

Prompt 1 (1a): "Translate into [TG]: [Input]"

Prompt 1 (1b): "Translate the following text into [TG]: [Input]"

On the one hand, having the whole document available offers more context, which should be useful for its translation. On the other, recent research has shown that translation performance decreases with the length of the input text (Peng et al., 2024).

After the evaluation of the creativity score for these two outputs (see Section 4.1), we proceeded with the better granularity method for each target language to further experiment with different temperatures and prompts. Namely, in CA and ES, the document-level translations were assessed to be better, thus the following experiment in CA and ES were conducted at document level (1b), while the opposite was the case for ZH and NL (1a).

3.3.2 Phase 2. Temperature

The temperatures selected are 0.0 (2a) and 1.0 (2b). In the API setup, ChatGPT's temperature can range from 0.0 to 2.0. The default temperature setting of

[3]Version 1.54.4, i.e. the latest when we started out experiments.

[4]https://openai.com/index/hello-gpt-4o/

[5]https://www.softcatala.org/traductor/

[6]https://translate.google.com/

ChatGPT is officially stated to be set to 1.0[7]. However, according to our experience the default value seems to be 0.0. Namely, we used the automatic evaluation metric chrF (Popović, 2015) to examine how similar the translations produced using different temperature values are to the translations produced in phase 1, which used the default setting for temperature. For NL, for example, the chrF result decreased as the temperature went up: 88 (temperature=0.0), 86 (0.5), 83 (1.0), 82 (1.1) and 81 (1.2). This means that ChatGPT is not deterministic, even with temperature 0.[8] We then chose the value 0 in phase 2 to see the effect of non-determinism.

On the other hand, the higher the temperature is, the more creative the text is expected to be. However, the highest value we chose was not the maximum offered by the API (2.0) but 1.0. This is because we noticed that at higher temperature values, there are more instances of "word vomit" in the output which makes the text incoherent and impossible to read.[9] Therefore, at temperature 1.0 the system is most likely to generate more creative content while not suffering from word vomit.

The same prompt as in phase 1 was used and we proceeded with the best temperature setting after evaluation (see Section 4.2). For ES, NL, and ZH, we proceeded with temperature 1.0 (2b), while for CA we proceeded with 0.0 (2a).

3.3.3 Phase 3. Prompting Strategies

The zero-shot prompting strategies we designed included prompting with the specific domain information, i.e. author and genre (3a), and prompting with direct instructions to generate creative outputs (3b). The final prompts are as follows:

- **Prompt 2** (more info about genre and author, 3a): Translate the following text into [TG] taking into consideration that this is (from) a science fiction story by Kurt Vonnegut: [input]

- **Prompt 3** (request of creativity, 3b): Translate the following text into [TG] creatively: [input]

[7] https://platform.openai.com/docs/api-reference/chat/create

[8] If it was deterministic, then chrF's score when comparing phase 1's translation with phase 2's translation with temperature=0 would have been 100,

[9] We tried three values of temperature higher than 1.0 (1.1, 1.2 and 1.5) and noticed severe issues with values 1.1 (CA), 1.2 (ES), 1.5 (NL and ZH). We speculate that the reason why ChatGPT has issues in CA and ES at a lower temperature values than NL and ZH is because for the former the document is translated at once.

3.4 Data Annotation

Following data collection, each sentence of each translation was manually annotated in terms of acceptability and novelty, as discussed in section 2. The annotators were blind to the specific setting they were evaluating.

Acceptability was measured according to the number and severity of errors in the TTs based on the harmonized DQF-MQM Framework (Lommel et al., 2014). The severity of each error was marked as: Neutral (0 points for repeated errors or preferences), Minor (1 point), Major (5 points), and Critical (15 points). Minor refers to errors that do not lead to loss of meaning and do not confuse or mislead the reader but are noticeable, hence they decrease stylistic quality, fluency or clarity, or make the content less appealing. Major refers to errors that may confuse or mislead the reader or hinder the understanding of the text due to significant change in meaning or because errors appear in a visible or important part of the content. Critical refers to errors that may misrepresent or damage the reputation of the author or publishing house, causes the text to stop working as a literary artefact and affect the communicative flow, or if the language is perceived as offensive (when unintended), but also, if the text departs from the source text in such a significant way that has a large impact on the understanding of the entire story.

For example: the title of the story, "2BR02B", is a play on words on the famous quote *To be or not to be* in Shakespeare's *Hamlet*. If left in English in ES and CA, the understanding of the entire story is compromised, and it is therefore considered a Critical error. If the word *business* is translated literally in a context where it does not refer to a commercial activity but to a person's concern, then this can be considered a Major error because the entire text is understood albeit with certain difficulties. Finally, a spelling mistake would be considered Minor.

For novelty, the translation solutions to the UCPs selected were annotated in the TTs. All translations that deviate from the ST that are neither the exact reproduction of the ST nor an omission nor an error count as CS and are classified in the following manner: Abstraction refers to instances when translators use more vague, general or abstract solutions. Concretization refers to instances when the TT evokes a more explicit, more detailed, and more precise idea or image. Modification refers to instances when translators use a different solution in

the TT (e.g. express a different metaphor without the image becoming more abstract or concrete).

For example: if the title of the story, *2BR02B*, is translated into CA as *C-O-N-O-C*, a play on words that evocates *Ser o no ser*, the standard Catalan phrase, this would be considered a Modification and classified as CSM. While in NL, the title *2BR02B* is left as is, as the standard phrase in Dutch remains *To be or not to be*. This is then considered a Reproduction, and classified as such. As this exemplifies, Reproductions are not errors by default, although some UCPs that are not translated might be considered as containing an error.

In the process of annotation, the translation of the 54 UCPs was assessed. For each translated UCP, the annotator decided if the resulting TT was a CS, an omission (O), a reproduction (R), or if it was impossible to classify (E). The CSs were further classified into abstraction (CSA), concretization (CSC), and modification (CSM). Each of the 48 sentences were annotated for errors according to the severity criteria described. The total number of CSs and Error points was used for the creativity index, introduced next.

3.5 Data Evaluation

Acceptability and novelty are combined into a single score using the creativity index (CI) formula:

$$\text{CI} = \left(\frac{\#CSs}{\#UCPs} - \frac{\#\text{error points}}{\#\text{words in ST}} \right) \times 100$$

The index considers both novelty (CSs) and appropriateness or acceptability (errors), enabling a quantifiable comparison between different translation modalities (Guerberof-Arenas and Toral, 2020, 2022).

Apart from the creativity index, we used a number of automatic evaluation metrics (AEMs): BLEU, chrF, TER (Snover et al., 2006), COMET and COMET-Kiwi (Rei et al., 2022). The first three are string-based[10] while the last two are based on multilingual language models.[11] Another distinction is that the first four evaluate a translation with respect to a reference translation (see Section 3.6),[12] while the last one does so with respect to the source text. Since we do not have a reference translation for ZH, only COMET-Kiwi was used.

[10]We compute them with sacrebleu 2.5.1

[11]We used models wmt22-comet-da and wmt22-cometkiwi-da, respectively.

[12]COMET takes into account also the ST.

3.6 Human Reference

Since this experiment utilizes a dataset from the Guerberof-Arenas and Toral (2022) project, we had access to translations created by professionals in EN⇒CA, EN⇒NL and EN⇒S,[13] but unfortunately not for EN⇒ZH. Table 1 shows the scores for the selected UCPs for these languages.

	# CSs	# Errors	Error Points	CI
ENCA	21	2	2	40
ENES	22	6	6	40
ENNL	29	17	25	50

Table 1: Creativity Index in Human Reference

The results for ENCA and ENES were annotated by a professional literary translator, while the ENNL was annotated by a different one for that language pair, and this could account for the differences in judgement, although, of course, this could also mean that there are differences in the quality provided by the translators. One aspect to note here is that while annotating the UCPs, the reviewers also remarked that the entire segments contained other CSs. For example, in ES and CA, the translators changed the name of the characters to be able to create meaningful play on words that were present in the ST.

4 Results

The following subsections contain the results obtained in each of the phases explained in the methodology. Detailed annotations per language are provided in Appendix A.

4.1 Phase 1. Text Granularity

Table 2 shows the results for phase 1, ChatGPT outputs at paragraph (1a) and document level (1b) were compared.

In this instance, the best solutions for ES and CA are at document level, as we would expect since the context of the sentences is considered. However, for NL and ZH the best performance is at paragraph level, mainly due to error points.

In the case of NL, the version at document level included more grammatical errors (such as missing articles *van drieling* ("of triplet"), incorrect subject-verb agreement, e.g. *wat je zaken was* ("What your

[13]The Spanish translation was not analyzed in the previous project, but was translated by the researcher to be used as a reference. This version was then annotated for errors by a professional literary translator.

	Paragraph (1a)				Document (1b)			
	# CSs	# Errors	Error Points	Score	# CSs	# Errors	Error Points	Score
ENCA	**5**	**51**	199	-23.80	**5**	**51**	**158**	**-16.99**
ENES	6	59	253	-31.00	**10**	**52**	**219**	**-18.00**
ENNL	**15**	**61**	**108**	**9.84**	12	68	174	-6.68
ENZH	**11**	**51**	**187**	**-10.69**	10	56	298	-30.98

Table 2: Creativity score for two different text granularities: paragraph and document. The best score per language and criterion is shown in bold.

business is", where *zaken* is plural but *was* singular), hallucinations ("formidable" became *ontslagbaar*, a non-existent word, meaning something like *unfireable*), and typos than the version at paragraph level. The paragraph level version does lack consistency at times ("orderly" is translated differently three times), but still has fewer errors.

For ZH, the document-level translations tend to make more grammatical and factual errors, too, especially towards the end of the document. For example, "sheave-carrier" is translated to " 运载屁股的人 (ass-carrier)", which is a critical error that disrupts the narrative significantly. "He was seven feet tall" is translated to " 两个高大的人 (two tall men)", perhaps in an attempt to convert seven feet to two meters. "He said to her as she fell" is translated to " 他对她说，落 (he said to her, falls)" and is not coherent in the target language. Such examples suggest that ChatGPT tends to perform progressively worse for ZH as it processes the whole document. This is in line with previous findings by Wang et al. (2024) that LLMs demonstrate short-comings in long-text translations, and their performance diminishes as document size increases.

For CA, the difference (in quantitative terms) between the paragraph and the document levels is largely accounted for by the fact that, in the latter, all the fanciful sobriquets[14] for an institution (the Federal Bureau of Termination) are translated, whereas at paragraph level only 6 (out of 14) are. In other respects, differences between the two CA versions are not that pronounced.

For ES, the paragraph and document level translations are not that dissimilar quantitatively. However, the document level resolves certain translations problems better. For example, the expression "seven feet tall" is converted at document-level into meters while it remains in feet at sentence level, and "trick telephone number" is translated as *número de teléfono trampificado* which does not exist as a term, while the document-level uses *número de teléfono con truco* that is correct in Spanish.

4.2 Phase 2. Temperature

Table 3 shows the results of ChatGPT outputs when the temperature was set at 0.0 (2a) and at 1.0 (2b).

The best performance for ES, NL and ZH are at a temperature of 1.0, but for CA the best output is at temperature 0.0. For most languages, a temperature value of 1.0 outputs more CSs but also more errors–only in ES does a temperature of 0.0 have more errors–as was expected.

In NL, for instance, the output at temperature 1.0 translates "triplets" as *drieën* (*threes*)–this is more creative and it could work in some contexts, but not when talking about three babies born at the same time. Still, weighing the CSs against the errors in the creativity index reveals that a temperature of 1.0 has a better output for ES, NL and ZH, despite the errors in the last two.

The general trend is observable for CA too–a higher temperature yields both more CSs and more errors. What sets CA apart is that the higher number of CSs does not compensate for the number of errors because of their severity. At temperature 0.0, for example, "Chicago Lying-in Hospital" is adequately translated, whereas at temperature 1.0 the "Lying-in" segment is left untranslated. Other segments are translated in both settings, but the rendering provided at temperature 1.0 is not acceptable. For example, "Kiss this sad world toodle-oo" is translated as *donaré adéu* ('I will give goodbye'), a collocation that does not exist in CA. Also, "Good gravy", used as an interjection, is adequately translated at temperature 0.0 and wrongly rendered as *Bona sort* ("Good luck") at 1.0.

4.3 Phase 3. Prompting Strategies

Table 4 shows the results for ChatGPT outputs when prompting with more information about

[14]These are nicknames given to the gas chambers in this dystopian world, e.g. Weep-no-more, Good-by, Mother or Easy-go

	T-0.0 (2a)				T-1.0 (2b)			
	# CSs	# Errors	Error Points	Score	# CSs	# Errors	Error Points	Score
ENCA	4	**57**	**200**	**-25.82**	**6**	69	248	-30.08
ENES	8	55	216	-21.00	**9**	**50**	**164**	**-11.00**
ENNL	11	**49**	**99**	3.93	**12**	52	108	**6.13**
ENZH	11	**45**	**155**	-5.38	**14**	51	165	**-1.48**

Table 3: Creativity score for two different temperature values: 0.0 and 1.0.
The best score per language and criterion is shown in bold.

	Prompt 2 (3a)				Prompt 3 (3b)			
	# CSs	# Errors	Error Points	Score	# CSs	# Errors	Error Points	Score
ENCA	**4**	57	204	-26.48	**4**	**55**	**202**	**-26.15**
ENES	12	57	244	-18.31	**13**	**43**	**166**	**-3.50**
ENNL	10	43	89	3.73	**18**	**30**	**76**	**20.71**
ENZH	14	39	171	-2.48	**15**	**37**	**161**	**1.03**

Table 4: Creativity score for two different prompting strategies.
The best score per language and criterion is shown in bold.

genre and author (Prompt 2, 3a) or a request of creativity (Prompt 3, 3b).

For all our languages, Prompt 3 (3b) has better solutions than Prompt 2 (3a) as it generates more CSs and fewer errors. When compared to the results of the other phases, we also see that Prompt 3 has the best performance overall for ES, NL and ZH, with the most number of CSs and the least number of errors. However, for CA, the best performance was in Phase 1 (1b), with Prompt 1 at the document level. The explanation for this lies again in the translation of sobriquets, which are left untranslated in both 3a and 3b. In fact, the only settings in which sobriquets are translated at all are paragraph level (6 out of 14, as said above) and document level (all of them). Since 4 sobriquets are UCPs classified independently, their translations impact the formula. If the sobriquets were excluded, 3a and 3b would be the best-performing settings for CA. The sobriquets were also problematic for ZH and ES: for ES, 3a kept all sobriquets in English and 3b did not translate 2 out of 14 sobriquets; for ZH, it was 3b that did not translate the words but kept them in English, although 3b output had better performance than 3a or any of the other outputs. Surprisingly, for NL, both 3a and 3b translated the sobriquets into Dutch, although 3a retained one sobriquet in English. This might explain the relative high score for NL with 3b compared to the other languages.

4.4 ChatGPT vs Others

We also compare the performance of ChatGPT with that of DeepL and since CA is not available in the latter, we use GT (ENCA-G) and Softcatalà (ENCA-S). At the time we ran DeepL its new-gen version was available for ZH but not for ES nor NL. Therefore we used DeepL new-gen for ZH and DeepL classic for ES and NL. The creativity index of these baseline systems are shown in Table 5.

	# CSs	# Errors	Error Points	CI
ENCA-S	1	83	393	-63.43
ENCA-G	1	66	261	-41.50
ENES	9	58	237	-22.70
ENNL	6	51	103	-6.00
ENZH	11	42	152	-4.88

Table 5: Creativity Index in Others (3c). S stands for Softcatalà's *Traductor* and G for Google Translate.

In all languages, the selected NMT system (3c) performs worse than the best setting of ChatGPT. In ZH, NL and ES, DeepL performs better than some of the other settings in ChatGPT, while in CA the two NMT systems perform worse than all ChatGPT outputs. This shows that ChatGPT, with an appropriate prompting strategy, has the potential to outperform its NMT counterparts in literary text in terms of creativity.

5 Analysis

We wanted to further analyse the MT performance in all the phases, and also compare the best performing setting with the professional translation

described in Section 3.6. Firstly, Figure 2 and Figure 3 show the comparison between ChatGPT in the different settings in terms of CSs and Error points (including Others as in Section 4.4).

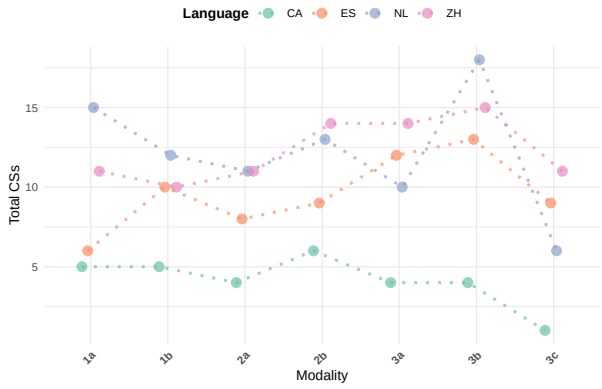

Figure 2: Total CSs per Modality and Language

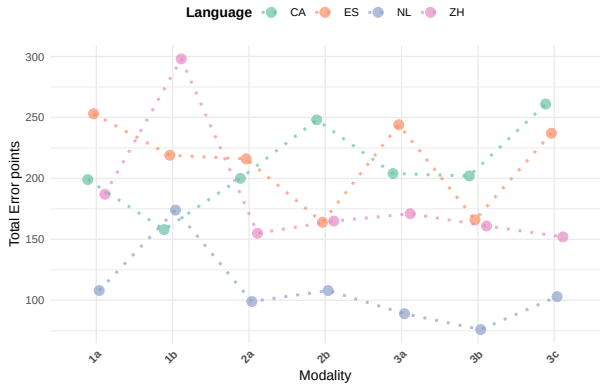

Figure 3: Total Error points per Modality and Language

Figure 2 and 3 illustrate the results already illustrated in Tables 2, 3, 4, 5 more clearly. To assess the effect of Modality and Language on CSs, an Aligned Rank Transform (ART) ANOVA was conducted for non-parametric data. Results show a significant main effect of Language, F(3, 1431) = 30.26, p = .000. However, there are no effects of Modality or the interaction of Modality and Language. Pairwise comparisons using Bonferroni correction show that CSs was significantly lower in CA than ES, NL and ZH p = .000. This is somewhat logical as the number of CSs is very low in all settings, and even lower in CA. We then assess the effect of Modality and Language on Error Points, the results show a significant effect of Modality, F(6, 1431) = 4.63, p = .000, and Modality × Language interaction, F(18, 1431) = 1.7, p = .03. The pairwise comparisons show that Error points was significantly higher in 1b when compared to 2a, p = .000, and to 3b, p = .001, and 2b was significantly higher than 3b, p = .025. This shows again that 1b and 3b were the best performing settings for these languages. The interaction analysis shows only a significant result between 3b/NL and 3c/CA.

Secondly, Figures 4 and 5 illustrate the comparison of the best performing setting with the professional translations in terms of CSs and Error points. To assess the effect of Modality and Language on CSs, we created a subset by grouping the best performing setting under the variable MT to compare it to HT. The ANOVA indicates a significant main effect of Modality (only HT and MT in this case), F(1, 265) = 31.70, p = .000, and Language, F(2, 265) = 4.26, p = .015. There was no effect of the interaction of Modality and Language. A pairwise comparison shows that CS was significantly higher in HT than in MT (p = .00). The effect of Language was only significant for CA and NL (p = .02), but not for CA and ES or ES and NL.

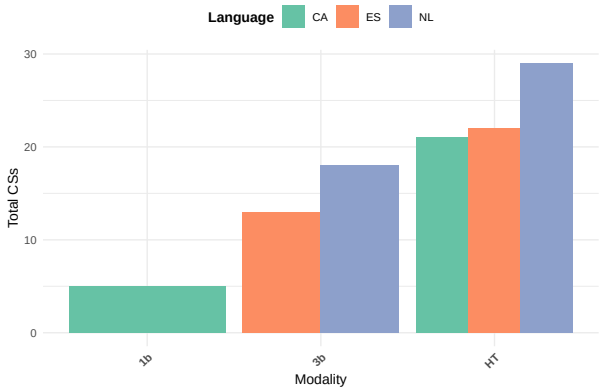

Figure 4: Total CSs per best ChatGPT Modality and HT

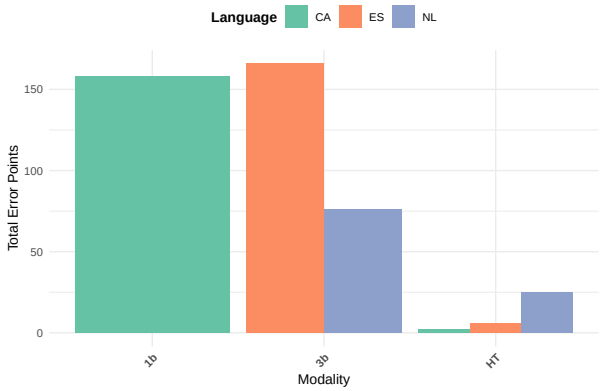

Figure 5: Total Error points per best ChatGPT Modality and HT

For Error points the results show a significant main effect of Modality, F(1, 265) = 46.27, p = .00,

Language, F(2, 265) = 5.21, p = .006, and Modality × Language interaction, F(2, 265) = 10.66, p = .00. A pairwise comparison shows that Error points was significantly lower in HT than in MT (p = .00). The effect of Language was only significant for ES and NL (p = .005), but not for CA and ES or CA and NL. When looking at the interactions, the comparison of HT/Languages and MT/Languages, there is significance in all the combinations of HT and MT (p = .00) except in the interaction between HT/NL and MT/NL.

5.1 Analysis of AEMs

Our main interest in running a set of representative AEMs (see Section 3.5), is to find out whether any of them correlates significantly with any of the metrics used in the human annotation (i.e. CSs, error points, CI). A limitation in this regard is that the number of instances[15] is very small, which is why we use a non-parametric correlation metric (Spearman). We find, as expected, significant correlations between pairs of AEMs, which occur most often between pairs of string-based metrics. Only for one language (NL) do we find significant correlations between one human metric (CSs), and two AEMs: chrF ($p < .001$) and COMET-Kiwi ($p < 0.05$). Given, again, the small sample size, and that they occur only in two cases, we refrain from drawing any strong conclusion.

We also calculated detailed scores for the TER metric. Namely, the number of operations per operation type (insertions, deletions, shifts and substitutions), system and language. The main observation is that across all languages and systems, the number of substitutions (range $[960, 1286]$) is considerably higher than the number of the other operation types put together: insertions ($[149, 290]$), deletions ($[102, 225]$) and shifts ($[95, 129]$). All the scores with AEMs are reported in Appendix B.

6 Conclusions

We wanted to explore ChatGPT MT for the best possible setting for creativity. The results show that there is indeed variability per configuration and per language. The first observation, perhaps obvious for a translator but not so obvious for others, is that creativity is seriously affected by using ChatGPT in any setting. Not only is the number of CSs in the TTs provided by all ChatGPT models (but also DeepL, GT and Softcatalà) significantly lower than in HT, but the number of errors is also significantly higher. Even the most creative setting does not come close in three out of the four languages analysed (for ZH we did not have an HT reference). Further, it is important to note that the CI for HT is not only higher but it might also not be representative of the overall creativity of the HT TTs, since we are only analysing the solutions provided by the translators to the annotated UCPs but not the entire segment where translators use other techniques, e.g. compensations, to create the desired overall effect of the text.

The second observation is that less appears to be more when prompting ChatGPT to output a creative translation. Overall the best result is the one provided by **Prompt 3**: "Translate the following text into [TG] creatively". Although this prompt still yields a very high number of errors and very modest CSs, it still outperforms the others in ES, NL and ZH while in CA even less information is needed as **Prompt 1**: "Translate the following text into [TG]" outperforms the others. These results are in line with the previous results obtained for Chinese in Du (2024).

The different prompts have somewhat similar results across different languages, with better outputs for temperature 1.0 (2b) and with Prompt 3 (3b) for ES, NL and ZH, although there were differences when providing ChatGPT with paragraphs or the whole document and between CA and the other languages. Moreover, it is interesting to see that there is a level of randomization in the output that is quite unpredictable and that requires many iterations to find the optimal solution. We wonder how this fits in a context where MT is supposed to be used to increase translator's performance. Trying these different alternatives and still obtaining a sub-optimal result does not seem the best solution for practicing translators, although it is impossible to predict if some MT suggestions might spark creativity.

As this case study is of an exploratory nature, there are limitations, notably, we selected a reduced number of UCPs that were annotated by one single annotator, with a limited number of prompts. However, the striking differences in the performance in literary translation in comparison to what is reported in the media, i.e. singularity (Translated, 2025), merits urgent attention.

[15] i.e. number of modalities per language: $n = 8$ for CA and $n = 7$ for the other three languages.

Acknowledgments

This project has received funding from the EU ERC Consolidator Grant 101086819; a Beatriz Galindo senior fellowship (BG23/00152) from the Spanish Ministry of Science and Innovation; and Grant PID2023-150711OB-I00 funded by MICIU/AEI/10.13039/501100011033 and ERDF/EU.

Sustainability statement

All in all, we submitted 2,586 API requests to ChatGPT, leading to the processing of 436,054 tokens (combining inputs and outputs). To the best of our knowledge, the average CO^2 emissions of GhatGPT models is not disclosed. A calculation by a third party estimates that each message sent to ChatGPT produces approximately 4.32g CO^2 (Wong, 2024). Asking ChatGPT we obtain the range $[2.5, 23.75]$, depending on the electricity source and assuming 50 Wh per query. Using the 4.32g CO^2 figure above, our experiments would have emitted 11.2kg CO^2.

It is also worth taking into account that we submit two rather different types of queries: paragraph- and document-based. For a paragraph-based translation we submit 125 queries, which take around 2 minutes and 50 seconds, i.e. 1.36 seconds per query. For a document-based translation only 1 query is sent, which takes around 2 minutes and 7 seconds.

References

Gerrit Bayer-Hohenwarter. 2009. *Translational creativity: how to measure the unmeasurable*, volume 37. Samfundslitteratur Copenhagen.

Gerrit Bayer-Hohenwarter. 2011. "Creative Shifts" as a Means of Measuring and Promoting Translational Creativity. *Meta Journal des traducteurs*, 56(3):663–692.

Shuxiang Du. 2024. *Optimizing Creative Translations through ChatGPT: An analysis of the Creative Potential of Machine Translation in Literary TextsCommunication and Information Studies*. Master's thesis, University of Groningen.

Margot Fonteyne, Arda Tezcan, and Lieve Macken. 2020. Literary machine translation under the magnifying glass: Assessing the quality of an nmt-translated detective novel on document level. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3790–3798.

Yuan Gao, Ruili Wang, and Feng Hou. 2024. How to design translation prompts for chatgpt: An empirical study. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops*, pages 1–7.

Ana Guerberof-Arenas and Antonio Toral. 2020. The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces*, 9(2):255–282.

Ana Guerberof-Arenas and Antonio Toral. 2022. Creativity in translation. *Translation Spaces*, 11(2):184–212.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Katikapalli Subramanyam Kalyan. 2023. A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, 6:100048.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica tecnologies de la traducció*, (12):455–463.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the*

40th Annual Meeting on Association for Computational Linguistics, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? *arXiv preprint arXiv:2405.00492*.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards Making the Most of ChatGPT for Machine Translation.

Ziqian Peng, Rachel Bawden, and François Yvon. 2024. Investigating length issues in document-level machine translation. *arXiv preprint arXiv:2412.17592*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC de Souza, Taisiya Glushkova, Duarte Alves, Luísa Coheur, et al. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 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*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Arda Tezcan, Véronique Hoste, and Lieve Macken. 2017. Scate taxonomy and corpus of machine translation errors. *Trends in E-tools and resources for translators and interpreters*, 45:219–244.

Translated. 2025. Discover How Close We Are to AI Singularity.

Kurt Vonnegut. 1999. *Bagombo Snuff Box*. Putnam Adult.

Longyue Wang, Zefeng Du, Wenxiang Jiao, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek Wong, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking and improving long-text translation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7175–7187, Bangkok, Thailand. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661.

Vinnie Wong. 2024. Gen AI's Environmental Ledger: A Closer Look at the Carbon Footprint of ChatGPT. https://piktochart.com/blog/carbon-footprint-of-chatgpt/. Accessed: 2025/02/07.

A Detailed Human Annotations

Tables 6, 7, 8 and 9 show the detailed human annotations of all languages. The best condition per language is shown in bold.

ENZH	Para	Doc	T-0.0	T-1.0	Prompt2	Prompt3	DeepL
Abstraction	1	1	1	**2**	0	0	1
Concretization	5	5	5	5	4	**8**	4
Modification	5	4	5	7	**10**	7	6
Reproduction	35	**30**	39	35	35	32	36
Omission	4	5	4	2	2	**1**	**1**
Error in UCPs	4	9	**0**	3	3	6	6
#CSs	11	10	11	14	14	**15**	11
#Errors	51	56	45	51	39	**37**	42
Error Points	187	298	155	165	171	161	**152**
Score	-10.69	-30.98	-5.38	-1.48	-2.48	**1.03**	-4.88

Table 6: Detailed human annotation - ENZH

ENNL	Para	Doc	T-0.0	T-1.0	Prompt2	Prompt3	DeepL
Abstraction	2	**5**	3	4	3	2	1
Concretization	**4**	3	2	2	3	**4**	2
Modification	9	4	6	7	4	**12**	3
Reproduction	**33**	37	40	38	42	**33**	44
Omission	0	0	0	0	0	0	1
Error in UCPs	**2**	5	3	3	**2**	3	3
#CSs	15	12	11	13	10	**18**	6
#Errors	61	68	49	51	43	**30**	51
Error Points	108	174	99	108	89	**76**	103
Score	9.84	-6.68	3.93	6.13	3.73	**20.71**	-6.00

Table 7: Detailed human annotation - ENNL

ENES	Para	Doc	T-0.0	T-1.0	Prompt2	Prompt3	DeepL
Abstraction	1	**2**	1	1	**2**	1	1
Concretization	2	2	3	2	**4**	**4**	1
Modification	3	6	4	6	6	**8**	7
Reproduction	42	43	44	42	40	**38**	**38**
Omission	1	**0**	1	1	**0**	**0**	2
Error in UCPs	5	**1**	**1**	2	2	3	5
#CSs	6	10	8	9	12	**13**	9
#Errors	59	52	55	50	57	**43**	58
Error Points	253	219	216	**164**	244	166	237
Score	-31.00	-18.00	-21.00	-11.00	-18.31	**-3.50**	-22.70

Table 8: Detailed human annotation - ENES

ENCA	Para	Doc	T-0.0	T-1.0	Prompt2	Prompt3	Softcatalà	Google Translate
Abstraction	0	0	0	0	0	0	0	**1**
Concretization	**1**	**1**	0	**1**	0	0	0	0
Modification	4	4	4	**5**	4	4	1	0
Reproduction	42	46	47	45	47	47	**39**	45
Omission	1	1	1	1	1	1	1	1
Error in UCPs	6	2	2	2	2	2	13	7
#CSs	5	5	4	**6**	4	4	1	1
#Errors	**51**	**51**	57	69	57	55	83	66
Error Points	199	**158**	200	248	204	202	393	261
Score	-23.80	**-16.99**	-25.82	-30.08	-26.48	-26.15	-63.43	-41.50

Table 9: Detailed human annotation - ENCA

System	**BLEU**			**chrF**			**TER**			**COMET**			**COMET-Kiwi**			
	ENCA	ENES	ENNL	ENCA	ENES	ENNL	ENCA	ENES	ENNL	ENCA	ENES	ENNL	ENCA	ENES	ENNL	ENZH
1a	24.8	23.0	28.9	51.9	51.1	55.3	61.4	65.0	55.8	0.781	0.7641	0.8254	0.7857	0.8033	0.8223	0.8075
1b	23.1	22.3	26.3	50.4	50.3	53.2	63.0	64.7	58.1	0.7687	0.7482	0.8101	0.7804	0.7918	0.8037	0.6098
2a	25.6	23.2	29.5	52.0	51.0	56.3	60.6	64.5	55.2	0.7744	0.7567	0.8281	0.7935	0.8000	0.8252	0.8089
2b	23.4	22.0	29.0	50.8	50.4	55.9	63.0	65.2	55.7	0.7693	0.7650	0.8248	0.7753	0.8049	0.8252	0.8093
3a	26.0	22.5	28.6	52.3	49.9	55.5	60.6	65.8	56.8	0.7736	0.7569	0.8253	0.7908	0.7983	0.8276	0.8092
3b	25.4	22.7	25.0	51.7	50.2	53.8	61.3	65.2	61.6	0.7703	0.7604	0.8238	0.7880	0.7955	0.8163	0.8017
3c	21.7	25.1	31.8	47.9	51.7	55.9	65.4	62.5	54.3	0.7158	0.7714	0.8257	0.7495	0.8078	0.8301	0.8077
3d	25.3			51.3			61.2			0.7576						0.7714

Table 10: Scores with a set of AEMs for each system and language pair.

B Scores with Automatic Evaluation Metrics

Table 10 shows the scores for each system and target language with a set of representative automatic evaluation metrics (see Section 3.5), while Figure 6, Figure 7 and Figure 8, show TER's number of operations per operation type for each system and target language.

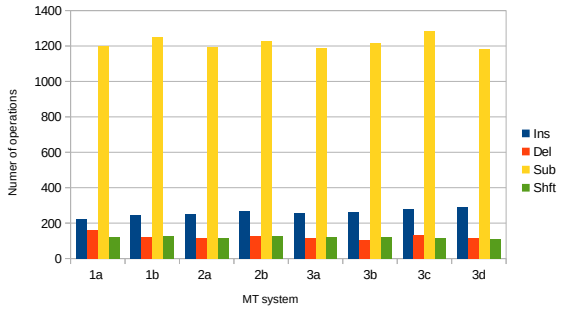

Figure 6: TER's number of operations per operation type (insertions, deletions, substitutions and shifts) for English⇒Catalan

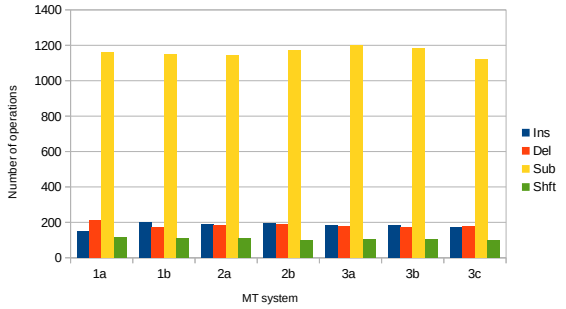

Figure 7: TER's number of operations per operation type (insertions, deletions, substitutions and shifts) for English⇒Spanish

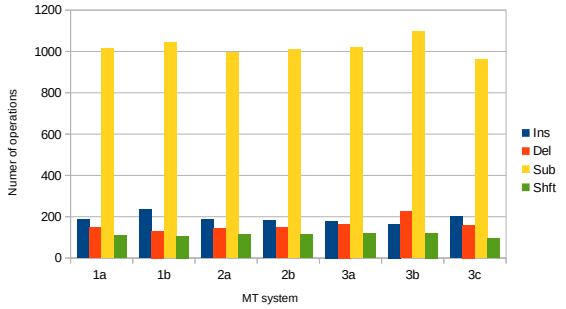

Figure 8: TER's number of operations per operation type (insertions, deletions, substitutions and shifts) for English⇒Dutch