Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3790–3798
Marseille, 11–16 May 2020

Literary Machine Translation under the Magnifying Glass: Assessing the Quality of an NMT-Translated Detective Novel on Document Level

Margot Fonteyne, Arda Tezcan, Lieve Macken
LT^3, Language and Translation Technology Team, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
{mafontey.fonteyne, arda.tezcan, lieve.macken}@ugent.be

Abstract

Several studies (covering many language pairs and translation tasks) have demonstrated that translation quality has improved enormously since the emergence of neural machine translation systems. This raises the question whether such systems are able to produce high-quality translations for more creative text types such as literature and whether they are able to generate coherent translations on document level. Our study aimed to investigate these two questions by carrying out a document-level evaluation of the raw NMT output of an entire novel. We translated Agatha Christie's novel *The Mysterious Affair at Styles* with Google's NMT system from English into Dutch and annotated it in two steps: first all fluency errors, then all accuracy errors. We report on the overall quality, determine the remaining issues, compare the most frequent error types to those in general-domain MT, and investigate whether any accuracy and fluency errors co-occur regularly. Additionally, we assess the inter-annotator agreement on the first chapter of the novel.

Keywords: literary machine translation, quality assessment, document-level evaluation

1. Introduction

It is widely accepted that with the advent of neural machine translation (NMT) translation quality has made a big leap forward. Several studies, covering many language pairs and translation tasks, using both human and automatic evaluation methods, have demonstrated that NMT systems outperform (the previous state-of-the-art) statistical MT (Bentivogli et al., 2016; Burchardt et al., 2017; Toral and Sánchez-Cartagena, 2017; Klubička et al., 2018; Van Brussel et al., 2018; Shterionov et al., 2018; Jia et al., 2019; Daems and Macken, 2019). Hassan et al. (2018) even claimed that, based on sentence-level evaluation protocols, Microsoft's NMT systems achieved human parity on the translation of Chinese to English news texts. Läubli et al. (2018), however, carried out document-level evaluations on the same data set and found that professional translators preferred human over machine translations, emphasizing the need to shift to document-level evaluations as machine translation quality improves. NMT's quality improvements have also aroused interest in the use of MT for more creative text types such as literature (Toral and Way, 2018; Kuzman et al., 2019; Matusov, 2019). This study aims to assess whether a general-domain NMT system is able to produce high-quality translations for literary translation taking into account document-level aspects of translation such as coherence and cohesion.

We evaluate the MT output of Agatha Christie's novel *The Mysterious Affair at Styles*, which was translated by Google's NMT system (GNMT) from English into Dutch. For the case study by Tezcan et al. (2019), which focuses on how the MT differs from the published human translation of the book in terms of stylistic features, the first chapter of the novel had already been annotated by one annotator using an adapted version of the SCATE error taxonomy, a hierarchical, fine-grained error taxonomy based on the well-known distinction between fluency and accuracy errors. For this study, we had the complete novel annotated by a second annotator using the same error taxonomy and report on the analysis of these annotations. We also determined the agreement between the two annotators on the annotations made in the first chapter to check the reliability of the annotations and thus the validity the annotation scheme. The error analysis shows us that 44% of the annotated sentences do not contain any errors. Within the issues that remain, the accuracy error 'mistranslation' is most frequent, which is in line with findings for general-domain MT. Fluency errors with respect to 'coherence' and 'style & register' complete the top three. These two error types were specifically added to the SCATE taxonomy to evaluate literary NMT on document level. Within this top three, coherence and mistranslation errors co-occur regularly.

In Section 2, we first highlight some of the relevant work that has been done on MT quality assessment, in particular for literary MT. We explain the reasoning behind our text selection, discuss our approach to classify and annotate errors in literary NMT, and expand on how we assessed the inter-annotator agreement (IAA) in Section 3. In Section 4, we present our results on the IAA and our analysis of the error annotations. Lastly, Section 5 comprises our conclusion and outlook on future work.

2. Related research

The quality of MT output can be judged either automatically or manually. In case of automatic evaluation the MT output is typically compared to a reference translation as is the case with metrics such as BLEU (Papineni et al., 2002) or TER (Snover et al., 2006). In case of manual evaluation human evaluators are typically asked to assess two different aspects of the MT output: fluency, which is concerned with the well-formedness of the target language, and accuracy, also referred to as adequacy, which is concerned with the correct transfer of the source content. Recently, critical comments have been made concerning the use of reference-based evaluation metrics to assess the quality of NMT sys-

tems. Shterionov et al. (2018) claimed that automatic metrics such as BLEU and F-measure tend to underestimate the quality of NMT systems and Hassan et al. (2018) found that, as translation quality has dramatically improved "automatic reference-based evaluation metrics have become increasingly problematic." Another point of criticism that was recently raised by Läubli et al. (2018) is that most evaluation methods are still sentence-based. They state that "as machine translation quality improves, translations will become harder to discriminate in terms of quality, and it may be time to shift towards document-level evaluation, which gives raters more context to understand the original text and its translation, and also exposes translation errors related to discourse phenomena which remain invisible in a sentence-level evaluation." Already in 2012, Voigt and Jurafsky pointed out that especially in the case of literary translation larger-scale textual features beyond the sentence level should be addressed. In their study they compared referential cohesion in literary texts and news texts and found that in literary texts more dense reference chains were used to create a higher level of cohesion and that MT had difficulty in conveying this referential cohesion in literary texts. In recent years, research has been carried out on the usefulness of NMT for literary translation. Toral and Way (2018) conducted a rank-based manual evaluation in order to compare human translations of literary texts with the output of a phrase-based and a neural system. They found that a considerable number of NMT sentences (between 17 and 34%) were judged to be of equivalent quality compared to the human translations. In the evaluation process the source sentences were displayed in context, but the evaluation itself was still sentence-based.

A less commonly used translation quality assessment method is error annotation and error analysis. Although time-consuming to create, fine-grained error annotations of MT output provide a rich data set that is extremely valuable to gain a better understanding of quality issues in MT. Small-scale error analyses on literary MT have been carried out by Matusov (2019) for English-Russian and German-English, and by Kuzman et al. (2019) for English-Slovene. Matusov (2019) noted that MT quality can be improved by using more contextual information especially for the translation of character names, places, as well as pronoun resolution and translation style (formal vs. informal). To annotate translation errors several error taxonomies have been proposed, ranging from coarse-grained (Vilar et al., 2006) to fine-grained categorisation schemes, such as Multidimensional Quality Metrics (MQM) (Lommel et al., 2014). In this study, we make use of an adapted version of the SCATE error taxonomy (Tezcan et al., 2019), which is very similar to MQM, but allows multiple annotations on the same text span and links both source and target words in case of accuracy errors.

3. Method

3.1. Text selection

Our data set comprises Agatha Christie's detective novel *The Mysterious Affair at Styles*, translated into Dutch by GNMT. The original English text counts 58,110 words, the MT version 58,039. Both contain 5,276 sentences. The translation was generated in May 2019. As Tezcan et al. (2019) already mentioned, the main reason for choosing this fictional text as our case study is that it was also used to compile the Ghent Eye-Tracking Corpus (GECO) (Cop et al., 2017). This corpus contains eye movement data from Dutch speakers reading both the original English text and the human translation (HT) of the novel. In future work, we want to compare the reading process of the HT to that of the MT and use the error annotation to find out which errors in the MT output have the greatest impact on it.

3.2. Error classification

The SCATE taxonomy (Tezcan et al., 2017) represents the hierarchy of errors in a maximum of three vertical levels as shown in Figure 1 (e.g. fluency → coherence → verb tense). It contains different subcategories based on the well-known distinction between fluency and accuracy errors. As fluency errors only concern the target language, the MT suffices to detect such errors. Accuracy errors, on the contrary, can only be unveiled when comparing the MT to the original text.

FLUENCY
- coherence
  - logical problem
  - non-exisiting word
  - cultural reference
  - discourse marker
  - co-reference
  - inconsistency
  - verb tense
- lexicon
  - lexical choice
  - wrong preposition
- grammar & syntax
  - agreement
  - verb form
  - word order
  - extra word(s)
  - missing word(s)
- style & register
  - disfluency
  - repetition
  - register
  - untranslated
- spelling
- other

ACCURACY
- mistranslation
  - multiword
  - word sense
  - semantically unrelated
  - part-of-speech
  - partially translated
  - other
- do not translate
- untranslated
- addition
- omission
- capitalisation & punctuation
- other

Figure 1: Overview of the extended SCATE taxonomy. On level 1, a main distinction is made between fluency and accuracy errors. The categories preceded by a black bullet are situated on level 2 in the hierarchy, and the categories preceded by an indented white bullet on level 3.

The original taxonomy has already been adapted for NMT by Van Brussel et al. (2018). They added two extra NMT-specific categories, one for unnecessarily repeated words and one for mistranslations that are semantically unrelated

to the source word. The adaptation by Van Brussel et al. (2018) has in turn been tailored by Tezcan et al. (2019) to annotate literary MT on document level. To gain insights into the quality of MT for this specific text type, they added two fluency categories to the taxonomy: 'style & register' and 'coherence'. Both comprise multiple subcategories. Moreover, as considerable improvements in quality have been observed since the arrival of NMT, the fluency category 'multiple errors' was deemed unnecessary and removed from the taxonomy. Lastly, Tezcan et al. (2019) also split the grammar category 'word form' into 'agreement' and 'verb form'. In our study, we make use of this updated version of the SCATE taxonomy.

3.3. Annotation process

For the actual annotation, the WebAnno[1] annotation tool was used. To avoid interference of the English source text when the annotator is marking the fluency errors, only the Dutch target text was visible in the first step of the annotation process (Figure 2). Both source and target texts and the annotated fluency errors were displayed in the second step (Figure 3), in which the annotator marks all the accuracy errors. Accuracy errors are labelled in both target and source text (except for omissions and additions, which are only marked in the source and target texts, respectively) and subsequently linked to enrich the data set even more. Also, if a text span contains more than one error type, it can be annotated multiple times, each time with a different label (Figure 4).

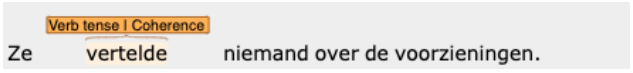

Figure 2: An example MT translation in WebAnno, after the fluency error annotation has been carried out in step 1. The annotator does not yet have access to the corresponding source text.

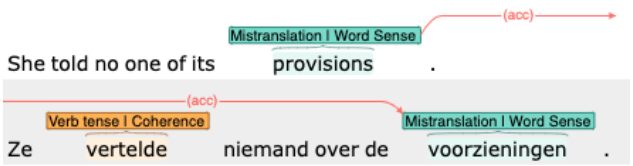

Figure 3: An example MT translation and its now visible corresponding source text in WebAnno, after the accuracy error annotation and linking has been carried out in step 2.

In total, two annotators worked independently on the data set. Both of them have a background in linguistics, one annotator has a master's degree in translation. To ensure consistency and a higher IAA, we provided annotation guidelines. The first annotator only annotated the first chapter of the novel (2,560 words), the second annotator annotated

[1]https://webanno.github.io/webanno/

Figure 4: An example MT translation in which several text spans that were annotated by the same annotator overlap.

the entire novel (58,039 words), which took approximately 159 hours. The fact that the annotators had access to entire chapters at once enabled the evaluation of the MT output on document level. They were allowed to return to earlier sections to review the annotations they had already made.

3.4. Inter-annotator agreement

Before reporting on the analysis, we assess whether the annotations scheme is appropriate for literary NMT and yields a high IAA. We can check the validity of our scheme by assessing how much the two annotators agree on their annotations in the first chapter of the novel. However, as annotating comprises three different steps (detect the errors, mark the error spans, and classify the errors within the taxonomy), it is a complex task.

Usually, the IAA is measured with Cohen's kappa coefficient (Cohen, 1960). This statistic takes into account the fact that the agreements could have arisen by chance and can be interpreted with the help of the scale by Landis and Koch (1977). However, for annotation schemes such as the SCATE approach, in which one's error annotations are allowed to overlap, it is not entirely clear how this coefficient should be calculated. The solution that Tezcan (2018) came up with, is to separately determine the agreement on and calculate kappa for error detection (the degree to which the annotators mark the same errors) and error categorization (the extent to which the annotators classify the errors in the same way). To calculate the IAA, we base our method on the one proposed by Tezcan (2018).

We first considered the problem of error detection as a binary task, deciding for each word whether it was part of an annotation span or not, and calculated Cohen's kappa at word level. Additionally, we determined the agreement on error detection at annotation level, which depends less on the agreement between the two annotators in regard to the error spans (it suffices when two annotation spans simply overlap), but for which Cohen's kappa can not be calculated (as the length of an annotation span is variable, it is impossible to count the number of unannotated instances in a text). To determine the IAA on error detection at annotation level, we collected all overlapping annotations made by the two annotators and build annotation pairs out of them. First, we paired all annotations that overlap with a single annotation from the other annotator. Then, we looked at additional criteria to pair the annotations that overlap with multiple annotations. There are two criteria that should be taken into account when pairing those annotations: their classification within the taxonomy and their text span. In the first place, we looked at the number of taxonomy levels they agree on. In the second place, we considered the

degree to which their text spans overlapped. Together with the annotations that never overlap, the originally overlapping annotations that are left after this pairing procedure form the group of isolated annotations.
To gain insight into the agreement on error categorization, we grouped the annotation pairs with exactly matching text spans (for those pairs we can be sure the annotators fully agree on their error detection) according to the number of taxonomy levels they agree on. We also counted the number of agreements and disagreements for each error category on level 1 and 2 of the error hierarchy to detect which error types the two annotators tend to disagree on. This kind of information is useful to detect confusion between categories, which can be used to revise the error taxonomy, error definitions and/or annotation guidelines in future work.

4. Results

4.1. Inter-annotator agreement

4.1.1. Error detection

	Annotator		
Words	1	2	**All**
Paired	371	371	742
Isolated	98	497	595
All	469	868	1337

Table 1: Paired and isolated annotated words per annotator.

To determine the degree to which the annotators agree on error detection, there are two options. The first option is to assess the agreement on word level. In that case, we divide the total number of paired annotated words (742), which we obtained by using the methodology detailed in Section 3.4., by the total number of annotated tokens (1337) (Table 1). This results in an observed agreement of 55.5%, which is quite low. The same goes for Cohen's kappa coefficient, which is only moderate (0.45) according to the interpretation scale by Landis and Koch (1977). There are two possible explanations for these low scores on word level. The first is that the two annotators simply do not agree in terms of which instances are considered to be errors or not. The second is that the two annotators do agree on error detection most of the time, but that they tend not to agree on the length of the error annotation span.

	Annotator		
Annotations	1	2	**All**
Paired	233	233	466
Isolated	64	109	173
All	297	342	639

Table 2: Paired and isolated annotations per annotator.

Assessing the agreement on annotation level is the second option. In that case, we divide the total number of paired annotations (466) by the total number of annotations (639) (Table 2). This results in an observed agreement of 72.9%. The fact that the agreement on annotation level is decent implies that the two annotators do agree on error detection most of the time, but that they tend not to agree on the length of the error annotation span. A further analysis on the annotated data provided additional evidence to this, as we observed that in 98% of the annotation pairs that do not have exactly matching text spans, one annotation was a subspan of the other annotation. Moreover, in most those cases (83%), the second annotator annotated a larger text span than the first. If we calculate the average annotation length of the two annotators, it becomes also clear that the annotations made by the second annotator are often longer than those made by the first. On average, the annotations of the second annotator are 0.96 tokens (61%) longer.

4.1.2. Error categorization

To assess the agreement between the two annotators on error categorization, we grouped the annotation pairs with exactly matching text spans according to the number of taxonomy levels they agree on in Table 3.

Levels	**Annotation pairs**
Total	131 (100%)
Level 1	130 (99.2%)
Level 2	118 (90.1%)
Level 3	95 (72.5%)

Table 3: Subdivision of all annotation pairs with exactly matching text spans based on the number of taxonomy levels they agree on. Both the absolute numbers and percentages in relation to the total number of annotation pairs with exactly matching text spans are given.

Table 3 shows us that, out of the 131 annotation pairs with exactly matching text spans, 95 pairs (72.5%) agree on all taxonomy levels, 118 pairs (90.1%) at least on level 2, and 130 pairs (99.2%) at least on level 1. To detect whether the observed disagreement is caused by confusion between any specific categories, we analyze the disagreements on error categorization on the top two levels of the taxonomy.

	Fluency	Accuracy
Fluency	67	1
Accuracy	0	63

Table 4: Error categorization matrix on level 1 for all annotation pairs. Error categories in the first column represent the annotations from annotator 1, those in the first row the annotations from annotator 2.

As can be seen in Table 4, there is almost no disagreement between the two annotators on the main distinction between fluency and accuracy. On level 2 (Table 5), the most pronounced disagreements can be observed between the following error categories: 'coherence' vs. 'style & register' (total of 4.6%), 'lexicon' vs. 'style & register' (total of 1.5%), and 'mistranslation' vs. 'untranslated' (total of 1.5%). Zooming in on level 3 of the annotation pairs

	M	C	S	L	U	G	A
M	52	0	0	0	0	0	0
C	0	36	3	1	0	1	0
S	0	3	12	0	0	0	0
L	0	1	0	5	0	0	0
U	2	0	0	0	5	0	0
G	0	0	1	0	0	4	0
A	0	0	0	0	0	0	4

M — Mistranslation, C — Coherence, S — Style & Register, L — Lexicon, U — Untranslated, G — Grammar, A — Addition

Table 5: Error categorization matrix on level 2 for annotation pairs agreeing on level 1. Error categories in the first column represent the annotations from annotator 1, those in the first row the annotations from annotator 2. Level 2 categories that do not occur in this data set were not included in this table due to size constraints.

disagreeing in 'coherence' vs. 'style & register', shows us that 4 out of 6 pairs are a combination of 'coherence – logical problem' and 'style & register – disfluent sentence/construction'.

4.2. Error analysis

4.2.1. Overall quality

A quick glimpse into the number of sentences with and without errors gives us an indication of the overall quality of the MT translation. 2,316 of the 5,276 sentences (43.9%) in our data set do not contain any errors. This percentage is equal to the one in the case study on the first chapter of the novel (Tezcan et al., 2019). It is higher than the percentages found in studies on literary GNMT for other language pairs. Kuzman et al. (2019) worked with an 929-word excerpt taken from the romance novel *Something about you* by Julie James, translated by GNMT from English into Slovene. They point out that none of the sentences they analyzed were free of errors and that all of them would require post-editing. Matusov (2019) carried out a document-level error analysis of an excerpt taken from the story *The Lift* by Arthur Conan Doyle, translated by GNMT from English into Russian. He found that only 22 out of the 129 analyzed segments (17.1%) did not contain any errors. He also reports on the document-level analysis of an excerpt from *Die Verwandlung* by Franz Kafka, translated by GNMT from German into English. In this case, 36 out of the 125 analyzed segments (28.8%) were considered as acceptable. Thus, NMT for the translation of literary texts seems to be more promising for English into Dutch. Yet, it should be pointed out that Kuzman et al. (2019) and Matusov (2019) worked with differently sized data sets consisting of excerpts written by different authors.

Table 6 shows that, on average, the length of erroneous source sentences is longer than the length of those without errors, but also that there is more variation in length for the sentences with errors. As can be seen in Figure 5, performance decreases nonetheless from a sentence length of 10 words onward.

	With errors	**Without errors**
Mean	17.11	9.56
Standard deviation	10.09	5.55

Table 6: Mean and standard deviation of the source sentence length of sentences with and without errors.

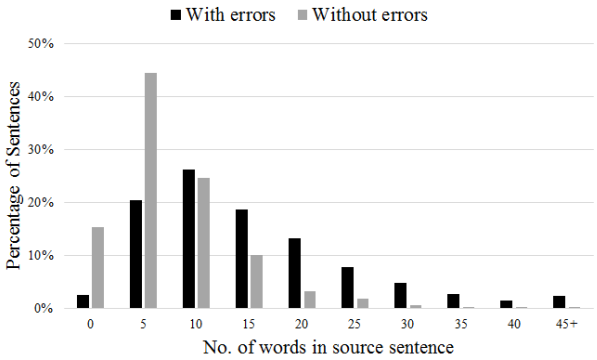

Figure 5: Distribution of sentences with and without errors per source sentence length.

4.2.2. Error distribution

To reveal the main issues that remain for literary NMT, we inspect the frequency distribution of the different error types. In total, the second annotator indicated 3,980 fluency errors (58.8%) and 2,784 accuracy errors (41.2%). These results are consistent with Van Brussel et al. (2018), who found that NMT for English-Dutch contains more fluency than accuracy errors.

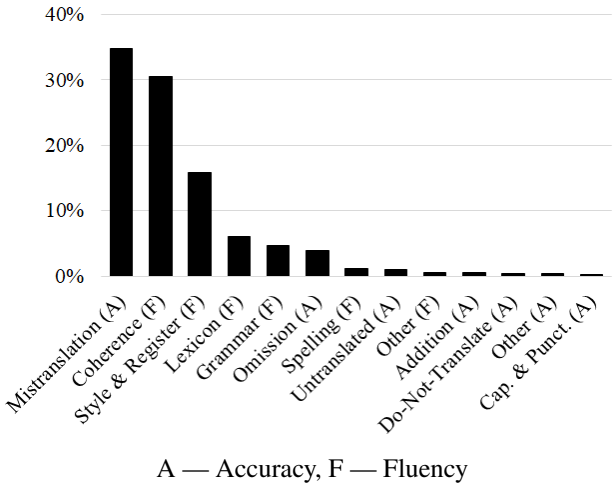

Figure 6: Frequency of error types expressed as percentage of all errors.

The distribution of the errors annotated across the entire novel by the second annotator (Figure 6) is very much in line with the one in the case study on the first chapter of the novel (Tezcan et al., 2019), which was based on the annotations by the first annotator. The most common error

type in the data set is 'mistranslation' (34.7%).

Most mistranslation issues relate to word sense issues (38.3%), issues without a specific subcategory (33.5%), and multi-word expressions (16.6%). These findings correspond with those by Kuzman et al. (2019) for English-to-Slovene and Matusov (2019) for English-to-Russian. Both report that a large number of semantic errors in connection with idioms ('multi-word expression') and ambiguous words ('word sense') were found.

The second most common error type is 'coherence' (30.5%). Within the data set of fluency errors, 'coherence' errors are in the majority (51.8%). The main 'coherence' issue appears to be the category 'logical problem' (82.8%). Second comes the category 'verb tense' (7.3%). Style and register issues, the third most common error type, consist mainly of disfluent sentences and constructions (89.5%).

4.2.3. Co-occurring fluency and accuracy errors

In this section, we focus on the relationship between fluency and accuracy errors. If certain accuracy errors tend to cause certain fluency errors, we can hypothesize that those particular fluency and accuracy error categories will co-occur regularly. To test this hypothesis, we examine the annotations from the second annotator, who annotated the whole document, and measure how often certain fluency and accuracy errors co-occur, on level 2 and level 3.

For this analysis we only considered the annotations with identically matching spans. Out of the 6,764 error annotations that the second annotator made, 1,089 annotation pairs consisting of a fluency and an accuracy error can be made. This means that 27.4% of the fluency and 39.1% of the accuracy error spans fully overlap with the span of an error belonging to the other category.

Fluency	Accuracy	Overlaps
Coherence	Mistranslation	919
Style & Register	Untranslated	50
Lexicon	Mistranslation	43
Other	Other	21
Coherence	Do-Not-Translate	20

Table 7: Most common co-occurrences on level 2.

Table 7 shows that the vast majority of co-occurrences on level 2 are a combination of a coherence and a mistranslation error. No more than 919 annotation pairs (84.4% of all co-occurring accuracy and fluency annotations) belong to this category. All other co-occurrence types make up for less than 5%. When inspecting the five most common co-occurrences on level 3, 4 out of 5 co-occurrence types consist of a 'coherence – logical problem' error with some kind of mistranslation error. An example of such a co-occurrence is given in Figure 7.

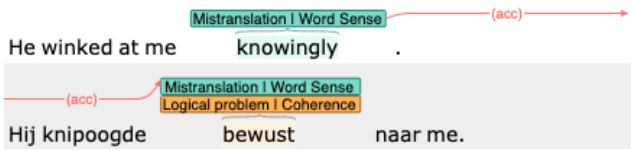

Figure 7: Co-occurrence of a logical problem and word sense issue

In this example, the adverb 'knowingly' has been translated to 'bewust' ('consciously'). However, although 'bewust' is indeed one of the possible translations for 'knowingly', in this context the adverb should have been translated into 'veelbetekenend' ('meaningfully'). Such errors are classified as 'word sense' errors. As the rest of the text does not give a reason for the character to wink consciously, understanding the target sentence within the given context becomes difficult. In other words, a coherence issue (or to be more precise, a logical problem) arises and disturbs the reading comprehension of the machine translated novel.

4.2.4. Error comparison with general-domain MT

To find out how literary NMT differs from NMT for a more general domain, we compare our error distribution with the distribution in Van Brussel et al. (2018). The data set Van Brussel et al. (2018) used consists of 665 GNMT-translated sentences belonging to three different text types: external communication, non-fiction literature and journalistic texts. It is important to stress that the MT output generated by Van Brussel et al. (2018) is two years older than the MT output in our data set, which means that some differences could also be due to the fact that GNMT has improved over time.

Since our adapted taxonomy does not differ from the SCATE taxonomy used by Van Brussel et al. (2018) in regard to accuracy errors (with the exception of the addition of the category 'other'), we can easily compare each subcategory in this main category.

Error type	Literature	General
Addition	1.3%	0.4%
Cap. & Punct.	0.4%	2.5%
Do-Not-Translate	1.0%	4.7%
Mistranslation	85.1%	69.9%
Omission	9.8%	13.1%
Untranslated	2.4%	9.3%

Table 8: Accuracy error distributions in literary and general-domain MT.

A comparison of the two data sets shows us that mistranslations are without a doubt the biggest accuracy issue for both literary and general-domain MT (Table 8). By carrying out our error annotation on document level, it should have also been easier to spot these mistranslations. Läubli et al. (2018) argue for instance that "document-level evaluation unveils errors such as mistranslation of an ambiguous

word, or errors related to textual cohesion and coherence, which remain hard or impossible to spot in a sentence-level evaluation.". This argumentation is based on the fact that they observed the phenomena above where sentence-level evaluation led to mixed judgements, but where the HT was strongly preferred in document-level evaluation.

Error type	Literature	General
Multi-word Expression	16.6%	26.4%
Other	33.5%	17.3%
Semantically Unrelated	8.0%	13.3%
Part-Of-Speech	2.9%	5.8%
Partial	0.6%	1.8%
Word Sense	38.3%	35.5%

Table 9: Mistranslation error distributions in literary and general-domain MT.

Table 9 presents the percentages of the different mistranslation subcategories. For both text genres, 'word sense' errors are the biggest issue within the category 'mistranslation'. The subcategories 'MWE' and 'other', however, have switched places within the top three of most common mistranslation issues.

A comparison between our results and those of Van Brussel et al. (2018) in regard to fluency errors is more difficult, because the taxonomy was adapted considerably within this category. For example, existing subcategories were moved into the newly added categories 'coherence' and 'style & register'. Yet, it is possible to compare the ranking of the error types within the accuracy subcategory 'grammar', as the error types within this subcategory have largely remained the same (Section 3.2.).

Error type	Literature	General
Word Form	20.3%	31.9%
Extra Word	9.2%	13.5%
Missing Word	28.9%	36.2%
Word Order	41.6%	18.3%

Table 10: Grammar error distribution in literary and general-domain MT. In the literature-domain evaluation 'agreement' and 'verb form' have been merged into 'word form' for a proper comparison with the general-domain evaluation.

As can be seen in Table 10, the most common 'grammar' subcategory in NMT-translated literary texts is 'word order', followed by 'missing word', 'word form', and, finally, 'extra word'. Compared to the ranking in Van Brussel et al. (2018), the subcategory 'word order' has jumped from the third place to the top.

5. Conclusions and future work

In this case study, we aimed to assess the potential of using a general-domain NMT system for literary translation. We did this by conducting a fine-grained document-level error analysis on an entire novel, *The Mysterious Affair at Styles* by Agatha Christie, which was translated by GNMT from English into Dutch. The SCATE MT error taxonomy, as tailored by Tezcan et al. (2019) for document-level evaluation of literary NMT, was used to perform the error annotation. The annotated data set has been made publicly available.[2]

We first assessed the validity of our annotation scheme by determining the IAA on the first chapter on the novel, which was annotated independently by two annotators. The two annotators usually agreed with respect to error detection and error categorization, yet it has to be noted that one of the annotators often annotated longer text spans than the other. Since we carried out a fine-grained error analysis, we were also able to detect the categories on which the annotators sometimes disagreed, such as 'coherence' vs. 'style & register', 'lexicon' vs. 'style & register', and 'mistranslation' vs. 'untranslated'. Especially 'coherence – logical problem' and 'style & register – disfluent sentence/construction' were often confused. This might be due to a cause-effect relationship between these two error types: a sentence that is very hard to read (i.e. disfluent), often ends up being illogical or confusing (i.e. results in a logical problem).

In regard to the potential of GNMT for literary translation from English into Dutch, the fine-grained error analysis on document level has helped us gain a better understanding of:

- The overall quality of the NMT translation. 44% of the sentences do not contain any errors. The percentage of correct sentences for the language pairs English-Slovene, English-Russian, and German-English was found to be remarkably lower. Thus, compared to these three language pairs, English-Dutch seems to be the more promising language pair for literary MT.
- The main remaining issues. The most frequent issue for literary MT is the accuracy subcategory 'mistranslation'. Next are the fluency subcategories 'coherence' and 'style & register', which were specifically added to the taxonomy for evaluating literary NMT on document level.
- Which accuracy errors typically co-occur with fluency errors in literary MT. We observed on level 2 of the taxonomy that the vast majority of co-occurrences are a combination of a coherence and a mistranslation error. It is a question of future research to investigate whether this type of co-occurrence is characteristic for literary MT or not.
- The similarities and differences with general-domain MT. The adjusted taxonomy and the fact that we used an NMT system of a later date made the comparison difficult. Nonetheless, the comparison still revealed that mistranslations are the biggest accuracy issue for general-domain MT as well.

In future work, we aim to determine the generalizability of this case study by annotating and then analyzing the first

[2]https://github.com/margotfonteyne/StylesNMT

chapters of some other literary works by different authors. We will also study the similarities and differences with general-domain MT in more detail. Additionally, we would like to explore the potential of NMT for literary translation in more detail with the help of an even richer data set. We plan to enrich our data set by annotating and linking the cohesive and stylistic devices throughout and across all versions (the source, the HT, and the MT). This would allow us to compare the different versions on document level in regard to cohesion (both local and global) and style.

We will also expand the Ghent Eye-Tracking Corpus (GECO) with eye-tracking data for the GNMT version of the novel. This will allow us to compare the reading process of MT with that of manually translated text. The error annotation will enable us to investigate the impact of different categories of MT errors on the underlying reading comprehension process.

6. Acknowledgements

This study is part of the ArisToCAT project (Assessing The Comprehensibility of Automatic Translations), which is a four-year research project (2017-2020) funded by the Research Foundation – Flanders (FWO) – grant number G.0064.17N.

7. Bibliographical References

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas, November. Association for Computational Linguistics.

Burchardt, A., Macketanz, V., Dehdari, J., Heigold, G., Peter, J.-T., and Williams, P. (2017). A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. *The Prague Bulletin of Mathematical Linguistics*, 108:159–170, June.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Cop, U., Dirix, N., Drieghe, D., and Duyck, W. (2017). Presenting geco : an eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2):602–615.

Daems, J. and Macken, L. (2019). Interactive adaptive SMT versus interactive adaptive NMT: a user experience evaluation. *Machine Translation*, 33(1):117–134, Jun.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (2018). Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Jia, Y., Carl, M., and Wang, X. (2019). Post-editing neural machine translation versus phrase-based machine translation for English–Chinese. *Machine Translation*, 33(1):9–29, Jun.

Klubička, F., Toral, A., and Sánchez-Cartagena, V. M. (2018). Quantitative fine-grained human evaluation of machine translation systems: a case study on English to Croatian. *Machine Translation*, 32(3):195–215, Sep.

Kuzman, T., Špela Vintar, and Arčan, M. (2019). Neural machine translation of literary texts from English to Slovene. In *Proceedings of the Qualities of Literary Machine Translation*, pages 1–9, Dublin, Ireland, 19 August. European Association for Machine Translation.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Läubli, S., Sennrich, R., and Volk, M. (2018). Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium, October-November. Association for Computational Linguistics.

Lommel, A. R., Burchardt, A., and Uszkoreit, H. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0(12):455–463.

Matusov, E. (2019). The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland, 19 August. European Association for Machine Translation.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shterionov, D., Superbo, R., Nagle, P., Casanellas, L., O'Dowd, T., and Way, A. (2018). Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation*, 32(3):217–235, May.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*.

Tezcan, A., Hoste, V., and Macken, L. (2017). Scate taxonomy and corpus of machine translation errors. In Gloria Corpas Pastor et al., editors, *Trends in E-tools and resources for translators and interpreters*, volume 45 of *Approaches to Translation Studies*, pages 219–244. Brill — Rodopi.

Tezcan, A., Daems, J., and Macken, L. (2019). When a 'sport' is a person and other issues for NMT of novels. In *Proceedings of the Qualities of Literary Machine Translation*, pages 40–49, Dublin, Ireland, 19 August. European Association for Machine Translation.

Tezcan, A. (2018). *Informative quality estimation of machine translation output*. Ph.D. thesis, Ghent University.

Toral, A. and Sánchez-Cartagena, V. M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Vol-*

ume 1, Long Papers, pages 1063–1073, Valencia, Spain, April. Association for Computational Linguistics.

Toral, A. and Way, A. (2018). What level of quality can neural machine translation attain on literary text? In Joss Moorkens, et al., editors, *Translation Quality Assessment*, Machine Translation: Technologies and Applications, pages 263–287. Springer International Publishing AG.

Van Brussel, L., Tezcan, A., and Macken, L. (2018). A fine-grained error analysis of NMT, SMT and RBMT output for English-to-Dutch. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3799–3804, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Vilar, D., Xu, J., D'haro, L., and Ney, H. (2006). Error analysis of statistical machine translation output. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy. European Language Resources Association (ELRA).

Voigt, R. and Jurafsky, D. (2012). Towards a literary machine translation: The role of referential cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 18–25, Montréal, Canada, June. Association for Computational Linguistics.