ELIA
Estudios de Lingüística Inglesa Aplicada

ROUND-TRIP TRANSLATION AS A WRITING TOOL FOR ENGLISH AS A SECOND LANGUAGE

LA TRADUCCIÓN DE IDA Y VUELTA COMO HERRAMIENTA DE ESCRITURA EN EL INGLÉS COMO SEGUNDA LENGUA

Juan Rafael Zamorano-Mansilla
Universidad Complutense de Madrid
jrzamora@ucm.es

Abstract

The potential of the technique known as round-trip translation to detect errors in language use has been exploited in the design of programs for automatic error detection (Hermet & Désilets, 2009; Madnani et al., 2012), but to my knowledge, no study has explored the potential of translators as a tool that learners themselves can use to correct their writing in a second language. Consequently, there is no information as to how many of the transformations introduced by round-trip translation are useful for learners, how many simply rephrase the original text, or how many actually make it worse. Hermet and Désilets (2009) report a "repair rate" of 66.4% working with prepositions in French, while Madnani et al. (2012) report 36% successful changes, 33% paraphrasing and 31% changes for the worse in 200 sentences in English. The present study found a significant improvement in the number of corrections in texts written in English by Spanish students (97%) at the cost of generating an excessive number of false positives

(34%). The most reliable transformations are those affecting spelling or word morphology, which correct errors in 88.33% and 78.57% of cases, respectively. These results show the progress made in machine translation and the reliability of the round-trip translation technique for correcting errors and inform which transformations are most useful.

Keywords: round-trip translation; English L2 writing; error detection; machine translation; language learning

Resumen

El potencial de la técnica conocida como traducción de ida y vuelta para detectar errores se ha empleado en el diseño de programas para la detección automática de errores (Hermet y Désilets, 2009; Madnani et al., 2012), pero hasta ahora ningún estudio ha explorado el potencial de los traductores como herramienta que los propios estudiantes pueden utilizar para corregir su producción escrita. En consecuencia, no hay información sobre cuántas de las transformaciones introducidas por la traducción de ida y vuelta son útiles para los estudiantes, cuántas se limitan a reformular el texto original o cuántas lo empeoran. Hermet y Désilets (2009) dan una "tasa de reparación" del 66,4% aplicado a las preposiciones en francés, mientras que Madnani et al. (2012) reportan un 36% de cambios exitosos, un 33% de parafraseos y un 31% de cambios a peor en 200 oraciones en inglés. En el presente estudio se ha descubierto una mejora significativa en el número de correcciones en textos escritos en inglés por hablantes nativos de español peninsular (97%) a costa de generar un número excesivo de falsos positivos (34%). Las transformaciones más fiables afectan a la ortografía o la morfología de las palabras, que corrigen los errores en un 88,33% y un 78,57 de los casos, respectivamente. Estos resultados muestran el avance de la traducción automática y la fiabilidad de la traducción de ida y vuelta para corregir errores e informan de qué transformaciones son más útiles.

Palabras clave: traducción de ida y vuelta; escritura en inglés L2; detección de errores; traducción automática; aprendizaje de idiomas

1. Introduction

The development of technology has brought a wide range of tools for language instruction. Among them, there is the family of tools designed to find problems with language use, such as Grammarly, Quillbot or Microsoft's spell and grammar checker. Although translators are not designed for this purpose, they will correct errors when performing what is known as round-trip translation (RTT), which involves translating a text into a target language and back again into the source language – for example, English-Spanish-English.

The potential of RTT to find mistakes – mainly those concerning spelling and grammar – has been exploited in the design of programs for automated grammatical error detection (Hermet & Désilets, 2009; Madnani et al., 2012), but to my knowledge, no research has explored the potential of translators as a tool that students can use themselves to correct their writing in a second language. Previous studies have focused on the effects of using translators freely – normally, to translate a text from the student's L1 to L2 – or to translate fragments of text (e.g., Cancino & Panes, 2021; Chon et al., 2021; Chung & Ahn, 2021; Kol et al., 2018; Mujtaba et al., 2022; O'Neill, 2016). Less frequently, previous studies have investigated the use of translators as a reference for comparisons between the student's translation into the L2 and the machine output, as in Lee (2020) and Tsai (2019).

This study intends to investigate the strengths and weaknesses of using RTT as a tool that students can use to correct their writing in English as an L2. To that end, a sample of 4,871 words written by Spanish students of English as an L2 was translated into Spanish and back into English. The 723 differences between the original texts and the result of RTT were analyzed to determine the extent to which these differences provided corrections or useful suggestions for learners. The results can inform teachers and students' decisions regarding the use of RTT for pedagogical purposes.

2. Literature Review

2.1. Machine Translation and Foreign Language Instruction

Since the popularization of machine translation (MT) with the advent of free web-based online translators, there has been an interest in exploring its relation to language instruction. While some instructors have expressed their reservations about the ethics of using this technology (e.g., McCarthy, 2004; Stapleton, 2005; Steding, 2009; Correa, 2011, 2014), it seems that the prevailing attitude today is embracing MT as one more tool available to learners, provided they are informed of its advantages and limitations (e.g., Benda, 2013; Clifford et al., 2013; Ducar & Schocket, 2018; Groves & Mundt, 2015; Stapleton & Leung Ka Kin, 2019; Wallwork, 2016;).

Niño (2009) summarizes the main applications of MT in language instruction: (a) MT as a "bad model," where students are asked to correct the mistakes made by the machine; (b) MT as a "good model," which involves the use of translation memories as a reference tool for autonomous learning ; (c) MT as a vocational tool, which is more relevant to translation students; and (d) MT as a CALL tool, where teachers design activities with the aim of making students interact with the translator and make decisions about the output.

Several studies have investigated the effects of using translators as resources for writing in a second language. O'Neill (2016) compared the writing performance of three groups: one with no access to technology; one with access to a translator but no previous training on how to use it; and one with access to a translator and previous training. The students in O'Neill (2016) were native speakers of English who studied French as an L2. The results showed that all three groups had similar performance, with the last group outscoring the other two in grammar accuracy and spelling, while the difference in areas like syntax, vocabulary and content was not statistically significant. Five years later, in a similar study involving Spanish-

speakers learning English, (Cancino & Panes, 2021) reported-that students with access to a translator not only wrote more words and with more grammatical accuracy, but also displayed more syntactic complexity. Working with students in Pakistan who learned English as an L2, Mujtaba et al. (2022) reported that students improved in lexical variety and sentence complexity. Accuracy also improved, but the effect was more visible in less advanced writers. Chon et al. (2021) also reported an improvement in lexical variety and sentence complexity, as well as a reduction in grammar errors. However, a taxonomy of errors allowed the researchers to identify certain grammar points where MT was especially beneficial (articles and prepositions), and some problems exacerbated by MT (mistranslations and poor word choice). Chung and Ahn (2021), also working with Korean-speaking students, found that MT had greater advantages to offer for less proficient writers. Accuracy improved for all levels of proficiency, but clause subordination and the use of sophisticated vocabulary tended to decrease, which the authors attributed to MT providing grammatically correct but simple structures – at least when translating from Korean into English. In fact, it seems that more advanced students pay less attention to accuracy than they do to overall structural and lexical complexity when analyzing and editing the results of MT, while the reverse is true of less advanced students (Chung, 2020). In Kol et al. (2018), the same group of students had to write compositions with and without access to a translator. The students were from Israel, and English was their L2. The authors found that students wrote more when assisted by a translator, and that their vocabulary profile improved. However, the authors did not find a statistically significant difference in grammar accuracy. More noteworthy, the authors found that, although the majority of students regularly used online translators, they said they did so to translate words or phrases, and rarely to translate whole sentences or paragraphs. Some of the reasons the students proffered were that they wanted to practice their English or that they did not want to become too dependent] on MT.

A slightly different approach uses MT as a substitute for peer or teacher revision. In Lee (2020) and Tsai (2019), students still wrote the original text in their first language (Korean and Mandarin Chinese respectively), but then compared their own translation into English with the computer-generated translation. Both studies concluded that MT helped students write more words, produce fewer grammar errors, and use more complex vocabulary and more idiomatic language.

In conclusion, greater grammatical accuracy in the L2 is the effect of MT most often reported in these studies, with the qualification that this effect is more marked in less advanced learners. Increased vocabulary and syntactic complexity are also often mentioned in these studies, although the results in these areas are more contradictory.

2.2. The Concept of Error

What constitutes an error and how errors can be classified are crucial aspects in this study, as will become clear in the methodology section. A review of the literature reveals that the notion of error is not a discrete category. Influential works like Corder (1971), Lennon (1991) and Ellis (1994) recognize that some examples are easier to judge than others, and that teachers and native speakers may differ in their judgement. Clear instances of error are often associated in these works with grammar rules, while problematic cases are associated with semantics and pragmatics. However, this is only a tendency. Consequently, judging what is an error depends to some extent on a subjective element, especially for what Lennon (1991) calls the "middle ground". In fact, this author proposes placing the concept of error on a continuum, and Pawlak (2013) extends its definition to cover not just any divergence from the native model, as is the traditional definition, but also whatever the teacher perceives as requiring correction.

In spite of the elusiveness of the notion, different taxonomies have been proposed for the classification of errors. In the 1980s, a review of the literature by Dulay et al. (1982) identified four bases commonly used in error classifications: (a) the linguistic unit involved; (b) the surface strategy, which describes how the error differs from the target form; (c) a comparative analysis with the L1 or with interlanguage development; and (d) the communicative effect of the error. The first two have been employed more frequently, as can be seen in James (1998) and in Díaz-Negrillo and Fernández-Domínguez (2006), who reviewed several error taxonomies designed for corpus annotation. Taxonomies can vary in the tags they contain and how they are organized. For example, Dagneaux et al.'s (1998) taxonomy, applied to the International Corpus of Learner English, consists of areas or levels of analysis (Formal, Grammatical, Lexical, Word, Register, Style) which can be further refined through various layers of subtags specifying the linguistic unit or phenomenon involved (such as *article* or *gender*) and occasionally, the target modification (*word missing*). By contrast, Nicholls' (2003) taxonomy, designed for the Cambridge Learner Corpus, combines a list of target modifications (*wrong form*, *something is missing*, *something must be replaced, etc.*) with the word-class involved (verb, noun, etc.) and punctuation. The resulting combinations are supplemented with special error types, such as *incorrect argument structure, incorrect word order*, *collocation error*, etc.

2.3. RTT in Automated Grammatical Error Detection

The identification of errors by a machine is the goal of the discipline known as automated error detection, and MT translation is one of the methods that have been proposed for that task (Heift & Schulze, 2007; Leacock et al., 2010; Rauf et al., 2017). The most common approach sees error detection and correction as a translation process

from L2 English to standard, native-model English (e.g., Brockett et al., 2006; Park & Levy, 2011; West et al., 2011; Yuan & Felice, 2013).

RTT, the approach explored in the present study, is much less common. Hermet and Désilets (2009) observed how many preposition errors in French L2 were corrected when translating a set of sentences into English and back into French. The authors reported a "repair rate" (i.e., a percentage of errors corrected) of 66.4%. However, they also reported a repair rate at "clause level" to account for the fact some sentences underwent wider changes. Some of these changes yielded more idiomatic results (e.g., *sur la scène du crime* > *sur le lieux du crime*), resulting in a clause repair rate of 44.8%.

Madnani et al. (2012) applied RTT to 200 English sentences containing a wide range of errors. The authors generated six corrections for each sentence corresponding to six different methods based on the output of RTT. The evaluation was based on the judgement of two experts, who classified each correction as a success, a draw or a failure. A success is a grammatical improvement of the original that preserves the meaning. A draw is a correction that provides no grammatical improvement, but still preserves the original meaning. Corrections that are less grammatical than the original or change the original meaning are considered failures. Of the six methods, the one with the best results provided 36% successes, 33% draws and 31% failures.

3. Methodology

3.1. Research Questions

In order to assess the usefulness of RTT as a tool for writing in English L2, the differences between original texts and the result of performing RTT were analyzed. In what follows, these differences will be referred to as *RTT modifications*. These are the research questions the present study attempted to answer:

— RQ1: How many of the RTT modifications correct errors?

— RQ2: Are some RTT modifications more useful than others?

RQ1 will reveal if the advances in MT since Hermet & Désilets (2009) and Madnani et al. (2012) yield better results for RTT, and RQ2 will allow a comparison with the results reported for translators as writing tools in previous studies.

3.2. Data Collection

The data for this study were obtained applying the RTT technique to 30 short writings from 30 Spanish-speaking learners of English. The 30 writings were between 100-230 words each (totalling 4,871 words) and were part of three different course activities. The learners belonged to the second year of the Modern Languages degree at a Spanish university, which is officially described as a B2 level according to the Common European Framework of Reference. The activities were completed outside the classroom and had to be submitted in electronic format.

3.3. Translator

The translator used in the present study was DeepL, mainly because it is a good representative of the deep neural network translators currently available online. A comparison between different free web-based translators falls outside the scope of this paper. The RTT technique was applied by translating the students' writings into Spanish and then back into English. The Spanish intermediary text was not revised, so there was no reason for using this particular language other than the fact that one intermediary language had to be selected. Again, a comparison between the results obtained with different intermediary languages would require a separate study.

4. Data Analysis

Evaluating the usefulness of RTT modifications requires defining: (a) what counts as a modification; (b) what is a useful modification.

The quantification of RTT modifications was done manually. Unlike string-based algorithms, the quantification used in the present study is inspired by how humans process differences between texts, which is in terms of the syntactic structure and hierarchy of constituents rather than strings. This will be illustrated with an example of how MS Word's revision tool analyzes differences between texts and how humans process them.

In the pair *It could be about this last one, because…* : *It could be the latter, since…*, MS Word analyzes this as one deletion (*about this last one, because*) and one insertion (*the latter, since*). However, a typical human interpretation of the differences between the two versions is more likely to be like this:

— The preposition *about* has been deleted in the second version.

— The phrase *this last one* has been replaced with *the latter*.

— The conjunction *because* has been replaced with the conjunction *since*.

In other words, humans are apt to respond to changes in the syntactic structure and make connections or alignments between the constituents that have been deleted and those that have been inserted. Furthermore, from a language learner's point of view, these changes raise the following questions:

— Is the preposition *about* incorrect in this sentence or simply optional?

— What is the difference between *this last one* and *the latter*? Are they both equally acceptable?

— What is the difference between *because* and *since*? Do they have exactly the same meaning?

This quantification method makes it possible to isolate modifications so they can be counted and evaluated individually. Furthermore, the method is closer to describing how hypothetical learners would interpret the differences between their original writing and the RTT version. However, it has the disadvantage of being potentially subjective and inconsistent. For this reason, the differences between the original writings and their RTT counterpart were also quantified using the online tool *String similarity test*[1], which is based on an algorithm for string comparison that identifies chunks of diverging text and computes the operations necessary to make it identical to each other in terms of deletions and insertions (Myers, 1986). The results, given in similarity percentages, were used to test how different the manual quantification of modifications was from an automated, string-based comparison between the original texts and the RTT versions.

For the qualitative analysis of RTT modifications, a double coding system was used: an evaluative coding and a taxonomic coding.

The evaluative coding classifies RTT modifications into four categories:

— Correction. Modifications that correct errors.

— Suggestion. Modifications that improve the naturalness or fluency of the text but cannot be considered error corrections.

— Alternative. Modifications that simply offer an equally normative and natural alternative to express the same meaning.

— Blunder. Modifications that fail to correct an error, introduce an error where there was none, or change the meaning of the original.

[1] https://www.tools4noobs.com/online_tools/string_similarity/

The taxonomic coding primarily describes the change and language level involved:

— Spelling: A change in how a word is written.

— Morphology: A change in the inflection or derivation of a word.

— Addition: The insertion of a word that is absent in the original text.

— Deletion: The removal of a word from the original text.

— Replacement: The substitution of a word for another word of the same class, or a prepositional phrase for an adverb with the same function.

— Word order: The relocation of a word within a sentence without any rearrangement of the syntactic structure.

— Rearrangement: The modification of the syntactic structure of a sentence or a phrase. For instance, from the active voice to the passive voice; from the nominal construction 'noun *of* noun' to 'noun *'s* noun', etc.

This was complemented with part-of-speech (POS) tagging, except for word order and rearrangement, which involve more than one word.

5. Results

The comparison of the 30 compositions (4,871 words) with their RTT counterpart produced 723 modifications. Table 1 shows the number of modifications identified in each writing, the number of modifications per word for normalization, and the percentage of string-based similarity between the original and the RRT version. A strong correlation between the normalized frequencies of the manual

quantification and the string-based similarity percentages was found, $r\,(28) = -.78$, $p < .001$. Such correlation would suggest that the manual quantification applied in this study has a degree of consistency and reliability similar to that of computerized string-based methods.

Table 1: Number of modifications manually identified (n), number of modifications per word (w/n), and string similarity percentage (%)

Text id	n	w/n	%
1	16	8.56	89,45
2	13	13.23	89,34
3	44	3.11	81,95
4	26	5.73	85,24
5	39	5.33	85,48
6	26	5.69	84,36
7	23	6.17	85,79
8	25	8.24	85,77
9	8	16.62	90,66
10	36	3.69	81,13
11	40	3.92	82,06
12	33	5.36	82,95
13	23	6.87	87,96
14	30	6.77	87,14
15	32	4	82,45
16	30	6.87	85,73
17	27	8.04	91,5
18	49	3.53	79,31
19	33	6.48	88,68
20	26	8.96	86,74
21	10	9.9	88,65
22	25	6.4	85,15
23	10	14.7	92,97
24	32	4.75	87,02
25	12	10.5	85,9

26	9	13	90,37
27	14	11.28	93,13
28	8	16	93,35
29	18	12.89	91,38
30	6	20.17	89,25
	723		

Remarkably, while the number of modifications per word varies widely (from one modification every 20 words to one every 3 words), string similarity is never lower than 79.31%. This is because texts with a high number of modifications contain more spelling and morphology changes (such as *wich* > *which*; *womans* > *women*), and these pairs still have many characters in common.

Table 2 shows the effect of the RTT modifications according to the evaluative coding defined in the methodology, whereas Table 3 shows the effect of each modification type defined in the taxonomic coding.

Table 2: Effect of RTT modifications

Corrections	Suggestions	Alternatives	Blunders	Total
248 (34.30%)	112 (15.49%)	338 (46.75%)	25 (3.46%)	723

Table 3: Effect of RTT by modification type

Modification	Corrections		Suggestions		Alternatives		Blunders		Total
	n	%	n	%	n	%	n	%	
Spelling	53	88.33	0	0.00	6	10.00	1	1.67	60
Morphology	44	78.57	3	5.36	6	10.71	3	5.36	56
Replacement	60	17.60	60	17.60	211	61.88	10	2.93	341
Addition	29	43.28	7	10.45	25	37.31	6	8.96	67
Deletion	37	45.12	8	9.76	29	35.37	8	9.76	82
Rearrangement	25	22.52	33	29.73	48	43.24	5	4.50	111
Word order	3	23.08	3	23.08	7	53.85	0	0.00	13

Two modification types are particularly likely to be corrections: when a word is changed in its morphology and when a word is changed in its spelling. The least corrective modification is replacement, with only 17.6% of the cases actually correcting an error. However, not all modification types are equally represented in the data. In particular, the narrow criteria employed for word-order modifications make it anecdotal in the study.

In terms of precision and recall, which are common metrics for the evaluation of error detection systems (Chodorow et al., 2012), the results are:

— Recall = 0.97 (248 true positives for 255 critical errors identified in the dataset)

— Precision = 0.34 (468 false positives)

The sum of true positives and false positives is 716, not 723 as expected, because the seven blunders where the RTT version modified an error but failed to fix it were excluded. Such examples cannot be considered either false negatives, since they were flagged in a way, or true positives, since the result was not successful. The blunders computed as false positives here are cases where the RTT modification was less correct than the original.

The following sections provide more information about the effects of each modification type.

5.1 Spelling

The modifications classified as *spelling* in the taxonomic coding were not classified further since the extra information was not expected to provide any insights. When RTT changed the spelling of a word it was virtually always to correct a spelling mistake. The only six alternatives are cases where capitalization is possible but not obligatory (*emperor Claudius* vs *Emperor Claudius*) or where

different national variants exist (*honor* vs *honour*). In addition, RTT also handled adequately all typos (*realted* > *related*; *vey positive* > *very positive*) and repetitions (*200 hundred people* > *200 people*). The only blunder is the incorrect transformation of the verb *singed* into *sung*: the problem in the original writing was the spelling of the verb *sign*, not the participle of the verb *sing*.

5.2 Morphology

Table 4: Effect of morphological modification by POS

POS	Correction	Suggestion	Alternative	Blunder
Pronoun	1 (100%)			
Determiner	3 (100%)			
Adjective	1 (100%)			
Noun	13 (81.25%)	1 (6.25%)	2 (12.5%)	
Verb	26 (74.29%)	2 (5.71%)	4 (11.43%)	3 (8.57%)

Because English is not a heavily inflected language, this category contains only 56 examples distributed mainly between verbs and nouns. Corrections in this area include problems with agreement (*this problems* > *these problems*; *it coincide with* > *it coincides with*), conjugation (*costed; didn't accepted*) and comparison (*biggest than* > *bigger than*), but the most frequent cases concern irregular plural nouns (n =12) and the use of tenses (n =17). The three suggestions are cases where the selection of a singular/plural noun or a verb tense made the original more natural or idiomatic but could not be said, in my judgement, to correct a serious problem: *some comments said* > *some comments say; explain any side effect* > *explain any side effects*. The six alternatives were all cases where usage and prescription accept competing forms: *the data is* > *the data are*; *began to give*> *began giving; in every country* > *in all countries*. Finally, the three blunders are examples of tense selection (*it was necessary to have the help of volunteers* > *it has been necessary to have the help of volunteers*).

5.3 Addition and Deletion

Table 5: *Effect of adding a word by POS*

POS	Corrections	Suggestion	Alternative	Blunder
Pronoun	4 (80%)		1 (20%)	
Conjunction			3 (100%)	
Determiner	16 (61.54%)		4 (15.38%)	6 (23.08%)
Adjective	2 (100%)			
Noun		1 (100%)		
Preposition		1 (16.67%)	5 (83.33%)	
Verb	1 (14.29%)	4 (57.14%)	2 (28.57%)	
Adverb	1 (50%)	1 (50%)		
Punctuation	5 (33.33%)		10 (66.67%)	

Table 6: *Effect of deleting a word by POS*

POS	Corrections	Suggestion	Alternative	Blunder
Pronoun	5 (35.71%)	2 (14.29%)	5 (35.71%)	2 (14.29%)
Conjunction	1 (100%)			
Determiner	25 (64.10%)	2 (5.13%)	7 (17.95%)	5 (12.82%)
Adjective			1 (50%)	1 (50%)
Noun	1 (33.33%)		2 (66.67%)	
Quantifier	1 (50%)	1 (50%)		
Preposition	1 (50%)		1 (50%)	
Verb	1 (8.33%)	2 (16.67%)	9 (75%)	
Adverb		1 (50%)	1 (50%)	
Punctuation	2 (40%)		3 (60%)	

Most corrections in both addition and deletion involve the use of determiners, mainly *the*. Pronouns are the second most common correction due to examples of ellipsis (*the experiment had positive results but (it) had different effects*) or to complete the syntactic structure of the sentence (*Valentine's day (it) is celebrated on; But these days the festivity is not religious or pagan anymore, (it) is just a capitalist celebration*). Punctuation also stands out from the rest of categories, particularly for addition.

The cases classified as suggestions are examples where the addition or deletion of a word could not be said to correct a grammar mistake, but made the text more natural, cohesive or concise: *and a tradition was born > and thus a tradition was born; the festivity at first comes from a pagan celebration > the festivity comes from a pagan celebration.*

Alternatives represent a high percentage of additions and deletions, almost as high as corrections. These are cases where (a) punctuation is optional, as in the Oxford comma; (b) English grammar allows two constructions (*helped (to) form; say (that); (in) this way*); or (c) the transformation could not be said to improve the original (*which earned him a reputation > earning him a reputation*).

Finally, most of the blunders (11 out of 14 cases) for addition and deletion involve the use of *the*. Two of the blunders ignore the rule for ellipsis of subjects after a comma (*, and has > , and it has*), perhaps reflecting widespread practice among native speakers in non-academic writing. The only blunder that does not involve a determiner is a case where information was omitted from the original for no apparent reason: *such as marijuana and the emergence of rampant STDs > such as marijuana and the emergence of STDs.*

5.4 Replacement

Table 7: *Effect of substituting a word for another by POS*

POS	Corrections	Suggestion	Alternative	Blunder
Pronoun	6 (24%)	3 (12%)	15 (60%)	1 (4%)
Conjunction	1 (9.09%)		10 (90.91%)	
Determiner	3 (30%)	4 (40%)	3 (30%)	
Adjective	5 (17.24%)	8 (27.59%)	14 (48.28%)	2 (6.90%)
Noun	5 (5.81%)	13 (15.12%)	63 (73.26%)	5 (5.81%)
Quantifier	1 (14.29%)	2 (28.57%)	4 (57.14%)	
Preposition	20 (32.26%)	13 (20.97%)	29 (46.77%)	
Verb	13 (18.57%)	14 (20%)	41 (58.57%)	2 (2.86%)
Adverb	4 (10.26%)	3 (7.69%)	32 (82.05%)	
Punctuation	2 (100%)			

The results show that when a whole word is replaced by another in the RTT version, it is unlikely to provide a correction. Functional words, like prepositions and determiners, present higher percentages of correction, but even these are relatively low (32.26% and 30%). Blunders are very infrequent (2.93% out of 341 cases), but some of them produced surprising results: *one-eyed people > dyed people; the chart > the table; the one-eyed sight > blindsight).*

5.5 Word Order

Changes affecting word order in a sentence without any further modification were infrequent in the study (only 13 cases), and most of them are alternatives because word order is often a flexible characteristic of languages, even in English. The only corrections (*n* =3) are cases where English has more rigid patterns in how words must be arranged in the sentence: *were chosen two hundred people > two hundred people were chosen*. Such modifications tend to reflect different patterns in word order between English and Spanish.

5.6 Rearrangement

Making a generalization about this type of modification is hard because it covers a wide range of transformations in the syntactic structure of sentences or phrases. Some of the transformations are easy to classify, such as active vs passive voice, Noun *of* Noun vs Noun*'s* Noun or nominalizations, but as a rule, any modification that did not fit in the rest of categories was classified as a rearrangement.

Only 25 of the 111 cases identified (22.52%) can be considered corrections. Unfortunately, while these transformations have the positive effect of making any input more acceptable, they also often conceal any problem in the original. For instance, the fragment *use*

of drugs which becoming more popular was changed into *increasingly popular use of drugs*, which is much better but prevents the learner from noticing the problem with the verb phrase. The examples classified as corrections are very varied and resist classification, normally reflecting an imperfect knowledge of lexicon or grammatical structures: *are more disagree > disagree more; still preparing > continued to prepare; it was made an animal sacrifice > an animal sacrifice was made; are less tend to > are the least likely*.

Suggestions account for 33 out of 111 cases (29.73%) and are mostly changes from the active to the passive voice. The reason for this is that the use of the passive voice often improves the coherence of a text by reinforcing topic continuity, but few teachers will mark as incorrect a well-constructed sentence in the active voice.

Alternatives occurred in 48 out of 111 cases (43.24%), with examples such as: *in honor of > honoring; before he was executed > before being executed; your affection for them > the affection you have for them*.

Blunders occurred on five occasions (4.05%), and these are all cases of poorly written passages that the machine failed to make sense of: *This is a negative point because instead they showed us that drugs can have a recreational use, they are still very dangerous > This is a negative point because instead of showing us that drugs can be used recreationally, they are still very dangerous; for men to be gathered in war > so that men could meet in war.*

6. Discussion

Compared to the results reported in Hermet and Désilets (2009) and Madnani et al. (2012), this study reveals a clear improvement in MT over the last decade. Madnani et al. (2012) reported an almost equal distribution between successes, draws and failures (roughly corresponding to corrections, alternatives and blunders in the

present study), with successes slightly outnumbering the other two (36%). In this study, a similar percentage was found for corrections (34%). The number of blunders, however, decreases drastically – only 3.46% – while the number of draws increases (46.76%). Furthermore, the percentage of successes or draws would be higher if we added what was labelled as *suggestions* in the present study to account for middle ground cases.

The improvement is also evident in comparison to the results reported in Hermet and Désilets (2009) for prepositions in French. The repair rate or number of errors corrected was 66.4%, whereas in the present study, every single error involving a preposition (n =35) was corrected. A clearer indication of the improvement of MT over the past years is that typos and spelling mistakes that result in non-existing words, which had to be removed beforehand in the experiments reported in Madnani et al. (2012) because the machine could not handle them adequately, are the errors most efficiently corrected by RTT today.

A clearer indication of the improvement of MT over the past years is that Madnani et al. (2012) explained that the input fed to the translator in their experiment had to be revised beforehand because the machine could not adequately handle typos and spelling mistakes that resulted in non-existing words, whereas in the present study those are precisely the errors most efficiently corrected by RTT.

Any performance comparison with tools specifically designed for error detection must be taken with caution, as the pedagogical use of the RTT technique has characteristics that set it apart from those tools. However, a comparison with the results reported for applications like Grammarly or ProWritingAid (Sahu et al., 2020) shows that RTT outperforms existing application in correcting the errors present in a text (recall = 0.97) at the cost of generating a considerable number of false positives (precision = 0.34).

Some modifications, however, are more reliable than others. Changes in the spelling or morphology of a word are very likely to

correct mistakes, which is in line with the greater grammar accuracy reported in studies on the use of translators in a conventional way. The rest of modifications provide a paraphrase of a correct sentence roughly as often as they correct mistakes, and when a word is replaced by another word of the same class a correction happens only 17.6% of the time. Some subtypes of modification offer promising results, like the addition of pronouns, the replacement of punctuation, the addition of adjectives or the addition or deletion of determiners. However, some of these subtypes are not sufficiently represented in a comprehensive study that attempted to examine the result of applying RTT to real compositions, and they would require specific studies.

7. Conclusion

This study set out to investigate the strengths and weaknesses of using RTT as a writing assistant in English L2 by examining how many of the differences between the original text and the RTT version correct errors and if some of these modifications are more useful than others.

The results show that RTT's main strength is the elimination of most of the errors in a text (97%), while it has the disadvantage of generating too many false positives (only 34% of the modifications actually correct errors). Changes in the spelling or morphology of a word are much more likely to be a correction, while the rest of changes, such as deleting or adding a word, replacing a word for another word of the same class or paraphrasing the structure of a sentence only correct errors between 45% and 17% of the time.

How useful is the RTT technique for writing in English L2? To answer this question, one must be aware of two facts. First, RTT is not the most efficient way of producing a text written in English. If a user who does not feel confident about their English skills is only interested in the final product, using a translator in the conventional way demands less effort. RTT is advisable only for users who regard

the process of writing as part of their training to improve their English. Second, RTT is not an error detection program. Because it generates a high proportion of false positives and provides no explanations, RTT is advisable only for learners who approach the machine output critically and understand that it must be coupled with research. Used in this way, RTT can be an excellent tool to expand the students' writing skills and raise their awareness of the relation between words and constructions.

8. Limitations of the Study

The results in this study are based on the writing of Spanish-speaking learners, which means that the effectiveness of the RTT modifications could be influenced by that language and its proximity to English. It was not possible to investigate in this study the extent to which the linguistic background of the learner could have an impact on the quantity and quality of RTT modifications. The use of Spanish as an intermediary language also must influence the result of RTT, as translators are known to perform differently depending on the languages involved. It would be worthwhile to investigate what languages offer more useful results when performing RTT.

Finally, because this study attempted to analyze the output a hypothetical learner would obtain from applying the RTT technique to whole compositions, some phenomena are not sufficiently represented to draw strong conclusions. Separate studies focusing on specific aspects of the language, such as prepositions, determiners, punctuation, etc., would be necessary to address this gap.

References

Benda, J. (2013). Google Translate in the EFL classroom: Taboo or teaching tool? *Writing and Pedagogy*, *5*(2), 317–332. https://doi.org/10.1558/wap.v5i2.317

Brockett, C., Dolan, W. B., & Gamon, M. (2006). Correcting ESL errors using phrasal SMT techniques. *Proceedings of the 21st International Conference on Computational Linguistics (COLING) and 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, 249–256. https://doi.org/10.3115/1220175.1220207

Cancino, M., & Panes, J. (2021). The impact of Google Translate on L2 writing quality measures: Evidence from Chilean EFL high school learners. *System, 98*, 1–11. https://doi.org/10.1016/j.system.2021.102464

Chodorow, M., Dickinson, M., Israel, R., & Tetreault, J. (2012). Problems in evaluating grammatical error detection systems. In *Proceedings of COLING 2012,* 611–628. https://aclanthology.org/C12-1038.pdf

Chon, Y. V., Shin, D., & Kim, G. E. (2021). Comparing L2 learners' writing against parallel machine-translated texts: Raters' assessment, linguistic complexity and errors. *System, 96*, 1–12. https://doi.org/10.1016/j.system.2020.102408

Chung, E. S. (2020). The effect of L2 proficiency on post-editing machine translated texts. *Journal of Asia TEFL, 17*(1), 182–193. https://doi.org/10.18823/asiatefl.2020.17.1.11.182

Chung, E. S., & Ahn, S. (2021). The effect of using machine translation on linguistic features in L2 writing across proficiency levels and text genres. *Computer Assisted Language Learning*, 1–26. https://doi.org/10.1080/09588221.2020.1871029

Clifford, J., Merschel, L., & Munné, J. (2013). Surveying the Landscape: What is the Role of Machine Translation in Language Learning? *@tic Revista d'Innovació Educativa, 10*, 108–123. https://www.redalyc.org/pdf/3495/349532398012.pdf

Corder, S. P. (1971). Idiosyncratic dialects and error analysis. *IRAL: International Review of Applied Linguistics in Language Teaching, 9*(2), 147–160. https://doi.org/10.1515/iral.1971.9.2.147

Correa, M. (2011). Academic Dishonesty in the Second Language Classroom: Instructors' Perspectives. *Modern Journal of Language Teaching Methods, 1(1)*, 65–79. https://www.academia.edu/download/7863470/correa2011-MJLTM.pdf

Correa, M. (2014). Leaving the "peer" out of peer-editing: Online translators as a pedagogical tool. *Latin American Journal of Content and Language Integrated Learning, 7(1)*, 1–20. https://doi.org/10.5294/laclil.2014.7.1.1

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System*, *26*(2), 163–174. https://doi.org/10.1016/S0346-251X(98)00001-3

Díaz-Negrillo, A., & Fernández-Domínguez, J. (2006). Error tagging systems for learner corpora. *Revista Española de Lingüística Aplicada*, *19*, 83–102. http://www4.ujaen.es/~svalera/Research/informatizacion/RESLA%202006.pdf

Ducar, C., & Schocket, D. H. (2018). Machine translation and the L2 classroom: Pedagogical solutions for making peace with Google translate. *Foreign Language Annals, 51(4)*, 779–795. https://onlinelibrary.wiley.com/doi/10.1111/flan.12366

Dulay, H., Burt, M., & Krashen, S. (1982). *Language two*. Oxford University Press.

Ellis, R. (1994). *The study of second language acquisition.* Oxford University Press.

Groves, M., & Mundt, K. (2015). Friend or foe? Google Translate in language for academic purposes. *English for Specific Purposes, 37*, 112–121. https://doi.org/10.1016/j.esp.2014.09.001

Heift, T., & Schulze, M. (2007). *Errors and intelligence in computer-assisted language learning: Parsers and pedagogues*. Routledge. https://doi.org/10.4324/9780203012215

Hermet, M., & Désilets, A. (2009). Using first and second language models to correct preposition errors in second language authoring. *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics*. https://doi.org/10.3115/1609843.1609853

James, C. (1998). *Errors in language learning and use: Exploring error analysis*. Pearson.

Kol, S., Schcolnik, M., & Spector-Cohen, E. (2018). Google translate in academic writing courses? *The EUROCALL Review, 26(2)*, 50–57. https://doi.org/10.4995/eurocall.2018.10140

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). *Automated grammatical error detection for language learners*. Springer. https://doi.org/10.1007/978-3-031-02137-4

Lee, S.-M. (2020). The impact of using machine translation on EFL students' writing. *Computer Assisted Language Learning, 33(3)*, 157–175. https://doi.org/10.1080/09588221.2018.1553186

Lennon, P. (1991). Error: Some problems of definition, identification, and distinction. *Applied Linguistics, 12(2)* , 180–196. https://doi.org/10.1093/applin/12.2.180

Madnani, N., Tetreault, J., & Chodorow, M. (2012). Exploring grammatical error correction with not-so-crummy machine translation. *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP,* 44–53. Association for Computational Linguistics. https://aclanthology.org/W12-2005.pdf

Myers, E. W. (1986). AnO(ND) difference algorithm and its variations. *Algorithmica, 1*, 251–266. https://doi.org/10.1007/BF01840446

McCarthy, B. (2004). Does online machine translation spell the end of take-home translation assignments? *CALL-EJ Online, 6(1)*, 26–39. http://www.callej.org/journal/6-1/Mccarthy2004.pdf

Mujtaba, S. M., Parkash, R., & Reynolds, B. L. (2022). The effects of language proficiency and online translator training on second language writing complexity, accuracy, fluency, and lexical complexity. *Computer Assisted Language Learning, 23(1)*, 150–167. https://repository.um.edu.mo/handle/10692/97205

Nicholls, D. (2003). The Cambridge learner corpus: Error coding and analysis for lexicography and ELT. *Proceedings of the Corpus Linguistics 2003 conference. Vol. 16.*, 572–581. https://www.academia.edu/download/43303478/CL2003_Nicholls.pdf

Niño, A. (2009). Machine translation in foreign language learning: Language learners' and tutors' perceptions of its advantages and disadvantages. *ReCALL, 21(2)*, 241–258. https://doi.org/10.1017/S0958344009000172

O'Neill, E. M. (2016). Measuring the Impact of Online Translation on FL Writing Scores. *The IALLT Journal, 46(2)*, 1–39. https://doi.org/10.17161/iallt.v46i2.8560

Park, Y. A., & Levy, R. (2011). Automated whole sentence grammar correction using a noisy channel model. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 934–944. https://aclanthology.org/P11-1094.pdf

Pawlak, M. (2013). *Error correction in the foreign language classroom: Reconsidering the issues*. Springer. https://doi.org/10.1007/978-3-642-38436-3

Rauf, S. A., Saeed, R., Khan, N. S., Habib, K., Gabrail, P., & Aftab, F. (2017). Automated grammatical error correction: A comprehensive review. *NUST Journal of Engineering Sciences*, *10*(2), 72–85. https://doi.org/10.24949/njes.v10i2.219

Sahu, S., Vishwakarma, Y. K., Kori, J., & Thakur, J. S. (2020). Evaluating performance of different grammar checking tools. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(2), 2227–2233. https://doi.org/10.30534/ijatcse/2020/201922020

Stapleton, P. (2005). Using the web as a research source: Implications for L2 academic writing. *The Modern Language Journal, 89*, 177–189. https://doi.org/10.1111/j.1540-4781.2005.00273.x

Stapleton, P., & Leung Ka Kin, B. (2019). Assessing the accuracy and teachers' impressions of Google Translate: A study of primary L2 writers in Hong Kong. *English for Specific Purposes, 56*, 18–34. https://doi.org/10.1016/j.esp.2019.07.001

Steding, S. (2009). Machine translation in the German classroom: Detection, reaction, prevention. *Die Unterrichtspraxis, 42(2)*, 178–189. https://doi.org/10.1111/j.1756-1221.2009.00052.x

Tsai, S.-C. (2019). Using google translate in EFL drafts: a preliminary investigation. *Computer Assisted Language Learning*, *32*, 1–17. https://doi.org/10.1080/09588221.2018.1527361

Wallwork, A. (2016). *English for academic research: A guide for teachers*. Springer.

Wallwork, A. (2016). *English for academic research: A guide for teachers*. Springer. https://doi.org/10.1007/978-3-319-32687-0

West, R., Park, Y. A., & Levy, R. (2011). Bilingual random walk models for automated grammar correction of ESL author-produced text. *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, 170–179. https://aclanthology.org/W11-1421.pdf

Yuan, Z., & Felice, M. (2013). Constrained grammatical error correction using statistical machine translation. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, 52–61. https://aclanthology.org/W13-3607.pdf

First version received: June, 2023
Final version accepted: October, 2023