The Riddle of (Literary) Machine Translation Quality

Gys-Walt van Egdom
Onno Kosters
Christophe Declercq

Gys-Walt van Egdom
Utrecht University;
g.m.w.vanegdom@uu.nl;
ORCID: 0000-0002-7521-6792

Onno Kosters
Utrecht University;
o.r.kosters@uu.nl;
ORCID: 0000-0002-6155-4379

Christophe Declercq
Utrecht University;
c.j.m.declercq@uu.nl;
ORCID: 0000-0002-6687-120X

Abstract

This study aims to gauge the reliability and validity of metrics and algorithms in evaluating the quality of machine translation in a literary context. Ten machine translated versions of a literary story, provided by four different MT engines over a period of three years, are compared applying two quantitative quality estimation scores (BLEU and a recently developed literariness algorithm). The comparative analysis provides an insight not only into the quality of stylistic and narratological features of machine translation, but also into more traditional quality criteria, such as accuracy and fluency. It is found that evaluations are not always in agreement and that they lack nuance. It is suggested that metrics and algorithms cover only parts of the notion of "quality", and that a more fine-grained approach is needed if potential literary quality of machine translation is to be captured and possibly validated using those instruments.

Keywords: literary machine translation, quality, literariness, automated metrics, machine learning.

Resumen

Este estudio pretende calibrar la fiabilidad y la validez de métricas y algoritmos para evaluar la calidad de la traducción automática en un contexto literario. Se comparan diez versiones traducidas automáticamente de una historia literaria, proporcionadas por cuatro motores de traducción automática diferentes a lo largo de un periodo de tres años, aplicando dos puntuaciones cuantitativas de estimación de la calidad (BLEU y un algoritmo de literariedad desarrollado recientemente). El análisis comparativo ofrece una visión no sólo de la calidad de los rasgos estilísticos y narratológicos de la traducción automática, sino también de criterios de calidad más tradicionales, como la precisión y la fluidez. Se constata que las evaluaciones no siempre coinciden y que carecen de matices. Se sugiere que las métricas y los algoritmos sólo cubren una parte de la noción de «calidad», y que es necesario un enfoque más detallado si se quiere captar la calidad literaria potencial de la traducción automática y, posiblemente, validarla mediante esos instrumentos.

Rebuda: 31 de març de 2023 / Acceptació: 20 de novembre de 2023 / Publicació avançada: 20 de desembre de 2023

Revista Tradumàtica. Número 21 (Desembre 2023)
Cita recomanada: VAN EGDOM, GYS-WALT; KOSTERS, ONNO; DECLERCQ, CHRISTOPHE (2023). «The Riddle of (Literary) Machine Translation Quality: Assessing Automated Quality Evaluation Metrics in a Literary Context.». Revista Tradumàtica. Tecnologies de la Traducció, 21, 129-159. <https://doi.org/10.5565/rev/tradumatica.345>

http://revistes.uab.cat/tradumatica
ISSN 1578-7559

Palabras clave: traducción automática literaria, calidad, literalidad, métricas automatizadas, aprendizaje automático.

Resum

Aquest estudi pretén calibrar la fiabilitat i la validesa de mètriques i algoritmes per avaluar la qualitat de la traducció automàtica en un context literari. Es comparen deu versions traduïdes automàticament d'una història literària, proporcionades per quatre motors de traducció automàtica diferents al llarg d'un període de tres anys, aplicant dues puntuacions quantitatives d'estimació de la qualitat (BLEU i un algoritme de literarietat desenvolupat recentment). L'anàlisi comparativa ofereix una visió no només de la qualitat dels trets estilístics i narratològics de la traducció automàtica, sinó també de criteris de qualitat més tradicionals, com la precisió i la fluidesa. Es constata que les avaluacions no sempre coincideixen i que no posseixen matissos. Se suggereix que les mètriques i els algoritmes tan sols cubreixen una part de la noció de «qualitat», i que és necessari un enfocament més detallat si es pretén captar la qualitat literària potencial de la traducció automàtica i, possiblement, validar-la per mitjà d'aquests instruments.

Paraules clau: traducció automàtica literària, qualitat, literalitat, mètriques automatitzades, aprenentatge automàtic

Introduction

In recent years, firm claims about the progress of neural machine translation (NMT) have been made (Wu et al., 2016, Castilho et al., 2017). Initially, these statements focused on the quality of LSP translation (Chu et al. 2017, Speerstra 2018, Jia et al. 2019, Kosmaczewska and Train, 2019, among others). Recently, acknowledgements of possibilities for integrating machine translation (MT) into the production of translations of literary texts have been tested and increasingly research acknowledges that, owing to the improvements in quality of MT output, MT can also be leveraged in literary contexts in the near future (Voigt and Jurafski, 2012, Toral and Way, 2014, Toral and Way, 2018, Guerberhof and Toral, 2022). Moreover, for the language pair English into Dutch, research increasingly covers the grounds for establishing a fully integrated MT-supported or even MT-driven approach to literary translation production (Tezcan et al., 2019, Webster et al., 2020, Macken et al., 2022).

However, as Félix Do Carmo (2022) recently pointed out, claims about output quality often rely on misleading conceptual constructs of "quality". This is not surprising, given that "translation quality" has always been "an elusive and indeterminate" concept (see Holmes 1988 and Moorkens et al. 2018). However, the problematic conceptualisation and operationalisation of "quality" in the context of MT has led to some peculiar developments and misplaced claims. Automatic metrics, such as BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007) and COMET (Rei et al., 2020), have gained serious popularity in recent years. These systems seem to measure only specific aspects of quality, which the developers themselves acknowledge, yet, these metrics are widely used, primarily because of their (alleged) correlation with human evaluation. Frequent use of these metrics has led to misplaced confidence, with a good many scholars assuming they effectively do measure quality (see Do Carmo 2022, who refers to the highly insightful work of Marie 2022a, 2022b). This misplaced belief is reinforced both by scholars who rely on the metrics without taking a critical stance and scholars who present research

on said automatic metrics in a highly selective manner (see Do Carmo, 2022). Additionally, in translation technology research, evaluative judgments are often passed by individuals who lack the required expertise to judge the quality: "Software engineers have judged MT like a blind person judging colour" (Van Egdom et al., 2017 [own translation, from Dutch]). Evaluation is often performed by people that were simply available and willing to help out.

This article contributes to the ongoing debate on quality in literary translation when MT is involved. The study aims to provide a fillip to the ongoing methodological discussion about the operationalisation of the notion of "quality" through the use of automatic metrics and algorithms. After a qualitative analysis of a literary short story ("Wrote a letter…" by the American writer Donald Barthelme), the quality of 10 first output translations produced by 4 different engines is analysed. First, these translations are assessed from a linguistic and stylistic vantage point by human evaluators with varying translation experience. In the second and third phase of the experiment, BLEU scores and literariness scores are calculated. The goal is to determine the extent to which human and automatic judgments effectively seem to correlate in the context of literary translation. This research takes into account NMT's potential for self-learning, and tries to do justice to developments in custom MT, by drawing a customised engine into the analysis.

Method

For this experiment, "Wrote a letter…" (524 words) was selected, a short story published by Donald Barthelme in 1980, which had not been translated into Dutch prior to this study. Subsequently, a literary translator with 15 years of professional experience, agreed to translate the story into Dutch, as if it were a proper assignment aimed at publication, thus creating a human translation (HT) that will act as a reference. The source text (ST) was then machine translated ten times. Three different open access NMT engines – DeepL, Systran and Google NMT – were used to provide one translation each. An annual translation snapshot was produced during a period of three years to track progress to determine if the capabilities of MT engines would improve over time as a result of self-learning as well as added data and algorithmic capacity. As customisation is considered to provide a major impetus to MT output quality, one more version of the ST was produced using a custom MT trained on literary data (Koehn and Knowles, 2017, Matusov, 2019, Saunders, 2022). The customised system used for this experiment was built by Toral et al. (see Toral et al. 2020, Toral et al 2021). The system, dubbed "S3Big", is based on a Transformer model and trained with in-domain data, in this case a parallel corpus of 500 English novels and Dutch translations, plus additional monolingual data, consisting of 4,435 literary works (see Toral et al., 2020, Toral et al., 2021). The custom MT not necessarily features as an additional aspect of the developments over time but as a testing ground of state-of-the-art customisation of engines for literary purposes (vs-a-vs the main baseline engines).

Comentado [RP1]: Incoherent i difícil de resoldre - possiblement volen dir "The custom MT does not necessarily feature".

	HT	DeepL	Systran	Google	Custom MT
July 2020		√	√	√	
July 2021		√	√	√	
June 2022	√	√	√	√	√

Table 1. Overview of research materials

Quality Assessment

A preliminary literature review was conducted to investigate the literary characteristics of the ST. For the qualitative analysis, a total of 28 text items were selected from the ST. Item-related methods are rooted in a tradition of analytical evaluation, while in human evaluation, a broad distinction can be made between holistic and analytical assessment. Both methods have their advantages and disadvantages (see Van Egdom et al. 2018). Holistic evaluation methods, on the one hand, consider the text as a whole, but the evaluation hinges mainly on the impression the assessor has of a target text. On the other hand, the analytical method starts from characteristics of the source or target text but pays little attention to the text as a whole. Analytical methods simply aim at detecting errors in the target text (e.g. DQF, MQM framework). In this research, a preselection of items was used. Preselection methods are aimed at identifying difficulties in a translation task, taking into account the characteristics of the source text, the contrast between source and target languages and the translation brief. Items typically used in research on translation quality include words and fragments that are likely to pose a problem to translators and/or translation engines. (see Nord 1988).[1] For this study, the items or "rich points" were related to three criteria:

1. Accuracy
2. Fluency
3. Style.

The first two criteria are often used to assess general text quality of MT output (Wu et al. 2016, Castilho et al. 2017), while the third was employed to shed light on the literary characteristics of the MTs (Castilho and Resende, 2022). The text items were verified by two native Dutch assessors with profound literary expertise and near-native knowledge of the English language and the cultural frames involved. The same two

[1] The selection of criteria for binary error analysis was influenced by this awareness of the translation task's inherent embedding within a specific historical framework. In other words, the selection was based on the fact that the TT was intended for a Dutch-speaking audience in the early 2020's, wishing to read a literary text that reflects the style and content of the ST.

assessors then evaluated the TT solutions, classifying them as either correct, undesirable or incorrect solutions. Undesirable solutions were later discussed by the assessors and categorised as either correct or incorrect. This evaluative classification formed the basis for a qualitative analysis of the MTs. A third assessor of equivalent profile in terms of source/target language, source/target culture proficiency, reviewed the quality estimations by the prime two assessors.

In the second part of the study, BLEU scores of the respective MT were calculated. BLEU is a metric that has been used to make statements about machine translation quality (Papineni et al., 2002). BLEU scores are said to correlate with human evaluation of MT output, and BLEU metrics basically provide an indication of the formal similarities between several texts. The 10 MTs used for this study were compared separately with the HT, then the BLEU scores were compared with the findings of human assessors, more specifically, their findings related to the representation of accuracy and fluency.

In the third part, the literariness of the target texts (TTs) was determined using a literariness algorithm recently developed by members of the "The Riddle of Literary Quality" project (Van Cranenburgh et al., 2019; see also, Koolen et al., 2020; Van Dalen-Oskam, 2021). Based on a large reader survey, which yielded almost 14,000 responses, the project members have rolled out a supervised model that can predict (human-informed) literary quality ratings from textual factors quite successfully (see Van Cranenburgh et al., 2019). Textual features that laid the foundation of this literariness algorithm are basic stylometric variables such as word frequency, density and positioning of specific sets of words in relation to their context and sentence length (see Koolen et al., 2020). For this study, the literariness scores were calculated for the MTs as well as for the HT. By dint of comparison (with findings from the qualitative analysis), the results of this analysis were scrutinised with a view to evaluating the usefulness of the algorithm.

Results

In order to assess the quality of the machine output, an attempt was made to pinpoint the literary qualities in the ST. A short literary review of literature on Donald Barthelme's writings (see Gordon, 1981, Couturier & Durand, 1982, Molesworth, 1982, Roe, 1992) showed that there is a general consensus about the textual features that make up the literary quality of his short stories. His postmodern short stories feature elements of the absurd to comment on the *condition humaine.* His stories are described as playful and unconventional in their narrative structure, while his characters are often disillusioned (see Taylor, 1977 and McCaffery, 1980).

"Wrote a letter…" fits perfectly within Barthelme's oeuvre. The premise of the story is whimsical: the protagonist of the story corresponds with the President of the moon, using a range of unconventional communication means ("moonbeaming", "flights of angels"). The tone employed in the story is light and humorous, and the topics that are discussed in the correspondence are far from abnormal (mental health, a Honda that has been towed away, apartment rentals). This constant juxtaposition of or clash between the

surreal and the banal add to the otherworldly atmosphere of the story. The clash is also reflected in colloquial uses of languages ("You ever seen them…"). The story's thematic absurdity is carefully constructed: the various means of communication used for correspondence become increasingly outlandish, which builds up a sense of disorientation or unease in the reader.

If a translation is to mimic Barthelme's sharp wit and the text's ability to challenge the reader, a high-quality literary rendition of this ST should reflect these characteristics. At the same time, the criteria that apply in the context of more general quality assessment also apply in this context: it is imperative that the translation accurately reflects the meaning of the original, while making it understandable to its new audience.

Table 2 shows the ST elements that not only highlight the literary features that require attention in a qualitative analysis, but also textual features that may pose problems to the accuracy and fluency of a TT. In total, 28 ST features were selected, but it is important to note that there is a striking imbalance in the distribution of selected items: stylistic features account for the vast majority of the items (16 out of 28); while only 7 were used to assess fluency, and 5 to gauge accuracy. This disparity was due to the study's primary focus on literariness. Nonetheless, the evaluators acknowledged that some items that were categorised as stylistic features were also inextricably linked to either fluency or accuracy. The second criterion presented in Table 2 serves only as an indication of the conceptual complexity associated with literariness, but they were not taken part of the dichotomous assessment of MT items. Moreover, the table provides an overview of how ST features have been rendered in the HT (later used as a yardstick for the BLEU score calculation).

	ST	Rich point	HT
1	, asked him	Style - colloquialism	en vroeg 'm
2	towaway zones	Accuracy	wegsleepzones
3	and I didn't like it	Fluency	en daar had ik de pest over in
4	Cost me ..., plus	Style - colloquialism	Kostte me ..., nog afgezien van
5	tiny little cars	Style - colloquialism	kleine autootjes
6	You ever notice ...? You ever seen...? No you haven't.	Style - colloquialism [Fluency]	Nooit opgevallen...? Ooit gezien...? Precies.
7	, and to keep some mental health warm ..., and could I interest him...	Style - colloquialism [Fluency]	, en of hij wat van dat hoognodige mentale welbevinden ... apart kon houden, en of ik 'm blij kon maken...
8	a bucket of ribs in red sauce	Accuracy	een grote portie spareribs in rode saus
9	Which I would gladly carry up there...	Fluency - idiom	Die ik ... graag voor hem zou meebrengen...
10	I cabled him	Style - absurdism [Accuracy]	Ik telegrafeerde hem
11	and, by the way, what was the apartment situation up there?	Style - colloquialism [Fluency]	en, tussen twee haakjes, hoe was eigenlijk de situatie op de appartementenmarkt daarboven?
12	It was bad,	Fluency - idiom	Die was dramatisch,

Table 2. Categorisation of ST items, flanked by HT solutions.

In Tables 3, 4, 5 and 6, the overall results of a critical interrogation of the automated rendition of the same features are presented (for a full overview of items, see Annex A). Evaluators were asked to perform a dichotomous assessment of the translated items, considering the criteria used for the assessment of the source text items. Evaluators were asked to indicate whether the item had been translated correctly or incorrectly (i.e., dichotomous evaluation). In cases where evaluators were experiencing doubts as to the correctness of a solution (undesirable but not necessarily incorrect solutions), they were

instructed to flag the respective item. Next, the evaluators engaged in a discussion until they reached a consensus regarding the correctness of solutions.

The tables below provide an overview of how well the 4 MT engines performed on the task. They do so in a highly insightful way, because they reveal the strengths and weaknesses of each engine and demonstrate whether any considerable progress has been made in terms of output quality. In other words, these tables allow for the reader to gauge the extent to which the MT engines have been able to capture accurately relevant features of the ST and provide a solid basis for a qualitative comparison of engines.

	Accuracy (/5)	Fluency (7)	Style (16)	Total (/28)
DeepL 2020	1 (20%)	1 (14%)	4 (25%)	6 (21%)
DeepL 2021	1 (20%)	2 (29%)	6 (38%)	9 (32%)
DeepL 2022	1 (20%)	2 (29%)	6 (38%)	9 (32%)

Table 3. Quality evaluation of DeepL, based on human assessment

The results of the DeepL evaluation presented in Table 3 show that DeepL's overall performance is far from ideal, but it improved considerably from 2020 to 2021, and remained stable, without noticeable improvement, in 2022. The breakdown into categories provides additional insights into the system's strengths and weaknesses. It is abundantly clear that there is quite some room for improvement for accuracy, fluency and style, with scores of 1 out of 5 for accuracy, 1 and (later) 2 out of 7 for fluency, and 4 and (later) 6 out of 16 for style. Despite room for improvement, the performance in rendering the literary style of Barthelme is worthy of note. The stylistic prowess of DeepL is predominantly linked to the representation of the absurd. It is important to note that accuracy, as can be inferred from Table 2, has proven to be crucial for conveying absurdism effectively (see item 10 ["I cabled him"] and item 19 ["Drumming fiercely on… the moon frequency"]). In other words, the scores for accuracy are somewhat skewed. One evaluator even stated that DeepL's ability to convey meaning remains relatively acceptable compared to other systems. Still, in all three years, output quality was consistently lower than anticipated; it was expected that scores for accuracy and fluency would be somewhat acceptable, i.e., that the system would attain percentages of >50% for both categories.

	Accuracy (/5)	Fluency (7)	Style (16)	Total (/28)
Systran 2020	0 (0%)	0 (0%)	1 (7%)	1 (4%)
Systran 2021	0 (0%)	0 (0%)	1 (7%)	1 (4%)

Systran 2022	1 (20%)	0 (0%)	2 (13%)	3 (11%)

Table 4. Quality evaluation of Systran, based on human assessment

The scores Systran received can be aptly described as "abysmal": the MT engines performed extremely poorly in all three criteria, with a total score of 1 out of 28 for output produced in 2020 and 2021. In 2022, Systran performed slightly better, with a score of 3. Nevertheless, the output is still a far cry from being usable, even with extensive editing.

Systran failed miserably at producing accurate, fluent, and stylistically appropriate solutions. What stands out most in Table 4, is the fact that the system never came up with a single correct solution for items that tested fluency. Poorly translated items that stood out in this category were item 9 ("Which I would gladly carry up there") and 26 ("it looked … hand up there"). Scores for accuracy were hardly better: in 2022, Systran opted for a borrowing of the source element "Space Shuttle Hurry-up Fund", which made sense to the evaluators. Still, the consistently low scores in all categories clearly suggest that producing a high-quality target text is something that is not within immediate reach, let alone producing a high-quality literary translation.

	Accuracy (/5)	**Fluency (7)**	**Style (16)**	**Total (/28)**
Google 2020	1 (20%)	2 (29%)	5 (31%)	8 (29%)
Google 2021	0 (20%)	1 (14%)	7 (44%)	8 (29%)
Google 2022	0 (20%)	1 (14%)	7 (44%)	8 (29%)

Table 5. Quality evaluation of Google NMT, based on human assessment

The output of Google was also evaluated on three criteria. Although Table 5 shows that, generally speaking, the engine performed poorly in 2020, with no more than 8 items solved correctly, it was actually top of the class in that year. In the following two years, the overall output quality did not get any better, but there are some differences between the three Google versions that are worthy of note. In 2021 and 2022, the engine demonstrated consistently better performance within the domain of style (7 out of 16). Google NMT appears to have managed to gain a firmer handle on the literariness of the source text. Particularly interesting is the rendition of colloquialism: good cases in point are the translations of the ellipsis in item 4 ("Cost me …, plus") and the redundancy in item 5 ("tiny little cars"). Absurdism is also a stylistic feature that the system managed to highlight in the automated output in 2021 and 2022 (see items 19 ["Drumming fiercely on… the moon frequency"] and 23 ["by means of … my Apple computer"]). As mentioned earlier, these items also require an accurate understanding of the source text. Again, this could lead us to believe that scores for accuracy are somewhat skewed. Paradoxically, however, the scores for accuracy dropped from 1 to 0. Google NMT never

came up with a correct solution for seemingly simple items like items 2 ("towaway zones") and 8 ("A bucket of ribs"). In sum, the total score of 8 out of 28 indicate that the system is competitive with DeepL, that the system even seems better at prioritising style, but also that it still has a very long way to go, if it is to produce output usable for literary translation.

	Accuracy (/5)	Fluency (7)	Style (16)	Total (/28)
Custom LMT 2022	2 (40%)	2 (29%)	7 (44%)	11 (39%)

Table 6. Quality evaluation of DeepL, based on human assessment

An interesting new player on the market is the Custom-built MT engine. The table above indicates that the system, although it was still found lacking, performed moderately well on the translation task, with a total score of 11 out of 28. With this score, it even outperformed all the other systems. Errors related to the basic output quality measures "accuracy" and "fluency" were still quite common in the text. A score of 2 out of 7 for fluency suggests that the text contains some awkward or unnatural phrasing, as in the case of items 9 ("Which I would gladly carry up there") and 12 ("it was bad"). However, it produced a couple of very fluent solutions, as in the case of item 3 ("and I didn't like it"). The same can be said about accuracy: items like item 2 ("towaway zones") prompted a completely incomprehensible noun in the Dutch version ("Daarsleepzones), but the Custom MT was the only system that produced an accurate solution for item 15 ("root cellar"). Comparable to Google, the system was able to capture the literariness of the ST to a certain extent, with 7 items that were solved correctly. Interestingly, it proved to be inconsistent in its rendering of colloquialism, but it did a splendid job depicting Barthelme's absurdism. The customised system therefore shows potential as a tool for translating literary texts, but the future will tell whether further training and tweaking will yield better results or whether progress will be brought to a halt.

The qualitative analysis evaluation of MT engines for literary translation clearly revealed that there is still significant room for improvement. An impetus is not only required for the literariness of output, but also even the accuracy and the fluency of the automated renditions continue to be problematic. Based on this case study, it can safely be stated that automation in literary translation cannot be said to be an immediate prospect.

BLEU

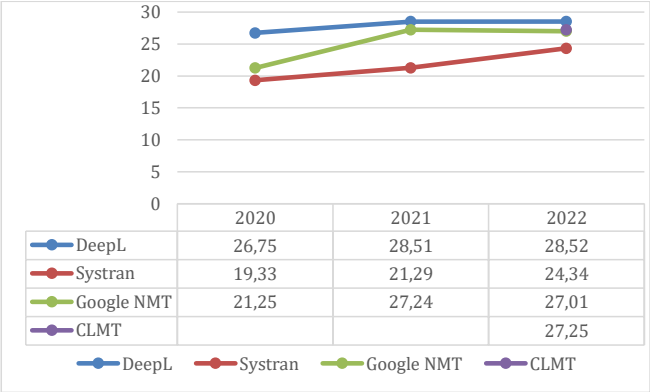
In the next phase of the study, the performance of the MT engines was compared, using BLEU scores as a means of comparison. BLEU is a widely used metric for evaluating the quality of MT output. It is designed to measure the similarity between one or more MTs and one or more reference translations made by human translators. BLEU scores run from 0 to 100, where a higher score is indicative of a more accurate and fluent

translation (see Table 7, see, Rekha et al. 2022: 45). Scores are grouped into ranges: scores of less than 10 awarded to texts that are considered useless; scores between 10 and 19 denote a higher likelihood that the gist of the target text is difficult to grasp; scores between 20 and 29 denote that the gist is clear but there are considerable errors in the output, and so on. As a rule of thumb, MTs with scores of 30 and upwards are generally seen as understandable or even good translations. BLEU scores are usually presented as useful since they provide a quantitative way of assessing the quality of MT systems and they correlate with human judgment. BLEU scores were calculated for the 10 MTs used for this study.

BLEU score	Explanation
< 10	Almost useless
10-19	Hard to get the gist
20-29	The gist is clear, but has significant errors
30-39	Understandable or even good translation
40-49	High quality translation
50-59	Very high-quality translation (adequate and fluent)
≤ 60	Quality comparable to or better than human translation

Table 7. Rough guideline for interpreting BLEU scores

As can be inferred from Table 8, DeepL clearly outperformed the other two off-the-shelf systems in 2020, scoring nearly 4 percentage points higher than Google NMT and 5 percentage points higher than Systran. In the following two years, DeepL stayed in the lead, but both Google NMT and Systran managed to narrow the gap. Systran showed some improvement: in 2020, it produced a TT that, according to BLEU score interpretation, did not even allow readers to glean the gist, with a meagre score of 19.33, but in the following years it performed marginally better, narrowing the gap with DeepL to 4 percentage points. Google NMT also seemed to be lagging in the first year (21.25), but it made steady progress in 2021 (with a score of 27.24). It fell back a bit in 2022 (27.01). In 2022, the customised engine was also put to the test for the first time: with a BLEU score of 27.25, it did not seem to challenge DeepL just yet, but it did a splendid job outperforming the other systems.



	2020	2021	2022
DeepL	26,75	28,51	28,52
Systran	19,33	21,29	24,34
Google NMT	21,25	27,24	27,01
CLMT			27,25

Table 8. BLEU scores of MTs

Looking at these results, it is tempting to state that the qualitative evaluation and the BLEU scores are fairly consistent. As in the qualitative analysis, Systran clearly seems to be missing the mark in many respects. Still, the observed improvements in the output quality of Systran in 2021 and 2022 are clearly at odds with the findings in the qualitative analysis: in terms of accuracy and fluency Systran made no significant strides. Another thing that stands out, is the striking disparity between the qualitative and the quantitative analysis of Google NMT. Based on the qualitative evaluation, one would expect Google to be more or less on a par with DeepL, at least in 2020 (see Tables 3 and 5). However, according to the BLEU calculation, Google only managed to catch up with DeepL in 2021, at a time when scores for accuracy and fluency dropped. A possible explanation for sudden improvements of estimated output quality in 2021 and 2022 seems to lie in Google's rendition of lengthy stylistic items (e.g., items 19 and 23). A system that has been passed over in silence in the diachronic comparison, is the custom-built engine. As in the qualitative analysis, the system outpaced Systran and Google, but it remains remarkable that the custom MT system lost out against DeepL, which clearly carries the day in this first quantitative analysis of "overall" output quality. According to the qualitative evaluation, no MT system came near the custom MT in terms of overall quality.

At first glance, there seems to be no striking disparity between the results of the qualitative and the first quantitative analysis, which might be interpreted as an indication that BLEU does correlate with human judgment. This can be accounted for quite easily: both analyses did show that all 4 MT systems still struggle to meet the passing threshold for an "understandable or even good translation". However, upon closer inspection it becomes clear that BLEU does nothing but measure similarities: therefore, it fails to come to grips with subtle but meaningful changes in output quality, and in no way does it reflect improvement or deterioration over time.

Literariness Prediction

BLEU is a tool primarily used to evaluate the MT quality in a general setting and has been primarily used to gauge the quality of more business-like TTs (see "Research context"). As such, it was never designed to measure the literariness of MT output. In the next phase, an attempt was made to evaluate the literary qualities of MT versions of Barthelme's text, using automatic metrics developed to predict the literariness of a text. One such metric system ensued from the "The Riddle of Literary Quality" project (see Koolen et al., 2020). The literariness algorithm relies on word embeddings and attempts to evaluate (or predict) the "literary quality" of Dutch texts by looking at frequency, density and positioning of specific sets of words in relation to their context and sentence length. The biggest advantage of this algorithm is that it does not require a reference text to calculate the literariness. In other words, this specific literariness algorithm could be used to measure the literary qualities of the MTs as well as of the HT that was used as a reference translation in the BLEU experiment. The results of this measurements are shown in Table 9.

What stands out in Table 9, is that DeepL MTs consistently score the highest, although their literariness does decline slightly over time (4.41 in 2020 and 4.37 in 2022). Despite the decline, DeepL remains unthreatened: in terms of literariness, the MT engine even structurally outperforms the HT (3.8), according to the algorithm. The customised engine also seems to deliver when it comes to literariness: with a score of 4.17, it is the runner up in the (2022) series. The literary quality scores of Google NMT are remarkable: they show good prospects in 2020 (4.08), beating HT with almost 0.3, but the system experiences a free fall in the following two years (3.27 in 2020). A similar effect is seen in the Systran scores; however, the decline in literariness is limited to less than 0.2 and the system seems to keep pace with the HT.

	HT	DeepL	Systran	Google	CLMT
July 2020		4.41	3.77	4.08	
July 2021		4.39 (↓)	3.8 (↑)	3.37 (↓)	
June 2022	3.8	4.37 (↓)	3.63 (↓)	3.27 (↓)	4.17

Table 9. Predicted literariness of all TTs

These results are remarkable, but they become more peculiar when considered in relation to the results obtained in previous paragraphs. In the qualitative analysis, it was found that while DeepL does seem to represent stylistic elements of the ST in a reasonably accurate manner, the system completely missed the mark when it came to colloquialism and fluency. Therefore, the high rating of the literary quality was quite unexpected. Google, on the other hand, proved to be quite capable when representing the colloquial characteristics of the ST, and it even showed improvement in 2021 and 2022. However, the literariness algorithm offers a rather bleak perspective on the literary qualities of Google's output in 2021 and 2022. The relative stability of Systran is also highly problematic: in terms of literary quality, Systran fell short in all versions, yet its

abysmal quality never truly affected the literariness scores of its output. The predicted literariness score of the custom MT is the only score that seems somewhat acceptable in this case: at times, the system represented absurdism and colloquialism fairly well. A reasonably high literary score therefore seems appropriate. Conversely, the score of the HT raises a few eyebrows: it can be assumed that the HT preserves literary quality (almost) optimally, but this clearly does not show in the table. The results in this section therefore cast serious doubt on the ability of the literary algorithm to accurately assess literary quality.

Discussion

This first part of the study attempts to evaluate the quality of machine-generated translations by identifying literary attributes in the ST and its corresponding machine translation (MT) produced by different translation engines. The study emphasised the importance of translations capturing Barthelme's style, while ensuring accuracy and fluency. Among all MT systems that were examined, Systran exhibited notably inferior performance, registering low scores across all evaluated criteria. The engines DeepL and Google NMT demonstrated comparable proficiency, albeit manifesting progress in different ways and to varying degrees: DeepL's strength predominantly lay in its ability to represent the absurd present in Barthelme's writing, whereas Google NMT exhibited a slightly superior capability in capturing Barthelme's colloquial style. Surprisingly, the household baseline MT engines, DeepL and Google NMT, were outperformed by the Custom-built Literary MT engine. While this engine displayed inconsistency in rendering colloquialism, it successfully conveyed Barthelme's absurdism. This outcome suggests that custom-built MT solutions do seem to hold potential, even in the literary domain, but this potential should never be overstated.

In this study, human evaluation served as a foundation for the analysis of the automated metrics. Initially, the BLEU scores appeared to align reasonably well with the assessments recorded in the human evaluation: scores consistently remained below 30 ("the gist is clear, but [the MT] has significant errors"). However, it is noteworthy that the Custom LMT engine, which ranked the best in the human evaluation, did not hold the top position according to BLEU scoring. Additionally, the distinction between the weakest system, Systran, and the best-performing systems was not very pronounced. What is more, BLEU failed to capture subtle or pronounced differences in quality of MT versions across the different years. Based on these findings, it can be asserted that BLEU does not provide sufficient insights into MT quality, particularly in a literary context.

The discussion then transitioned towards assessing the literariness of MT output, a task not originally within the BLEU scope. This part of the study attempted to evaluate the literary qualities of MT's as well as the (human) translator's version of Barthelme's text, using automated metrics designed to predict literariness. Interestingly, the human translation did not receive the highest score. The literariness algorithm identified DeepL as the most literary translation, with the custom-built engine securing the second position.

The literariness of Google NMT, which human assessors found reasonably adept at capturing Barthelme's colloquial style, was paradoxically rated very low by the algorithm. Surprisingly, the Google NMT versions from 2021 and 2022 even scored much lower on literary quality than Systran, which, in terms of literariness, performed similarly to human translation. These findings cast doubt on the literariness algorithm's ability to accurately gauge literary quality.

Conclusion

While there has been increasing optimism about the improvements in MT quality, including in literary contexts (Chu et al. 2017, Speerstra 2018, Jia et al. 2019, Kosmaczewska and Train 2019, Voigt and Jurafski 2012, Toral and Way 2014, Toral and Way 2018, Guerberhof and Toral 2022), claims about the effectiveness of MT seem to be based on inadequate measures of quality. This study has set out to critically interrogate the possibilities and limitations of using metrics (such as BLEU and the recently developed literariness algorithm) to (re)assess the quality MT output in a literary context, with emphasis on the evaluation of stylistic and narratological features of said translation. This study has shown that BLEU was unable to capture subtle changes in output quality. This became all the more apparent when BLEU was assessed diachronically. The literariness algorithm also fell short, as it failed to pick up on the literariness of the HT and shamelessly overpraised the literary qualities of MT output. This study therefore serves as a corrective countervoice to potential hypes surrounding MT, especially for the use in literary translation contexts, and might encourage experts in MT to reconsider the use of metrics to assess (literary) MT quality or, at least, (further) refine existing models that purport to "measure" MT quality (see Do Carmo, 2022, Marie, 2022a, 2022b).

At the same time, it should be noted that this study has some limitations. It does not seem superfluous to point out that research was limited to no more than one single source text and 10 MT versions. In other words, far more research is needed to gain a more complete understanding of the ability of MT to represent accuracy, fluency, style – and, more specifically, literary style – or the lack thereof. Adding more source material and automated renditions to the equation might help. Additionally, research with multilingual material and texts from different genres and literary movements seems in order so as to further expand our comprehension of the state of MT. However, the main goal of this article is to draw attention to the inherent reductionism of automated measurements and their limitations in capturing the full complexity of literary translation. Its research design may serve as a blueprint for further research and help fuel the meta-methodological debate.

References

Barthelme, Donald. (1980). I Wrote a Letter. In: Kim Herzinger (ed.). *The Teachings of Don B.* Berkeley: satires, parodies, fables, illustrated stories and play of Donald Barthelme. Berkely, CA: Counterpoint Press.

Castilho, Sheila; Moorkens, Joss; Gaspari, Federico; Calixto, Iacer; Tinsley, John; Way, Andy (2017). Is Neural Machine Translation the New State of Art? *The Prague Bulletin of Mathematical Linguistics*, n. 108 (June), pp. 109-120. <https://core.ac.uk/download/pdf/195384513.pdf>. [Accessed: 20231211].

Chu, Chenhui; Dabre, Raj; Kurohashi, Sadao (2017). An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation. In: P*roceedings of the 55th Annual Meeting of the Association for Computational Linguistics,* vol. 2, pp. 385-391. <https://doi.org./10.18653/v1/P17-2061>. [Accessed: 20231211].

Couturier, Maurice; Durand, Regis (1982). *Donald Barthelme.* London [etc.]: Methuen. (Contemporary writers).

Do Carmo, Felix (2022). Debunking a few machine translation myths: from 'zero-shot translation' to 'human parity' and 'no language left behind. *University of Surrey* (03 November). <https://www.surrey.ac.uk/news/convergence-lecture-series-debunking-few-machine-translation-myths-zero-shot-translation-human>. [Accessed: 20231211].

Gordon, Lois (1981). *Donald Barthelme*. Boston: Twayne. (Twayne's United States authors series; TUSAS 416).

Guerberhof Arenas, Ana; Toral Antonio (2022). Creativity in translation; Machine translation as a constraint for literary texts. *Translation Spaces,* v. 11, n. 2, pp. 1-31. <https://doi.org/10.48550/arXiv.2204.05655>. [Accessed: 20231211].

Koehn, Philipp; Knowles, Rebecca (2017). Six challenges for neural machine translation. In: Thang Luong; Alexandra Birch; Graham Neubig; Andrew Finch (eds.). *Proceedings of the First Workshop on Neural Machine Translation*, Vancouver. Association for Computational Linguistics, pp. 28-39. <https://doi.org//10.18653/v1/W17-3204>. [Accessed: 20231211].

Koolen, Corina; Dalen-Oskam, Van Dalen-Oskam, Karina; Van Cranenburgh, Andreas; Nagelhout, Erica (2020). Literary quality in the eye of the Dutch reader: The National Reader Survey. *Poetics,* v. 97 (April). <https://doi.org/10.1016/j.poetic.2020.101439>. [Accessed: 20231211].

Kosmaczewska, Kasia; Train, Matt (2019). Application of Post-Edited Machine Translation in Fashion eCommerce. In: Mikel Forcada; Andy Way; John Tinsley; Dimitar Shterionov; Celia Rico; Federico Gaspari (eds.). *Proceedings of MT Summit XVII: Translator, Project and User Tracks: august 2019, Dublin, Ireland.* European Association for Machine Translation, pp. 167-173. <https://aclanthology.org/W19-6730>. [Accessed: 20231211].

Lavie, Alon; Agarwal, Abhaya (2007). METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In: Chris Callison-Burch; Philipp

Koehn; Cameron Shaw Fordyce; Christof Monz (eds.). *Proceedings of the Second ACL Workshop on Statistical Machine Translation: June 2007,* Prague, pp. 228-231. <https://aclanthology.org/W07-0734>. [Accessed: 20231211].

Marie, Benjamin (2022a). Science Left Behind. *Medium.* <https://medium.com/@bnjmn_marie/science-left-behind-ca0a58231c20>. [Accessed: 20231211].

Marie, Benjamin (2022b). An Automatic Evaluation of the WMT22 General Machine Translation Task. *ArXiv online* <https://arxiv.org/pdf/2209.14172.pdf>. [Accessed: 20231211].

Matusov, Evgeny (2019). The Challenges of Using Neural Machine Translation for Literature. In: James Hadley; Maja Popovic; Haithem Afli; Andy Way. Proceedings of the Literary *Machine Translation: August 2019,* Dublin. European Association for machine Translation, pp. 10-19. <https://aclanthology.org/W19-7302.pdf>. [Accessed: 20231211].

McCaffery, Larry (1980). Donald Barthelme and the Metafictional Muse. *Current Trends in American Fiction*, v. 9, n. 27, pp. 78-88. <https://doi.org/10.2307/3683881>. [Accessed: 20231211].

Molesworth, Charles (1982). *Donald Barthelme's Fiction: The Ironist Saved from Drowning*. Columbia: University of Missouri Press.

Nord, Christiane (1988). *Textanalyse und Übersetzen: Theoretische Grundlagen, Methode und didaktische Anwendung einer übersetzungsrelevanten Textanalyse*. Heidelberg: Groos.

Papineni, Kishore; Roukos, Salim; Ward, Todd; Zhu, Wei-Jing (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL):* Philadelphia, July 2002, pp. 311-318. <https://aclanthology.org/P02-1040.pdf>. [Accessed: 20231211].

Roe, Barbara Louise (1992). *Donald Barthelme: A Study of the Short Fiction*. New York: Twayne; Toronto: Maxwell Macmillan Canada; New York: Maxwell Macmillan International.

Saunders, Danielle (2022). Domain Adaptation and Multi-Domain Adaptation for Neural Machine Translation: A Survey. *Jair: Journal of Artificial Intelligence Research*, v. 75, pp. 351-424. <https://doi.org/10.1613/jair.1.13566>. [Accessed: 20231211].

Speerstra, Nander (2018). A Comparison of Statistical and Neural MT in a Multi-Product and Multilingual Software Company: User Study. In: Juan Antonio Pérez-Ortiz, *et al.* (eds.). *Proceedings of the 21st Annual Conference of the European Association for Machine Translation: 28-30 May 2018*, Universitat d'Alacant: Alacant, pp. 315-321. <https://aclanthology.org/2018.eamt-main.34>. [Accessed: 20231211].

Taylor, David Michael (1977). *Donald Barthelme: an approach to contemporary fiction [PhD Dissertation].* Florida State University.

‹https://www.proquest.com/docview/3028533672pq-origsite=gscholar&fromopenview=true›. [Accessed: 20231211].

Tezcan, Arda; Daems, Joke; Macken, Lieve (2019). When a 'sport' is a person and other issues for NMT of novels. In: James Hadley; Maja Popovic; Haithem Afli; Ande Way (eds.). Proceedings of t*he Qualities of Literary Machine Translation: Aug. 19-23, Dublin.* European Association for Machine Translation, pp. 40-49. ‹https://aclanthology.org/W19-7306/›. [Accessed: 20231211].

Toral, Antonio; Van Cranenburg, Andreas; Nutters, Tia (2021). Literary-Adapted Machine Translation in a Well-Resourced Language Pair: Explorations with More Data and Wilder Context. In: *Book of abstracts 7th Conference of The International Association for Translation and Inter-Cultural Studies (IATIS), Barcelona.*

Toral, Antonio; Oliver, Antoni; Ribas Ballestín, Pau. (2020). Machine Translation of Novels in the Age of Transformer. In: Jörg Porsiel (ed.). *Maschinelle Übersetzung für Übersetzungsprofis. Berlin:* BDÜ-Fachverlag, pp.276-295. ‹https://arxiv.org/ftp/arxiv/papers/2011/2011.14979.pdf›. [Accessed: 20231211].

Toral, Antonio; Way, Andy (2018). What level of quality can neural machine translation attain on literary text? In: Joss Moorkens; Sheila Castilho; Federico Gaspari; Stephen Doherty (eds.). *Translation Quality Assessment: From Principles to Practice. 1st ed.* Cham: Springer International Publishing. ‹https://arxiv.org/pdf/1801.04962.pdf›. [Accessed: 20231211].

Toral, Antonio; Way, Andy (2014). Is Machine Translation Ready for Literature? *Proceedings of Translating and the Computer,* v. 36, pp. 174-176. ‹https://aclanthology.org/2014.tc-1.23/›. [20231211]

Van Cranenburgh, A.; van Dalen-Oskam, K.; van Zundert, J. (2019). Vector space explorations of literary language. *Language Resources and Evaluation*, v. 53, n 4 (December), pp. 625-650. ‹https://doi.org/10.1007/s10579-018-09442-4›. [20231211].

Van Dalen-Oskam, Karina (2021) *Het raadsel literatuur. Is literaire kwaliteit meetbaar?* Amsterdam: Amsterdam University Press.

Van Egdom, Gys-Walt; Bloemen, Henri; Segers, Winibert (2017). Machinevertaling, singularity et prometheische Scham. *Filter: tijdschrift over vertalen*, v. 24, n. 2, pp-19-26. ‹https://www.tijdschrift-filter.nl/jaargangen/2017/242/machinevertaling-singularity-et-prometeische-scham-19-26/›. [20231211].

Voigt, Rob; Jurafski, Dan (2012). Towards a Literary Machine Translation: The Role of Referential Cohesion. In: David Elson; Anna Kazantseva; Rada Mihalcea; Stan Spakowicz (eds.). *Workshop on Computational Linguistics for Literature: June 2012, Montréal, Canada.* Association for Computational Linguistics, pp. 18-25. ‹https://aclanthology.org/W12-2503/›. [20231211].

Webster, Rebecca; Fonteyne, Margot; Tezcan, Arda; Macken, Lieve; Daems, Joke. (2020). Gutenberg Goes Neural: Comparing Features of Dutch Human Translations with Raw Neural Machine Translation Outputs in a Corpus of English Literary Classics. *Informatics,* v. 7, n. 3, 32 p. <https://doi.org/10.3390/informatics7030032>. [20231211].

Wu, Yonghui; *et al.* (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv online.* <https://doi.org/10.48550/arXiv.1609.08144>. [20231211].

Annex A

		DeepL 2020	DeepL 2021	DeepL 2022
1	, asked him	, vroeg hem	, vroeg hem	, vroeg hem
2	towaway zones	Sleepzones	Wegsleepzones	wegsleepzones
3	and I didn't like it	ik vond het niet leuk	Ik vond dat niet leuk	Ik vond dat niet leuk
4	Cost me ..., plus	Het kostte me... , plus...	Het kostte me..., plus...	Het kostte me..., plus...
5	tiny little cars	kleine auto's	kleine auto's	kleine auto's
6	You ever notice ...? You ever seen...? No you haven't.	Is het je ooit opgevallen ... ? Heb je ze ooit ... zien? Nee, dat heb je niet gezien.	Is het je ooit opgevallen ...? Heb je ze ooit...? Nee, dat heb je niet.	Is het je ooit opgevallen ...? Heb je ze ooit...? Nee, dat heb je niet.
7	, and to keep some mental health warm for me who needed it ...,	, en om wat geestelijke gezondheid warm te houden voor mij die het nodig had,	, en om wat geestelijke gezondheid warm te houden voor mij die het nodig had,	, en om wat geestelijke gezondheid warm te houden voor mij die het nodig had,
8	a bucket of ribs	een emmer met ribben	een emmer ribbetjes	een emmer ribbetjes
9	Which I would gladly carry up there...	Wat ik graag voor hem zou doen	Die ik graag naar hem toe zou brengen	Die ik graag naar hem toe zou brengen
10	I cabled him	Ik belde hem	Ik telegrafeerde hem	Ik telegrafeerde hem
11	and, by the way, what was the apartment situation up there?	, en trouwens, wat was de situatie in het appartement daarboven?	, en, tussen haakjes, hoe was de woonsituatie daar?	, en, tussen haakjes, hoe was de woonsituatie daar?
12	It was bad,	Het was slecht.	Het was slecht.	Het was slecht.

		DeepL 2020	DeepL 2021	DeepL 2022
13	he replied by platitudinum plate	antwoordde hij per platitudinaplaat	antwoordde hij per platitudinaal bord,	antwoordde hij per platitudinaal bord,
14	but what could he do?	maar wat kon hij doen?	maar wat kon hij doen?	maar wat kon hij doen?
15	root cellar	wortelkelder	kelder	kelder
16	'cause of me being a friend of the moon.	omdat ik een vriend van de maan ben	omdat ik een vriend van de maan ben	omdat ik een vriend van de maan ben
17	pretty nice place	een mooie plek	een mooie plek	een mooie plek
18	the Space Shuttle Hurry-Up Fund	Space Shuttle Hurry-Up Fund	Space Shuttle Hurry-Up Fonds	Space Shuttle Hurry-Up Fonds
19	Drumming fiercely on a hollow log with a longitudinal slit tuned to moon frequencies	Ik trommelde hevig op een holle boomstam met een longitudinale spleet, afgestemd op de maanfrequenties	Ik trommelde hevig op een holle boomstam met een spleet in de lengterichting, afgestemd op de maanfrequenties	Ik trommelde hevig op een holle boomstam met een spleet in de lengterichting, afgestemd op de maanfrequenties
20	employment, medical coverage, retirement benefits, tax shelterage, convenience cards, and Christmas Club accounts	werkgelegenheid, medische dekking, pensioenuitkeringen, belastingopvang, gemakskrantenkaarten en de rekeningen van de kerstclub	werk, medische dekking, pensioenuitkeringen, belastingaftrek, gemakskaarten en kerstclubrekeningen	werk, medische dekking, pensioenuitkeringen, belastingaftrek, gemakskaarten en kerstclubrekeningen
21	That's a roger,	Dat is een roger	Dat is een groot voordeel	Dat is een groot voordeel

		DeepL 2020	DeepL 2021	DeepL 2022
22	he moonbeamed back	hij straalde terug.	maant hij terug.	maant hij terug.
23	by means of curly little ALGOL circuits I had knitted myself on my Apple computer	door middel van gekrulde kleine ALGOL-circuits die ik zelf had gebreid op mijn Apple-computer	door middel van krullerige kleine ALGOL schakelingen die ik zelf had gebreid op mijn Apple computer	door middel van krullerige kleine ALGOL schakelingen die ik zelf had gebreid op mijn Apple computer
24	that ticktacktoe was about as far as they'd got in that direction	dat ticktacktoe ongeveer zo ver was als ze die kant op waren gegaan	dat ticktacktoe ongeveer zo ver was als ze in die richting waren gekomen	dat ticktacktoe ongeveer zo ver was als ze in die richting waren gekomen
25	via flights of angels with special instructions	via vluchten van engelen met speciale instructies	via vluchten van engelen met speciale instructies	via vluchten van engelen met speciale instructies
26	it looked to me like he had things pretty well in hand up there	dat het er voor mij op leek dat hij de dingen daarboven vrij goed in de hand had	dat het er volgens mij op leek dat hij de zaken daarboven goed onder controle had	dat het er volgens mij op leek dat hij de zaken daarboven goed onder controle had
27	Part-time if need be?	Deeltijd als dat nodig is?	Desnoods parttime?	Desnoods parttime?
28	a shower of used-car asteroids with blue-and-green bumper stickers	in een douche van tweedehands-auto asteroïden met blauw-groene bumperstickers	in een regen van gebruikte auto asteroïden met blauw-groene bumperstickers	in een regen van gebruikte auto asteroïden met blauw-groene bumperstickers

		SYSTRAN 2020	SYSTRAN 2021	SYSTRAN 2022
1	, asked him	waarin ik hem vroeg	waarin ik hem vroeg	, vroeg hem
2	towaway zones	daar naar toe moesten	daar naar toe moesten	daarboven wegzones
3	and I didn't like it	kleine auto's	kleine auto's	kleine auto's
4	Cost me ..., plus	en ik vond het niet leuk.	en ik vond het niet leuk.	en ik vond het niet leuk.
5	tiny little cars	Ik heb ... gekost ..., plus...	Ik heb ... gekost ..., plus...	Het kostte me..., plus...
6	You ever notice ...? You ever seen...? No you haven't.	Heb je ooit gemerkt...? Heb je ze ooit ...? Nee, dat heb je niet.	Heb je ooit gemerkt...? Heb je ze ooit ...? Nee, dat heb je niet.	Heb je ooit gemerkt...? Heb je ze ooit ...? Nee, dat heb je niet.
7	, and to keep some mental health warm for me who needed it	, en om een geestelijke gezondheid warm te houden voor wie het nodig had	, en om een geestelijke gezondheid warm te houden voor wie het nodig had	, en dat ik een beetje geestelijke gezondheid warm zou houden voor wie het nodig had
8	a bucket of ribs	een emmer van ribben	een emmer van ribben	een emmer ribben
9	Which I would gladly carry up there...	Welke zou ik graag naar hem doorzetten	Welke zou ik graag naar hem toe zetten	Wat ik daar graag mee zou willen doen
10	I cabled him	Ik gaf hem de opdracht	Ik gaf hem de opdracht	Ik telegrafeerde hem
11	and, by the way, what was the apartment situation up there?	, en trouwens, wat was de appartementsituatie daar?	, en trouwens, wat was de appartementsituatie daar?	en trouwens, tussen haakjes, wat was de appartementsituatiesituatie daar in het appartement

		SYSTRAN 2020	SYSTRAN 2021	SYSTRAN 2022
12	It was bad,	Het was slecht,	Het was slecht,	Het was erg,
13	he replied by platitudinum plate	hij antwoordde op de plaat van platitudinum,	hij antwoordde op de plaat van platitudinum,	hij antwoordde met een platitudinumplaat,
14	but what could he do?	maar wat kon hij doen?	maar wat kon hij doen?	maar wat kon hij doen
15	root cellar	wortelkelder	wortelkelder	wortelkelder
16	'cause of me being a friend of the moon.	want ik ben een vriend van de maan	want ik ben een vriend van de maan	omdat ik een vriend van de maan was
17	pretty nice place	mooie plek	mooie plek	mooie plek
18	the Space Shuttle Hurry-Up Fund	Space Shuttle-Up Fonds	Space Shuttle-Up Fonds	Space Shuttle Hurry-Up Fund
19	Drumming fiercely on a hollow log with a longitudinal slit tuned to moon frequencies	Ik trommelde hard op een holle log met een langsscheepse snede op maanfrequenties.	Ik trommelde hard op een holle log met een langsscheepse snede op maanfrequenties.	Ik drumde heftig op een hol logboek met een longitudinale slit afgestemd op de maan frequenties.

		SYSTRAN 2020	SYSTRAN 2021	SYSTRAN 2022
20	employment, medical coverage, retirement benefits, tax shelterage, convenience cards, and Christmas Club accounts	werk, medische dekking, pensioenuitkeringen, belastingonderdak, goedkope kaarten en kerstclubs	werk, medische dekking, pensioenuitkeringen, belastingonderdak, goedkope kaarten en kerstclubs.	werkgelegenheid, medische dekking, pensioenuitkeringen, belastingbescherming, goedkope kaarten, en de rekeningen van de Kerstclub.
21	That's a roger,	Dat is een roger,	Dat is een roger.	Dat is een roger.
22	he moonbeamed back	hij heeft zich teruggeschroefd.	hij heeft zich teruggeschroefd.	hij heeft weer gestraald.
23	by means of curly little ALGOL circuits I had knitted myself on my Apple computer	naar krullende ALGOL circuits die ik op mijn Apple computer had gekend	naar krullende ALGOL circuits die ik op mijn Apple computer had gekend.	door middel van krullende kleine ALGOL circuits die ik zelf had gebreid op mijn Apple computer
24	that ticktacktoe was about as far as they'd got in that direction	dat de tikkapper ongeveer zo ver ging als ze in die richting hadden gestaan	dat de tikkapper ongeveer zo ver ging als ze in die richting hadden gestaan.	dat de ticktacktoe ongeveer zo ver ging als ze in die richting waren gekomen
25	via flights of angels with special instructions	via engelenvluchten met speciale instructies	via engelenvluchten met speciale instructies	vluchten van engelen met speciale instructies

		SYSTRAN 2020	SYSTRAN 2021	SYSTRAN 2022
26	it looked to me like he had things pretty well in hand up there	dat het naar mij leek alsof hij daar behoorlijk goed in de hand had	dat het naar mij leek alsof hij daar behoorlijk goed in de hand had	dat het leek alsof hij daar vrij goed in de hand had
27	Part-time if need be?	Deeltijd indien nodig?	Deeltijd indien nodig?	Deeltijd indien nodig?
28	a shower of used-car asteroids with blue-and-green bumper stickers	in een douche van asteroïden uit tweedehands auto's met blauw-en-groene bumperstickers	in een douche van asteroïden uit tweedehands auto's met blauw-en-groene bumperstickers	in een douche van tweedehands auto asteroïden met blauw-en-groen bumperstickers

		GOOGLE NMT 2020	GOOGLE NMT 2021	GOOGLE NMT 2022
1	, asked him	vroeg hem	, vroeg hem	, vroeg hem
2	towaway zones	weggedoken zones	xleepzones	xleepzones
3	and I didn't like it	dat beviel me niet	ik vond het niet leuk	ik vond het niet leuk
4	Cost me ..., plus	Kostte me... plus	Kostte me... plus	Kostte me... plus
5	tiny little cars	kleine autootjes	klleine autootjes	kleine autootjes
6	You ever notice ...? You ever seen...? No you haven't.	Heb je ooit opgemerkt ...? Heb je ze ooit ...? Nee, dat heb je niet.	Is het je ooit opgevallen ...? Heb je ze ooit ...? Nee, dat heb je niet.	Is het je ooit opgevallen ...? Heb je ze ooit ...? Nee, dat heb je niet.
7	, and to keep some mental health warm for me who needed it	en om wat geestelijke gezondheid warm te houden voor mij die het nodig had	en om wat geestelijke gezondheid warm te houden voor mij die het nodig had	en om wat geestelijke gezondheid warm te houden voor mij die het nodig had
8	a bucket of ribs	een emmer ribben	een emmer spareribs	een emmer spareribs
9	Which I would gladly carry up there...	Wat zou ik hem graag verder vertellen	Wat ik graag bij hem zou voortzetten	Wat ik graag bij hem zou voortzetten
10	I cabled him	Ik heb hem bekabeld	Ik telefoneerde hem	Ik telefoneerde hem
11	and, by the way, what was the apartment situation up there?	wat was de situatie daarboven	hoe was de situatie in het appartement daarboven	hoe was de situatie in het appartement daarboven
12	It was bad,	Het was slecht.	Het was slecht.	Het was slecht.

		GOOGLE NMT 2020	GOOGLE NMT 2021	GOOGLE NMT 2022
13	he replied by platitudinum plate	antwoordde hij door platitudinum plaat.	antwoordde hij met een platitudinum-plaat.	antwoordde hij met een platitudinum-plaat.
14	but what could he do?	maar wat kon hij doen?	maar wat kon hij doen?	maar wat kon hij doen?
15	root cellar	wortelkelder	kelder	kelder
16	'cause of me being a friend of the moon.	omdat ik een vriend van de maan ben	omdat ik een vriend van de maan ben	omdat ik een vriend van de maan ben
17	pretty nice place	een mooie plek	een mooie plek	een mooie plek
18	the Space Shuttle Hurry-Up Fund	Space Shuttle Hurry-Up Fund	Space Shuttle Haast-Up Fund	Space Shuttle Haast-Up Fund
19	Drumming fiercely on a hollow log with a longitudinal slit tuned to moon frequencies	Hevig trommelend op een holle stam met een longitudinale sleuf afgestemd op de maanfrequenties.	Hevig trommelend op een holle boomstam met een spleet in de lengterichting afgestemd op maanfrequenties,	Hevig trommelend op een holle boomstam met een spleet in de lengterichting afgestemd op maanfrequenties,
20	employment, medical coverage, retirement benefits, tax shelterage, convenience cards, and Christmas Club accounts	werk, medische dekking, pensioenuitkeringen, belastingvrijstelling, gemakskaarten en Christmas Club-accounts.	werk, medische dekking, pensioenuitkeringen, fiscale onderdak, gemakskaarten en kerstclubrekeningen	werk, medische dekking, pensioenuitkeringen, fiscale onderdak, gemakskaarten en kerstclubrekeningen
21	That's a roger,	Dat is een roger.	Dat is een roger.	Dat is een roger.

		GOOGLE NMT 2020	GOOGLE NMT 2021	GOOGLE NMT 2022
22	he moonbeamed back	hij straalde terug.	straalde hij terug.	straalde hij terug.
23	by means of curly little ALGOL circuits I had knitted myself on my Apple computer	door middel van gekrulde kleine ALGOL-circuits die ik zelf op mijn Apple-computer had gebreid	door middel van krullende kleine ALGOL-circuits die ik zelf had gebreid op mijn Apple-computer	door middel van krullende kleine ALGOL-circuits die ik zelf had gebreid op mijn Apple-computer
24	that ticktacktoe was about as far as they'd got in that direction	dat ticktacktoe ongeveer zover was als ze in die richting waren gekomen.	dat tiktakteen ongeveer zo ver was als ze in die richting waren gekomen.	dat tiktakteen ongeveer zo ver was als ze in die richting waren gekomen.
25	via flights of angels with special instructions	via vluchten van engelen met speciale instructies	via engelenvluchten met speciale instructies	via engelenvluchten met speciale instructies
26	it looked to me like he had things pretty well in hand up there	dat het mij leek alsof hij het daar redelijk goed voor elkaar had	dat het me leek alsof hij de zaken daar behoorlijk goed in de hand had	dat het me leek alsof hij de zaken daar vrij goed in de hand had
27	Part-time if need be?	Parttime indien nodig?	Eventueel parttime?	Eventueel parttime?
28	a shower of used-car asteroids with blue-and-green bumper stickers	in een regen van asteroïden voor gebruikte auto's met blauw-en-groene bumperstickers)	in een regen van gebruikte asteroïden met blauwe en groene bumperstickers	in een regen van gebruikte asteroïden met blauwe en groene bumperstickers

		CUSTOM LTM
1	, asked him	en vroeg hem
2	towaway zones	daarsleepzones
3	and I didn't like it	dat stond me niet aan
4	Cost me ..., plus	Het kostte me ..., plus ...
5	tiny little cars	Kleine autootjes
6	You ever notice ...? You ever seen...? No you haven't.	Heb je ooit gezien ... ? Heb je ze ooit ... ? Nee, dat heb je niet .
7	, and to keep some mental health warm ..., and could I interest him...	om een geestelijke gezondheid warm te houden voor mij die dat nodig had
8	a bucket of ribs	een emmer ribbetjes
9	Which I would gladly carry up there...	Wat ik graag naar hem toe zou brengen
10	I cabled him	Ik telegrafeerde hem
11	and, by the way, what was the apartment situation up there?	en trouwens, hoe was de flat daar ?
12	It was bad,	Het was erg ,
13	he replied by platitudinum plate	antwoordde hij op een platitudinium plaat,
14	but what could he do?	maar wat kon hij doen?
15	root cellar	voorraadkelder
16	'cause of me being a friend of the moon.	omdat ik een vriend van de maan was
17	pretty nice place	een mooie plek
18	the Space Shuttle Hurry-Up Fund	Space Shuttle Hurry-Up Fund
19	Drumming fiercely on a hollow log with a longitudinal	Heftig trommelend op een holle boomstam met een gleuf in de lengte die was afgestemd op de maanfrequentie,

		CUSTOM LTM
	slit tuned to moon frequencies	
20	employment, medical coverage, retirement benefits, tax shelterage, convenience cards, and Christmas Club accounts	werk, ziektekostenverzekering , pensioenuitkeringen, belastingheffing , convenience cards en de rekeningen van de Christmas Club
21	That's a roger,	Dat is een roeger,
22	he moonbeamed back	kaatste hij terug,
23	by means of curly little ALGOL circuits I had knitted myself on my Apple computer	aan de hand van krullerige ALGOL-circuitjes die ik zelf op mijn Apple-computer had gebreid
24	that ticktacktoe was about as far as they'd got in that direction	dat ze niet verder in die richting waren gekomen
25	via flights of angels with special instructions	via engelenvluchten met speciale instructies
26	it looked to me like he had things pretty well in hand up there	goed voor elkaar had
27	Part-time if need be?	Parttime, als dat nodig is?
28	a shower of used-car asteroids with blue-and-green bumper stickers	In een regen van tweedehands asteroïden met blauwgroene bumperstickers