When a 'sport' is a person and other issues for NMT of novels

Arda Tezcan, Joke Daems, Lieve Macken
LT^3, Language and Translation Technology Team
Ghent University
Belgium
firstname.lastname@ugent.be

Abstract

We report on a case study in which we assess the quality of Google's Neural Machine Translation system on the translation of Agatha Christie's novel *The Mysterious Affair at Styles* into Dutch. We annotated and classified all MT errors in the first chapter of the novel making use of the SCATE error taxonomy, which differentiates between fluency (well-formedness of the target language) and accuracy errors (correct transfer of source content). We modified the SCATE MT error taxonomy to be able to annotate text-level phenomena such as textual coherence (e.g. anaphora and coreference) and textual cohesion (e.g. lexical consistency) and literature-specific issues such as cultural references. Apart from annotating the errors in the MT output, we investigate how the machine translated version differs from the published human translated Dutch version of the book. We look at stylistic features such as lexical richness, cohesion, and syntactic equivalence.

1 Introduction

In literary translation, unlike in most other types of translation, the goal is not just to offer an adequate translation that preserves the meaning of the original, but rather to offer the reader a comparable reading experience (Toral and Way, 2015b). What makes this particularly difficult is the presence of cultural references (Besacier and Schwartz, 2015), the fact that literary texts are lexically richer than other texts (de Camargo, 2004) and the frequent use of idiomatic expressions. While, intuitively, these aspects make literary texts poor candidates for machine translation (MT), researchers have looked into the use of statistical MT (SMT) and, more recently, neural MT (NMT) for literary translation and found it to have a potential use. Still, as the research in this field is limited, "a thorough investigation of [MT's] utility in this space [...], both from the point of qualitative and quantitative evaluation" (Toral and Way, 2015b) is needed. Our goal with this study is to get a better understanding of raw NMT quality for literary translation by comparing the Dutch NMT translation of an English novel with its original Dutch translation. To the best of our knowledge, it is the first study into the usability of generic NMT for Dutch literary translation. We place particular emphasis on features that might impact the reading experience. In the following sections, we first highlight some of the relevant work that has been done on SMT and NMT for literary translation, we then discuss how we adapted the SCATE MT quality assessment approach to cover coherence issues relevant for the study of literary translation, followed by our analysis of the raw MT quality and a comparative analysis of key features (lexical richness, cohesion, and syntactic equivalence) between MT and the original human translation (HT).

2 Related research

Voigt and Jurafsky (2012) were some of the first to question whether statistical MT at the time was sufficiently developed to start thinking about using it for the translation of literary works, looking at Chinese to English translations. They were particularly interested in literary cohesion, and found lit-

© 2019 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

erary texts to contain more dense reference chains (a higher number of mentions per entity) than non-literary texts. More importantly, they discovered that, while human translators manage to maintain this density, MT does not capture literary cohesion as well. If we are to apply MT to literature, they argue, we should think beyond the sentence level and incorporate discourse features in our analysis. Rather than looking at MT quality as such, Besacier and Schwartz (2015) studied the potential use of SMT for the post-editing of a literary text (an essay by Richard Powers from English into French). They found the process to be faster than manual translation would have been, and a group of readers found the product to be of acceptable quality. Still, Powers' official French translator found the post-edited product to be lacking in a few specific ways, such as source language structure being preserved in the target text, and cultural references or idiomatic language not being taken into account. Toral and Way (2015a; 2015b) looked into MT quality for literary translation by building a literature-specific MT systems for Spanish into Catalan, and French into English and Italian. Interestingly, they found that MT translation quality was comparable to human quality for 60% of the sentences (2015a), although they did work on closely related languages (Catalan and Spanish). Some of the main issues in the MT output were lexical choice, verbal tense, particles, and (gender) agreement (2015b).

MT quality improved even more in 2016, with the arrival of neural machine translation (NMT). Toral and Way (2018) argue that its increased quality (Junczys-Dowmunt, Dwojak, and Hoang, 2016) and the fact that NMT can handle lexically rich texts (Bentivogli et al., 2016) make it better suited for literary translation than SMT systems. By training an NMT and SMT system on literary texts and comparing the output, they indeed found that NMT quality outperformed SMT quality. Up to 34% of the NMT sentences were perceived to be of equal quality to human translations (compared to 20% for SMT). Professional translators, however, still preferred human translation over post-editing for literary texts (Moorkens et al., 2018), listing the following as the main limitations of MT: in literary translation, it is important to preserve the reading experience and in particular context is important, while MT has a fragmented view working on a sentence level; though NMT translates less literal than SMT, it is still not good with certain vocabulary and uses the wrong level of politeness; figurative language and cultural items remain difficult for both MT paradigms.

3 Method

3.1 Text selection

We use the Dutch MT translation of The Mysterious Affair at Styles, a 56000-word detective novel by Agatha Christie, as a case study. This book was specifically chosen as it is also the book used in the Ghent Eye-Tracking Corpus (GECO) (Cop et al., 2017), which contains eye movement data from Dutch speakers reading the human-translated version. As such, it offers a great reference for the reading process of manually translated literary text, which, in the future, can be compared to the reading process of MT. As the goal of literary translation is to preserve the reading experience, this will give us a way to establish which features in MT output (as discovered in this case study) have the greatest impact on said reading experience. In addition to this pragmatic choice, the novel contains key stylistic elements common to other literary works, which have been found to be potentially problematic for MT, such as the use of idioms, incomplete sentences in dialogue, and fragments in different languages, making our findings likely transferable to other literary works. The MT was generated by Google Translate (NMT), a freely available neural MT translation system, in May 2019.

3.2 Translation quality annotation process

To get an idea of the quality of neural MT for literary translation of English into Dutch, we first adapted the SCATE taxonomy (Tezcan et al., 2017) for literary translation, then annotated the first chapter of *The Mysterious Affair at Styles*. The SCATE taxonomy was selected because it was specifically developed to annotate MT output and it studies two distinct aspects of MT quality: fluency and accuracy. Fluency relates to all errors that can be spotted when looking at the target text only, such as grammar, lexicon, and orthography. The second aspect is accuracy, where source and target text are compared to discover potential issues such as omissions, additions, and mistranslations. As coherence was found to be such a crucial aspect of literary MT translation evaluation (Voigt and Jurafsky, 2012; Moorkens et al., 2018), we

added a category ‘coherence’ for fluency. Subcategories were ‘logical problem’, if information made no sense when looking at the rest of the text, ‘non-existing words’ for words that did not exist in Dutch and as such made no sense, ‘discourse marker’, where a linking word expressed a strange relationship, ‘co-reference’, when there was a mismatch between entities that was not grammatically incorrect in the sentence itself, for example, a feminine pronoun referring to a male person mentioned in a previous sentence, ‘inconsistency’, when a term or notation was used inconsistently throughout the text, and ‘verb tense’, where the tense was grammatically correct, but it was illogical or wrong when compared to the rest of the sentence or surrounding sentences. In addition to coherence, we added a category for ‘style & register’, which consisted of the subcategories ‘disfluency’, for fragments or sentences that , though grammatically correct, were difficult to read or not quite idiomatic, ‘repetition’, when the same or a similar word is used more than once in a sentence, ‘register’, when the register (formal/informal) or regional variety did not match the target audience, and ‘untranslated’, where an English word for which a Dutch translation exists was left untranslated. An overview of the extended SCATE taxonomy can be seen in Figure 1.

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.

We annotated the first chapter of the novel according to this extended SCATE error taxonomy using the web-based annotation tool WebAnno[1] (Yimam et al., 2013). The chapter is 351 sentences and 4358 words long, with an average sentence length of 11.5 words. Annotation was performed by one of the authors, who has over twenty years of experience in translation technology and translation quality evaluation. Fluency and accuracy were annotated in two distinct steps. For fluency, the annotator only had access to the target text, for adequacy, the annotator could compare source and target text. It is therefore possible for more than one annotation to be attached to the same word or phrase.

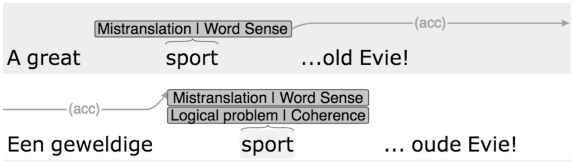

Figure 2: Annotation example.

An example of a double annotation can be seen in Figure 2. The English word ‘sport’ (in this context, a person), was translated in Dutch as ‘sport’ (an actual sport). From a fluency perspective, this is a logical problem, as the reader has no way of understanding why the word ‘sport’ would appear in this sentence. From an adequacy perspective, however, this word is a mistranslation of the type ‘word sense’, as the wrong sense of the word ‘sport’ was used here.

3.3 Textual feature analysis

In addition to the quality annotation of the first chapter, we compared some key textual features between MT and the original human translation of the novel. For these analyses, the entire novel was used.

As NMT is said to be able to handle lexically rich texts better than SMT (Bentivogli et al., 2016), we wanted to get an idea of the lexical richness of the novel and compare how well NMT manages to capture this richness as opposed to the original human translation. To do so, we look at the word frequency distribution, lexical density, and translation entropy.

To calculate lexical density, we used a variety of type-token ratio measures. The idea is that the more types there are in comparison to the number of tokens, the greater the lexical variety in a

[1] Version 3.4.5.

text. The following standard measures were used, where *t* is the number of types, and *n* is the number of tokens:

TTR (type-token ratio):

$$TTR = t/n \quad (1)$$

RTTR (root type-token ratio):

$$RTTR = t/\sqrt{n} \quad (2)$$

CTTR (corrected type-token ratio):

$$CTTR = t/\sqrt{(2n)} \quad (3)$$

A possible critique of the above formulas is that standard TTR-measures are sensitive to text length (Torruella and Capsada, 2013). We therefore also calculated Mass index and the mean segmental type-token ratio (MSTTR) as follows:

Mass index:

$$MASS = (\log(n) - \log(t))/\log^2(n) \quad (4)$$

MSTTR (Johnson, 1944): The text to be analysed is divided into equal segments of 100 words. MSTTR is calculated as the arithmetic mean of the TTR values for each segment.

Word translation entropy indicates the degree of uncertainty to choose a correct translation from a set of target words $t_i...t_n$, for a given source word s. If the probabilities are distributed equally over a large number of items, the word translation entropy is high and there is a large degree of uncertainty regarding the outcome of the translation process. If, however, the probability distribution falls unto just one or a few items, entropy is low and the certainty of target words to be chosen is high (Schaeffer et al., 2016).

Word translation entropy has often been analyzed as an indicator of cognitive effort in the context of human translation, by collecting translations for a given sentence from multiple translators (Carl et al., 2017; Vanroy et al., 2019a). In this study, however, we use it to measure average word translation entropy (AWTE) on document level, by making the calculation using all the words that appear in the source text and its translated versions, both automatically and manually. After calculating word translation entropy for each document pair (source-HT and source-MT), we take the arithmetic average of all entropy values to obtain AWTE.

For each unique source word s in the given source text, word translation entropy is defined as the sum over all observed word translation probabilities into target text words $t_i...t_n$, multiplied with their information content (Carl et al., 2017). For each source word s translation entropy is calculated as follows:

$$E(s) = \sum_{i=1}^{n} p(s \rightarrow t_i) * I(p(s \rightarrow t_i)) \quad (5)$$

where $p(s \rightarrow t_i)$ stands for the word translation probabilities of a source word s and its possible translations $t_i...tn$, which is calculated as the number of alignments $s \rightarrow t_i$ divided by the total number of observed translations $t_i...t_n$:

$$p(s \rightarrow t_i) = count(s \rightarrow t_i)/translations \quad (6)$$

The information I that is present in a distribution with equal probability of an event p can be formulated as in Equation (7).

$$I(p) = -\log_2(p) \quad (7)$$

While the probability p expresses the expectation for an event, the information I indicates the minimum amount of bits with which this expectation can be encoded.

In order to obtain translation options and calculate word translation probabilities, we used a freely available implementation of IBM models GIZA++ (Och and Ney, 2003) on source-HT and source-MT sentence pairs, respectively. While IBM-style models dominate the field of statistical word-alignment, they are also prone to overfitting the data and often propose many incorrect word alignments for rare words, a phenomenon called *garbage collection* (Moore, 2004). Furthermore, we can expect additional word alignment errors when this technique is used to align words between a source text and its machine translated version, which potentially contains translation errors. In order to be more confident about the differences between the AWTE values for HT and MT, we repeat the calculations by increasing the minimum frequency threshold for the set of source words we take into consideration. While a minimum threshold frequency of 1 covers all the source words in the source text (as each word occurs at least once), a threshold of n calculates AWTE only for the subset of source words that appear at least n times.

A key aspect of literary translation is the importance of cohesion (Voigt and Jurafsky, 2012) and looking beyond the sentence level (Moorkens et

al., 2018). While the error annotation already covers cohesion, that analysis was limited to the first chapter and it is also very time-consuming. Inspired by previous work on local cohesion indices (McNamara et al., 2002; Crossley et al., 2016), we therefore measure local cohesion in terms of lexical and semantic overlap between a given sentence and the succeeding sentence(s) (up to two sentences). According to Crossley et al. (2016), looking at the lexical overlap between a sentence and the upcoming two sentences is a "significant indicator of perceived human text organization". While lexical overlap is measured by comparing lemmas of content words (nouns, verbs, adjectives and adverbs), semantic overlap uses WordNets in the NLTK package[2] and further compares the shared synsets (sets of cognitive synonyms each expressing a distinct concept) of content words. We report both the number of sentences that overlap with succeeding sentence(s) (with at least one overlapping lemma) and the total number of overlapping lemmas, summed over all sentences[3].

A final feature we studied was syntactic equivalence. One of the issues Besacier & Schwartz (2015) discovered for SMT was the fact that it followed the syntactic structure of the source text too closely. As we can expect NMT to translate less literal (Moorkens et al., 2018) and to lead to fewer word order issues than SMT (Bentivogli et al., 2016), the question is whether the syntactic structures found in the NMT output still closely resemble the source text structures or not.

As proposed by Vanroy et al. (2019b), we calculate syntactic equivalence between a source sentence and its translation in terms of their *cross value*, the number of times word-alignment links cross each other, averaged by the number of alignment links. Similar to Vanroy et al. (2019b), we calculate cross values in two ways: by looking at how (1) each individual word moves with respect to other words in the sentence, and (2) sequential words move together as a group. The second approach seeks the longest possible word sequence alignments between the source and target sentences with the following criteria:

- each word in the source sequence is aligned to at least one word in the target sequence and vice versa,
- each word in the source word sequence is only aligned to word(s) in the target word sequence and vice versa,
- none of the alignments between the source and target word sequences cross each other.

For both methods, we use GIZA++ to obtain word alignments between the source and target sentences automatically.

Vanroy et al. (2019b) argue that, cross value based on sequence alignments is a better representation of the clashing syntactic shifts that a source sentence has to go through to become the target sentence, as it indicates crossing groups of words rather than single entities. These two approaches are illustrated in Figure 3 for a source sentence in English and its translation in Dutch.

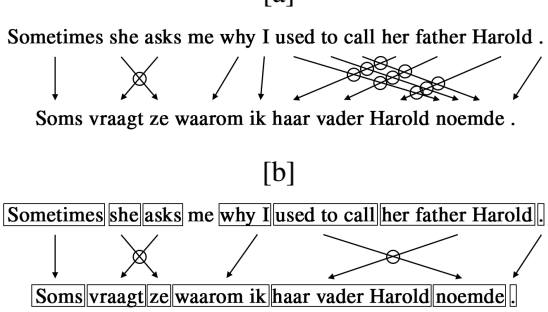

Figure 3: A visual representation of (a) word and (b) sequence alignments, and *crosses*, indicated by circles.

In these examples, each arrow indicates an alignment link between a source and target word or a word sequence. Please note that the source word "*me*" is not aligned to a target word in this example. In Figure 3a, we count ten *crosses*. This value is then averaged by the number of alignments to get average cross value of the whole sentence. In this case that is $^{10}/_{12} = 0.8\overline{33}$. In Figure 3b, the cross value based on sequences is calculated as $^{2}/_{7} = 0.286$.

4 Results

4.1 Quality

Looking at the NMT quality for the first chapter of the novel, we see that 44% of the sentences did not contain any errors. This is interesting in a number of ways. Firstly, earlier work comparing NMT to SMT and RBMT for English-Dutch general texts (newspaper articles and nonfiction) found that 33% of NMT sentences contained no errors (Van Brussel et al., 2018), which

[2] https://www.nltk.org/

[3] In a given sentence, each lemma is checked for overlap only once.

is lower than the 44% found here. Secondly, Toral & Way (2018) built a custom NMT system tailored to literary translation and found that up to 34% of sentences was perceived by native speakers as being of equal quality to a human translation, which is again lower. While 'not containing any errors' is in no way equal to 'comparable to human quality', this already gives some indication of the potential of NMT for the translation of literary texts from English into Dutch. As can be expected and as can be seen in Figure 4, performance decreases with sentence length. Most of the sentences without errors were shorter than 15 words. The maximum length for a sentence without errors was 37 words, which seems to align with findings that NMT quality decreases with sentence length (Bentivogli et al., 2016), to the extent that it might be outperformed by SMT for sentences longer than 40 words (Toral and Snchez-Cartagena, 2017).

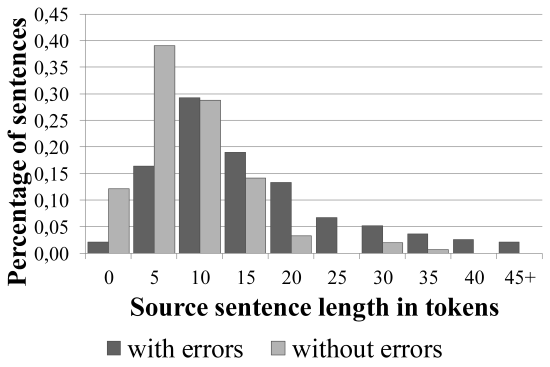

Figure 4: Distribution of sentences with and without errors per sentence length.

In total, 278 fluency errors and 205 accuracy errors were found in the dataset, which is in line with findings by Van Brussel et al. (2018) that NMT for English-Dutch contains more fluency issues than adequacy issues. Figure 5 shows how common the different subtypes are.

Coherence indeed seems to be a crucial addition to the taxonomy for literary translation, making up more than 50% of all fluency errors. Most coherence issues relate to logical problems. For accuracy, the most common error type is mistranslation, which makes up around 80% of all accuracy errors. Most mistranslation issues relate to multiword expressions, word sense issues, and issues without a specific subcategory. Style and register issues consisted mostly of disfluent sentences or constructions, indicating that this might still be an issue for NMT as it was for SMT (Besacier and

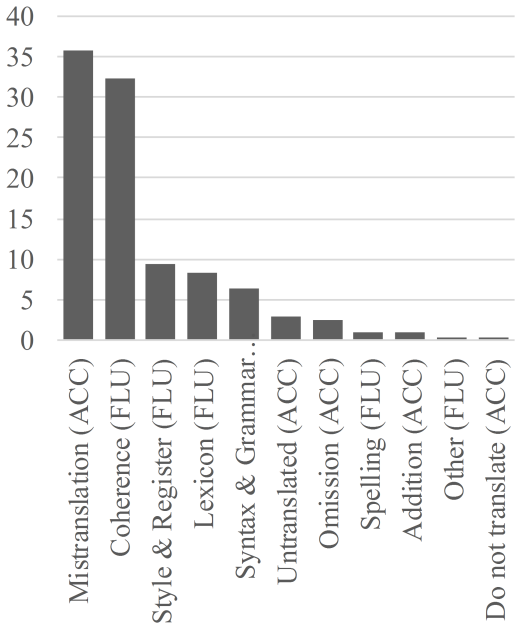

Figure 5: Frequency of error types expressed as percentage of all errors.

Schwartz, 2015). Issues found to be problematic in SMT by Toral and Way (2015b) such as lexical choice, verbal tense, and agreement, only occurred a few times in our NMT output, although it must be stressed that many cases of what we currently label as coherence issues might in other taxonomies be labeled as lexical choice issues. Indeed, Van Brussel et al. (2018) found lexical choice to be the most common fluency issue in Dutch NMT.

4.2 Key features

Lexical richness

Compared to the source text, both human translation and NMT have a higher number of unique words (5907 and 5948, respectively, as opposed to 5320 in the source). This difference is greatest for the number of singletons, i.e., words occurring only once, which is almost 500 words higher for HT and NMT as compared to the source. At first sight, this seems to indicate that both HT and NMT are lexically richer than the source text, with NMT the richer of the two. When comparing the number of unique words to the total number of words in Figure 6, this effect becomes even stronger: despite having the lowest number of total words, MT also has the highest number of unique words. A possible explanation for the higher number of unique words lies in the differences between both languages. In Dutch, compound nouns

are often written as one word, whereas they consist of two words in English. For example, in Dutch, you can have the words 'eet' ('to eat'), 'kamer' ('room'), and the compound 'eetkamer' ('dining room') as three unique words, whereas in English there would be only two words: 'dining' and 'room'.

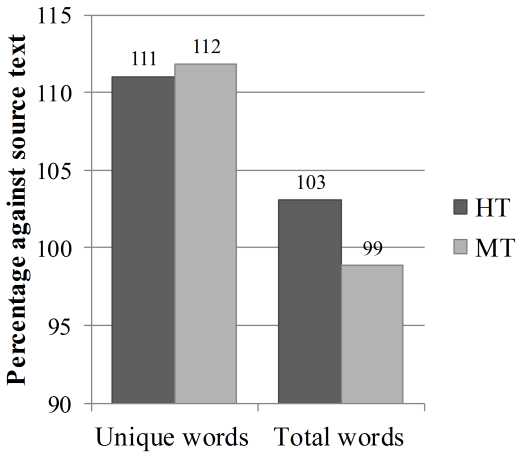

Figure 6: Unique words and total words as compared to the source text.

To further verify this claim, we studied lexical density by looking at a variety of type-token ratio measures, which we summarize in Table 1.

	Source	HT	MT
TTR	0.073	0.079	**0.083**
Root TTR	19.71	21.56	**22.17**
Corr. TTR	13.94	15.24	**15.68**
Mass index	**0.021**	0.020	0.020
MSTTR	0.648	**0.670**	0.660

Table 1: Summary of lexical density measures.

Most measures show comparable trends, with MT having a somewhat higher TTR than both the source text and the human translation. The measures for which this does not hold, however, are Mass index (highest in the source) and MSTTR (highest for HT), which have been argued to be better measures of lexical density than some of the other measures. As the differences between the three texts are rather small, we would argue that this seems to confirm that NMT can be at least as lexically rich as the original literary text and corresponding human translation. Still, in NMT, judging by the abundance of mistranslations and logical issues we found in the first chapter, it is possible that this lexical richness is in fact caused by translation errors. We therefore did not only look at the number of words in isolation, but we calculated translation entropy for HT and MT to gain a better understanding of what happens in translation. Figure 7 gives an overview of the translation entropy for words with different frequencies in the source text. It can be seen that translation entropy is always higher for HT, regardless of source word frequency.

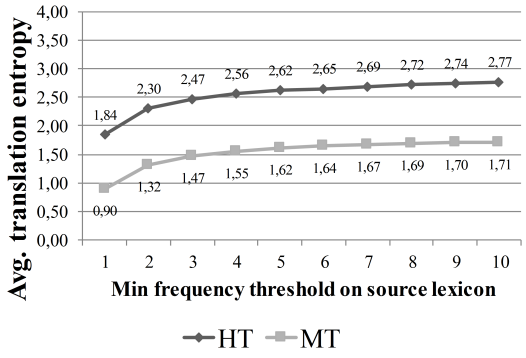

Figure 7: Average word translation entropy at different frequency thresholds.

This indicates that, in human translation, there is a higher level of uncertainty for the potential translations of a word than in MT, which, in turn, supports the theory that lexical richness in MT is potentially caused by erroneous translations, although a closer look at the data would be necessary to further substantiate that claim.

Cohesion

To study cohesion, we looked at the overlap of lemmas between a sentence and the following two sentences as these are a proxy for textual organisation, and we compare the overlap in the source text with that in HT and NMT for the number of sentences as well as the number of lemmas (as there can be more than one lemma overlapping in one sentence). Figure 8 shows lexical overlap (comparing lemmas of content words) and Figure 9 shows semantic overlap (comparing synonyms of lemmas of content words).

Looking at lexical overlap, it is clear that there is a greater level of overlap between sentences in the original than in either human translation or MT. The overlap for MT is somewhat higher than HT on a sentence level (a difference of 15 sentences), but quite a bit lower than HT on a lemma level (a differences of 92 lemmas). It is possible that English and Dutch have a different degree of lex-

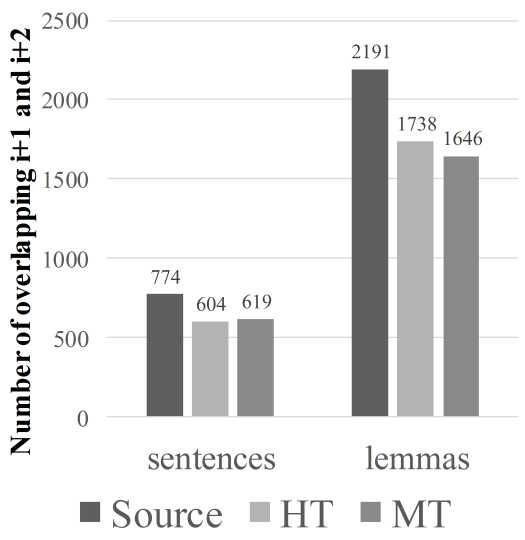

Figure 8: Local lexical cohesion.

ical richness, or this could also be caused by differences between original and translated text, with the latter generally exhibiting less variation than the first (Baker, 1996).

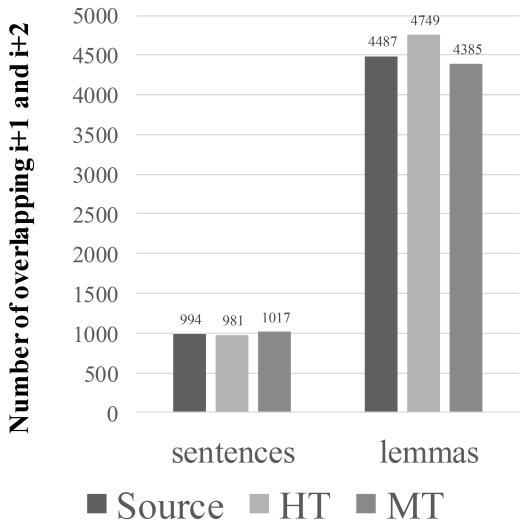

Figure 9: Local semantic cohesion.

Taking synonyms into account, the trend changes. On a sentence level, MT has a greater number of overlap than either the source text or HT (a difference of 23 sentences and 36 sentences, respectively), on a lemma level, HT has the greatest number of overlap (262 lemmas more than in the source, 364 lemmas more than in MT). A possible explanation, combining the information in Figure 8 and Figure 9, would be that, where the original author often reused the exact same word(s), the Dutch translator introduced synonyms more often. This is supported by research into literary translator style, where the avoidance of repetition in literary translation is considered to be a 'translation universal' (Ben-Ari, 1998). Looking at the number of exact or semantically-related overlapping lemmas, MT exhibits the least overlap. This could be an indication of MT being less coherent, possibly caused by errors in the MT output, as erroneous words would not be identified as semantically related. As for translation entropy, further analysis of the data would be needed to verify this.

Syntactic equivalence

Looking at syntactic variation between source and target text in Figurue 10, we clearly see that the cross values for human translation are much higher than those for MT. It is striking that 80% of all MT sentences have a cross value in the range $0 - 0.5$, indicating that MT follows the structure of the source text closely. The human translator introduced much more variation. There are 334 instances of cross values greater than 2.5 in human translation, compared to 16 in MT. The highest cross value for an MT sentence was 4, whereas for HT, there were 93 cases with a cross value over 4.

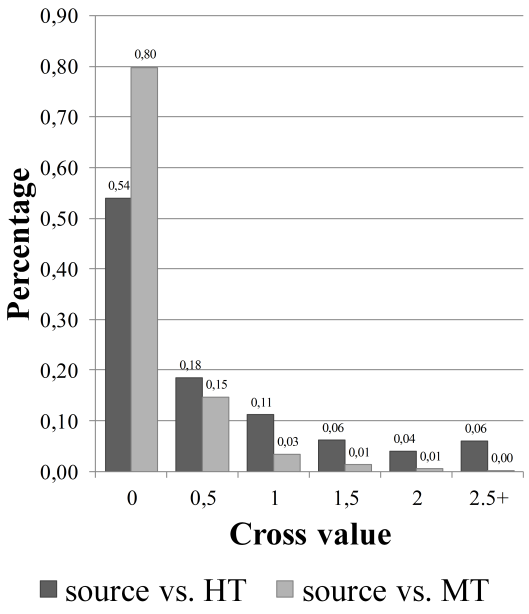

Figure 10: Frequency distribution of cross values (word).

Sequence cross values showed a very similar trend, with 78% of all sentences in MT having a cross value of zero, as compared to 52% in HT. This seems to indicate that the issue of MT closely following the source text structure leading

to potentially unidiomatic language (Besacier and Schwartz, 2015) has not entirely been solved in NMT yet.

5 Conclusion

We conducted the very first case study into the potential of NMT for literary translation for the English–Dutch language pair. Our goal was to get an idea of the current quality of NMT for literary translation in this language pair and to identify likenesses and differences between source, HT, and MT for three key features: lexical richness, cohesion, and syntactic equivalence. In particular for shorter sentences, NMT quality seems promising. 44% of the sentences we studied contained no errors, which is impressive for a general-domain MT system. On the other hand, the MT output still contained many coherence issues and mistranslations. Despite MT containing the highest number of unique words, measures of lexical density did not confirm that it was lexically richer than the source text or HT. The higher translation entropy in HT further confirms that there is a difference between MT and HT, despite their TTR scores being comparable, a difference that might be caused by the many mistranslations found in the MT output. Looking at local cohesion, we found that it seems strongest in the source text, with human translation favouring synonyms over exact repetition and MT being the least cohesive of the three when considering overlapping lemmas. Our analysis of syntactic equivalence further shows that MT generally remains faithful to the source text structures, whereas HT shows a greater diversity compared to the source text. It remains to be seen to what extent these issues impact the quality of the output or the reading experience. Word order issues were rare in our dataset, but disfluency issues were more common. In the future, our goal is to annotate the rest of the novel and have a second independent annotator perform the same work, so we can compare the inter-annotator agreement and generate a gold standard annotation for the whole novel. We will then compare the textual feature analysis with the quality evaluation in more detail, to learn if and how they influence each other. This knowledge could then be applied to build quality estimation systems that use textual features as a proxy for quality. A second future goal is to use eyetracking to measure the readability of the raw MT output. As the GECO corpus contains information on the reading of the original English source and Dutch target text, we can use them as a reference to see to what extent MT impacts the reader's experience, and which features or errors impact this reading experience the most.

Acknowledgments

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.

References

Baker, Mona 1996. Corpus-Based Translation Studies: The Challenges That Lie Ahead. In Harold Somers (ed.) *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, pp. 175-186.

Ben-Ari, Nitsa 1998. The Ambivalent Case of Repetitions in Literary Translation. Avoiding Repetitions: a "Universal" of Translation? *Meta*, 43(1), 68–78.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 257–267.

Besacier, Laurent and Lane Schwartz. 2015. Automated translation of a literary work: a pilot study. In *Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature*, pp. 114–122.

Michael Carl, Srinivas Bangalore, and Moritz Schaeffer 2017. Why translation is difficult: A corpus-based study of non-literality in post-editing and from-scratch translation. *HERMES-Journal of Language and Communication in Business*, (56), 43-57.

Cop, Uschi, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49(2), 602–615.

Crossley, Scott A., Kristopher Kyle, and Danielle S. McNamara. The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing 32*: 1–16.

de Camargo, Diva Cardoso. 2004. An investigation of a literary translator's style in a novel written by Jorge Amado. *Intercâmbio. Revista do Programa de Estudos Pós-Graduados em Linguística Aplicada e Estudos da Linguagem*, ISSN 2237-759X, 13.

Johnson, Wendell. 1944. Studies in Language Behavior. *Psychological Monographs*, 56, 1.

Junczys-Dowmunt, Marcin, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? A case study on 30 translation directions. In *Proceedings of the Ninth International Workshop on Spoken Language Translation (IWSLT)*.

McNamara, Danielle S., Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. Automated evaluation of text and discourse with Coh-Metrix. *Cambridge University Press.*

Moore, Robert C. 2004. Improving IBM word alignment model 1. *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)* pp. 518–525.

Moorkens, Joss, Antonio Toral, Sheila Castilho, and Andy Way. 2018. Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*. 7(2), 240–262.

Och, Franz Josef and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* volume 29, number 1, pp. 19–51.

Schaeffer, Moritz, Barbara Dragsted, Kristian Tangsgaard Hvelplund, Laura Winther Balling, and Michael Carl. 2016. Word translation entropy: Evidence of early target language activation during reading for translation. In *New directions in empirical translation process research*, pp. 183–210. Springer, Cham.

Tezcan, Arda, Véronique Hoste, and Lieve Macken. 2017. SCATE taxonomy and corpus of machine translation errors. In Gloria Corpas Pastor and Isabel Durán-Muñoz (Eds), *Trends in e-tools and resources for translators and interpreters*, pp. 219-244. Brill, Rodopi.

Torruella, Joan and Ramón Capsada. 2015. Lexical statistics and tipological structures: a measure of lexical richness. *Procedia-Social and Behavioral Sciences*, 95, 447–454.

Toral, Antonio and Victor M. Sánchez-Cartagena. 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 (EACL), vol 1.*, pp. 1063–1073.

Toral, Antonio and Andy Way. 2015. Translating Literary Text between Related Languages using SMT. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature, NAACL*, pp. 123–132.

Toral, Antonio and Andy Way. 2015. Machine-assisted translation of literary text: A case study. *Translation Spaces*, 4(2): 240–267.

Toral, Antonio and Andy Way. 2015. What Level of Quality Can Neural Machine Translation Attain on Literary Text?. In J. Moorkens, S. Castilho, F. Gaspari, and S. Doherty (Eds.), *Translation Quality Assessment. Machine Translation: Technologies and Applications, vol 1.*, Springer, Cham.

Van Brussel, Laura, Arda Tezcan, and Lieve Macken. 2018. A fine-grained error analysis of NMT, PBMT and RBMT output for English-to-Dutch. In *Eleventh International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA)*, pp. 3799-3804.

Vanroy, Bram, Orphée De Clercq and Lieve Macken. 2019a. Correlating process and product data to get an insight into translation difficulty. *Perspectives-studies in Translation Theory and Practice*, 1–18.

Vanroy, Bram, Arda Tezcan, and Lieve Macken. 2019b (*submitted*). Predicting syntactic equivalence between source and target sentences. *CLIN Journal.*

Voigt, Rob and Dan Jurafsky. 2012. Towards a literary machine translation: The role of referential cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pp. 18-25.

Yimam, Seid Muhie, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*: System Demonstrations, pp. 1-6.