Machine translation as a form of feedback on L2 writing

Abstract: With advances in artificial intelligence (AI), many language teachers have started exploring the classroom implications of AI-powered technology, including machine translation (MT). To examine the usefulness of MT technology in writing instruction, we conducted a mixed-methods study comparing two types of written feedback: comprehensive direct Teacher Corrective Feedback (TCF), and MT feedback. Participants were 23 Japanese university students in an intact L2 writing classroom. Sample size adequacy was confirmed through a priori power analysis. Participants were instructed to describe a picture prompt in L2 English and then in L1 Japanese. Half the participants received first TCF then MT on their L2 English text, while the order was reversed for the other half. Participants in both conditions were then asked to study the feedback and describe the same picture prompt without the feedback. In the following phase, both groups completed the same tasks in reverse order. Participants also responded to a survey exploring their engagement with the feedback. Results reveal that: 1) TCF improved complexity; 2) MT improved accuracy and fluency; and 3) variation in outcomes may be explained by the different ways in which participants engaged with both TCF and MT. Implications for appropriate classroom use of MT are discussed.

Keywords: teacher corrective feedback; machine translation; L2 writing revision; engagement; artificial intelligence

1 Introduction

The evolution of second language (L2) writing instruction has been characterized by a quest for effective feedback mechanisms. Historically, teacher corrective feedback
(TCF) has been central to this endeavor, supporting both text production and language development (Hyland and Hyland 2019). Yet despite its enduring significance, the exclusivity of TCF in writing pedagogy has been progressively challenged (Atkinson and Tardy 2018). The limitations of TCF, including the intensive time commitment required from educators (e.g., Lee 2019) and the risk that it may lead to text appropriation by the teacher (Sommers 1982) have prompted a search for alternative strategies. In pursuit of this goal, the potential of machine translation (MT) as a supplementary or alternative feedback mechanism is an area yet to be fully explored.

Earlier versions of MT had accuracy problems at lexical, syntactic, and pragmatic levels, thus limiting its use as an instructional tool (Lee 2020). However, recent advances in artificial intelligence (AI), especially following the advent of Neural MT, have made MT a viable tool for facilitating both writing and language development (Wu et al. 2016). MT’s promise lies in its efficiency as well as its ability to reflect students’ intended meaning by translating their text from their first language (L1). This study therefore investigates the impact of TCF versus MT feedback, a novel approach that conceptualizes MT as reverse reformulation (from L1 to L2; see below).

Being an early attempt to compare the effects of TCF and MT as feedback, the study employs a mixed-methods approach, exploring not only quantitative changes in three linguistic aspects (complexity, accuracy, and fluency, or CAF) in participants’ pre- and post-revision texts but also their cognitive and affective engagement with feedback on their initial revision (Han and Hyland 2015).

2 Literature review

TCF can be classified in terms of: (1) how it is provided (direct correction vs. indirect, e.g., underlining); and (2) how comprehensively it is provided (A. focus on a particular item; B. comprehensive correction; and C. reformulation, i.e., the teacher’s rewriting the student’s entire L2 text). As mentioned above, TCF has been intensively studied from various perspectives since the formative years of L2 writing instruction (Hyland and Hyland 2019). However, we only focus here on studies closely related to research questions dealing with content and methodology.

2.1 Evolution of TCF in L2 writing: effects and drawbacks

Effects of TCF have been controversial in the field of L2 writing since Truscott (1996) claimed its uselessness. However, subsequent empirical studies and recent meta-analyses and comprehensive reviews (e.g., Li and Vuono 2019) have shown that TCF is
generally useful. For example, Kang and Han (2015) conducted a meta-analysis of 21 empirical studies and concluded that TCF does improve grammatical accuracy in L2 writing. These researchers also compared the effectiveness of various types of TCF, including direct versus indirect and focused versus comprehensive. Although the results were statistically inconclusive, they found that direct TCF is more effective than indirect TCF and that focused TCF is more effective than comprehensive TCF in terms of “substantial” (Li and Vuono 2019: 102) effect sizes. Despite these findings, we opted to use direct comprehensive TCF as the counterpart to MT because it is more similar to MT feedback than other types of TCF (e.g., indirect or focused) as it covers each learner’s entire L2 text.

Although TCF has been regarded as one of L2 “writing teachers’ most important tasks” (Hyland and Hyland 2019: xiii), it has its own drawbacks. Of these, as mentioned earlier, two of the most relevant to our study are its inefficiency and the possibility of appropriation. First, comprehensive feedback “can be exhausting for teachers to give and overwhelming for students to receive” (Ferris and Hedgcock 2023: 341), which explains why focused feedback is often recommended as a more realistic approach (e.g., Lee 2019). Yet, comprehensive TCF continues to be widely used as a result of students’ preferences, teachers’ beliefs, or institutional requirements (e.g., Lee 2008), especially in EFL contexts.

A second concern is the risk of appropriation, which has evolved over time since it was first problematized by Sommers (1982) as the teacher taking ownership of the students’ voice by rewriting their texts. In contrast, in addition to introducing five recent studies of dialogic appropriation (i.e., between teacher and student writer), Tardy (2019) demonstrated in a case study how appropriation can emerge from multiple dialogic interactions between mentor and writer, the latter being not only an L2 student but a writer with sociocultural obligations (e.g., having to fulfill a senior co-author’s intention outside the writing class). Though such appropriation processes may enable students to reflect their own voice in their writing, such interactions take up a great deal of time for both teachers and students.

2.2 Machine translation (MT) as a learning and teaching tool in L2 writing

Before 2016, when the field entered the era of Neural MT, MT output contained so many lexico-grammatical errors that most of it was categorized as “ill-formed examples of L2” (García and Pena 2011: 473) to be corrected in post-editing when used as teaching or learning aids in L2 writing. However, after 2016, many researchers started to view MT outputs as good models. In fact, Yulianto and Supriatnaningsih (2021) reported that free and fast MT platforms such as Google Translate and DeepL
were arguably comparable with human translation if the two languages belonged to the same broad family (e.g., French and English). Among those studies examining MT as a tool for helping L2 writers after 2016, most relevant to our study are those that require students to write an L1 text, translate it manually into the L2 (Own L2 Text), translate their initial L1 text using MT (MT L2 Text), and finally edit their Own L2 Text by looking at the MT L2 Text. We operationally call this approach “comparative analysis” because in the final stage, students are asked to edit their Own L2 Text by comparing it with their MT L2 Text (Table 1).

One of these earlier and exemplary studies using comparative analysis was conducted by Lee (2020), who had 34 Korean university students with 75–90 TOEFL iBT scores describe a video clip in their L1, translate their text into English, compare their output by MT of their text into English, and finally edit the English text by comparing it with the MT version. Lee’s results showed a positive effect on the edited version of the English texts, which contained fewer lexico-grammatical errors. However, Lee’s qualitative data (interviews and reflection papers) indicated that the students were divided by their L2 proficiency in their appreciation of the value of MT as a learning tool. While lower-proficiency students found MT useful for improving their writing, higher-proficiency students only found it helpful for learning vocabulary but also found MT output inaccurate in longer texts and therefore inappropriate for academic purposes. However, this may be due to the still imperfect quality of MT when Lee conducted her study.

In a later study, Lee (2022) found that higher-level (university) students benefited more from such comparative analysis than lower-level students despite receiving help from (still imperfect though probably better) MT translation. In this mixed-methods study, Lee also reported that students at all proficiency levels found MT helpful to their revisions of various linguistic elements (e.g., vocabulary, grammar, expression) as well as to their efficient and individualized provision of translated texts.

Table 1: Comparative analysis design.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Source</th>
<th>Action</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Student’s intention</td>
<td>Students write text in L1</td>
<td>Own L1 text</td>
</tr>
<tr>
<td>2</td>
<td>L1 text</td>
<td>Students translate Own L1 text into Own L2 text manually</td>
<td>Own L2 text</td>
</tr>
<tr>
<td>3</td>
<td>L1 text</td>
<td>Students translate Own L1 text using MT</td>
<td>MT L2 text</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Revise Own L2 text by comparing it with MT L2 text (using other resources such as dictionaries, as needed)</td>
<td>Revised Own L2 text</td>
</tr>
</tbody>
</table>
Lee’s (2020) study was replicated in another mixed-methods study by Yang et al. (2023) with 31 Chinese EFL university students who were comparable to those in Lee (2020). Yang et al.’s results were similar to Lee’s in that the MT L2 texts improved in quality, especially in terms of lexical complexity. In addition, 90% of the students found MT helpful in improving their writing ability even though they thought that MT translation was not perfect (see Klimova et al. 2023 for more studies using a comparative design).

2.3 MT as a feedback tool: rationale for the study

As explained in the previous section, studies using a comparative analysis design not only used MT like traditional TCF but also addressed the two drawbacks of TCF mentioned earlier. First, unlike traditional reformulation feedback, MT L2 texts more closely reflect L2 writers’ intentions as expressed in their original L1 writing. Further, the provision of MT is time-efficient as well as customized to the student’s needs because it is based on their own L2 Text, thus better reflecting each student’s voice. Nevertheless, one critical difference between the comparative analysis design and past TCF studies is that effects of TCF are measured without reference to feedback itself during the revision session, thus making it impossible to ascertain whether any knowledge or skills gained from the TCF has become internalized, especially when framed by a cognitive perspective, as were most past TCF studies (see, for example, Peng et al. 2023). In other words, though past MT studies using the comparative analysis design showed that MT was useful as a learning and teaching tool, its effects as L2 writing feedback have not been adequately assessed. Further, though most studies using comparative analysis collect qualitative data from participating students regarding their perceptions of MT’s usefulness along with their affective reaction, none of them considered these questions in relation to other types of TCF. However, As Kahneman and Tversky (1979) suggest, people’s decision-making may vary when an object is considered in isolation versus when compared with alternatives. Thus, it is crucial to systematically compare MT with other types of feedback when we consider L2 writing classrooms where multiple feedback options may be offered.

2.4 Complexity, accuracy, and fluency (CAF) as quantitative measures of the effects of feedback

Recently, an increasing number of studies have treated TCF not simply as a grammar correction tool but as a way of facilitating broader language development through
“learners’ engagement with feedback” (Manchón 2020: 14). This position is supported by studies that used not only accuracy as their measure (e.g., Benson and DeKeyser 2019) but also fluency (e.g., Criado et al. 2022), complexity (e.g., Valizadeh 2022), or all three CAF measures (e.g., Zhang and Cheng 2021). Among these measures, TCF was shown to have significant effects on accuracy, whereas the results of studies that investigated its effects on fluency and complexity have been mixed (e.g., Li and Vuono 2019).

2.5 Participant engagement as qualitative measures of the effects of feedback

As many scholars (e.g., Peng et al. 2023) have pointed out, engagement is one of the crucial determinants of TCF effectiveness. Following Ellis (2010: 342), we operationally define engagement as “how learners respond to the feedback they receive.” Compared to the effects of various types of TCF, how L2 writers engage with TCF has not been adequately investigated (e.g., Crosthwaite et al. 2022). Given Han and Hyland’s (2015) systematic and detailed classification of Ellis’s (2010) multidimensional framework for investigating engagement with TCF, the results of many subsequent studies support Ellis’s perspective but also reveal complex interactions between cognitive, behavioral, and affective engagement processing. However, the methods they employ (e.g., retrospective interviews) have their own demerits (e.g., memory loss, difficulties with quantification), which need to be overcome (e.g., Mackey and Gass 2022).

2.6 The study

On the basis of previous literature, we asked the following questions:
RQ1: Do direct TCF and MT have different effects on revised L2 texts in terms of complexity, accuracy, and fluency?
RQ2: Do the two feedback types have different effects on L2 writers’ engagement when examined from a multi-dimensional perspective?
RQ3: Which feedback type do participants find more useful or motivating and why?

These three questions will be first answered quantitatively, and the quantitative results will be supplemented by qualitative data to help us interpret the quantitative data.
3 Method

3.1 Participants

A sample of 23 participants were recruited from a second-year university English L2 writing class in Japan. Written consent was obtained from participants after they were informed that they were free to withdraw from the study at any time and that this would not affect their grades in any way. To confirm that sample size was adequate, we performed an a priori power analysis with G*Power version 3.1.9.6 (Faul et al. 2007) to estimate the number of participants needed to reach the recommended threshold of 0.80 with an alpha level of 0.05 (Cohen 1988). Using effect size ($f$) converted from Hedge's $g$ (0.68) as reported in Kang and Han's (2015) meta-analysis of TCF, the analysis confirmed that our sample size reached the desired statistical power threshold because it was over 14, the necessary number of participants in the repeated measure ANOVAs (within-by-within design) employed here. The participants’ mean TOEFL ITP score was 511.3 (SD = 27.6), indicating that their English proficiency was at about CEFR B1 level.

3.2 Procedure

We used a within-participant crossover design (Lui 2016), in which participants were randomly divided into two conditions: MT first-TCF second (MT-TCF; 11 students: 7 females), and TCF first-MT second (MT-TCF; 12 students: 2 females). We chose this design because it would help minimize individual variation caused by two feedback types while addressing the ethical issue of providing treatment to only one group, which can be a problem in a simple experiment-control design.

Data collection took place over four weeks in the same classroom. The treatment was divided into four phases (Table 2):

3.3 Types of feedback

3.3.1 Machine translation (MT)

MT feedback consisted of the Japanese text the participants had written the previous week in addition to its translation into L2 English using DeepL (Figure 1). We chose the DeepL platform (https://www.deepl.com/translator) because several studies (e.g., Yulianto and Supriatnaningsih 2021) have reported that it was one of the best MT applications available at the time. We also added the comparison between the
students’ original English text from the previous week and the MT version, with modified segments marked in light blue to make them readily comparable to the TCF feedback.

### 3.3.2 Direct comprehensive teacher corrective feedback (TCF)

For English texts in TCF condition, errors were jointly coded following the 40-category classification in Sachs and Polio (2007). The two coders were the first author and an ESL instructor with a doctorate in applied linguistics and 20 years of EFL teaching experience. Inter-coder agreement for Sessions 1 and 2 combined was 0.92, and each disagreement was resolved through discussion. Once agreed, the number of errors was also used to measure accuracy in the L2 texts. As in Sachs and Polio, direct and comprehensive feedback with insertions marked in blue and deletions marked in red (Figure 2) was returned in that format to each TCF participant.
3.4 Quantitative measures

To quantitatively compare the effects of TCF and MT (RQ1), we used the three above-mentioned CAF measures (complexity, accuracy, and fluency). For complexity, we adopted an aggregated component score that included clause complexity, phrase complexity, and syntactic sophistication (Kyle 2016). This was chosen over conventional measures such as mean length of T-units, which have been criticized as being unsuitable for measuring various stages of writing development (Norris and Ortega...
To calculate this aggregated component score, we used the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) developed by Kyle (2016). For the accuracy measure, we used the number of errors per 100 words, following Zhang and Cheng (2021), who evaluated the effects of comprehensive direct TCF on their participants’ academic writing development. We selected this measure because it better reflects accuracy in the revised texts than measures such as percentage of error-free clauses (Zhang and Cheng 2021) when many items may require correction in a single clause. Lastly, to gauge fluency, we measured the total word count produced within the allotted time.

3.5 Qualitative data

3.5.1 Post-TCF/MT engagement questionnaires

We developed two post-TCF/MT questionnaires to address RQ2, which compared the effects of the two feedback types (TCF and MT) on participants’ engagement. In crafting our questionnaires, we closely followed Dörnyei’s (2003) step-by-step procedure. First, we created a pilot version of each of the Post-TCF/MT Engagement Questionnaires based on the two dimensions of engagement used in Han and Hyland (2015: 33) and drawn from Ellis’s (2010) multi-dimensional perspective on engagement processes with feedback:

(1) Cognitive: (a) Awareness: noticing; (b) Awareness: understanding; (c) Metacognitive operations; (d) Cognitive operations deployed to process written TCF and generate revisions.

(2) Affective: (a) Immediate emotional reaction; (b) Attitudinal response.

Based on examples in Han and Hyland’s Table 1 and Appendix F along with the findings of their case study, we initially devised six items for Post-TCF Cognitive engagement, eight for Post-MT Cognitive engagement, and five each for Post-TCF and Post-MT Affective engagement, including a single open-ended item designed to explore information not elicited by the close-ended items. All items in the questionnaires were written in Japanese and translated into English by the first author, with the translations checked by the second author for validity and accuracy.

We tested these items with four volunteers from L2 writing classes similar to the one in our study and had them do exactly the same task as described above (Table 2). Based on the results of the pilot study, we revised the original items to make them fit the task at hand in three ways. First, we merged items related to the categories of “Awareness: noticing” and “Awareness: understanding” because we were unable to distinguish between them. Instead, we created a distinction between “Immediate
problem-solving operations” and “Future-oriented operations,” the former being directly related to the feedback or the text and the latter unrelated to it (Table 6a). Lastly, because the results of the pilot study showed that the participants were unable to identify meta-cognitive operations regulating mental processes in the 15 min allotted for reviewing the feedback, we removed metacognitive items. As regards “Affective engagement,” our pilot study participants used all four original close-ended items but added no additional information. We therefore kept the original four items while retaining one open-ended item for the final version. The final version of the Post-TCF/MT Engagement Questionnaire is presented in Tables 6a (Cognitive) and 6b (Affective).

3.5.2 Feedback comparison questionnaire

To answer RQ 3, we asked the following two questions, which were also tested as part of the pilot study, in order to elicit open-ended responses.

(1) Which do you think will be more useful: MT or TCF?
(2) Which do you think will be more motivating: MT or TCF?

3.5.3 Complexity-accuracy-fluency (CAF) retrospective improvement questionnaire

In Phase 4 (Table 2), an open-ended question was prepared to elicit our participants’ own account of their greatest improvement in one aspect of CAF (e.g., complexity) in their revision based on one specific feedback type (TCF or MT). For example, as Table 5 shows, Ken (all names are pseudonyms) made the greatest improvement in complexity (from 0.40 to 1.10) of all the participants when given TCF, and was therefore asked why he made such an improvement. This question was accompanied by detailed explanations using examples of the most improved parts in the participants’ original versus revised texts so as to jog their memory (10 min).

3.6 Data analysis

Before comparing the effects of the two feedback types in terms of CAF (RQ1), we examined data normality using a Shapiro–Wilk test as well as histograms (Figure 3). Due to non-normal variables, we used a robust 20 % trimmed-mean alternative to a two-way repeated measures ANOVA using Wilcox’s (2021) function. This analysis was accompanied by bootstrapping with 599 replicates for fair evaluation (Plonsky et al. 2015). The number of bootstrap samples (\( B \)) was 599, the default value appropriate for a 0.05 alpha level (Wilcox 2021). We selected these methods because no
non-parametric alternatives to two-way within-by-within ANOVA are available for non-normal data (see Larson-Hall and Herrington 2010 for robust statistics in applied linguistics).

R version 4.2.1 was used for all analyses. For reproducibility, all data and the R code used in this study are available in OSF (https://osf.io/emny7).

4 Results

4.1 Research question 1: effects of TCF versus MT on CAF

Table 3 shows the descriptive statistics for the CAF measures before and after each feedback session. Due to non-normal variables ($p < 0.05$, Shapiro-Wilk), the table also

Figure 3: Histograms of all variables.
Table 3: Descriptive statistics for CAF measures before and after feedback.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Condition</th>
<th>Time</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>20% trimmed mean</th>
<th>Shapiro-Wilk test (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity</td>
<td>TCF</td>
<td>Before</td>
<td>0.937</td>
<td>0.220</td>
<td>1.000</td>
<td>0.947</td>
<td>0.223</td>
</tr>
<tr>
<td></td>
<td></td>
<td>After</td>
<td>1.011</td>
<td>0.136</td>
<td>1.034</td>
<td>1.023</td>
<td>0.513</td>
</tr>
<tr>
<td></td>
<td>MT</td>
<td>Before</td>
<td>0.850</td>
<td>0.259</td>
<td>0.864</td>
<td>0.877</td>
<td>0.433</td>
</tr>
<tr>
<td></td>
<td></td>
<td>After</td>
<td>0.951</td>
<td>0.177</td>
<td>0.933</td>
<td>0.956</td>
<td>0.508</td>
</tr>
<tr>
<td>Accuracy</td>
<td>TCF</td>
<td>Before</td>
<td>11.777</td>
<td>4.789</td>
<td>11.340</td>
<td>11.011</td>
<td>0.081</td>
</tr>
<tr>
<td></td>
<td></td>
<td>After</td>
<td>9.427</td>
<td>5.649</td>
<td>8.540</td>
<td>8.629</td>
<td>0.063</td>
</tr>
<tr>
<td></td>
<td>After*</td>
<td></td>
<td>10.500</td>
<td>6.363</td>
<td>8.335</td>
<td>9.500</td>
<td>0.118</td>
</tr>
<tr>
<td></td>
<td>MT</td>
<td>Before</td>
<td>13.260</td>
<td>6.441</td>
<td>12.370</td>
<td>12.621</td>
<td>0.309</td>
</tr>
<tr>
<td></td>
<td></td>
<td>After</td>
<td>9.427</td>
<td>5.649</td>
<td>8.540</td>
<td>8.629</td>
<td>0.063</td>
</tr>
<tr>
<td>Fluency</td>
<td>TCF</td>
<td>Before</td>
<td>93.478</td>
<td>34.200</td>
<td>96.000</td>
<td>93.800</td>
<td>0.090</td>
</tr>
<tr>
<td></td>
<td></td>
<td>After</td>
<td>100.652</td>
<td>35.652</td>
<td>91.000</td>
<td>99.067</td>
<td>0.046</td>
</tr>
<tr>
<td></td>
<td>MT</td>
<td>Before</td>
<td>96.087</td>
<td>36.883</td>
<td>89.000</td>
<td>91.200</td>
<td>0.101</td>
</tr>
<tr>
<td></td>
<td></td>
<td>After</td>
<td>109.478</td>
<td>35.230</td>
<td>113.000</td>
<td>109.333</td>
<td>0.154</td>
</tr>
</tbody>
</table>

Note. n = 23 except for *, for which three cases are missing (n = 20).

reports the median and 20% trimmed mean. Figure 3 shows histograms for each variable before and after the TCF and MT sessions.

Table 4 summarizes the results of the 20% robust trimmed-mean test and its bootstrap-t method corresponding to a two-way repeated measures ANOVA for main effects and interactions (i.e., TCF or MT and time). Figure 4 visually represents these interactions, revealing no significant main effects for TCF or MT feedback.

Figure 4: Interaction plots of two main effects. Note. Means and standard errors were calculated based on 20% trimmed means. Error bars show 95% confidence intervals.
<table>
<thead>
<tr>
<th>Source</th>
<th>Complexity</th>
<th></th>
<th></th>
<th></th>
<th>Accuracy</th>
<th></th>
<th></th>
<th>Fluency</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Q</td>
<td>p</td>
<td>bootstrap</td>
<td>Q</td>
<td>p</td>
<td>bootstrap</td>
<td>Q</td>
<td>p</td>
<td>bootstrap</td>
</tr>
<tr>
<td>Condition (TCF/MT)</td>
<td>1.927</td>
<td>0.165</td>
<td>0.179</td>
<td>0.051</td>
<td>0.822</td>
<td>0.835</td>
<td>0.574</td>
<td>0.449</td>
<td>0.442</td>
<td></td>
</tr>
<tr>
<td>Time (before/after)</td>
<td>7.017</td>
<td>0.008</td>
<td>a</td>
<td>0.022</td>
<td>5.143</td>
<td>0.024</td>
<td>0.043</td>
<td>7.859</td>
<td>0.005</td>
<td>a</td>
</tr>
<tr>
<td>Interaction (condition × time)</td>
<td>0.001</td>
<td>0.971</td>
<td>a</td>
<td>0.970</td>
<td>0.620</td>
<td>0.431</td>
<td>0.446</td>
<td>3.882</td>
<td>0.049</td>
<td>a</td>
</tr>
</tbody>
</table>

*Note.* a *p < 0.05*. Twenty percentage trimmed means were used for both tests.
Additionally, interaction between condition and time was not significant except for the bootstrap-trimmed means of accuracy \((p = 0.049)\), which was close to the alpha level of 0.05. This suggests that neither condition significantly outperformed the other in either the pre- or post-CAF measures. However, significant main effects of time were observed in all CAF measures and were corroborated by a parametric two-way ANOVA.

To better understand these time effects, we conducted post-hoc tests using paired-sample robust \(t\)-tests with 20% trimmed means. Figure 5 summarizes the trimmed-mean differences for each CAF measure in TCF and MT conditions.

As regards complexity, a significant mean difference was observed in TCF condition, \(M_{\text{diff}} = 0.075 [0.004, 0.148], Y_t(14) = 2.260, p = 0.004\), with a medium effect size \((\xi)\) of 0.347 [0.000, 0.688]. In contrast, MT condition showed no such difference, \(M_{\text{diff}} = 0.078 [-0.190, 0.033], Y_t(14) = 1.504, p = 0.154, \xi = 0.284 [0.000, 0.732]\). Effect sizes ranged from 0 to 1 (see raw trimmed-mean differences for easier interpretation in Figure 5).

For accuracy, defined as number of errors per 100 words with lower values indicating better performance, MT condition showed significant medium effect size change post-feedback, \(M_{\text{diff}} = -3.99 [-7.043, -0.941], Y_t(14) = -2.807, p = 0.014, \xi = 0.453 [0.090, 0.807]\). In contrast, TCF condition showed no significant change, \(M_{\text{diff}} = -1.58 [-4.534, 1.378], Y_t(11) = -1.175, p = 0.265, \xi = 0.231 [0.000, 0.672]\).

In terms of fluency, MT condition exhibited a significant change, \(M_{\text{diff}} = 18.133 [4.501, 31.765], Y_t(14) = 2.853, p = 0.013, \xi = 0.303 [0.000, 0.633]\). However, TCF condition showed no significant difference, \(M_{\text{diff}} = 5.267 [-3.248, 13.781], Y_t(14) = 1.327, p = 0.206, \xi = 0.082 [0.000, 0.460]\).

In summary, while neither TCF nor MT feedback proved more effective than the other, post-hoc analysis revealed significant gains in complexity for TCF and in accuracy and fluency for MT when comparing pre- and post-feedback texts.

### 4.2 High-achieving participants’ responses to the CAF retrospective improvement questionnaire

As explained above, we administered a CAF retrospective improvement questionnaire to find out why one type of feedback (TCF or MT) facilitated improvement in certain CAF aspects of L2 texts. Even though we collected responses to this questionnaire from all participants, due to space limitations, we focus here on the responses of the top six participants (Top 2 for each aspect of CAF; Table 5). When a participant ranked high on more than two measures, the higher-ranking participant of the two candidates was chosen as the respondent for that variable.
Table 5 shows that these high-achieving students all actively engaged with the feedback. Furthermore, more than half of all responses (4 out of 6) show that their behavioral engagement sometimes went beyond the text they were reviewing (i.e., the underlined segments).

**Figure 5**: Post-hoc trimmed-means difference for each CAF measure in TCF and MT conditions. *Note.* Error bars show 95% confidence intervals.
Table 6a presents percentages of participants who chose different types of cognitive engagement in answer to the Post-TCF/MT Engagement Questionnaire. First, more participants (91.3 %; Item No. 1 in green) noticed the TCF, while 45.5 and 72.7 % compared the MT with, respectively, the Japanese and English texts they wrote the previous week (Items No. 1 and 5 in green). Second, for both TCF and MT, approximately half to two-thirds of participants (TCF: Items No. 1–3, $M_{\%} = 69.6 \%$; MT: Items No. 1–7, $M_{\%} = 52.6 \%$) engaged with immediate problem-solving (IPS) operations, such as checking errors and improving writing.
Table 6a: Percentages of participants engaging with post-TCF/MT engagement questionnaire (max. \(n = 23\)).

<table>
<thead>
<tr>
<th>Feedback type</th>
<th>Item no.</th>
<th>Sub-construct</th>
<th>Item</th>
<th>Immediate problem-solving (IPS) or future-oriented (FO)</th>
<th>% (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCF</td>
<td>1</td>
<td>Awareness: noticing</td>
<td>I compared the corrected parts with my own errors</td>
<td>IPS</td>
<td>91.3 (21)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Cognitive operations</td>
<td>In addition to (1), I wondered why the item was corrected</td>
<td>IPS</td>
<td>47.8 (11)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Cognitive operations</td>
<td>In addition to (1), I memorized the correct forms</td>
<td>IPS</td>
<td>69.6 (16)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Cognitive operations</td>
<td>In addition to (1), I wonder how I would write the part that did not exist in the original text</td>
<td>FO</td>
<td>26.1 (6)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Cognitive operations</td>
<td>I wondered how to further improve the original text by adding things that did not exist before</td>
<td>FO</td>
<td>39.1 (9)</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Cognitive operations</td>
<td>I wondered how to include expressions from my Japanese text that did not exist in my English text into my revised English text</td>
<td>FO</td>
<td>17.4 (4)</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Others</td>
<td></td>
<td></td>
<td>4.3 (1)</td>
</tr>
<tr>
<td>MT</td>
<td>1</td>
<td>Awareness: noticing</td>
<td>I compared the MT with my own Japanese text</td>
<td>IPS</td>
<td>45.5 (10)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Cognitive operations</td>
<td>In addition to (1), I wondered how my Japanese text was translated</td>
<td>IPS</td>
<td>59.1 (13)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Cognitive operations</td>
<td>In addition to (2), I tried to memorize English words and expressions I did not know</td>
<td>IPS</td>
<td>50 (11)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Cognitive operations</td>
<td>In addition to (2), I tried to memorize English words and expressions in the MT that were better than mine</td>
<td>IPS</td>
<td>45.5 (10)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Awareness: noticing</td>
<td>I compared the MT with the English text I wrote last week</td>
<td>IPS</td>
<td>72.7 (16)</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Cognitive operations</td>
<td>In addition to (5), I tried to memorize English words and expressions in the MT I did not know</td>
<td>IPS</td>
<td>50 (11)</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Cognitive operations</td>
<td>In addition to (5), I tried to memorize English words and expressions in the MT that were better than mine</td>
<td>IPS</td>
<td>45.5 (10)</td>
</tr>
</tbody>
</table>
Table 6a: (continued)

<table>
<thead>
<tr>
<th>Feedback type</th>
<th>Item no.</th>
<th>Sub-construct</th>
<th>Item</th>
<th>Immediate problem-solving (IPS) or future-oriented (FO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 Cognitive operations</td>
<td>8</td>
<td>FO</td>
<td>In addition to (5), I wondered how I could write an English text better than the MT text given to me</td>
<td>22.7 (5)</td>
</tr>
<tr>
<td>9 Cognitive operations</td>
<td>9</td>
<td>FO</td>
<td>I wondered if I could add to the revision some expressions that did not exist in the Japanese text I wrote last week</td>
<td>9.1 (2)</td>
</tr>
<tr>
<td>10 Others</td>
<td>10</td>
<td></td>
<td></td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

Notes. Items in green cells refer to Awareness: noticing for immediate problem solving. Items in white cells are for Cognitive operations for immediate problem solving. Items in blue cells are for Future oriented cognitive operations.

as “I wonder why the item was corrected” for TCF and “I wondered how my Japanese text was translated” for MT. In contrast, a greater number of participants (Items No. 4–6 in blue; $M_{\%} = 27.5\%$) engaged with future-oriented (FO) cognitive operations; e.g., “I wonder how I would write the part that did not exist in the original text”) when given TCF compared to MT (Items No. 8 and 9 in blue: $M_{\%} = 15.9\%$).

Table 6b presents percentages of participants who chose items indicating affective engagement in response to the Post-TCF/MT Questionnaire administered after they were given either TCF or MT. When independently asked, more than half the participants reported feeling positive about both TCF and MT. However, many more felt much more positive and motivated about TCF because they felt they could learn more from it than from MT.

Table 6b: Percentages of participants engaging with the post-TCF/MT affective questionnaire (max. $n = 23$).

<table>
<thead>
<tr>
<th>Feedback type</th>
<th>Item no.</th>
<th>Sub-construct</th>
<th>Item</th>
<th>TCF % ($n$)</th>
<th>MT % ($n$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCF/MT</td>
<td>1</td>
<td>Emotional response</td>
<td>The feedback made me feel disappointed about my English proficiency</td>
<td>30.4 (7)</td>
<td>22.7 (5)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Attitudinal response</td>
<td>I felt positive about the feedback because I could learn something from it</td>
<td>69.6 (16)</td>
<td>50 (11)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Attitudinal response</td>
<td>The feedback made me feel motivated because I could learn something from it</td>
<td>43.5 (9)</td>
<td>13.6 (3)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Attitudinal response</td>
<td>I felt nothing</td>
<td>2 (8.7)</td>
<td>13.6 (3)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Others (give as many details as possible)</td>
<td></td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>
Table 7: Results of feedback comparison questionnaire.

<table>
<thead>
<tr>
<th>Item no.</th>
<th>Item</th>
<th>TCF</th>
<th>MT</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Which do you think will be more useful: TCF or MT?</td>
<td>78.3%</td>
<td>21.7%</td>
</tr>
<tr>
<td>7</td>
<td>Which do you think will be more motivating: TCF or MT?</td>
<td>87.0%</td>
<td>17.0%</td>
</tr>
</tbody>
</table>

4.4 Research question 3: comparison between MT and TCF

Table 7 presents the results of the Feedback Comparison Questionnaire, which addressed Research Question 3. Although Items 2 and 3 in Table 6b and Items 6 and 7 in Table 7, respectively, are similar, their percentages are quite different, probably because Items 2 and 3 in Table 6b were asked independently for each type of feedback (TCF and MT) whereas Items 6 and 7 in Table 7 comparatively. This is especially evident in Item 2, Table 6b, where both MT (50 %) and TCF (69.6 %) received relatively positive evaluations from participants, which contrasts sharply with their comparative choices in Items 6 and 7 in Table 7, where, when asked which of the two feedback types they preferred, about 80 % or more chose TCF as more useful and motivating. In fact, their responses to the open-ended questionnaire almost unanimously (17 out of 18) showed that they found TCF to be more useful because it allowed them to see their specific grammatical problems.

5 Discussion

We will now discuss our findings in light of previous studies. Given that very few studies have examined the effects of MT as feedback, we will contrast the results of our study with those of studies focusing on TCF. We know that the results of those past studies that used MT within a comparative analysis design should be treated with caution because they all allowed their participants to use resources, including MT, during the revision sessions. From a cognitive perspective (e.g., Peng et al. 2023), these results are not comparable with those of TCF studies as the improvement evident in those studies’ final revised texts may not reflect knowledge truly acquired by the writers because the improvement demonstrated in the final revised texts of MT studies may simply reflect the fact that writers copied material from the resources available while revising. We will therefore refrain from referring to the results of MT studies unless they are directly relevant.

Quantitatively, three findings emerge in our study: complexity tends to improve with TCF, and accuracy and fluency tend to improve with MT. First, TCF had a
positive effect on complexity measured as an aggregated component score that included clause complexity, phrase complexity, and syntactic sophistication. This is noteworthy since TCF was offered only once for a 15 min review before revision. Traditionally, complexity has not been extensively studied along with effects of TCF in L2 writing. Even when complexity was used as an exclusive tool to measure effects of TCF, as in Valizadeh (2022), where direct comprehensive TCF had positive effects on complexity in the post-test (but not in the delayed test), such mixed results were only explained from an outsider (etic) perspective.

In contrast, in our TCF condition, students writing more complex sentences could be explored from three perspectives. First, TCF has only one central focus of attention (the original English text containing error corrections) whereas MT has more than one such source (the Japanese text, its translation, and the original English text). Our questionnaire results (Table 6a), which indicate that TCF led a higher percentage of participants (91.3 %: Item 1) to notice the feedback compared to MT (45.5 %: Item 1; 72.7 %: Item 5), support this assumption. Moreover, when given TCF, more participants (27.5 %: Mean of No. 4–6) engaged with future-oriented cognitive operations compared to MT (15.9 %: Mean of No. 8 and 9). Recall that in both TCF and MT conditions, the revision was made after the writers were required to describe the same picture prompt from the beginning with no resources available after having done their best to review the given feedback and other resources (Table 2). In TCF condition, students may be better able to refine their English originals by “adding parts” (see Ken; Table 5) or by “trying to use more complex sentence structures” (Shou) compared to those who received MT feedback, who may have been busy making the most of a greater number of resources available during the review session.

Our second quantitatively notable result is that MT contributed to improvements in accuracy measured as the number of errors per 100 words. This contradicts the results of Zhang and Cheng (2021), who examined the effects of comprehensive direct TCF and found a significant difference on the same measure between the experimental and the control groups following a 16-week semester with four comprehensive TCF sessions plus three weeks for the delayed session in university English writing classes in China. In our study, in terms of engagement, 72.7 % of students (Item No. 5) in MT condition compared the MT feedback, i.e., the high accuracy level of L2 translations of the meaning they intended to convey, with their original English texts written at their own accuracy level in English, during the 15 min review session (Table 6a). This may explain the improvement in accuracy in their revised texts as both students who ranked high in accuracy (Table 5) mentioned that such a comparison when given MT was useful. However, there was a gap between MT’s contribution to accuracy and participants’ perceptions. A larger number of participants rated TCF positively and deemed it more motivating than MT
(Table 6b), especially when the two approaches were directly compared (Table 7), because they thought that TCF more closely catered for their specific grammatical errors.

Finally, MT also improved participants’ fluency between original and revised texts. This result can be viewed in differences in future-oriented cognitive engagement elicited by the TCF and MT conditions. As Shou’s Texts 1 and 2 before and after TCF show, when students in TCF condition were trying to use more complex sentence structures, that is, when they engaged in future-oriented cognitive operations, they tended to generate more new expressions related to the L2 original (e.g., “I felt that I needed to show development in my writing after receiving the feedback” – Shou; Table 5).

**Shou’s Text 1** (after TCF)*

```
Two *men* talked about a dinner party. A couple made some dishes like grilled fish, and another couple put on a dress or a suit. In *the* dinner party, somehow, the grilled fish *had disappeared*. A *man* rushed into a pizza shop and bought pizza. After that, *a the* dinner party started.
```

*Note: Shou’s fluency [No. of words = 55] is lower than that [M = 96.13, SD = 36.90] of the other participants in the MT condition*

**Shou’s Text 2** (Writing for Prompt 2 for the second time; green: fully corrected; blue: new errors, corrected in this new text)

```
Two *men* are talking about a dinner party. One couple is cooking a large fish dish. The other couple is wearing a suit and a dress. *When the dinner party starts*, the big fish dish has somehow disappeared. A *man* rushes to the pizza shop and *buys* a pizza. While eating the pizza, the two couples enjoy the *dinner party*.
```

However, as Shou’s Text 2 shows, this operation did not necessarily lead to more accurate or longer texts. In contrast, in MT condition, the product of writers’ future-oriented cognitive engagement with the feedback tended to relate to entirely new content. Based on the qualitative data presented in Table 5 and the information shown in Table 6a, we can speculate that during the review session, the MT participants focused on making the best of MT by memorizing expressions from the feedback so as to create a better text than their English original, as exemplified in Maya’s response to why her revised text became longer than her original:
The MT feedback gave me a greater variety of expressions I could use, and I could relate them to the expressions I had used in the original text in the revision session, which led to a longer text.

Together with their cognitive operations in comparing the MT with the English original (Table 6b), in MT condition, students improved in both accuracy and fluency by learning from the MT feedback, which provided near-perfect translations of what they originally wrote in Japanese (see also Yuki’s excerpt in Table 5).

Lastly, in terms of research design, the results of our study highlight several merits of the methods we applied and suggest areas that remain to be investigated. First, we used a crossover research design in which we could compare MT and TCF in a real L2 writing classroom so as to enable all participants to experience the respective merits of each type of feedback. This method minimizes ethical concerns as well as possible hidden influences of individual differences when comparing completely different groups (e.g., experimental vs. comparison; see Lui 2016). That said, because we used an intact L2 writing class, the sample size was inevitably small. Although we used a priori power analysis to test the sufficiency of the sample size for our statistical analyses, it would be useful to verify the results with a larger sample size. Second, insider (emic) data revealed by the mixed-methods design supplemented the quantitative results in ways that even a large-scale quantitative study may not reveal.

Our study is one of the first to use such data collected through theoretically framed surveys (e.g., Ellis 2010) and improved based on the results of a pilot study (e.g., through the addition of a number of “immediate problem-solving” vs. “future-oriented” items) or based on the quantitative results (CAF retrospective improvement questionnaire). Although questionnaires and post-hoc questions have demerits (Mackey and Gass 2022), if TCF continues to be viewed as a tool for L2 development (Manchón 2020), examining why certain types of TCF are facilitative for certain linguistic aspects of L2 texts should be useful.

In terms of targeted tasks, we chose picture prompts for the writing tasks to avoid unfamiliar content or language in the prompts themselves. To examine the impact of MT on student writing and learning, it will be important to conduct similar studies with a wider variety of genres and task types, both academic and non-academic. Another key factor, especially given the need for larger sample size in future studies, is students’ L2 proficiency level. In a recent MT study using a comparative analysis design, Lee (2022) reported that higher-level L2 proficiency students tended to benefit much more from MT than lower-level students in terms of improvement in the quality of their revised texts. Hence, the benefits received from TCF and MT may also differ depending on the learners’ L2 proficiency.
6 Conclusion

Comprehensive direct TCF and MT as feedback are similar in that they both provide L2 writers with possible canonical forms throughout the written text. Our quantitative results show that based on CAF measures, both types of feedback produced comparable improvements in participants’ writing. In addition, some participants found MT feedback both useful and motivating. Given these findings, MT feedback appears to be a viable alternative to TCF. However, this does not mean that MT can entirely replace TCF. Our analysis also indicates that TCF contributed more to complexity whereas MT helped improve accuracy and fluency. Furthermore, participants tended to feel more positive and much more motivated about TCF than about MT. Given different ways in which TCF and MT contributed to students’ language development and given students’ preferences for TCF, it may be advisable to use MT as a complement to TCF rather than as a replacement.

Using MT as an additional form of feedback has a number of advantages. First, it can reduce teachers’ burden. Instead of spending a large amount of time correcting student writing, teachers can provide fewer language-based comments, using MT for more comprehensive feedback. It also allows teachers to spend more time focusing on aspects of writing other than language – including writing purpose, audience, use of sources, and genre-specific features. MT feedback can also provide a wider variety of sentence structures, vocabulary, and expressions, especially in EFL contexts, where students’ exposure to the target language is relatively limited. Finally, MT can help students use the idea development and organization skills they have already developed in their L1. As we saw in Yuki’s case in Table 5, this benefit may be especially important for students with lower L2 proficiency levels, where the challenge of writing directly in English with limited L2 resources may prevent them from developing and expressing their ideas fully in the target language (e.g., van Weijen et al. 2009).

One challenge in implementing MT feedback is students’ perceptions. As noted above, most participants in this study found TCF more motivating than MT because they thought that TCF was more closely customized to their specific grammar issues even though MT contributed more to accuracy. To maximize student engagement, it may be necessary to explain to them the benefits of MT feedback as documented in Lee (2019) as well as in this study. Since our participants did value MT when asked independently, it should be possible to persuade students to see MT as a viable complement to teacher-provided feedback.

Research funding: This work was supported by MEXT/JPSP KAKENHI Grant No. 20H01286.
References


