4 Bilingual e-books via neural machine translation and their reception

Antoni Oliver, Antonio Toral, and Ana Guerberof Arenas

Introduction

InLéctor is a collection of bilingual e-books published with the aim of promoting the reading of novels in the language in which they were written for learners of that particular language. Books are published in several electronic formats (EPUB, MOBI, and HTML) and can be read on a range of devices (e.g., computer, tablet, e-book reader). The text of the original work and the translation is segmented; paragraphs are split into smaller units, usually sentences. Each segment of the work in the original language is linked to the corresponding segment in the target language. In this way, readers can switch from the original to its translation as they please.

This research targets a specific set of L2 readers, namely those with an intermediate or high level of proficiency in a foreign language and a desire to read books originally written in that language. Despite their language level, they might need to frequently stop reading to search for unknown words in dictionaries, and, even with this help, some passages may still remain hard to understand. This makes following the book difficult, and, after some time invested in this highly demanding cognitive activity, some of these readers might abandon the book and start reading a book in their mother tongue. Some researchers (Nuttall 1996) state that reading is a very important activity when studying a foreign language, and that the level of difficulty should be such as to not discourage the L2 reader. A bilingual e-book, such as those published in the InLéctor collection, aims to help these readers to access literary works in the original language.

InLéctor has already published several works in English, French, and Russian with existing translations in Spanish and Catalan. This collection only publishes original works in the public domain. Literary works enter the public domain a given time after the death of the author. This period depends on the country and typically ranges from 70 to 100 years. The same criterion applies to translations (List of Countries’ Copyright Lengths, n.d.). For this reason, finding good translations in the public
domain to create bilingual e-books has proven to be quite demanding
and has led to a bottleneck in the publication of works in the InLéctor
collection.

To speed up the process of publication, and given the great improve-
ment in quality achieved by the recent approach to machine translation
based on artificial neural networks (Bahdanau, Cho, and Bengio 2015;
Vaswani et al. 2017), we started to experiment with the use of neural
machine translation (NMT) systems for the creation of bilingual e-books.
The main hypothesis is that the quality of the output from an NMT
system specifically trained for novels would be sufficient to understand
difficult segments or paragraphs. We would expect the reader to read
most of the book in the original language (their L2) and then use the
machine translated version (their L1) to understand difficult sections in
the original version. In this way, the reading experience would come pri-
marily from the source text, while the target text would be used to help
the reader to understand without having to leave the book and look up
difficult words in a dictionary.

In the following sections, we present this process and the results of
a reading experiment. We published a short story written in English
by the American writer Kate Chopin in two versions: a monolingual
English e-book and a bilingual English-Catalan e-book, the latter using
our NMT system trained specifically in the literary domain. The aim
of the experiment was first and foremost to check our methodology
by including readers in the reception of texts mediated by NMT, but
also to explore whether the quality of the NMT system was sufficient
to help readers understand difficult segments and paragraphs in the
short story.

In the research presented in this chapter, we try to answer the following
research questions:

- **RQ1.** Is the reading experience different in monolingual and bilin-
gual e-books?
- **RQ2.** Is the translation quality provided by an NMT system
good enough to help the reader understand a difficult sentence or
paragraph?
- **RQ3.** Can bilingual e-books created using NMT help second-
language readers understand and enjoy a text?

**Bilingual books in language learning**

Reading is an important task in language learning. Two kinds of reading,
with different purposes, are usually considered in the context of language
learning (Elturki and Harmon 2020): intensive reading and extensive
reading. The goal of intensive reading is to find the main idea, supporting
details, and discrete information, whereas extensive reading, also known
as “for pleasure reading” (Bamford and Day 1997), has the goal of
reading as many texts as possible in a foreign language for enjoyment and with minimal mental effort. Extensive reading enhances both receptive and productive language skills (Hafiz and Tudor 1989). Prowse (2002) highlights ten principles for teaching extensive reading and cites Krashen (1993) to point out the importance of free voluntary reading. In addition, he does not recommend the use of dictionaries, as this prevents the extensive reader from developing valuable guessing skills. Chen et al. (2013) state that the integration of e-books in extensive reading activities helps to improve students’ reading attitude, reading comprehension, and vocabulary.

A parallel book published on paper is a bilingual book (Ernst-Slavit and Mulhern 2003) where the original is on one side and the translation on the other side. The texts are aligned, usually at paragraph level. With these parallel books, the reader can choose to read the original or the translation and can switch from one to the other simply by looking at the left or right page. Semingson, Pole, and Tommerdahl (2015) point out the importance of the quality of the translation in the creation of bilingual books. With this in mind, the first published books in the InLéctor collection were created using published human translations, avoiding the use of machine translation (MT) systems. As already mentioned, it is difficult to find good translations in the public domain; therefore, given the improvements in translation quality achieved by NMT systems, we decided to explore the use of this technology to create bilingual e-books.

A key challenge in MT for literary texts is preserving not only the meaning but also the reading experience (Toral and Way 2015). This is undoubtedly true when the MT text is produced in order to be read in its entirety. This is not, however, the case with the InLéctor collection, where the reader is expected to read the original text and only consult the translation sporadically, when a passage is not fully understood. In the books in the collection, the reading experience still comes primarily from the original work.

Methodology

This section offers a description of the MT engines used, the creation of the bilingual e-books, the literary text chosen, the reading experience survey, and, finally, a brief description of the participants in the survey. We have tried to give a clear and extensive overview of the technical process because we believe that it is necessary to understand the work that is required behind the scenes to create a tailored NMT system, and how these systems might be better at improving the reading experience as opposed to off-the-shelf solutions. This information may be useful to researchers outside the specific field of MT who use public engines for their experiments and who might not be fully aware of this specific translation process. Likewise, this process may be of interest to other researchers looking to replicate our methodology.
Description of engine and translation process

The MT system in this research uses the latest technology in this area: the Transformer architecture based on artificial neural networks (Vaswani et al. 2017), building on the English-to-Catalan MT system tailored to literary texts by Toral, Oliver, and Ribas Ballestín (2020). This represents the state of the art for current NMT systems for translating fiction in this language pair. In the rest of the chapter, we refer to two systems: the previous system, i.e. that by Toral, Oliver, and Ribas Ballestín (2020) (system A), and the new system trained for this study (system B). The main differences between system A and system B have to do with the training data used and the way the data are pre-processed.

In terms of pre-processing, in system A, following the typical conventions followed in MT research, data were tokenized, truecased, and normalized with a commonly used set of tools: Moses’ scripts (Koehn et al. 2007). In addition, system A made use of subword units by means of the byte-pair encoding algorithm (Sennrich, Haddow, and Birch 2016), performing 32,000 operations jointly on both the source and target languages. System B had considerably simpler pre-processing; it only used character normalization and Sentence Piece (Kudo and Richardson 2018) for joint tokenization and subword units (48,000 operations). More operations were performed than in the previous system (48,000 instead of 32,000) because the amount of training data was considerably greater (see Table 4.1).

With respect to the training data, system A was a purely in-domain system, i.e., it was trained solely on literary texts: 133 parallel novels (original in English and their human translation in Catalan), and around 1,000 monolingual books in the target language (Catalan). The latter dataset was back-translated (Sennrich, Haddow, and Birch 2015) into English using a phrase-based statistical MT system (see Toral and Way [2018] for additional details). System B added out-of-domain training data, namely parallel data collected by Softcatalà to the aforementioned in-domain datasets. In line with Caswell, Chelba, and Grangier (2019),

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of sentences or sentence pairs</th>
<th>Number of tokens*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>English</td>
</tr>
<tr>
<td>In-domain parallel</td>
<td>1,086,623</td>
<td>14,032,080</td>
</tr>
<tr>
<td>In-domain monolingual</td>
<td>4,194,347</td>
<td>57,209,987</td>
</tr>
<tr>
<td>Out-of-domain parallel</td>
<td>4,503,523</td>
<td>84,076,423</td>
</tr>
</tbody>
</table>

* These do not match the number of tokens reported by Toral and Way (2018), as they tokenized the corpora.
each source sentence in two corpora was preceded by a tag: <G> for the out-of-domain corpus and <B> for the in-domain back-translated corpus. We expected the addition of large amounts of out-of-domain data, when considered as a different subset by the NMT model (thanks to the use of a tag), to result in an improved performance.

Table 4.1 shows the number of sentences and tokens in each of the corpora used for training. As can be seen, system B was trained on roughly double the amount of data than system A.

We used the aforementioned training data to train three MT systems with different seeds until convergence using an early stopping criterion of 5 iterations, which means that the training process is stopped after 5 training iterations that do not achieve any improvement for a given automatic evaluation metric. We used the BLEU metric (Papineni et al. 2002) to evaluate the performance of the resulting model after each training iteration. To do so, the development set was translated with the resulting model for each iteration and the reference translation of this development set was used to calculate the BLEU score. Once the three MT systems were trained, they were fine-tuned—that is, the training process was resumed using the in-domain parallel data, again until convergence with the same criterion. The final MT system was an ensemble of the three fine-tuned models.

**Description and selection of the text**

To create a bilingual book and test the usability of this system, we needed to select a suitable text. Due to the experimental conditions and the budget, the following prerequisites were established:

- That it was publicly available so as not to infringe copyright laws.
- That the story was not included in or taken from a book used to train the customized engine (otherwise, the quality of the raw output would have been artificially augmented).
- That it had a reasonable text length so that participants in the usability test could read the text and complete a questionnaire in around 30 minutes.
- That the story was not from a specific genre and that it was engaging enough to reach a wide audience.
- That the English language used was not so complex as to prevent participants with lower levels of English as a second language from understanding the story.
- That the story could be included in a book with other stories from the same author for future projects.
- That the text would give visibility to women authors.

With these prerequisites in mind, we searched for a text in the Gutenberg Project Library (Project Gutenberg, n.d.) and finally opted for *A Pair*
of Silk Stockings by Kate Chopin (Saint Louis, USA, 1850–1904). The story was first published in Vogue in 1896, but we found the language and topic quite contemporary. The text tells the story of a woman (Mrs Sommers) who comes into a certain amount of money and decides to go shopping for her children; instead, she ends up treating herself to an afternoon in the city, during which she buys a pair of black silk stockings that serves as a trigger for the plot and gives the story its title. Through her shopping spree, the reader realizes that there is a certain uneasiness with her life choices and that she is, in fact, trying to escape her daily routine and, perhaps, longs for something different.

The text has 1901 words, 26 paragraphs, and 104 sentences. The readability scores in MS Word are Flesch Reading Ease 74.3 (Fairly Easy to Read) and Flesch-Kincaid Grade Level 6.7 (Easy to Read) for native speakers.

Description of the e-book creation process

Bilingual e-books are created by linking each segment of the work in the source language with the corresponding segment in the target language. The first books published in the InLéctor collection used published human translations, and the process of creation of the bilingual e-book involved the automatic alignment of source and target segments. The full process of creation of such bilingual e-books is described in Oliver (2017).

The creation of bilingual e-books using an NMT system, which is the case in this study, is much simpler, as the segment alignment step is not necessary. All the scripts for the creation of bilingual e-books using MT systems are released as free software and can be found on GitHub (Oliver 2021). The full process can be divided into the following steps:

- Obtaining the source text
- Conversion of the source text to DocBook
- Creation of the bilingual DocBook
- Creation of the bilingual e-book in the final format: EPUB and HTML

These steps are explained in the following subsections.

Obtaining the source text

Since the InLéctor collection only publishes works in the public domain, the main source for our books is the Project Gutenberg library. We have also used Wikisource (Wikisource, n.d.) to find works in the public domain. Both sites allow us to download the works in several formats. We can use two formats for the source text: an EPUB file (a common file format for e-books) and a plain text file. When using a text file, a minimum markup is needed to identify the different elements of the book.
Bilingual e-books via neural MT and their reception

(e.g., titles, author, chapters, etc.). When using EPUB files, some of these markup elements can be inferred from the EPUB file itself.

Conversion of the source text to DocBook

We use DocBook (DocBook.org, n.d.), an XML standard format for the representation of books and other types of documents, as the primary file format for working with books in the InLéctor collection. As it is a standard format, there are many tools available that handle DocBook files. We have two scripts to convert EPUB and text files into DocBook.\(^\text{11}\)

Creation of the bilingual DocBook

A bilingual DocBook has both source and target texts and each source segment is linked to the corresponding target segment, as in the following example:

```
<para>
<phrase id="ss-99"><link linkend="ts-99">The play was over, the music ceased, the crowd filed out.\</link></phrase>
<phrase id="ss-100"><link linkend="ts-100">It was like a dream ended.\</link></phrase>
</para>
...
<para>
<phrase id="ts-99"><link linkend="ss-99">L’obra s’havia acabat, la música havia cessat i la multitud s’havia dispersat.\</link></phrase>
<phrase id="ts-100"><link linkend="ss-100">Va ser com si s’acabés un somni.\</link></phrase>
</para>
```

Each paragraph is divided into segments—i.e., sentences—and each source segment (ss) is linked to its corresponding target segment (ts).

The project developed a script\(^\text{12}\) to create bilingual e-books using the MTUOC machine translation server (Oliver 2020) or a tab delimited file (tsv) containing the source text and the machine translated text.

Creation of the bilingual e-book in the final format: EPUB and HTML

Once we have the bilingual DocBook, we can create the e-book in the final format using the scripts\(^\text{13}\) provided. Alternatively, other freely available tools\(^\text{14}\) can be used for this conversion.

Figure 4.1 shows the EPUB edition of the short story by Kate Chopin. While reading the HTML version of the short story, clicking on a given
Mrs. Sommers selected a black pair and looked at them very long and closely. She pretended to be examining their texture, which the clerk assured her was excellent.

“A dollar and ninety-eight cents,” she mused aloud. “Well, I’ll take this pair.” She handed the girl a five-dollar bill and waited for her change and for her parcel. What a very small parcel it was! It seemed lost in the depths of her shabby old shopping-bag.

Mrs. Sommers after that did not move in the direction of the bargain counter. She took the elevator, which carried her to an upper floor into the region of the ladies’ waiting-rooms. Here, in a retired corner, she exchanged her cotton stockings for her new silk ones which she had just bought. She was not going through any acute mental process or reasoning with herself, nor was she striving to explain to her satisfaction the motive of her action. She was not thinking at all. She seemed for the time to be taking a rest from that laborious and fatiguing function and to have abandoned herself to some mechanical impulse that directed her actions and freed her of responsibility.

How good was the touch of the raw silk to her flesh! She felt like lying back in the cushioned chair and reveling for a while in the luxury of it. She did for a little while. Then she replaced her shoes, rolled the cotton stockings together and thrust them into her bag. After doing this she crossed straight over to the shoe department and took her seat to be fitted.

She was fastidious. The clerk could not make her out; he could not reconcile her shoes with her stockings, and she was not too easily pleased. She held back her skirts and turned her feet one way and her head another way as she glanced down at the polished, pointed-tipped boots. Her foot and ankle looked very pretty. She could not realize that they belonged to her and were a part of herself. She wanted an excellent and stylish fit, she told the young fellow who served her, and she did not mind the difference of a dollar or two more in the price so long as she got what she desired.

**Figure 4.1** EPUB version of the short story.
sentence will bring up the translated sentence as a pop-up text, as Figure 4.2 illustrates.

Description of the survey

To test the bilingual book against a monolingual version (our two experimental conditions), we created an online questionnaire consisting of four different sections with 37 items in total (each condition had a different number of items to respond to since some of the items were specific to each condition). It was distributed online to participants using Qualtrics software. The questionnaire was completely anonymous. We did not gather any personal information from the participants, which, in turn, limited our research somewhat, as will be explained in the Limitations of the Survey section. The participants were randomly assigned to one condition in Qualtrics, either to the Bilingual or the Monolingual version of the story. The other sections of the questionnaire were:

- Demographics and reading patterns (11 items). This section of the questionnaire contained five questions on demographics (gender, age, education, occupation, and mother tongue), five questions on reading patterns, e.g., “How much do you like reading?” or “How frequently do you read?”, and one question on their English level.
- Comprehension questions (10 items). After reading the text, the participants answered 10 four-choice questions to explore their basic comprehension of the story. We did not eliminate any participant according to their comprehension level, as we wanted to see if the reading condition (Bilingual or Monolingual) would influence their comprehension.
- Reading experience
  - Bilingual reading experience (8 items). Participants were then asked to answer questions designed to address their reading experience. The readers of the bilingual text were asked eight questions, e.g., “How much did you enjoy reading the text?” or “Would you recommend this text to a friend?”.
  - Monolingual reading experience (5 items). The readers of the monolingual text were asked five questions related to their experience, e.g., “How much did you enjoy reading the text?” or “Would you like to read a bilingual edition of this text?”.
There were two questions that were common to both versions. These were “How much did you enjoy reading the text?” and “How easy was the text to read?”, as we wanted to compare the two conditions regarding ease of reading and enjoyment.

Technical questions (3 items). All participants were asked which device they had used to read the text, if they had experienced any issues during the experiment, and if they wanted to make further comments.

The specific measurements and values will be explained fully in the Results section.

Participants

The criteria for the inclusion of the participants were that they were studying English as a second language in Catalonia and that they were willing to take part in the survey. We were looking for English readers or students who might read in English for pleasure, but also for language students that had to read as part of their English lessons, in line with our research goals. In Spain, learning English starts in primary school and can go on to university level. Of course, studying a language is not necessarily part of formal education—many language students just want to learn a language. Therefore, our criteria for participants were quite open in order to test the system.

To recruit the participants, we distributed the survey among the students of English with Catalan as a first language from the Centre for Modern Languages at the Universitat Oberta de Catalunya (UOC). A call for participation and the survey were also distributed in the mailing list of the APAC (Associació de Professors d’Anglés de Catalunya [Association of Teachers of English of Catalonia]). We also targeted the reading site Goodreads for readers in Catalan (Lectura en català). We posted the advertisement at intervals during December 2020 and January 2021.

The Qualtrics link to the questionnaire was opened by 146 participants, but only 57 of those fully completed the questionnaire. Of those 57 participants, we had to eliminate nine for the following reasons:

- The participants who said they had display problems. Most of them scored very low in comprehension, which seemed to indicate that they had not read the text. This could be because they were using mobile phones.
- All the participants who read the text and completed the survey in under 300 seconds, as this also seemed to indicate that the participants had not read the text.
- Spurious responses, i.e., responses of a humorous nature.
- The participants who completed the questionnaire without responding to the questions on the reading experience.
The participants who indicated having problems with the text links and who seemed to indicate that they had not read the text either.

The final analysis was done with 48 participants: 21 for the Bilingual and 27 for the Monolingual condition. Table 4.2 shows the distribution of data according to gender, mother tongue, education level, and age.

We asked the participants to estimate their English level according to the European Framework of Reference for Languages. These levels are A1, A2, B1, B2, C1, and C2, ranging from basic to very advanced. Table 4.3 shows these responses.

**Limitations of the survey**

The survey had several limitations. First, the survey was conducted online without direct access to the participants. When trying to measure completeness of a questionnaire and when measuring reading comprehension and enjoyment, direct access to the participants might be preferable. Nevertheless, the ease of online distribution and, most importantly, the fact that the survey was carried out during the COVID-19 pandemic meant that this was also the safest and only possible way of doing it.

Second, the participants were not paid to take part in the survey, and it was anonymous. This meant that some participants had little motivation to do it or to do it well, resulting in incomplete or inaccurate entries. The number of valid results was not as high as we had expected; this could also be, in part, due to the pandemic, as the participants might have had to prioritize their other online activities.
Third, since the participants were not paid, we felt constrained as to the amount of time we could ask the participants to spend on the questionnaire and, therefore, we did not include an English level test and limited the length of the questionnaire to the final 37 items.

Finally, there were more primary school students than we had anticipated, and perhaps the story was not as suitable to this age group as would be desirable, hence possibly resulting in lower levels of participation and enjoyment.

Results

**NMT evaluation results**

We conducted automatic evaluations with the BLEU metric (Papineni et al. 2002) on different variants of our previous and current NMT system (see the Description of Engine and Translation Process section). First, we will report results on the development set and then on the test set.

Table 4.4 shows the automatic evaluation scores on the development set of different variants of the MT systems developed. In the first two rows we compare the base versions of system A and system B. System B outperforms system A by almost 1 BLEU point (38.62 versus 37.86), which can be attributed to the different pre-processing and the addition of sizable amounts of out-of-domain training data (see Table 4.1).

The remaining three rows in the table present the results of adding additional features to the base variant of our current system. We can see that all three additions (tags, fine tuning, and ensembling, as explained in the Description of Engine and Translation Process section) result in improved scores of around 0.73, 0.7, and 1.4 points, respectively. As to the reasons why these additions result in improved scores, it seems that, first, adding tags to each sentence pair in the training data according to the corpus it comes from allows the system to treat those sentence pairs differently. Second, the fine tuning also yields high quality: once the training with the whole data is completed, we keep training with

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base system A (in-domain corpora from Table 4.1)</td>
<td>37.86</td>
</tr>
<tr>
<td>Base system B (all the data from Table 4.1)</td>
<td>38.62</td>
</tr>
<tr>
<td>Tags</td>
<td>39.35</td>
</tr>
<tr>
<td>Fine tuning</td>
<td>40.05</td>
</tr>
<tr>
<td>Ensembling</td>
<td>41.45</td>
</tr>
</tbody>
</table>
just the parallel in-domain data. Third, combining multiple independently trained systems by means of ensembling is known to lead to improvements.

We evaluated the best variant of system A (ensemble of 4 fine-tuned models) and the best variant of system B (last row in Table 4.4) on a test set of 12 novels previously used by Toral and Way (2018) and Toral, Oliver, and Ribas Ballestín (2020). The results are shown in Figure 4.3. Overall, the new MT system resulted in an improvement of 1.08 BLEU points: the average score for the 12 books was 31.91 with the new system compared to 30.83 with the previous one. We can observe that the scores with the new system are higher than with the previous one for all books except one: Rowling’s Harry Potter #7. This is an interesting case, since the previous 6 books in this series are contained in the in-domain training data. They amount to 50,000 sentence pairs (cf. Table 4.3 in Toral and Way 2018). Therefore, it is not surprising that having additional out-of-domain data does not improve the BLEU score for this book.

In addition to this automatic evaluation, the machine translated text for the short story used in our reading experiment (see the Description and Selection of the Text section) was post-edited by one of the authors of this chapter. The post-editing made minimal changes in order to obtain a correct translation that is fully understandable and fluent enough in Catalan, i.e., light post-editing. This post-editing was done solely for evaluation purposes, and the bilingual edition of the text was created using the raw machine translated text.
Each segment was classified into one of six levels of required post-editing effort, according to the number and type of errors encountered, as in Toral, Oliver, and Ribas Ballestín (2020). These levels of required post-editing effort are shown in Table 4.5.

Table 4.5 Levels of required post-editing effort

<table>
<thead>
<tr>
<th>Required post-editing</th>
<th>Perception of MT output</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>No post-editing required. No errors were found in the MT output.</td>
</tr>
<tr>
<td>Minimal</td>
<td>Easy to solve in parts of words, e.g., inflection and concordance errors.</td>
</tr>
<tr>
<td>Limited</td>
<td>Up to one word, e.g., lexical errors.</td>
</tr>
<tr>
<td>Moderate</td>
<td>More than one word, e.g., syntactical errors.</td>
</tr>
<tr>
<td>Considerable</td>
<td>More than one word, e.g., semantic, cognitive errors.</td>
</tr>
<tr>
<td>Retranslation</td>
<td>Rather than post-editing, a new translation is required.</td>
</tr>
</tbody>
</table>

Table 4.6 Results of the classification of the post-edited segments by required post-editing effort

<table>
<thead>
<tr>
<th>Required post-editing</th>
<th>Segments</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>47</td>
<td>42.99</td>
</tr>
<tr>
<td>Minimal</td>
<td>17</td>
<td>15.89</td>
</tr>
<tr>
<td>Limited</td>
<td>20</td>
<td>18.69</td>
</tr>
<tr>
<td>Moderate</td>
<td>13</td>
<td>13.08</td>
</tr>
<tr>
<td>Considerable</td>
<td>8</td>
<td>7.48</td>
</tr>
<tr>
<td>Retranslation</td>
<td>2</td>
<td>1.87</td>
</tr>
<tr>
<td>Total segments</td>
<td>107</td>
<td>100</td>
</tr>
</tbody>
</table>

Each segment was classified into one of six levels of required post-editing effort, according to the number and type of errors encountered, as in Toral, Oliver, and Ribas Ballestín (2020). These levels of required post-editing effort are shown in Table 4.5.

Analysis of the light post-editing led to the figures shown in Table 4.6. 42.99% of the segments (47 out of 107) required no post-editing. These segments account for 34.12% of the tokens in the source text (782 out of 2,290). Only two segments (1.87%) needed full retranslation, while 8 (7.48%) required a considerable amount of post-editing. Given that over 75% of the segments required from none to limited post-editing, we hypothesize that the quality of the translation is sufficient for L2 readers.

Using the post-edited text as reference, we calculated several automatic evaluation metrics. We will focus on HTER, as it is a common metric used in post-editing experiments to assess the use of raw MT output and hence its quality. The overall HTER for this text was 12, which means that few changes were required to bring the raw output to a comprehensible level. Table 4.7 shows the number and type of post-editing operations. As can be seen, the most common operation was substitution, representing 52.70% of the operations. This means that new words were required to achieve this level of acceptability.
Table 4.7 Number of post-edits per operation type. The percentage is with respect to the total number of operations

<table>
<thead>
<tr>
<th>Type of operations</th>
<th>Number of operations</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insertions</td>
<td>52</td>
<td>17.57</td>
</tr>
<tr>
<td>Deletions</td>
<td>50</td>
<td>16.89</td>
</tr>
<tr>
<td>Substitutions</td>
<td>156</td>
<td>52.70</td>
</tr>
<tr>
<td>Char shifts</td>
<td>18</td>
<td>6.08</td>
</tr>
<tr>
<td>Word shifts</td>
<td>20</td>
<td>6.76</td>
</tr>
</tbody>
</table>

Figure 4.4 Reading experience by condition.

Reception: survey responses

In this section, we analyse the responses from the 48 participants. We have grouped these into five different subsections to see the results more clearly: Reading experience, Comprehension, Reading patterns, English level, and Additional data.

Reading experience

From the 48 valid responses obtained, we plotted two common questions according to the Bilingual (21 participants) and Monolingual (27 participants) conditions. These questions were “How easy was the text to read?” (1 = Not at all easy to understand and 5 = Very easy to understand) and “How much did you enjoy reading the text?” (1 = I didn’t like it at all and 5 = I liked it a lot). Figure 4.4 shows the results by condition.

As we can see, the reading experience was better in the Bilingual condition (M = 3.38) than the Monolingual (M = 2.85), but the mean for both
reading experiences was above the overall mean value of the scale of 5. In order to see if there was a statistically significant difference between the two conditions, a nonparametric Mann-Whitney U test was used. The test showed significant differences between the two groups: \( U = 385, p < 0.03 \). The effect size \( d = 0.36 \) is medium. The readers of the bilingual version found it significantly easier to read and they enjoyed it significantly more than the readers of the monolingual version.

As mentioned in the Description of the Survey section, we asked the participants in the Bilingual condition further questions about their reading experience. There was a total of eight questions in this condition (see Table 4.9). The participants were asked to rate all of these questions on a 5-point Likert scale, with 1 always corresponding to an unfavourable value (e.g., 1 = the Catalan translation was not useful at all) and 5 to a favourable value (e.g., 5 = the Catalan translation was very useful). The maximum score for these eight questions was therefore 40. Table 4.8 shows the descriptive statistics for the sum of all values in the 8 questions.

Although the range of responses is wide, the mean value of 24.67 over the maximum possible value of 40 suggests that participants’ reactions were, on average, quite favourable towards the Bilingual condition.

Table 4.9 shows the descriptive values for each of these questions. From these questions, we can see that, although the participants did not always use the Catalan translation, their opinions were, on average, favourable towards this bilingual edition.

Further, we asked the monolingual readers five questions: 1) “How easy was the text to read?”, 2) “While reading, which of these tools did you use?”, 3) “How frequently did you use these tools?”, 4) “Would you like to read a text in a bilingual edition?”, and 5) “How much did you enjoy reading the text?”. Apart from questions 1 and 5, which were used to compare the reading experience according to the two conditions, we also looked at the responses to the remaining three questions. Out of the 27 respondents, 15 did not use any tools to read the text, 11 used Google Translate, 4 used online dictionaries, 1 asked a class colleague, and 1 used a printed dictionary, with an average frequency of 3.

It is interesting to see the number of participants that used Google Translate to understand the text even in this Monolingual condition. When asked if they would like to read a bilingual edition (1 = No, I wouldn’t like it at all and 5 = Yes, I would like it a lot), the average response was 2.92.

<table>
<thead>
<tr>
<th>Participants</th>
<th>Min.</th>
<th>1st Q</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Q</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>21</td>
<td>9</td>
<td>20</td>
<td>26</td>
<td>24.67</td>
<td>30</td>
</tr>
</tbody>
</table>
Comprehension

After reading the text, the participants were asked to answer 10 comprehension questions. Each of these questions had 4 options with only 1 possible correct answer so that the total number of points, if all the answers were correct, was 10 points. Figure 4.5 shows the comprehension results according to each condition.

As we can see, comprehension was very similar between the Bilingual condition ($M = 5.29$) and the Monolingual condition ($M = 5.37$), although the latter shows slightly higher values overall. There were no statistically significant differences between these two conditions. Therefore, even if the participants in the Bilingual condition found the text significantly easier to read, this was not reflected in their comprehension of the text. Obviously, participants may have the perception that it is easy to read and still have comprehension issues.

Reading patterns

Before reading the text, the participants were asked to answer 5 questions about their reading habits: 1) “How frequently do you read?” (a range between 1 = I do not read and 5 = Every day), 2) “How long do you read on these occasions?” (a range between 1 = Less than 15 minutes and 5 = More than 90 minutes), 3) “From all the books that you read in
a year, which percentage would you say are books in English?” (a range between 1 = 0% to 25% and 5 = 86% to 100%), 4) “Specify the genre or genres that you read most frequently in English” (to be selected from multiple options), and 5) “When you read a word that you do not understand in English, which tool do you use?” (to be selected from multiple options).

In order to rule out the possibility that the participants in the Bilingual condition showed a better reading experience (see the Reading Experience section) because they had different reading habits (read more frequently) to the Monolingual condition, we looked at the sum of questions 1, 2, and 3 (the maximum value being 15). Figure 4.6 shows the reading pattern results according to each condition.

As we can see, the reading patterns were very similar between the Bilingual condition (M = 7.43) and the Monolingual condition (M = 6.89).
There were no statistically significant differences between these two conditions. Therefore, the fact that participants in the Bilingual condition enjoyed the text significantly more than the Monolingual would not seem to be caused by their pre-existing reading habits.

**Self-reported English level**

We thought it was important to look at the participants’ self-reported English levels to see if this could be a factor that influenced the others. Before reading the text, the participants were asked to define their English level according to the Common European Framework of Reference for Languages (see the Participants section). Figure 4.7 shows the results per condition.

As we can see, the self-reported English level was higher for the participants in the Bilingual condition (M = 4.29) than in the Monolingual condition (M = 3.33). In order to see if there was a statistically significant difference between the two conditions, a nonparametric Man-Whitney U test was used. The test showed significant differences between the two groups U = 386.5, p < 0.03. The effects size d = 0.36 is medium. The readers in the Bilingual condition reported having a significantly higher English level than the readers in the Monolingual condition.

This is relevant as it could indicate that the participants in the Bilingual condition had a higher reading experience because their English level was also higher than that of the participants in the Monolingual condition, and not necessarily because they were provided with the Catalan translation. However, the comprehension level and reading patterns were not significantly different between the two conditions, and this means that we have to be cautious when considering the self-reported English levels.
It could be that the participants indicated a given English level when, in fact, their comprehension level was lower (as this was not significantly different from the one in the Monolingual condition). Nevertheless, it is an important outcome, and we would need further data (for example, an English test prior to the survey) to assess this more accurately.

To explore this potentially confounding variable further, we divided the participants into two levels: Low, including levels A1, A2, and B1; and High, including levels B2, C1, and C2. A total of 23 participants were classified as Low and 25 participants as High according to their English level. Table 4.10 shows the distribution per condition.

As we saw initially, there are more participants with a higher self-reported level in the Bilingual condition than in the Monolingual one. Figure 4.8 shows the reading experience results according to the participants’ English level and condition.

The participants in the Bilingual condition with different English levels have a similar reading experience if the average is considered (Low Bilingual $M = 3.14$ and High Bilingual $M = 3.5$), while the participants in the Monolingual condition with a low English level have a lower reading experience than the Monolingual readers with a higher English level (Low Monolingual $M = 2.53$ and High Monolingual $M = 3.23$).

<table>
<thead>
<tr>
<th>English level</th>
<th>Bilingual</th>
<th>Monolingual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>High</td>
<td>14</td>
<td>11</td>
</tr>
</tbody>
</table>

Figure 4.8 Reading experience according to ‘Low’ and ‘High’ English level, and condition.
This could indicate that the Bilingual books did indeed increase the reading experience of participants with a low level of English. With more participants, an English test, and inferential statistics, we could explore this relationship further. At the moment, we can say that the English level was a contributing factor to the reading experience. Likewise, we have sufficient data from the participants in the Bilingual condition to indicate that the experience with the Catalan translation was well received by the readers.

**Additional data**

To finalize this section, we would like to provide data on a couple of additional variables that we examined. First, we wanted to know if the time invested varied depending on the condition and, in turn, if the other variables might also be affected. Figure 4.9 shows the duration in seconds by condition.

As we can see, the duration, albeit with certain outliers, is higher in the Bilingual condition (M = 1868 seconds [31 minutes]) than in the Monolingual condition (M = 1395 seconds [23 minutes]). However, there are no statistically significant differences between these two conditions. Therefore, even if the participants in the Bilingual condition enjoyed the text significantly more than the monolingual, this does not appear to be because of the time they invested in reading the text.

We also looked at the type of device used by the participants to explore if this was a factor in the duration of the activity. Figure 4.10 shows the duration in seconds according to the device.

It is interesting to see that the participants who read on a tablet took much longer (M = 5127.33) to read than those on a computer screen.

*Figure 4.9* Duration in seconds by condition.
Antoni Oliver, Antonio Toral, & Ana Guerberof Arenas

(M = 1815.04) or mobile phone (M = 897.45). Although, as we saw before, the duration was not significantly different according to condition, we wonder if reading faster, for example, on a mobile phone might not have an indirect effect on the reading experience. Table 4.11 shows the distribution of devices per condition.

Table 4.11 Distribution of device per condition

<table>
<thead>
<tr>
<th>Device</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bilingual</td>
</tr>
<tr>
<td>Computer</td>
<td>9</td>
</tr>
<tr>
<td>Tablet</td>
<td>2</td>
</tr>
<tr>
<td>Mobile phone</td>
<td>10</td>
</tr>
</tbody>
</table>

Figure 4.10 Duration in seconds according to the device.

Table 4.11 shows that there were only 3 participants using a tablet, so it is difficult to generalize with these results; also, our sample size was relatively small. Nevertheless, the device seems to influence the time dedicated to the experiment. Perhaps it would be advisable to ask participants in the future to read on a tablet to make sure that the reading experience is longer and possibly more intense, leading to more reliable results.

Conclusions and future work

The analysis of the survey answers from readers of the bilingual and monolingual editions of the short story allowed us to answer the research questions presented at the beginning of this chapter.
Regarding RQ1 (Is the reading experience different in monolingual and bilingual e-books?), we can conclude that:

- Readers of the bilingual version had a significantly better reading experience. In other words, they found it easier to read and enjoyed it more. We ruled out the possibility that this could be attributed to their pre-existing reading habits. Readers of the bilingual version, on average, reacted quite favourably to the bilingual edition.
- English level is an important factor when looking at the reading experience, i.e., readers with a higher English level tended to enjoy reading more. However, bilingual versions appeared to level the reading experience between participants with low and high English levels.

We can clearly see that bilingual e-books do not hinder the reading experience of original work but instead improve it.

As for RQ2 (Is the translation quality provided by an NMT system good enough to help the reader understand a difficult sentence or paragraph?), we can see that:

- Regarding the readers of the monolingual version, a significant percentage of them used Google Translate to understand the text, and they expressed interest in reading a bilingual edition.
- However, there are no significant differences in comprehension between readers of the monolingual and bilingual editions.

Two questions from the survey answered by the readers of the bilingual edition can also help clarify this question:

- *If you used the Catalan translation, was it useful?* The mean was 3.44 (where 5 is the highest), so, in general, readers of the bilingual edition found the translation useful.
- *If you used the Catalan translation, what did you think of it?* The mean was 3 (where 5 is the highest), so the quality of the translation can be considered satisfactory, but there is still room for improvement.

Finally, with respect to RQ3 (Can bilingual e-books created using NMT help second-language readers understand and enjoy a text?), our answer to RQ1 shows that Bilingual readers had a better reading experience and enjoyed the text more, although our answer to RQ2 shows that they did not necessarily understand it better than the monolingual readers. The following questions for the readers of the bilingual edition also show that the readers liked this version:

- *Would you like to read more books in this bilingual edition?* The mean was 3.15 (where 5 is the highest), indicating a slightly positive predisposition to read such bilingual e-books.
Would you recommend this bilingual edition? The mean was 3.95 (where 5 is the highest), a clear indication of satisfaction with the bilingual editions.

Therefore, as an overall conclusion, it seems clear that the opinion of our participants concerning bilingual electronic editions created using NMT was positive and quite favourable.

As future work, we plan to extend this experiment to more readers, but we want to know more about the level of English of the participants. One idea to achieve this would be to collaborate with schools of English to distribute the survey to their students, grouped according to the level they are studying. Likewise, we would like to provide different bilingual texts suited to different age groups.

We also plan to publish several complete works using the NMT system and make them freely available in the InLéctor collection. When a user downloads the book, we will ask him or her to participate in a survey after reading the text. We also plan to keep improving the quality of the NMT system for English-Catalan and to train systems for other language pairs, such as English-Spanish, French-Spanish, and Russian-Spanish.

Notes

1 https://inlector.wordpress.com/
2 Split into smaller units, as words, numerical expressions, and punctuation marks.
3 Writing the words in their natural case (that is, whether the first letter is uppercase or lowercase) independently of their position in the text.
4 Converting different variants of a given character into a single equivalent symbol, e.g., all the variants of a double quote (“, ”, « and », etc.) are converted into “.
5 Text representation, smaller than a word and bigger than a character that has been shown to lead to better results in NMT.
6 Back-translation is a popular technique in NMT to increase the amount of parallel training data, whereby monolingual data in the target language are translated into the source language using an MT system. The resulting parallel data, whose source side is synthetic, can then be used for training an MT system that translates from the source language into the target language.
7 https://github.com/Softcatala/en-ca-corpus
8 The most widely used automatic metric in the MT research community to date. The higher the value, the better, in theory, the translation. Its lower bound is 0 and its upper bound is 100. This means that a value of 100 is a perfect translation as compared to a human reference.
9 The input given to the MT system is the text in the original language with one segment per line. The system then generates the translation, which contains the same number of segments, hence segment alignment is not necessary.
10 Markup is a set of symbols that indicate the formatting and structure of a text.
11 epub2docbook.py and text2docbook.py
For example, using dbtoepub and specifying a stylesheet to avoid having all the sentences underlined (as they are all links). For the creation of the HTML edition, xlstproc can be used (Oliver 2017).


token in NLP corresponds to a word or a punctuation mark.

HTER is an automatic score that reflects the number of edits performed on the MT output normalized by the number of words in the sentence. The closer HTER is to 0, the fewer the changes performed.

References


