Machine Translation (2021) 35:167–203
https://doi.org/10.1007/s10590-021-09266-0

An in-depth analysis of the individual impact of controlled language rules on machine translation output: a mixed-methods approach

Shaimaa Marzouk[1]

Received: 15 November 2019 / Accepted: 4 May 2021 / Published online: 11 June 2021

Abstract

Examining the general impact of Controlled Language (CL) rules in the context of Machine Translation (MT) has been an area of research for many years. The present study focuses on the following question: how do CL rules impact MT output *individually*? By analysing a German corpus-based test suite of technical texts that have been translated into English by different MT systems, this study endeavours to answer this question at different levels: the general impact of CL rules (rule- and system-independent), their impact at rule level (system-independent) as well as at rule and system level. The results of five MT systems are analysed and contrasted: a rule-based system, a statistical system, two differently constructed hybrid systems, and a neural system. For this, a mixed-methods triangulation approach that includes error annotation, human evaluation, and automatic evaluation was applied. The data was analysed both qualitatively and quantitatively in terms of CL influence on the following parameters: number and type of MT errors, style and content quality, and scores of two automatic evaluation metrics. In line with many studies, the results show a general positive impact of the applied CL rules on the MT output. However, at rule level, only four rules proved to have positive effects on the aforementioned parameters; three rules had negative effects on the parameters; and two rules did not show any significant impact. At rule and system level, the rules affected the MT systems differently, as expected. Rules that had a positive impact on earlier MT approaches did not show the same impact on the neural MT approach. Furthermore, neural MT delivered distinctly better results than earlier MT approaches, namely the highest error-free, style and content quality rates both before and after applying the rules, which indicates that neural MT offers a promising solution that no longer requires CL rules for improving the MT output.

Keywords Controlled language · Machine translation · Machine translation evaluation · Translation quality

✉ Shaimaa Marzouk
s.marzouk@uni-mainz.de

Extended author information available on the last page of the article

1 Introduction

Applying Controlled Language (CL) is a common pre-editing technique in the technical domain. As early as 1974, the Caterpillar Fundamental English CL was specifically developed to improve the comprehensibility and translatability of technical documentation (Caterpillar 1974). A CL refers to "an explicitly defined restriction of a natural language that specifies constraints on lexicon, grammar, and style" (Huijsen 1998). Through the different restrictions imposed by the CL, applying CL rules allows for the reduction of sentence length, the avoidance of complex sentence structures as well as the elimination of ambiguous vocabulary and constructions. Several studies found that the application of CL has a positive impact on different aspects of MT. As one of the earlier CLs, Caterpillar Technical English proved to have a significant positive effect on MT productivity (Kamprath et al. 1998). Reuther (2003) further concluded that implementing CL rules can improve readability and translatability of machine-translated texts. Another study found that the level of controlling the source text has a substantial impact on the accuracy of MT output (Nyberg and Mitamura 1996). In addition, more recent studies examined the impact of CL on post-editing effort. O'Brien (2006) found that CL reduces post-editing time. Another positive impact was linked to post-editing productivity (Aikawa et al. 2007). Bernth and Gdaniec (2001) introduced 26 rules for English as a source language that address different text characteristics aimed to increase machine translatability. They tested these rules with various commercially available MT systems and claimed that they were generalizable to different MT systems and language pairs.

Most CL studies have investigated the impact of CL on MT from a holistic perspective, i.e. the impact of comprehensive CL rule sets (e.g. Holmback et al. 1996; Spyridakis et al. 1997; Kamprath et al. 1998; Bernth 1999; Nyberg et al. 2003; Fiederer and O'Brien 2009). The results of this research provide an overall picture of the effect of CL, in which a positive effect of some rules may overshadow a negative effect of other rules, which leads to a biased end result. There have been a limited number of studies that focused on analysing the influence of individual CL rules. The results of these studies (O'Brien 2006; Roturier 2006; Roturier et al. 2012) showed that CL rules affected the MT in various ways and to different degrees. All of these studies were conducted on CL rules of the English language. Roturier et al. (2012) analyzed the impact of the CL rules on MT quality using automatic translation metrics. The experiments were conducted using a phrase-based system (Moses: Koehn et al. (2007)) for the target languages French and German. Roturier (2006) also focused on the same target languages, but he used a rule-based MT system (Systran) and was interested in analysing the impact of CL rules on comprehensibility. O'Brien (2006) examined the impact of CL on post-editing effort for German as a target language using a rule-based MT system (IBM WebSphere). In a recent study, Marzouk and Hansen-Schirra (2019) analysed the impact of a number of CL rules on different MT approaches for the language pair German-to-English. Comparing the results of rule-based (RBMT), statistical (SMT), hybrid (HMT), and neural (NMT) systems, they found that the earlier approaches (RBMT, SMT, and HMT) benefited from the CL rules, while the NMT system delivered mostly error-free output both before and after the application of the rules, but did show a decrease in quality after applying the rules. This paper reports on the same study of Marzouk and Hansen-Schirra (2019) shedding more light on the CL rules analysed and providing detailed insight into their individual impact on the different approaches.

Given that the MT quality differs depending on the language pair, translation direction, domain, and the applied MT system, the impact of each CL rule on the MT output should vary in accordance with these variables. Against this background, the present study focuses on the technical domain and on one language pair, German-to-English. The technical domain is the most common field of application of CL. Analysing the language pair German-to-English enables exploring CL rules of the German language; a language that has rarely been examined in CL research. In doing this the study has kept the variables language pair, translation direction, and domain constant in order to analyse and contrast the individual impact of nine CL rules on the MT output of four MT approaches (RBMT, SMT, HMT, and NMT). Exploring the NMT approach in the context of CL has—to the best of my knowledge—not yet been investigated in previous studies. The analysis was conducted at different levels: rule- and system-independent (general impact), at rule level (system-independent), at system level as well as both at rule and system level. The analysis at system level is presented in Marzouk and Hansen-Schirra (2019), while this paper covers the other analysis levels.

Identifying the individual impact of CL rules under the different systems allows for an effective implementation of those that have a positive impact. This in turn would help limit potential drawbacks of CL application (Lehrndorfer and Reuther 2008; Drewer and Ziegler 2014) such as the additional time both authors and translators need in order to consider the restrictions imposed by the rules, interruption of the writing flow, difficulty of implementing linguistically complex rules when the authors are domain experts with a limited linguistic background, and restriction of the authors' creativity.

The next section provides a description of the dataset. Section 3 outlines the applied methodology. Section 4 presents the results. The findings are summarised in the conclusion. Finally, the study limitations as well as ideas for future research are provided.[1]

2 Dataset

A test suite was created that consists of 216 source sentences (24 sentences per CL rule), extracted from a corpus of ten German user manuals for appliances, software, and machines. These sentences violated one of the nine analysed CL rules (*before CL* version). The CL rules were applied to each sentence (*after CL* version). Both

[1] This paper presents some of the results of a PhD thesis; the dissertation will be published soon (Marzouk in press).

versions were translated into English by five MT systems. Accordingly, the dataset consisted of 2160 MT sentences (216 source sentences * 2 versions * 5 systems). The entire dataset was analysed applying error annotation. 1100 MT sentences of the 2160 were evaluated by humans. Reasoning and selection criteria of the human-evaluated sentences are detailed under Human Evaluation in the Methodology section.

The source sentences were extracted using the CLAT CL checker. CLAT (Rösener 2010) is one of the most well-known CL checkers in Germany that has been developed by the Society for the Promotion of Applied Information Sciences (IAI) at Saarland University. Thanks to research cooperation with the IAI,[2] a license for CLAT was provided for research purposes in the present study. One consideration while creating the test suite was to have all user manuals represented as balanced as possible.

The rules investigated were taken from the tekom e. V. Guidelines for Technical Writing in the German Language "Leitlinie—Regelbasiertes Schreiben—Deutsch für die Technische Kommunikation" (tekom 2013). The tekom guidelines are widely implemented in Germany, both in research and industry. Thanks to close collaboration between academia, industry, service providers and software companies, the tekom rules provide a comprehensive set of rules across all language and documentation levels (tekom 2013). The nine rules analysed were selected based on three criteria: (i) rules that can be applied to just one sentence, as the analysis is conducted at sentence level; (ii) rules that can be applied in all respective sentences according to one fixed pattern (see Table 1 "How the rules were applied") in order to limit the number of independent variables; and (iii) rules for which the so-called CL position can be defined. Since different factors jointly influence the entire MT output at the same time (e.g. the MT system approach, training data together with the application or non-application of the CL rule etc.), it was necessary for the analysis to only focus on the word or word group directly related to the CL rule, referred to as the *CL position*. The CL position was defined as the part of the source sentence that has to be modified in order to apply the CL rule and its equivalence in the target sentence. A rule like "formulate short sentences" refers to the whole sentence. Therefore, it is not possible to define a specific CL position that can be analysed and compared in the error annotation or by human evaluators. Based on these criteria, the nine rules depicted in Table 1 were analysed.

The five MT systems examined were: the hybrid MT system "Bing" by Microsoft,[3] the neural MT system "Google Translate" (Wu et al. 2016),[4] the rule-based MT system "Lucy LT KWIK Translator" (cf. Alonso Martín and Serra 2014),[5] the statistical MT system "SDL Free Translation",[6] and another hybrid MT system "Systran".[7] Since hybrid systems are structured differently, the study uses two

[2] http://www.iai-sb.de/de/produkte/clat.

[3] https://www.bing.com/translator/.

[4] https://translate.google.de/.

[5] http://www.lucysoftware.com/english/machine-translation/lucy-lt-kwik-translator-/.

[6] https://www.freetranslation.com/de/.

[7] http://www.systranet.com/translate.

Table 1 Analysed CL rules and their application pattern

Analysed rules		How the rules were applied
Rule 1 (anz)	Using straight quotes for interface texts	By entering the interface text between straight quotes, see Example 1
Rule 2 (fvg)	Avoiding light-verb construction (Funktionsverbgefüge)	By using the meaning-bearing verb instead of the light verb construction, see Example 2
Rule 3 (kos)	Formulating conditions as 'if' sentences	By starting the conditional sentence by "if" instead of the verb, see Example 3
Rule 4 (nsp)	Using unambiguous pronominal references	By replacing the pronoun by its pronominal reference, see Example 4
Rule 5 (pak)	Avoiding participial constructions	By generating a subordinate clause based on the participial construction, see Example 5a and 5b
Rule 6 (pas)	Avoiding passives	By using the active voice, see Example 6
Rule 7 (per)	Avoiding the construction sein + zu + infinitive	By using the imperative instead of the construction sein + zu + infinitive, see Example 7
Rule 8 (prä)	Avoiding superfluous prefixes	By eliminating the superfluous prefix, see Example 8
Rule 9 (wte)	Avoiding omitting parts of the words	Missing parts of words were completed, see Example 9

hybrid systems: Bing is a statistical MT system with language-specific rule components, while Systran was originally a rule-based system and was later further developed into a hybrid system (cf. Werthmann and Witt 2014, p. 84). Accordingly, both yielded different outputs. The selection criteria of the systems were as follows: (i) to be an online freely available system, (ii) to offer the language pair German-to-English, and (iii) to cover different MT approaches. The systems examined are therefore generic black-box systems. A black-box system is "a system which has been trained and tuned a priori and for which we cannot access the model parameters or training data for fine-tuning or improvements" (Mehta et al. 2020, p. 1). Accordingly, the systems were not trained in advance with specific relevant corpora. Such training would have an impact on the results (i.e. better results in the controlled scenario if the corpora were controlled and vice versa) (cf. Reuther 2003). In addition, a reasonable comparison of the results of the different systems would not be feasible, as the corpus-based systems would have—depending on the degree of control—an advantage or disadvantage over the rule-based system. In order to overcome the expected difficulty of machine translating company-specific and specialist terms, such terms were replaced with common terms. A detailed description of the corpus-based test suite and its preparation steps is provided in Marzouk and Hansen-Schirra (2019). The dataset was machine-translated at the end of 2016.

3 Methodology

A triphasic mixed-methods triangulation approach was applied that incorporates three evaluation methods: error annotation, human evaluation, and automatic evaluation. These methods were carried out in the order shown below.

3.1 Error annotation

The goal of the error annotation is to identify the MT errors before and after applying the CL rules and compare them in terms of their number and type. The annotation was conducted by a qualified experienced German-English translator and checked by two professional German-English translators. Due to the large number of MT sentences (2160 sentences), each evaluator separately checked different halves of the set of sentences. Each evaluator had to indicate whether they agreed with the annotation or not. If not, they had to reannotate the translation. The percentage of reannotated sentences was 27% by the first evaluator and 31% by the second evaluator. In case of reannotation, the other evaluator checked both annotations and chose one.

Furthermore, based on the existence or non-existence of errors within the CL position, the data was divided into four groups, referred to as *annotation groups*.

These are: FF (for False–False)—translation contains error before and after CL; FR (for False-Right)—translation contains error only before CL; RF (for Right-False)—translation contains error only after CL; RR (for Right–Right): no errors before and after CL. Table 2 shows the error classification (cf. Vilar et al. 2006) applied.

The error taxonomy of Vilar et al. (2006) was used as a basis of the error annotation due to its explicitness, integrity and appropriate degree of granularity. However, further more extensive taxonomies, such as the Multidimensional Quality Metrics (MQM) framework (Lommel 2018) can be also used for the analysis. This would be particularly useful in case of examining fine-grained or more specific types of errors.

For the comparison of the number of errors before vs. after CL, the significance test Wilcoxon was used, since the analysed variables were ordinal. For the comparison of the error types before CL vs after CL, the McNemar significance test was used, which is designed for related dichotomous variables. A significant difference is realised at $p < 0.05$.

3.2 Human evaluation

The goal of the human evaluation is to compare the content and style quality of the MT within the CL position (not the quality of the entire sentence). Following the quality definition of Hutchins and Somers (1992):

- The *content quality* is the extent to which the translation reflects the information in the source text accurately; and the extent to which the translation is easy to understand. (ibid.)
- The *style quality* is the extent to which the translation sounds natural and idiomatic in Standard Written English, is appropriate for the intention of its content (ibid., Fiederer and O'Brien 2009) as well as presented clearly in terms of orthography. The definition covers orthography as an instrument for presenting the content in an adequate way that serves its intention.

Based on these definitions, the content quality covers the criteria accuracy and clarity; the style quality encompasses the criteria idiomaticity, appropriateness to the content intention as well as correctness and clarity of the orthographic presentation.

As the human evaluation aimed to compare the content and style quality of the MT within the CL position (not the quality of the entire sentence), it was necessary to initially correct all errors outside the CL position by applying the fewest possible edits to the MT output. This preliminary step was essential in order to keep the evaluators focused on the CL position. Otherwise, the participants would have evaluated the entire MT, commenting on all errors, although not all errors are related to the CL rule. As a result of this step, both versions of the MT (before and after applying the CL rule) were identical except for the CL position (see examples provided in the Results section).

Such corrections were not possible for all annotated sentences without affecting the CL position. Therefore, the study defined certain criteria that an MT has to fulfil in order to be included for human evaluation: For the MT sentences that *contained* errors within the CL position, only MT sentences with a maximum of *two* wrong words were included in the human evaluation. For the MT sentences that did *not* contain errors within the CL position, only MT sentences with a maximum of *three* wrong words were included in the human evaluation. The goal of setting these criteria was to avoid making too many corrections in the MT that may impact the evaluation of the CL position. For example, in the rule "Avoiding the construction sein + zu + infinitive", the sentence "Das Kaufdatum *ist* durch eine Kaufquittung *zu belegen*" was translated by Bing as "By a purchase receipt *to prove* [THE][8] date of purchase". In addition to the wrong translation of the CL position "ist zu belegen", the MT included other errors outside the CL position. Obviously, correcting these errors would have substantially changed the MT. The total number of excluded MT sentences according to these criteria was 595 (Excluded.1).

To assure the idiomaticity of the MT sentences after correcting the errors outside the CL position, two professional translators checked the MT output for stylistic acceptability. The focus of this acceptability check was to ensure that the MT outside the CL position remained stylistically acceptable after applying the fewest possible edits. If an MT sentence was evaluated as unacceptable by both translators, the MT sentence was excluded from the human evaluation. However, exclusion cases due to stylistic non-acceptance were rare; only 15 sentences (Excluded.2).

After undertaking these steps, the total number of MT sentences excluded from the human evaluation was 610 MT (sum of Excluded.1 + Excluded.2), see Table 3. The remaining 1550 (out of 2160) MT sentences included a total of 545 MT sentences that were identical across different systems, i.e. the source sentences were identically translated by different systems. It would have been ineffective to ask to the participants to evaluate 545 identical sentences. Therefore, for each source sentence, only one instance of the identical MT sentences was human-evaluated; 95 instances out of 545 are included in the 1100 human-evaluated MT sentences. The human evaluation scores of these 95 instances were then applied to the other repeated instances (450 out of 545). Thus, the results reported below are based on the total number of sentences of 1550 MT sentences (1100 + 450).

In the next step, the researcher verified whether the annotation groups within the 1550 MT sentences were comparable in the error annotation and human evaluation (Table 4). For example, in the error annotation 44.5% of MT sentences were error-free both before and after the CL application (group RR); in the human evaluation the analysed percentage of the group RR was comparable (44.2%).

The distribution of the 1550 MT sentences of the human evaluation across the MT systems is shown in Table 5:

The human evaluation (Fig. 1) consisted of:

- Evaluating the style and content quality of the MT (see (*) in Fig. 1) on two 5-point Likert scales ((1) in Fig. 1);
- Selecting the relevant quality criteria that justify the assigned quality scores: accuracy and clarity under the content quality; idiomaticity, appropriateness to

[8] [THE] was not included in the MT.

the content intention as well as correctness and clarity of the orthographic presentation under the style quality ((2) in Fig. 1);

- Providing the word or part of the translation relevant to each chosen criterion ((3) in Fig. 1);
- If many modifications were necessary, the participant had to enter an alternative translation for the whole sentence ((4) in Fig. 1).

Regarding the participants, different studies recommend recruiting more than 3–4 participants (Fiederer and O'Brien 2009). In this study, five participants initially carried out the tests and the number of participants was successively increased until the accumulated average of the quality values stabilised. After the eighth participant was added, the accumulated quality averages remained largely unchanged. Accordingly, the number of participants was not increased further. The participants are native English speakers and hold a bachelor's degree in translation. In addition, all participants were students in the last or penultimate semester of a master's degree program in translation. Each participant had to evaluate the entire set of 1100 MT sentences. Participation was remunerated.

Regarding the test procedure, the 1100 MT sentences were randomised and split into 44 tests. Each participant had the opportunity to choose whether to rate one, two or three tests per day, depending on his or her availability. The basic requirement was to evaluate at least one test daily, thus avoiding interruptions that could possibly have a negative effect on the intra-rater agreement. In addition, the participants were asked to take a break between the tests. The 44 tests were sent in a different randomised order to the participants, e.g. the 1st participant received test 40, test 8, test 5 consecutively. A decreasing motivation over a 3–4-week evaluation period is unavoidable. Therefore, this randomisation ensured that no particular sentences were evaluated by all participants at the end of the evaluation. The tester received the completed tests every day and checked them for completeness (i.e. all sentences were rated and commented if necessary). In case of any missing data, the participant was asked to complete them, then he or she received the new tests for the next day.

For the comparison of the style and content quality before vs. after CL, the Wilcoxon test was used, as not all quality variables were normally distributed. In order to measure the correlation between the error types and the quality, the Spearman correlation test was used because one of the analysed variables was ordinal.

3.3 Automatic evaluation

The alternative translation obtained from the human evaluation acted as a reference translation for the automatic evaluation metrics (AEMs) in order to compare their scores before and after applying each CL rule. Two reference translations per sentence were randomly selected for the comparison. The study applied the TERbase and hLEPOR evaluation metrics. The former is a basic edit distance metric that calculates the minimum number of edits needed to change the evaluated MT so that it exactly matches the reference translation and works without stemming, synonymy lookup and paraphrase support (Snover et al. 2006, Gonzàlez and Giménez 2014).

It was necessary to consider the use of synonyms as an edit, as the participants quite often recommended the use of a certain synonym while evaluating the translation accuracy. TERbase works with negative values; its score ranges between − 1 (worst value) and 0 (best value). At the same time, hLEPOR was applied as one of the advanced metrics that has proven to have a state-of-the-art correlation with human evaluation compared with metrics like BLEU (Papineni et al. 2002), TER (Snover et al. 2006), and METEOR (Banerjee and Lavie 2005) amongst others (Han et al. 2013). The calculation model of hLEPOR is based on three factors: an enhanced length penalty, an N-gram position difference penalty and the harmonic mean of precision and recall (Han et al. 2013). hLEPOR works with positive values; its score ranges between 0 (worst value) and 1 (best value). The impact of the CL application was measured on the basis of the difference of the "mean after CL" *minus* the "mean before CL". Therefore, a positive difference indicates an improvement in the AEM score, and conversely, a negative difference indicates a deterioration in the AEM score.

Finally, using the Spearman correlation test, the study investigated how the difference in the AEMs scores (after CL *minus* before CL) of TERbase and hLEPOR correlates with the difference in the overall quality.[9] The Spearman correlation test was used because not all variables were normally distributed.

4 Results

4.1 The general impact of CL application

The results of the error annotation showed that the number of errors decreased significantly across the rules by 23.5% (z (N = 1080) = − 5.589/p < 0.001). Based on the human evaluation, the style quality (SQ) increased by 1.7% (z (N = 775) = − 2.062/p = 0.039) and the content quality (CQ) improved even more by 2.9% (z (N = 775) = − 4.566/p < 0.001) after applying the rules.[10] With regard to the automatic evaluation, both AEMs scores rose slightly after the application of the rules. Furthermore, the Spearman correlation test showed a significant positive strong correlation between the difference in the overall quality and the differences in the scores of TERbase (ρ (N = 775) = 0.520, p < 0.001) and hLEPOR (ρ (N = 775) = 0.519, p < 0.001), which indicates that an increase in the overall quality (i.e. mean of SQ and CQ) was accompanied by an improvement in the AEMs scores.

Accordingly, the general impact is consistent with the results of previous studies that found that CL application improves MT output (cf. Nyberg and Mitamura 1996; Bernth 1999; Bernth and Gdaniec 2001, p 208; Drewer and Ziegler 2014, p

[9] The overall quality is the mean of the quality of style and quality of content, as analysing the correlation here requires no distinction between the quality parameters.

[10] As mentioned in Sect. 3.2, the results reported on the human evaluation are based on the total number of 1550 MT sentences, which is shown here as N = 775 referring to the comparison of 775 MT sentences of the "*before* CL scenario" with the 775 MT sentences of the "*after* CL scenario".

196). Nonetheless, the different changes in style and content quality after implementing the rules pose the question: which cases specifically displayed a marked increase in content quality over style? This can be answered at rule level.

4.2 The impact of individual CL rules

The analysis of the annotation groups (FF, FR, RF, RR) revealed that, based on the existence and non-existence of MT errors, the CL impact *cannot* be effectively considered positive. The only positive impact can be observed in the FR group (False before CL – Right after CL). This group ranges merely between 8% (rule "pas—Avoiding passives") and 31% (rule "anz—Using straight quotes for interface texts"), Fig. 2.

In the RF annotation group (Right before CL – False after CL), the CL impact is clearly negative. The most dominant annotation groups in all rules were RR and FF. Since the translations were error-free (RR group) or faulty (FF group) both before and after the CL application, a positive impact of a certain rule can only be justified, if the quality values of these two groups increased after rule application. A quality increase in the RR group would mean that the quality of a correct MT after CL is higher (e.g. stylistically better) than that of a correct MT before CL. Similarly, a quality increase in the FF group would imply that comparing two wrong translations before and after CL, the quality of the wrong MT after CL is higher (e.g. includes a less severe error type).

In order to explore quality changes in each annotation group, *a triangulation of the results of the error annotation and human evaluation* was performed. Table 6 summarises how the style and content quality changed after the application of each rule at annotation group level.

Only two CL rules proved to have a positive impact on MT quality:

- The first one is "anz—Using straight quotes for interface texts". This was the only rule in which the SQ and CQ of the FF and RR groups significantly increased after rule application. In addition, the highest percentage of the FR group (31%) and the lowest percentage of the RF group (2%) were represented in this CL rule. This shows a clear positive impact of using straight quotes for interface texts on MT quality.
- The second rule is "per—Avoiding the construction sein + zu + infinitive". Avoiding sein + zu + infinitive (comparable to the structure to be + to + base infinitive) improved the SQ in the FF and RR group significantly, whereas the CQ did not show a significant change. Furthermore, both the SQ and CQ increased in the FR group significantly. Hence, using an imperative (after CL) instead of sein + zu + infinitive (before CL) had a positive impact on the MT output, particularly on the SQ.

Negative impacts of CL rules can be observed in the following three rules:

Example 1 Rule “anz—using straight quotes for interface texts”

Before CL	Wählen Sie danach die Option **Software automatisch installieren**
	Then select the option ***software automatically*** **install**
After CL	Wählen Sie danach die Option **"Software automatisch installieren"**
	Then select the option ***"Install software automatically"***

The CL position is presented in bold black. Italic is used for correct parts of the translation; underlining for the wrong parts

- In the rule “pak—Avoiding participial constructions”, the RF group (22%) was more than twice as high as FR (9%). Further examination of the FF (42%) and RR (28%) groups shows that the SQ decreased significantly in FF and both the SQ and CQ decreased significantly in RR, i.e. when comparing two correct MT before and after applying the rule, the SQ and CQ were higher in cases where participial constructions were used (before CL). In the FR group, only the CQ increased significantly while the SQ increase was marginal. The results, accordingly, indicate the difficulty of the MT of participial constructions (before CL) and show at the same time that substituting it with a subordinate clause (after CL) was linked to quality deterioration—particularly regarding style.
- Likewise, in the case of the rule “pas—Avoiding passives”, the representation of the RF group (13%) was larger than that of FR (8%). However, unlike the rule “pak”, the RR group (49%) was much higher than FF (29%), which shows that the systems were able to translate nearly half of the sentences in both scenarios (passive and active) correctly. In the FF group, using active voice (after CL) resulted in a significantly lower SQ. Even in the FR group, the increases in the SQ and CQ were not significant. In the RR group, the quality values did not change significantly after using the active voice when compared to the passive voice.
- For the rule “wte—Avoiding omitting parts of the words”, in the FR group, only the CQ increased significantly. The application of this rule had a noticeable negative impact on the SQ, which was also detected in the FR, FF, and RR groups. This revealed that repetition associated with the usage of complete words instead of their reduced forms (see Example 9) was stylistically unacceptable.

The rest of the rules (“fvg—Avoiding light-verb construction (Funktionsverbgefüge)”, “kos——Formulating conditions as ‘if’ sentences”, “nsp—Using unambiguous pronominal references”, and “prä—Avoiding superfluous prefixes”) did not show—at this analysis level—a significant impact on MT quality. All quality values of these four rules in the FF and RR groups were insignificant.

Thus, in spite of the general positive impact of the CL application shown in Sect. 4.1 at annotation group level, only two of the nine rules demonstrate having a positive impact on MT output. This result urged the researcher to further examine the data from different perspectives in order to find out whether further rules can be recommended for improving MT output.

The analysis of the quality changes based on the human and automatic evaluations confirms to a large extent the results obtained at annotation group level (Table 6):

The rules "anz—Using straight quotes for interface texts" and "per—Avoiding the construction sein + zu + infinitive" showed again a significant positive impact on MT quality. The scores of TERbase and hLEPOR improved; at the same time, based on the human evaluation, a significant positive impact on both SQ and CQ was detected at rule level.

- Regarding "anz", the positive impact on the style was due to the clear orthographic presentation of the interface texts (see Example 1). In addition, using straight quotes enhanced the appropriateness of the translation to the intention of its content.
- Concerning "per", the evaluators found that using the imperative instead of the construction sein + zu + infinitive was stylistically better, as it addressed the reader directly and incites him or her to act (For example, "Before the parameterization, *configure* the controller" instead of "Before the parameterization, the controller *is to be configured*"). Regarding the CQ, both the accuracy and clarity increased after rule application, while the effect on clarity was higher.

The effect of the rules "pak—Avoiding participial constructions", "pas—Avoiding passives", and "wte—Avoiding omitting parts of the words" was negative:

- The rule "pak" was applied by generating a subordinate clause based on the participial construction. The human evaluation revealed that the evaluators found the MT of the participial construction more idiomatic than that of the subordinate clause (see Example 5b). Accordingly, the SQ decreased significantly while the CQ decrease was not significant. The automatic evaluation confirmed this result, showing a significant decrease in the quality scores of TERbase and hLEPOR.
- In the case of "wte", because of noun repetition (instead of using the reduced form, see Example 9), the evaluators judged the MT as unnatural. Thus, the SQ dropped significantly, while the CQ decreased but not significantly. In addition, the scores of both AEMs declined significantly.
- In "pas", all quality parameters (SQ, CQ, and both AEMs scores) decreased significantly after rule application. Avoiding the use of passive voice is a widely recommended CL rule. Several studies argue that avoiding the passive voice improves machine translatability, as it enables circumventing grammatical parsing issues (Bernth and Gdaniec 2001; Reuther 2003; Fiederer and O'Brien 2009). According to the human evaluation, stylistically, the evaluators considered that the active voice (after CL) is not ideal for the intention of the sentence. Concerning the content quality, the accuracy was judged to be lower after rule application (see Example 6).

Parallel to the triangulated results of the error annotation and human evaluation (Table 6), the AEMs scores (both TERbase and hLEPOR) as well as the human

Example 2 Rule “fvg—avoiding light-verb construction”

Before CL	Die **Reinigung** der Küchenmöbel sollten Sie mit einem leicht feuchten Tuch **vornehmen**
	The ***cleaning*** of the kitchen furniture you should **start** with a slightly damp cloth
After CL	Sie sollten die Küchenmöbel mit einem leicht feuchten Tuch **reinigen**
	You should ***clean*** the kitchen furniture with a slightly damp cloth

The CL position is presented in bold black. Italic is used for correct parts of the translation; underlining for the wrong parts

Example 3 Rule “kos—formulating conditions as ‘if’ sentences”

Before CL	**Steht** ein normierter Faktor zur Verfügung, kann dieser Faktor direkt in der Eingabemaske eingegeben werden
	XXX Is a standardized factor available, this factor can be entered directly in the input mask
After CL	**Wenn** ein normierter Faktor zur Verfügung **steht**, kann dieser Faktor direkt in der Eingabemaske eingegeben werden
	If a standardized factor ***is*** available, this factor can be entered directly in the input mask

The CL position is presented in bold black. Italic is used for correct parts of the translation; underlining for the wrong parts; XXX refers to an omission

scores (of SQ and CQ) did not reflect a significant impact of the rules “nsp—Using unambiguous pronominal references” and “prä—Avoiding superfluous prefixes” on MT quality:

- Regarding “nsp”, the decision of using a pronoun or substituting it with its reference (i.e. applying or rejecting the rule) is usually made on a case-by-case basis depending on the sentence and the formulation of the sentences that precede and follow it (Bernth and Gdaniec 2001). That may be the reason why no significant impact could be detected. When identifying the reference was difficult, the usage of the pronominal reference was advantageous, which resulted in an increase in MT clarity. However, in some cases, the repetition of the pronominal reference was criticised.
- Regarding the rule “prä”, the RR annotation group was very dominant (67.5%). If the MT systems were able to translate a verb with and without a superfluous prefix (e.g. *anbieten* and *bieten*) correctly, the translations were identical in both cases (correct translation: *offer*). Having a large number of correct identical translations led to achieving comparable quality before and after rule application.

For the last two rules “fvg—Avoiding light-verb construction (Funktionsverbgefüge)” and “kos—Formulating conditions as ‘if’ sentences”, while the analysis of the annotation groups did not reflect a substantial quality increase (only the quality in FR increased significantly, see Table 6), in the human and automatic evaluations, an improvement of MT quality was observed after rule application.

- With regard to the rule “fvg—Avoiding light-verb construction (Funktionsverbgefüge)”, a significant increase in MT quality (SQ, CQ, and both AEMs scores) was detected after rule application. Using the meaning-bearing verb instead of the light-verb construction enhanced the MT semantically and lexically, as not all German light verbs have a counterpart in English (see Example 2). This in turn made the translation more appropriate to its intention and easier to understand.

- Regarding "kos—Formulating conditions as 'if' sentences", the automatic evaluation (both TERbase and hLEPOR) showed only a slight quality increase after the rule application. However, based on the human evaluation, a significant improvement in the SQ and CQ was found. Leaving out the conditional conjunction 'wenn' (if), which is grammatically possible in German but not in English, caused grammatical parsing issues (see Example 3). The rule application enabled better parsing, therefore the quality scores increased with regard to accuracy, clarity, and idiomaticity.

As for the correlation between the difference in the overall quality and the differences in the AEMs scores, the Spearman correlation test showed a significant strong positive correlation for the rules "fvg", "kos" and "pas" ($\rho > 0.5$) and a significant moderate positive correlation in the remaining rules ($\rho > 0.3$). Accordingly, the quality changes detected in both analyses (human and automatic evaluation) were in line with each other.

4.3 The impact of individual CL rules at MT system level

So far, the results at rule level showed that four rules had a positive impact on the MT quality ("anz—Using straight quotes for interface texts", "per—Avoiding the construction sein + zu + infinitive", "fvg—Avoiding light-verb construction", and "kos—Formulating conditions as 'if' sentences") and three rules tended to have a negative impact on MT quality ("pak—Avoiding participial constructions", "pas—Avoiding passives", and "wte—Avoiding omitting parts of the words")—particularly regarding style quality. For these seven rules, the following analysis at MT system level explored which systems displayed the identified effect.

The impact of the two remaining rules, "nsp—Using unambiguous pronominal references" and "prä—Avoiding superfluous prefixes", was not conclusive. For these two rules, the following analysis at MT system level closely examined whether a significant impact was traceable within a certain MT system.

Rule 1 "anz—Using straight quotes for interface texts" was the only rule associated with a reduction in the number of errors (Fig. 3) as well as an improvement in SQ and CQ (Fig. 4) in all MT systems. The decrease in the number of errors after rule application was not significant in the case of the NMT system (Google Translate) and one hybrid system (Systran) for different reasons; in Google Translate, the number of errors was very small (5 errors before CL, 1 error after CL). The percentage of sentences translated correctly both before and after the usage of straight quotes was 83% in Google Translate, followed by only 17% in SDL. Accordingly, the SQ and CQ were the highest in Google Translate in both scenarios. In Systran, the number of errors was very high and barely changed (a decrease from 42 to 41 errors).

As Example 1 shows, separating the interface text *Software automatisch installieren* using quotes enabled the systems to identify the text as a caption. This in turn improved the parsing of the source sentence. Accordingly, two error types (see *install*) were corrected after the rule application: capitalising the interface text (OR.02) and correcting the word order (GR.10). In Lucy, the correction of these two error types after rule application was strongly correlated with the quality increase. In Bing, only the correction of the word order error proved to strongly correlate with the quality improvement. In the other systems, no correlations between any of the error types and the quality could be detected.

For the second rule "fvg—Avoiding light-verb construction (Funktionsverbgefüge)", the general positive impact on MT output at system level was as follows (Figs. 5 and 6): For the RBMT system (Lucy) and one hybrid system (Systran), this rule was very advantageous in reducing the number of errors and increasing SQ significantly. In the SMT system (SDL), the number of errors decreased significantly, but the increase in SQ and CQ was not significant. In the second hybrid system (Bing), the number of errors decreased and the SQ and CQ increased; however, these changes were not significant. The human evaluators found that using the meaning-bearing verb (after CL) instead of the light verb construction (before CL) makes the translation easier to understand and stylistically more attention-grabbing. Analysing the NMT system (Google Translate) showed distinct results: the number of errors was minimal (3 errors before CL, 1 error after CL). It was able to translate 88% of the sentences both before and after the rule application correctly, followed by 46% in Bing. This displayed the highest SQ and CQ among all systems both before and after rule implementation.

"Avoiding light-verb construction" is primarily related to sentence semantics. Since not all German light verbs have a counterpart in English, using the meaning-bearing verb (*reinigen* in Example 2, after CL) instead of the light-verb construction (*Reinigung vornehmen*, before CL) was associated with a correction of a number of semantic errors, particularly collocation (SM.13) and lexical errors. Lexical errors occurred when the systems translated the light verb literally (e.g. translating *zur Verfügung stellen* as *represent available* instead of *provide*). In Lucy, the correction of the semantic errors correlated with an increase in SQ and CQ. In Bing and SDL, a correlation was observed between the lexical errors (LX.03 and LX.04) and the quality. No further correlations were detectable in the other systems.

The application of Rule 3 "kos—Formulating conditions as 'if' sentences" was associated with a reduction in the number of errors in all systems except for the NMT system (Google Translate: 1 error before CL, 2 errors after CL), Fig. 7. The reduction in MT errors was only significant in one hybrid system (Bing) and the SMT system (SDL). Consequently, it was only in these two systems that a significant improvement in quality was achieved—in Bing both for the SQ and CQ, and in SDL only for the CQ, Fig. 8. This positive effect on the quality correlated strongly with the reduction in the error types LX.03 "Omission" and GR.10 "Wrong word order" (as noted in Example 3). In contrast, both SQ and CQ decreased in the RBMT system (Lucy) and the other hybrid system (Systran) after rule application. In Google Translate, the percentage of correct MT before and after the rule application (Group

Example 4 Rule "nsp—using unambiguous pronominal references"

Before CL	Fettreste müssen vollständig abgewaschen werden, da sich **diese** ansonsten in der Pfanne einbrennen können
	Grease residue must be completely washed off as ***it*** can otherwise burn in the pan
After CL	Fettreste müssen vollständig abgewaschen werden, da sich **diese Reste** ansonsten in der Pfanne einbrennen können
	Grease residue must be completely washed off as ***this*** **remains** can otherwise burn in the pan

The CL position is presented in bold black. Italic is used for correct parts of the translation; underlining for the wrong parts

RR) was 92%, followed by 71% in Lucy. This once again showed the highest SQ and CQ in both scenarios.

"Formulating conditions as 'if' sentences" mainly showed a positive lexical effect on the MT. Considering Example 3, the German verb can be used as a conditional word without the need for the conditional conjunction *wenn* (*if*). This is not the case in English, which was why omitting the conditional conjunction (before CL) was associated with two error types: the conditional conjunction *if* was missing (LX.03), and the verb used to formulate the conditional clause was placed incorrectly (GR.10), see *is* in Example 3. Therefore, a reduction in the number of errors in both error types was observed after the rule application. In Bing and SDL, the correction of these errors strongly correlated with the increase in quality. In the other systems, no correlations with specific error type were observed. After rule application, the evaluators found the MT more accurate, understandable, and idiomatic.

Rule 4 "nsp—Using unambiguous pronominal references" had a different impact on MT quality from one system to the other (Fig. 9). Only the RBMT system (Lucy) and the NMT system (Google Translate) exhibited significant quality changes: Lucy showed a slight increase in the SQ and a significant increase in the CQ, while Google Translate showed exactly the opposite—namely a significant decrease in the SQ and a slight decrease in the CQ. This could be explained by the various changes in the number of errors (Fig. 10). Google Translate, as opposed to the other systems, was mostly able to translate the pronouns correctly (before CL). However, using a pronominal reference (after CL) was stylistically criticised in some cases. Nevertheless, 83% of the translations by Google Translate were error-free before and after the rule application, followed by 67% in Lucy. In addition, Google Translate achieved the highest quality scores in both scenarios.

"Using unambiguous pronominal references" had two different effects on the MT: in Lucy, SDL, and Systran, the rule application was associated with a reduction in the "Confusion of sense" semantic error (SM.11). This error was especially apparent in the translation of demonstrative pronouns (*diese* and *dies*), as the MT systems found difficulties in identifying the reference and translating it correctly. Accordingly, after rule application, the human evaluators found the translation clearer.

However, the application of this rule was also associated with an increase in the lexical "Consistency" error (LX.06) in Bing and SDL, see Example 4. In order to implement this rule, a noun in the main clause should not be substituted

Example 5a Rule “pak—avoiding participial constructions”

Before CL	Durch Eingabe **der mit einem roten Sternchen gekennzeichneten Parameter** erfolgt die minimale Konfigurierung
	By entering **the marked with a red asterisk parameter**, the minimum configuration is performed
After CL	Durch Eingabe **der Parameter, die mit einem roten Sternchen gekennzeichnet sind**, erfolgt die minimale Konfigurierung
	By entering ***the parameters that are marked with a red asterisk***, the minimum configuration is performed

The CL position is presented in bold black. Italic is used for correct parts of the translation; underlining for the wrong parts

Example 5b Rule “pak—avoiding participial constructions”

Before CL	**Speziell auf diese Lautsprecher abgestimmtes Zubehör** erhalten Sie in unserem Webshop
	Special accessories for these speakers are available in our webshop
After CL	**Zubehör, das speziell auf diese Lautsprecher abgestimmt ist,** erhalten Sie in unserem Webshop
	Accessories, specially designed for these loudspeakers, are available in our webshop

The CL position is presented in bold black. Italic is used for correct parts of the translation; underlining for the wrong parts

by a pronoun in the subordinate clause; instead, a pronominal reference should be used (see *Reste* in Example 4). In some cases, the MT systems translated the second instance of the noun differently (*residue* in the main clause and *remains* in the subordinate clause), which resulted in a consistency error and hence reduced accuracy. However, a consistency error could be avoided if the terms used are maintained and managed in the system.

Rule 5 “pak—Avoiding participial constructions” had—in general—a negative effect on the SQ in all MT systems. Regarding the CQ, it only increased in two MT systems (Fig. 11): marginally in the SMT system (SDL) and significantly in one hybrid system (Systran). The number of errors increased in all systems except in SDL, where a slight decrease was found (Fig. 12). The NMT system (Google Translate) had no difficulty in translating participial constructions (only 4 errors before CL as opposed to 8 errors after CL). Furthermore, 71% of the translations by Google Translate were error-free before and after rule application, followed by only 29% in Bing. This demonstrated the highest quality rates in both scenarios. In all other systems, the number of errors was much higher both before and after rule application. Systran showed a significant increase in the number of errors after rule application.

Two different MT error types were associated with this rule: German participial constructions, especially lengthy ones, usually complicate sentence structure and consequently parsing, which results in word order errors (GR.10), see the participial construction *der mit einem roten Sternchen gekennzeichneten Parameter* in Example

Example 6 Rule “pas—avoiding passives”

Before CL	Durch diese Öffnung **kann** der Stecker mit dem Regler **verbunden werden**
	Through this opening, the plug ***can be connected*** to the controller
After CL	Durch diese Öffnung **können Sie** den Stecker mit dem Regler **verbinden**
	Through this opening, the plug **you can connect** to the controller

The CL position is presented in bold black. Italic is used for correct parts of the translation; underlining for the wrong parts

Example 7 Rule “per—avoiding the construction sein + zu + infinitive”

Before CL	Wenn ein mehrstufiges Modul parametriert ist, so sind die externen Kontakte **zu verriegeln**
	If a multi-stage module is parameterized, the external contacts **are to lock**
After CL	Wenn ein mehrstufiges Modul parametriert ist, **verriegeln Sie** die externen Kontakte
	If a multi-stage module is parameterized, **you** ***lock*** the external contacts

The CL position is presented in bold black. Italic is used for correct parts of the translation; underlining for the wrong parts

5a (before CL). This error type especially occurred in SDL and Bing in the translation of participial constructions and decreased after rule implementation.

However, after implementing the rule, all systems had difficulty with the comma placement in subordinate clauses, specifically in cases in which a distinction between the conjunction *which* and *that* needed to be made. Therefore, the number of punctuation errors (OR.01) increased. Consider Example 5b: unlike German, no commas are needed in English after the rule application. Nevertheless, it is important to note that the use of *which* vs. *that* is generally controversial and contextual information is usually required to decide on correct usage.

Regarding Rule 6 “pas—Avoiding passives”, all MT systems—except for Bing—demonstrated an increase in the total number of errors after rule application (Fig. 13). The two hybrid MT systems delivered contradictory results: in Bing, the total number of errors decreased significantly, while it increased significantly in Systran. In the other three systems, the increase in the number of errors was not significant. Bing and Google Translate were able to translate 71% of the sentences correctly in both passive and active voice, followed by 58% in Lucy. Both SQ and CQ decreased in all MT systems, except in Bing, where the CQ increased slightly (Fig. 14). In Systran, the SQ and CQ decrease was significant. In Lucy, the decrease in SQ was significant. Both before and after rule application, the highest SQ and CQ were seen in Google Translate. The rule application was associated with an increase in various error types across all systems; at the same time, none of the error types exhibited a significant increase after the rule application.

Example 6 shows how the use of active voice (after CL) was in some cases associated with a word order error (GR.10) (in *you can connect*), while the passive voice (before CL) was correctly translated.

Rule 7 “per—Avoiding the construction sein + zu + infinitive “ demonstrated a general positive impact on the MT output (Figs. 15 and 16): in one hybrid system

Example 8 Rule “prä—avoiding superfluous prefixes”

Before CL	**Schicken** Sie das Gerät originalverpackt an unsere Serviceadresse **ein**
	Please ***send*** the appliance in its original packaging to our service address **one**
After CL	**Schicken** Sie das Gerät originalverpackt an unsere Serviceadresse
	Please ***send*** the appliance in its original packaging to our service address

The CL position is presented in bold black. Italic is used for correct parts of the translation; underlining for the wrong parts

Example 9 Rule “wte—avoiding omitting parts of the words”

Before CL	Sogar **Soja- und laktosefreie Milch** lassen sich mit dieser Maschine perfekt aufschäumen
	Even **Soya-** ***and lactose-free milk*** can be perfectly frothed with this machine
After CL	Sogar **Sojamilch und laktosefreie Milch** lassen sich mit dieser Maschine perfekt aufschäumen
	Even ***soya milk and lactose-free milk*** can be perfectly frothed with this machine

The CL position is presented in bold black. Italic is used for correct parts of the translation; underlining for the wrong parts

Table 2 Error classification applied in the annotation

Error category	Error no.	Error type
Orthography	OR.01	Punctuation error
	OR.02	Capitalisation error
Lexis	LX.03	Omission
	LX.04	Addition
	LX.05	Untranslated
	LX.06	Consistency error (a word is repeated in the sentence and translated differently each time)
Grammar	GR.07	Wrong word class
	GR.08	Wrong verb tense/composition/person
	GR.09	Wrong agreement gender/number/person
	GR.10	Wrong word order
Semantics	SM.11	Confusion of sense (the output translation is possible, but not in the given context)
	SM.12	Wrong choice (the output translation is apparently wrong)
	SM.13	Collocation error

Table 3 Overview of the dataset

Number of annotated MT sentences	2160	(100%)
Number of human-evaluated MT sentences	1550	(72%)
Number of excluded MT sentences	610	(28%)

Table 4 Representation of the annotation groups in the error annotation and human evaluation

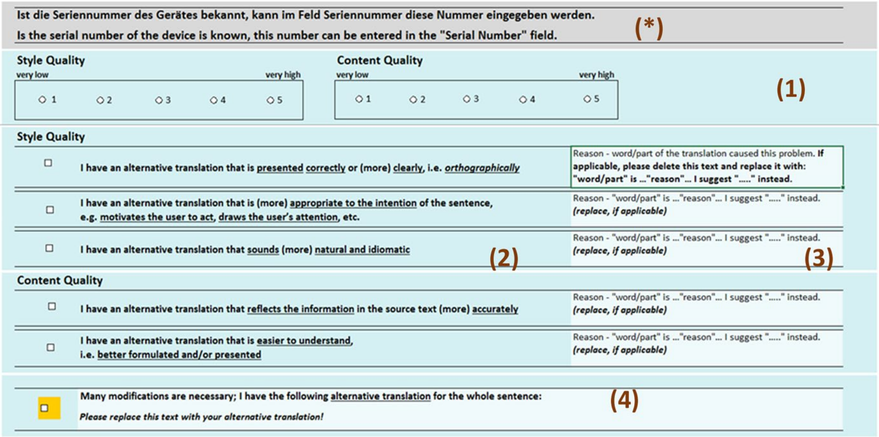
	FR	FF	RF	RR	
Error annotation	19.8%	25.8%	9.8%	44.5%	100.0%
Human Evaluation	20.4%	24.7%	10.7%	44.2%	100.0%

Error annotation N = 2160 MT sentences; Human evaluation N = 1550 MT sentences

Table 5 Percentages of MT sentences of each system in the human evaluation

Bing	Google	Lucy	SDL	Systran	
18%	25%	20%	20%	17%	100.0%

Human evaluation N = 1550 MT sentences



Ist die Seriennummer des Gerätes bekannt, kann im Feld Seriennummer diese Nummer eingegeben werden.
Is the serial number of the device is known, this number can be entered in the "Serial Number" field.

Style Quality
very low ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 very high

Content Quality
very low ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 very high

Style Quality

□ I have an alternative translation that is presented correctly or (more) clearly, i.e. *orthographically* — Reason - word/part of the translation caused this problem. **If applicable, please delete this text and replace it with: "word/part" is ..."reason"... I suggest "....." instead.**

□ I have an alternative translation that is (more) appropriate to the intention of the sentence, e.g. motivates the user to act, draws the user's attention, etc. — Reason - "word/part" is ..."reason"... I suggest "....." instead. ***(replace, if applicable)***

□ I have an alternative translation that sounds (more) natural and idiomatic — Reason - "word/part" is ..."reason"... I suggest "....." instead. ***(replace, if applicable)***

Content Quality

□ I have an alternative translation that reflects the information in the source text (more) accurately — Reason - "word/part" is ..."reason"... I suggest "....." instead. ***(replace, if applicable)***

□ I have an alternative translation that is easier to understand, i.e. better formulated and/or presented — Reason - "word/part" is ..."reason"... I suggest "....." instead. ***(replace, if applicable)***

□ Many modifications are necessary; I have the following alternative translation for the whole sentence:
Please replace this text with your alternative translation!

Fig. 1 Interface of the human evaluation

(Bing) and the SMT system (SDL), grammatical difficulties were observed when the construction sein + zu + infinitive (before CL) was used, although an equivalent construction exists in English (verb to be + to + infinitive). Accordingly, the rule application was associated with a significant reduction in two grammatical errors: incorrect verb tense (GR.08) and incorrect word order (GR.10) (see verb composition error in *are to lock* in Example 7, before CL). In Bing and SDL, correcting these error types correlated strongly with an increase in quality values.

The NMT system (Google Translate) was able to correctly translate 96% of the sentences both before and after rule application, followed by 42% in Systran. Therefore, the quality of MT output in both scenarios was the highest with a minimal increase in the SQ and a minimal decrease in the CQ.

However, applying the CL rule using the imperative (instead of the construction sein + zu + infinitive) was associated with the "Addition" lexical error (LX.04) (after CL) in the RBMT system (Lucy) and the other hybrid system

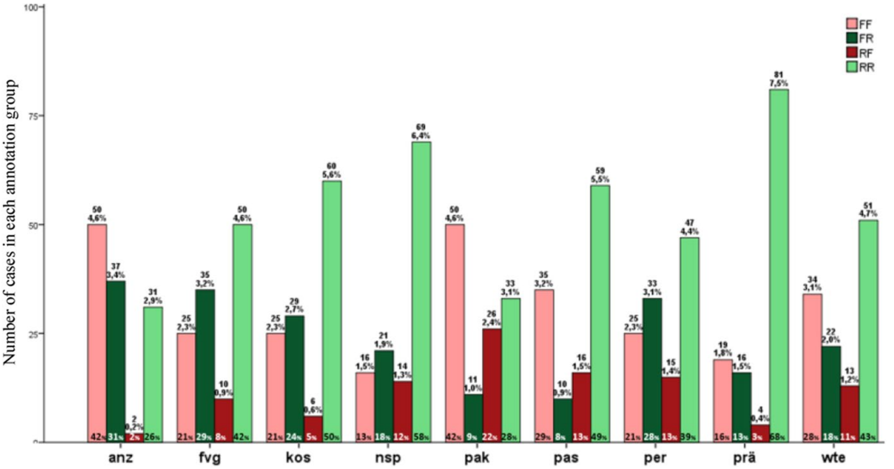


Fig. 2 Comparison of the annotation groups at CL rule level. The percentages displayed on the top of the bar are calculated based on the entire dataset of all rules (N = 1080). The percentages on the bottom are calculated at rule level (N = 120)

(Systran), as in some cases, the MT systems wrongly added the subject *you* (see Example 7, after CL). In Systran, the correction of this lexical error correlated strongly with the increase in the quality values.

Rule 8 "prä—Avoiding superfluous prefixes" was—in general—associated with a small number of errors both before (ranging between 4 errors in Google Translate and 12 errors in SDL) and after rule application (ranging between 3 errors in Bing and Google and 11 errors in SDL), Fig. 17. As these ranges show, the number of errors decreased after rule application. This reduction was only significant in one hybrid system (Bing). The SQ slightly improved in all systems except for Google Translate (Fig. 18). Also, the CQ increased minimally in Bing, Lucy, and Systran. In Google Translate, 88% of the sentences were correctly translated both before and after rule application, followed by 71% in Bing. The quality values in Google Translate showed a minimal decrease after rule application (SQ − 0.06, CQ − 0.03); at the same time, they were the highest among all MT systems both before and after the rule implementation.

In all systems, except for Google Translate, avoiding superfluous prefixes supported correct parsing of the verb. In particular, German "separable verbs" (similar to phrasal verbs in English) were often difficult to parse; depending on the sentence structure, prefixes should sometimes be placed at the end of the sentence—far from the rest of the verb. In such cases, the systems translated the prefix additionally, independently of the verb. Thus, avoiding superfluous prefixes resulted in correcting the "Addition" lexical error (LX.04) (see *one* in Example 8, before CL), which in turn improved the accuracy of the translation.

Finally, Rule 9 "wte—Avoiding omitting parts of the words" was associated with a marginal decrease in the number of errors in Google Translate, Lucy, and Systran (Fig. 19). Due to the differences in the orthographic rules in German and English

Table 6 Differences in style and content quality after the application of each rule at annotation group level

	FF			FR			RF			RR		
	Diff. SQ	Diff. CQ	Diff. Q	Diff. SQ	Diff. CQ	Diff. Q	Diff. SQ	Diff. CQ	Diff. Q	Diff. SQ	Diff. CQ	Diff. Q
anz	M = 0.47 p < 0.001 N = 30	M = 0.27 p = 0.003 N = 30	M = 0.37 p < 0.001 N = 30	M = 0.69 p < 0.001 N = 24	M = 0.84 p = 0.001 N = 24	M = 0.77 p < 0.001 N = 24	M = − 0.38 N = 1	M = 0 N = 1	M = − 0.19 N = 1	M = 0.32 p = 0.001 N = 19	M = 0.05 p = 0.273 N = 19	M = 0.18 p = 0.005 N = 19
fvg	M = 0.14 p = 0.420 N = 17	M = 0.14 p = 0.636 N = 17	M = 0.14 p = 0.421 N = 17	M = 0.85 p < 0.001 N = 25	M = 0.75 p < 0.001 N = 25	M = 0.80 p < 0.001 N = 25	M = − 0.16 p = 0.416 N = 7	M = − 0.50 p = 0.107 N = 7	M = − 0.33 p = 0.058 N = 7	M = 0.15 p = 0.140 N = 35	M = 0.03 p = 0.926 N = 35	M = 0.09 p = 0.190 N = 35
kos	M = 0.30 p = 0.090 N = 14	M = 0.38 p = 0.278 N = 14	M = 0.34 p = 0.157 N = 14	M = 0.71 p < 0.001 N = 23	M = 1.28 p < 0.001 N = 23	M = 0.99 p < 0.001 N = 23	M = − 0.71 p = 0.285 N = 3	M = − 1.04 p = 0. 285 N = 3	M = − 0.88 p = 0.285 N = 3	M = − 0.10 p = 0.620 N = 44	M = 0.01 p = 0.795 N = 44	M = − 0.04 p = 0.229 N = 44
nsp	M = − 0.15 p = 0.443 N = 8	M = − 0.08 p = 0.799 N = 8	M = − 0.11 p = 0.735 N = 8	M = 0.31 p = 0.011 N = 12	M = 0.68 p = 0.002 N = 12	M = 0.50 p = 0.002 N = 12	M = − 0.50 p = 0.012 N = 10	M = − 0.48 p = 0.012 N = 10	M = − 0.49 p = 0.012 N = 10	M = − 0.05 p = 0.220 N = 47	M = 0.10 p = 0.086 N = 47	M = 0.02 p = 0.795 N = 46
pak	M = − 0.19 p = 0.008 N = 37	M = 0.05 p = 0.411 N = 37	M = − 0.07 p = 0.507 N = 37	M = 0.44 p = 0.107 N = 9	M = 0.82 p = 0.021 N = 9	M = 0.63 p = 0.050 N = 9	M = − 0.73 p < 0.001 N = 23	M = − 0.52 p = 0.012 N = 23	M = − 0.63 p < 0.001 N = 23	M = − 0.38 p < 0.001 N = 28	M = − 0.13 p = 0.020 N = 28	M = − 0.25 p < 0.001 N = 28
pas	M = − 0.34 p = 0.010 N = 25	M = − 0.16 p = 0.177 N = 25	M = − 0.25 p = 0.019 N = 25	M = 0.54 p = 0.114 N = 6	M = 0.58 p = 0.080 N = 6	M = 0.56 p = 0.075 N = 6	M = − 0.99 p = 0.003 N = 12	M = − 1.93 p = 0.002 N = 12	M = − 1.46 p = 0.002 N = 12	M = − 0.13 p = 0.061 N = 40	M = 0.05 p = 0.427 N = 40	M = − 0.04 p = 0.383 N = 40
per	M = 0.33 p = 0.003 N = 20	M = 0.27 p = 0.136 N = 20	M = 0.30 p = 0.012 N = 20	M = 1.29 p < 0.001 N = 27	M = 1.59 p < 0.001 N = 27	M = 1.44 p < 0.001 N = 27	M = − 0.33 p = 0.016 N = 13	M = − 0.47 p = 0.007 N = 13	M = − 0.40 p = 0.001 N = 13	M = 0.24 p < 0.001 N = 37	M = 0.00 p = 0.522 N = 37	M = 0.12 p = 0.002 N = 37
prä	M = 0.10 p = 0.109 N = 11	M = 0.08 p = 0.765 N = 11	M = 0.09 p = 0.439 N = 11	M = 0.65 p = 0.002 N = 17	M = 0.78 p = 0.001 N = 17	M = 0.71 p = 0.001 N = 17	M = − 0.31 p = 0.180 N = 2	M = − 0.88 p = 0.180 N = 2	M = − 0.59 p = 0.180 N = 2	M = − 0.05 p = 0.148 N = 62	M = − 0.03 p = 0.828 N = 62	M = − 0.04 p = 0.122 N = 62
wte	M = − 0.27 p = 0.012 N = 24	M = − 0.29 p = 0.177 N = 24	M = − 0.28 p = 0.018 N = 24	M = 0.20 p = 0.266 N = 16	M = 0.52 p = 0.001 N = 16	M = 0.36 p = 0.010 N = 16	M = − 0.50 p = 0.028 N = 12	M = − 1.05 p = 0.004 N = 12	M = − 0.78 p = 0.004 N = 12	M = − 0.35 p < 0.001 N = 35	M = − 0.06 p = 0.078 N = 35	M = − 0.21 p < 0.001 N = 35

SQ style quality, *CQ* content quality, *Q* overall quality (mean of SQ and CQ); Diff. SQ = SQ before – SQ after CL application, similarly Diff. CQ and Diff. Q. White cells:

Table 6 (continued)

insignificant values $p \geq 0.05$; shaded cells: significant values $p < 0.05$. *M* mean

Bold shows rules that have a significant positive impact; Italics for a significant negative impact; For the rest of the rules, the results were insignificant

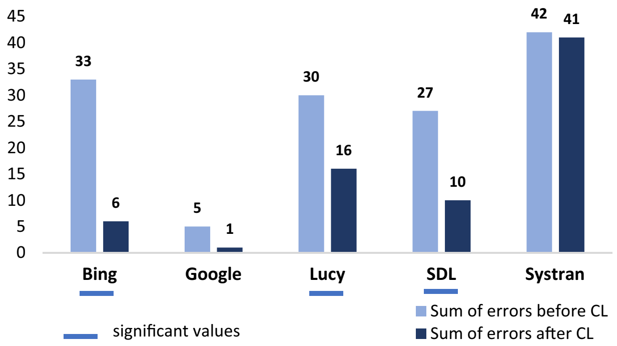


Fig. 3 Rule "anz—using straight quotes for interface texts"—number of MT errors before and after rule application

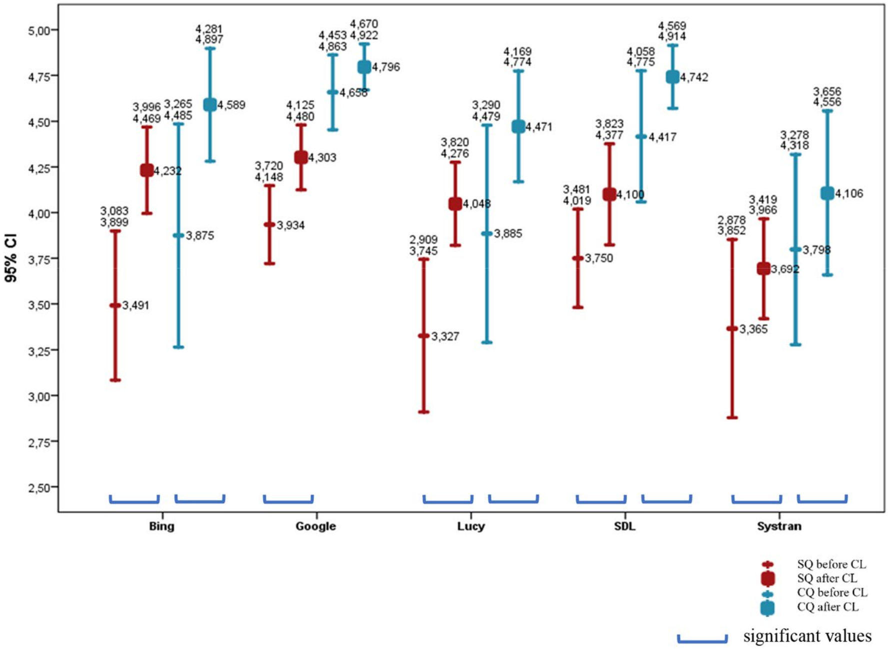


Fig. 4 Rule "anz—using straight quotes for interface texts"—style and content quality before and after rule application

with regard to hyphen usage, applying this rule was associated with a reduction in orthographic errors. Example 9 shows how the rule implementation was associated with the correction of punctuation and capitalisation errors (in *Soya*-).

However, the number of errors marginally increased in Bing and remained unchanged in SDL. Despite the different impacts on the number of errors, the SQ and CQ decreased in all systems except Systran, in which the CQ slightly increased (Fig. 20). The SQ decrease was significant in three systems (Google

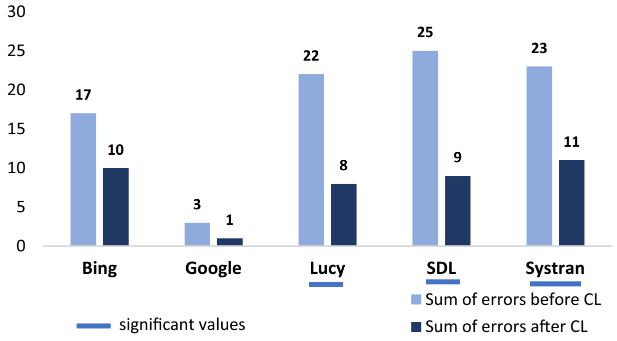


Fig. 5 Rule "fvg—avoiding light-verb construction"—number of MT errors before and after rule application

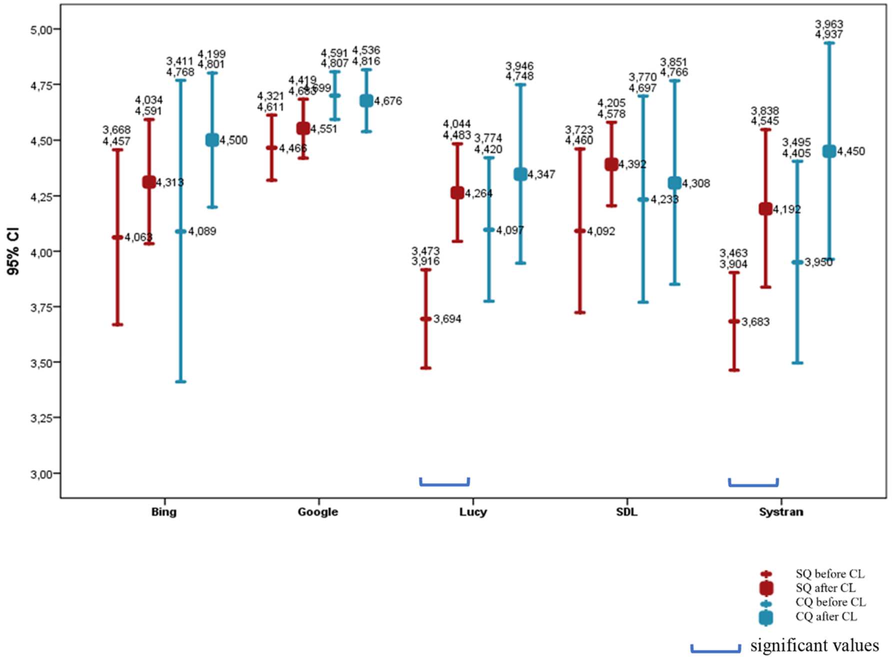


Fig. 6 Rule "fvg—avoiding light-verb construction"—style and content quality before and after rule application

Translate, Lucy, and SDL), as the human evaluators found the noun repetition unnatural (see milk in *soya milk and lactose-free milk* in Example 9, after CL). Again, Google Translate showed the lowest number of errors (75% of the translations were error-free both before and after CL, followed by 46% in SDL) and the highest quality values in both scenarios.

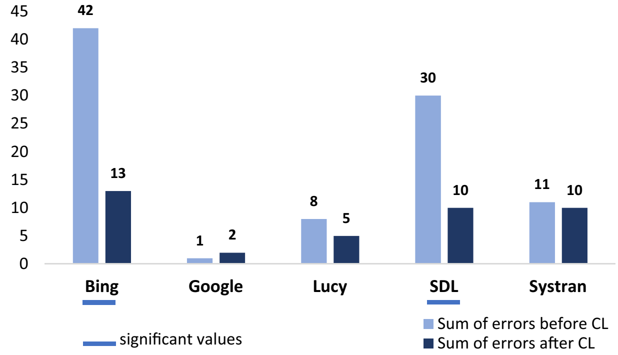


Fig. 7 Rule "kos—formulating conditions as 'if' sentences"—number of MT errors before and after rule application

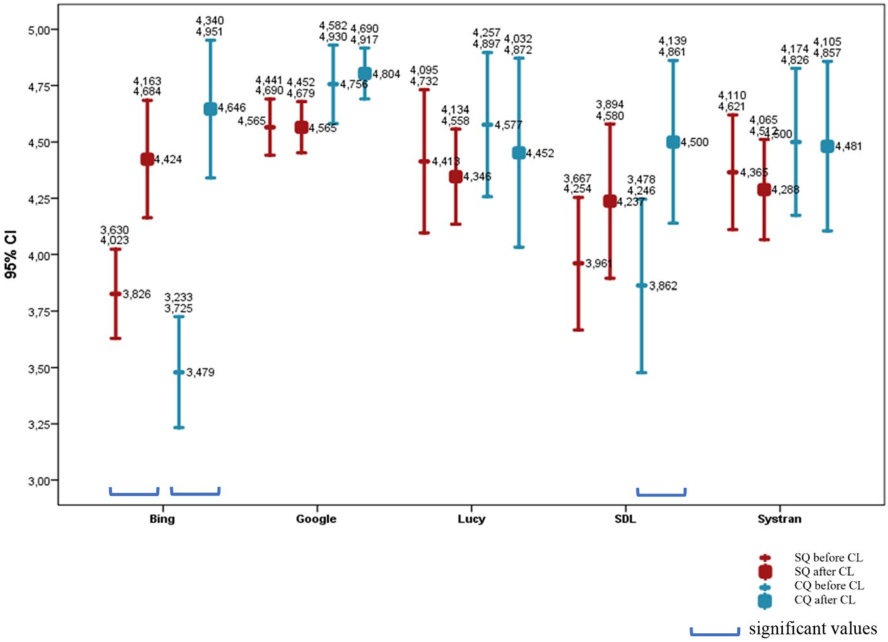


Fig. 8 Rule "kos—formulating conditions as 'if' sentences"—style and content quality before and after rule application

5 Conclusion

The study aims to analyse and contrast the impact of individual CL rules on MT output at different levels. In accordance with many previous studies, the results showed that CL application had in general, *at a rule- and system-independent level*, a significant positive impact on the MT output in terms of reducing the number of errors and

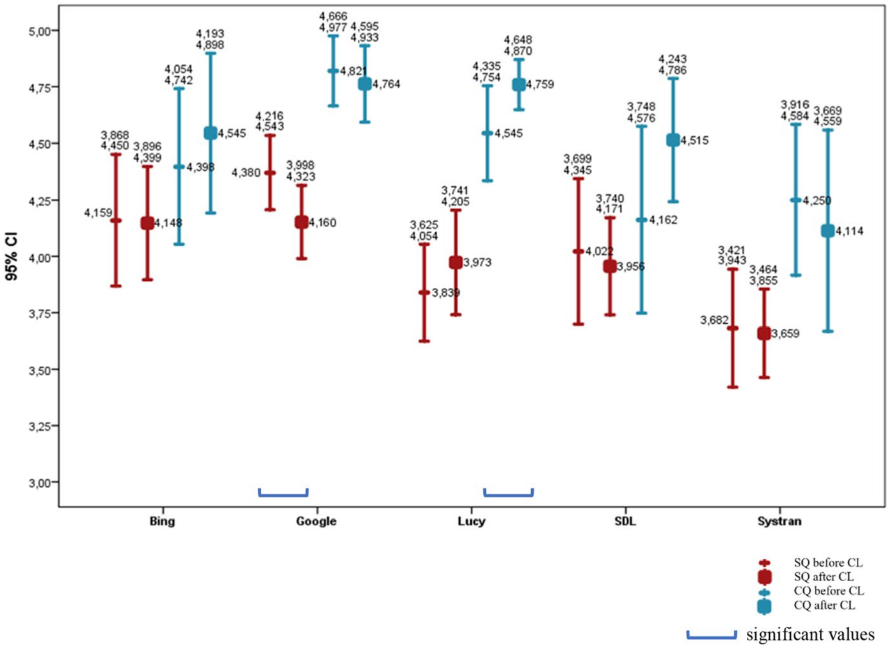


Fig. 9 Rule "nsp—using unambiguous pronominal references"—style and content quality before and after rule application

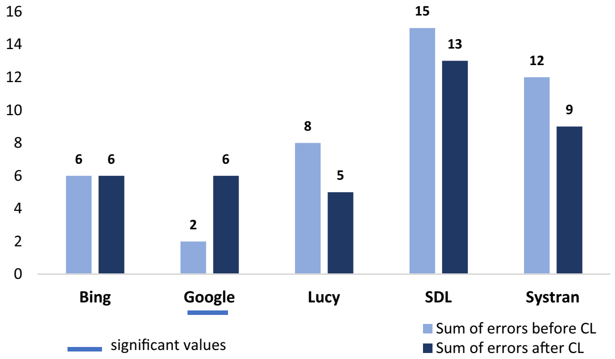


Fig. 10 Rule "nsp—using unambiguous pronominal references"—number of MT errors before and after rule application

increasing the style and content quality as well as the scores of two AEMs (TERbase and hLEPOR).

A closer analysis of the individual impact of the rules, *at a system-independent level*, revealed that only the rules "anz—Using straight quotes for interface texts", "per—Avoiding the construction sein + zu + infinitive", "kos—Formulating conditions as 'if' sentences", and "fvg—Avoiding light-verb construction" positively

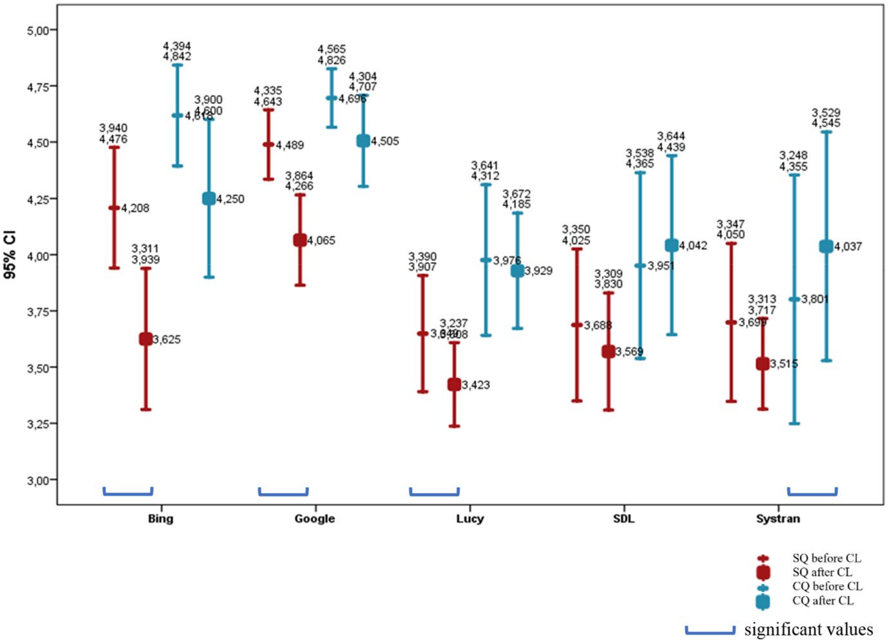


Fig. 11 Rule "pak—avoiding participial constructions "—style and content quality before and after rule application

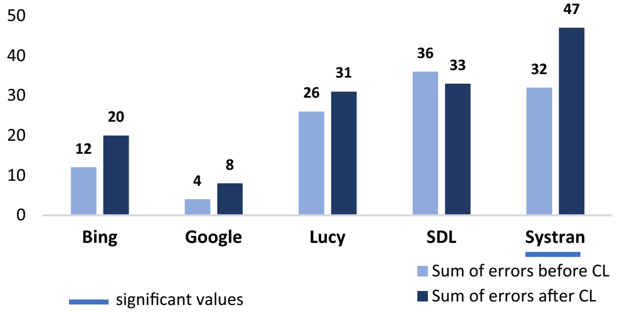


Fig. 12 Rule "pak—avoiding participial constructions"—number of MT errors before and after rule application

affected the MT output (fewer errors and better SQ, CQ, and AEMs scores). These rules enabled better parsing, which in turn contributed to getting a more accurate, comprehensible, stylistic, and attention-grabbing translation. On the contrary, the rule "pas—Avoiding passives" showed a significant negative impact on the SQ, CQ, and the AEMs scores. The human evaluators assessed the MT of the active voice to be less accurate and stylistically less adequate. In the rule "pak—Avoiding participial constructions", the AEMs scores and the SQ deteriorated significantly, as the MT of participial constructions was evaluated as more idiomatic. In the rule

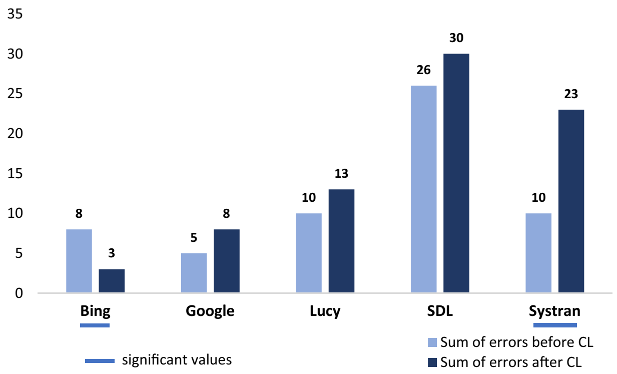


Fig. 13 Rule “pas—avoiding passives”—number of MT errors before and after rule application

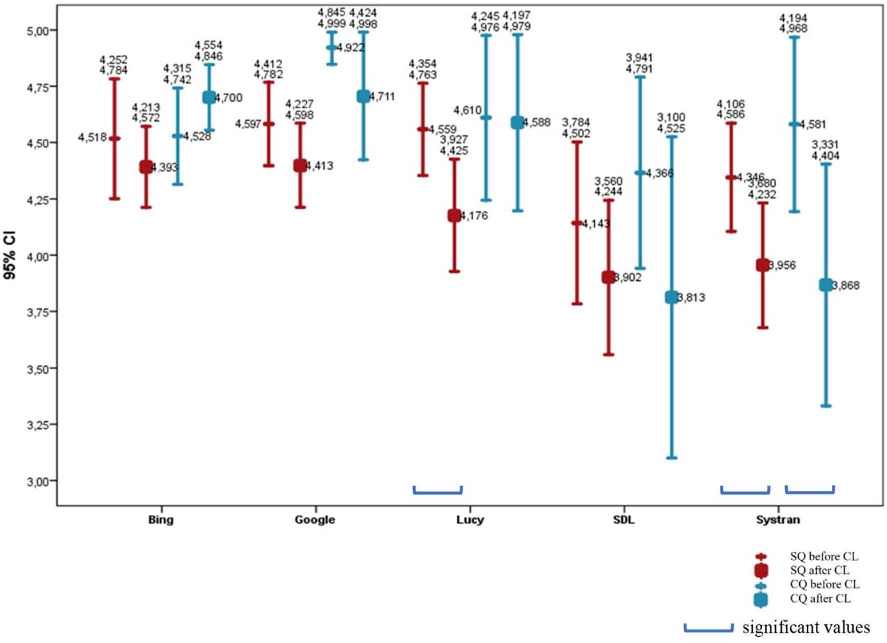


Fig. 14 Rule “pas—avoiding passives”—style and content quality before and after rule application

“wte—Avoiding omitting parts of the words”, only the SQ and both AEMs scores decreased, which showed that the MT sounded unnatural after rule application. For the rules “nsp—Using unambiguous pronominal references” and “prä—Avoiding superfluous prefixes”, no significant impact was found.

A more detailed examination of the impact of *each rule at MT system level* showed that when earlier MT approaches (RBMT, SMT, and hybrid systems) were applied, the impact of the individual rules varied to a large extent from one approach to the other. Since not all CL rules have a definite positive impact,

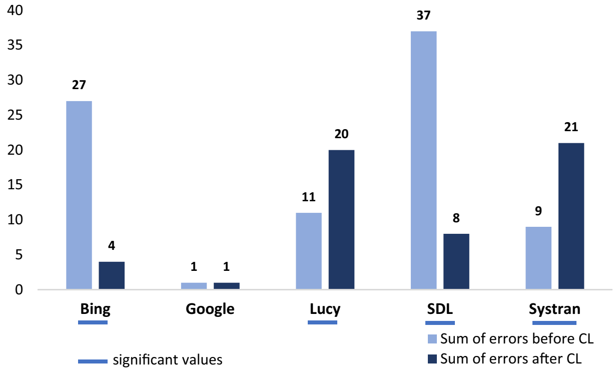


Fig. 15 Rule "per—avoiding the construction sein + zu + infinitive "—number of MT errors before and after rule application

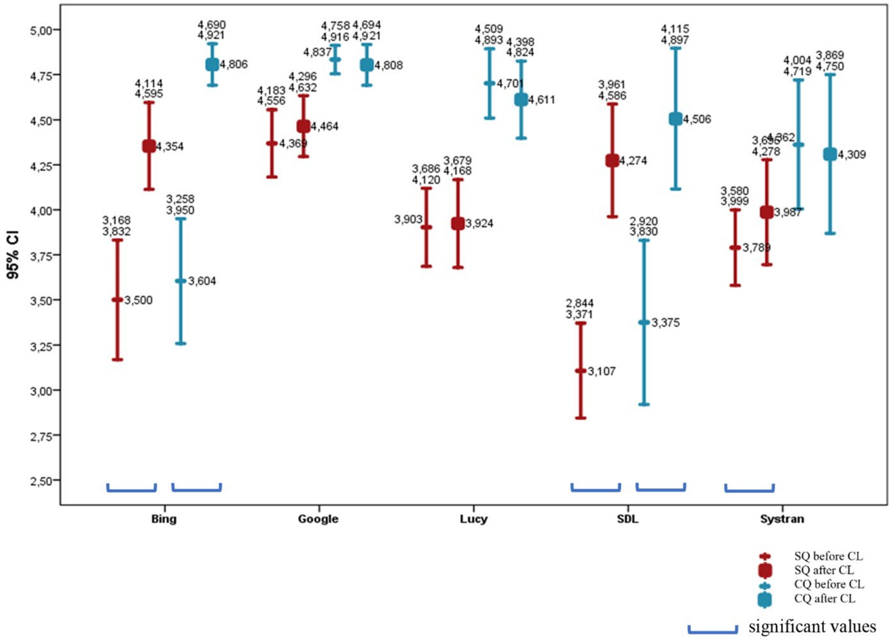


Fig. 16 Rule "per—avoiding the construction sein + zu + infinitive"—style and content quality before and after rule application

identifying effective rules in each implementation context (language pair, translation direction, domain, and MT approach) is necessary. Limiting the number of rules applied to those that are effective can be beneficial in avoiding drawbacks commonly associated with CL application (e.g. slowing down the authoring process and impacting it excessively). Comparing the earlier MT approaches to the recent NMT approach, the results revealed that while earlier MT systems

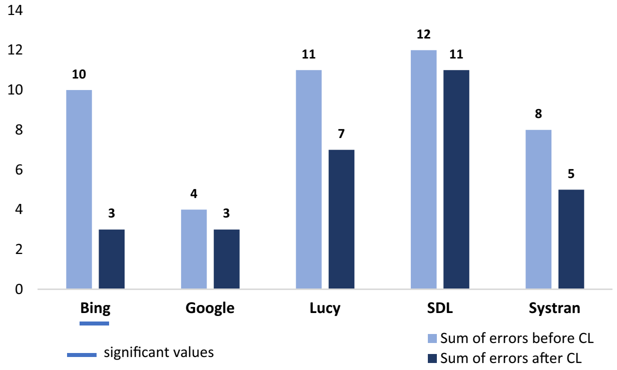


Fig. 17 Rule "prä—avoiding superfluous prefixes"—number of MT errors before and after rule application

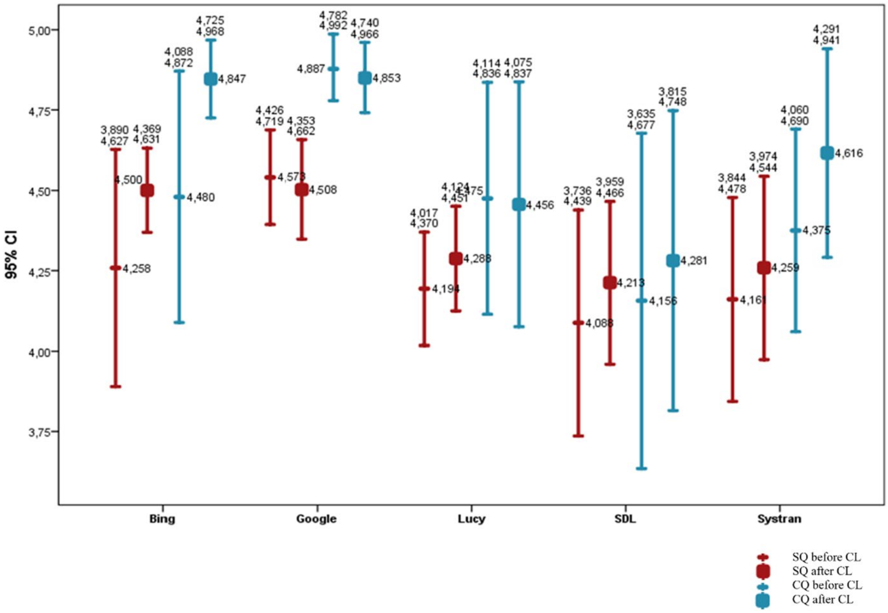


Fig. 18 Rule "prä—avoiding superfluous prefixes"—style and content quality before and after rule application

benefited in many cases from the CL rules in avoiding different MT errors and improving their output quality, the NMT system was able to translate most of the sentences before and after the application of all rules error-free (between 71% in the rules "pas" and "pak" and 96% in the rule "per"). Moreover, the NMT system recorded the highest style and content quality in both scenarios under all rules.

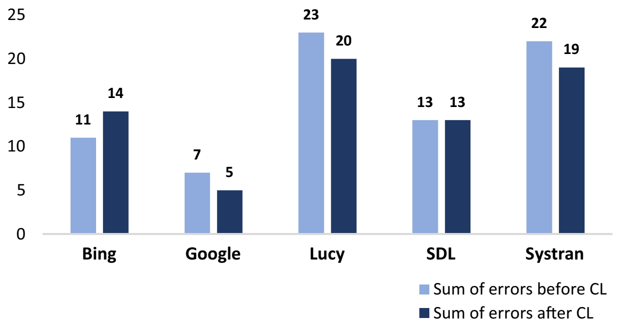


Fig. 19 Rule "wte—avoiding omitting parts of the words"—number of MT errors before and after rule application

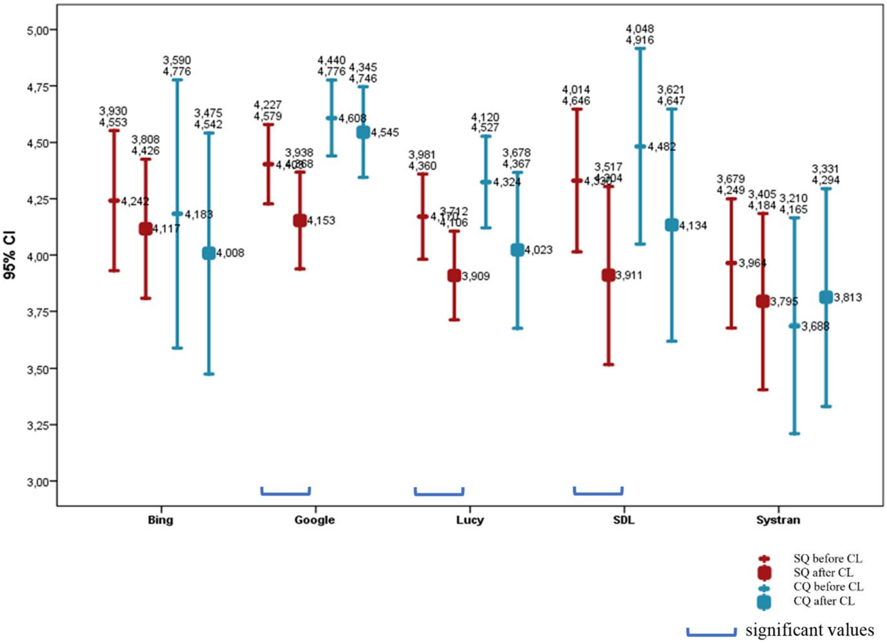


Fig. 20 Rule "wte—avoiding omitting parts of the words"—style and content quality before and after rule application

6 Limitations and future work

This study has explored the impact of a limited number of CL rules (nine rules) on the machine translatability of five different MT architectures, including NMT, which—to the best of my knowledge—has not yet been examined. For the analysed rules, the CL application with the aim to enhance the MT output is no longer necessary when neural MT technology is used. However, the analysis within the scope of this study was carried out at sentence level, so it examined only CL rules applied within the sentence. An analysis at an extended level (e.g. cross-sentential or document level) is of great interest in order to capture the impact of context-relevant CL rules (i.e. rules that affect several sentences) on the MT output. Contextual MT or MT at document level is one of the known challenging goals of MT (Zhang and Zong 2020). Recent NMT studies show that advances have already been made in the field of context-aware MT and MT at document level, even in the translation of literature, which is considered to be a particularly challenging domain for MT (cf. Toral and Way 2018; Matusov 2019). Furthermore, several studies have developed context-sensitive NMT models as well as strategies for NMT at document level, with which classic MT difficulties such as deixis, ellipses, co-reference resolution, coherence and lexical cohesion could be overcome (Müller et al. 2018, Stojanovski and Farser 2018, Voita et al. 2018, Stojanovski and Farser 2019, Voita et al. 2019). Based on the results of the present study as well as the rapid development progress of NMT, it can be expected that an application of CL for the purpose of machine translatability will be pushed into the background in the near future. To what extent CL can in the meantime support contextual MT or help to overcome other current NMT weaknesses is a question that needs to be answered empirically through the investigation of further CL rules across various NMT systems.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Aikawa T, Schwartz L, King R, Corston-Oliver M, Lozano M (2007) Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. In: Proceedings of the eleventh machine translation Summit 10–14 September, Copenhagen, Denmark, pp 1–7

Alonso Martin JA, Serra AC (2014) Integration of a machine translation system into the editorial process flow of a daily newspaper. Procesamiento Del Lenguaje Natural 53:193–196

Banerjee S, Lavie A (2005) METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: ACL 2005, Proceedings of the workshop on intrinsic and extrinsic evaluation measures for MT and/or summarization at the 43rd Annual meeting of the association for computational linguistics, Ann Arbor, Michigan, pp 65–72

Bernth A (1999) Controlling input and output of MT for greater user acceptance. In: Proceedings of the 21st conference of translating and the computer sponsored by ASLIB, 10–11 November 1999, London

Bernth A, Gdaniec C (2001) MTranslatability. In: Machine translation, December 2001, vol. 16, no. 3. Kluwer Academic Publishers, Dordrecht, pp 175–218

Caterpillar Corporation (1974) Dictionary for Caterpillar Fundamental English. Caterpillar Corporation, Peoria

Drewer P, Ziegler W (2014) Technische Dokumentation. Eine Einführung in die übersetzungsgerechte Texterstellung und in das Content Management, 2nd edn. Vogel, Würzburg

Fiederer R, O'Brien S (2009) Quality and machine translation: a realistic objective? J Spec Transl 11:52–74

Gesellschaft für Technische Kommunikation – tekom e. V. (2013) Leitlinie "Regelbasiertes Schreiben, Deutsch für die Technische Kommunikation". 2. Erweiterte Auflage. Stuttgart.

Gonzàlez M, Giménez J (2014) An open toolkit for automatic machine translation (meta-) evaluation. Technical manual v3.0. February 2014. Technical Report LSI-14-2-T. Departamento de Lenguajes y Sistemas Informáticos, Universitat Politècnica de Catalunya

Han ALF, Wong DF, Chao LS, He L, Lu Y, Xing J, Zeng X (2013) Language-independent model for machine translation evaluation with reinforced factors. In: Proceedings of the machine translation summit XIV (MT SUMMIT 2013), International Association for Machine Translation, Nice, France, pp 215–222

Holmback H, Shubert S, Spyridakis JH (1996) Issues in conducting empirical evaluations of controlled languages. In: Adriaens G, Havenith R (eds) Proceedings of the 1st international workshop on controlled language applications, (CLAW 1996), Leuven, Belgium, pp 166–177

Huijsen WO (1998) Controlled language: an introduction. In: Proceedings of the second controlled language application workshop (CLAW 1998), Pittsburgh, Pennsylvania, pp 1–15

Hutchins J, Somers HL (1992) An introduction to machine translation. Academic Press Limited, Cambridge

Kamprath C, Adolphson E, Mitamura T, Nyberg E (1998) Controlled language for multilingual document production: experience with caterpillar technical English. In: Mitamura et al (eds) Proceedings of the second international workshop on controlled language applications—CLAW '98, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, pp 51–61

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan C, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (2007) Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, Prague, Czech Republic, pp 177–180

Lehrndorfer A, Reuther U (2008) Kontrollierte Sprache—standardisierte Sprache? In: Muthig, Jürgen (ed) Standardisierungsmethoden für die Technischen Dokumentation. Schmidt-Römhild, Lübeck. (=tekom Hochschulschriften Nr.16), pp 97–121

Lommel A (2018) Metrics for Translation Quality Assessment: a case for standardising error typologies. In: Doherty S, Castilho S, Moorkens J, Gaspari F (eds) Human and machine translation quality and evaluation—from principles to practice. Springer, Berlin, pp 109–127

Marzouk S, Hansen-Schirra S (2019) Evaluation of the impact of controlled language on neural machine translation compared to other MT architectures. Mach Transl 33(1–2):179–203

Marzouk (in press) Sprachkontrolle im Spiegel der Maschinellen Übersetzung—Untersuchung zur Wechselwirkung ausgewählter Regeln der Kontrollierten Sprache mit verschiedenen Ansätzen der Maschinellen Übersetzung. Doctoral dissertation, Johannes Gutenberg University, Germersheim

Matusov E (2019) The challenges of using neural machine translation for literature. In: Proceedings of the workshop on qualities of literary machine translation, 17th MT Summit, Dublin, Ireland, pp 10–19

Mehta S, Azarnoush B, Chen B, Saluja A, Misra V, Bihani B, Kumar R (2020) Simplify-then-translate: automatic preprocessing for black-box translation. In: Proceedings of the 34th AAAI conference on artificial intelligence, New York, 8pp

Müller M, Rios A, Voita E, Sennrich R (2018) A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In: Proceedings of the third conference on machine translation: research papers", Brussels, Belgium, pp 61–72

Nyberg E, Mitamura T (1996) Controlled language and knowledge-based machine translation: principles and practice. In: Proceedings of the first controlled language application workshop (CLAW 1996), Leuven, Belgium, pp 74–83

Nyberg E, Mitamura T, Hujisen WO (2003) Controlled langauge for authoring and translation. In: Somers H (ed) Computers and translation: a handbook, benjamins translation library, vol 35. John Benjamins Publishing Company, Amsterdam, Philadelphia, pp 71–110

O'Brien S (2006) Machine translatability and post-editing effort: an empirical study using translog and choice network analysis. PhD dissertation. Dublin City University, Ireland

Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: ACL-2002: 40th annual meeting of the association for computational linguistics, Philadelphia, PA, pp 311–318

Reuther U (2003) Two in one—can it work? Readability and translatability by means of Controlled Language. In: Proceedings of the joint conference combining the 8th international workshop of the European Association for Machine Translation and the 4th controlled language applications workshop (CLAW 2003), Dublin, Ireland, pp 124–132

Rösener C (2010) Computational linguistics in the translator's workflow—combining authoring tools and translation memory systems. In: Proceedings of the NAACL HLT 2010 workshop on computational linguistics and writing: writing processes and authoring aids, Los Angeles, California, pp 1–6

Roturier J (2006) An investigation into the impact of controlled English rules on the comprehensibility, usefulness, and acceptability of machine-translated technical documentation for French and German users. PhD dissertation, Dublin City University, Ireland

Roturier J, Mitchell L, Grabowski R, Siegel M (2012) Using automatic machine translation metrics to analyze the impact of source reformulations. In: Proceedings of the tenth biennial conference of the association for machine translation in the Americas (AMTA-2012), San Diego, CA, 10pp

Snover M, Dorr BJ, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: AMTA 2006, Proceedings of the 7th conference of the association for machine translation in the Americas, Cambridge, MA, pp 223–231

Spyridakis JH, Holmback H, Schubert SK (1997) Measuring the translatability of simplified English in procedural documents. IEEE Trans Prof Commun 40(1):4–12

Stojanovski D, Fraser A (2018) Coreference and coherence in neural machine translation: a study using oracle experiments. In: Proceedings of the third conference on machine translation (WMT), volume 1: research papers, Belgium, Brussels, pp 49–60

Stojanovski D, Fraser A (2019) Improving anaphora resolution in neural machine translation using curriculum learning. In: Proceedings of the machine translation summit 2019, Dublin, Ireland, pp 140–150

Toral A, Way A (2018) What level of quality can neural machine translation attain on literary text? In: Moorkens J, Castilho S, Gaspari F, Doherty S (eds) Translation quality assessment: from principles to practice. Springer, Cham, pp 263–287

Vilar D, Xu J, D'Haro LF, Ney H (2006) Error analysis of machine translation output. In: LREC-2006: fifth international conference on language resources and evaluation, Proceedings, Genoa, Italy, pp 697–702

Voita E, Sennrich R, Titov I (2019) Context-aware monolingual repair for neural machine translation. In: Proceedings of proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), Hong Kong, China, pp 877–886

Voita E, Serdyukov P, Sennrich R, Titov I (2018) Context-aware neural machine translation learns anaphora resolution. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), Melbourne, Australia, pp 1264–1274

Werthmann A, Witt A (2014) Maschinelle Übersetzung—Gegenwart und Perspektiven. In: Stickel G (ed) Translation and interpretation in Europe. Contributions to the annual conference 2013 of EFNIL in Vilnius. Lang, Frankfurt am Main, pp 79–103

Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, Kaiser L, Gouws S, Kato Y, Kudo T, Kazawa H, Stevens K, Kurian G, Patil N, Wang W, Young C, Smith J, Riesa J, Rudnick A, Vinyals O, Corrado G, Hughes M, Dean J (2016) Google's neural machine translation system: bridging the gap between human and machine translation. CoRR abs/1609.08144

Zhang J, Zong C (2020) Neural machine translation: challenges, progress and future. arXiv:2004.05809v1

Authors and Affiliations

Shaimaa Marzouk[1]

[1] Johannes Gutenberg University Mainz, Mainz, Germany