# With or without you? Effects of using machine translation to write flash fiction in the foreign language

**Nora Aranberri**
IXA research group
University of the Basque Country UPV/EHU
nora.aranberri@ehu.eus

## Abstract

The improvement in the quality of machine translation (MT) for both majority and minority languages in recent years is resulting in its steady adoption. This is not only happening among professional translators but also among users who occasionally find themselves in situations where translation is required or MT presents itself as a easier means to producing a text. This work sets to explore the effect using MT has in flash fiction produced in the foreign language. Specifically, we study the impact in surface closeness, syntactic and lexical complexity, and edits. Results show that texts produced with MT seem to fit closer to certain traits of the foreign language and that differences in the use of part-of-speech categories and structures emerge. Moreover, the analysis of the post-edited texts reveals that participants approach the editing of the MT output differently, displaying a wide range in the number of edits.

## 1 Introduction

In recent years, the quality of machine translation (MT) has greatly improved, and as a consequence, increasingly more users are adopting the technology. These users can have varying profiles. On the one hand, we find professional translators, and on the other hand, we have users who do not belong directly in the translation industry but still, occasionally, need translations. Among the latter, we can distinguish scenarios where MT is used in professional settings and scenarios where MT is used to reduce the translation effort in the private sphere.

A good few studies have been conducted on the impact of MT for professional translators but still numerous questions remain unanswered. Among others, this research has focused on analysing how translating using MT differs from translating from scratch and on ways to optimally provide the automated translation to these professionals. However, little research has been carried out on non-professional translators, even when freely available online systems have been providing automated translations for a long time, since 2006 in the case of Google Translate. This situation leaves us with little insight into what happens when non-specialists avail of MT.

The scarce research carried out on regular users has mainly focused on measuring the usefulness of MT for assimilation, that is, to facilitate comprehension. Nurminen (2018) reported that people are using MT increasingly more for gisting purposes and that they are prepared to accept different quality levels for comprehension and for publication.

Bowker (2009) and Bowker and Ciro (2015) focused their efforts on the Canadian context. In the former study, the author examined the potential acceptance of MT output by minority communities. She reported a positive attitude towards output that had undergone rapid post-editing for assimilation purposes but the need for at least full post-editing for texts intended for cultural preservation. The latter study analysed the usefulness of machine translation to make the Ottawa Public Library website more accessible to Spanish speakers. Authors reported that users would be willing to ac-

© 2020 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

cept MT output, post-edited at different levels, for certain services.

Focusing on romance languages, another group of researchers studied MT in reference to the concept of intercomprehension, that is, the ability of speakers of different languages to understand one another (Martín, 2005; Martín Peris, 2011). Jordan-Nuñez, Forcada and Clua (2017) studied if users perceive MT output, non-native and native texts differently, and examined the usefulness of MT to improve comprehension in cases where native language texts would not be available. They highlighted that the efficiency seems to vary according to the level of specialisation of the texts, their domain and the MT system used.

Almost no research has been carried out on the effects of using MT to produce texts by regular users. The few efforts made in this area have mostly focused on the use of MT by non-native English speakers for academic publishing. Parra Escartín et al. (2017) studied five medical practitioners' papers and O'Brien, Simard and Goulet (2018) examined abstracts of ten scholars. Both studies found that whereas these professionals were able to correct and improve the MT output, their final versions still required further editing to be adequate for publication. Bowker and Ciro (2019) provide an overview of this user group and make a first attempt at establishing a framework for MT literacy for scholar communication. Further research in this line will prove essential to train different user groups in the optimal use of this technology, as non-language-specialists seem to be willing to accept low quality MT output when translating familiar topics (2014).

Within this context, the current work focuses on non-specialist users. We concentrate on using a series of metrics to compare texts produced by those users in the foreign language with and without MT. In particular, we aim to examine the effects of using MT in terms of accuracy, fluency and complexity. In the future this should be complemented with further qualitative analyses to account for word and word-sequence choices and editing.

## 2 Experimental set-up

### 2.1 Participants

A total of 40 participants from the Basque Country voluntarily got involved in the experiment, granting the permission to use their contributions for research purposes. All participants were students in the 19-25 age-range. As per the two official languages of the region, as can be seen in Figure 1, 85% report having Spanish as their mother tongue and Basque as a second language. The reported level of competence in both languages is similar, around 60% for Basque and 68% for Spanish, indicating a C1 level according to the CEFR[1]. The main difference is that while for Basque the remaining 40% report a B2 level, for Spanish, this is divided into B2 (25%) and C2 (7%). A clear difference between the languages is their reported use, which shows that while 75% report using Spanish more than 75% of the time, this range is only reported by 12% for Basque. Even so, it must be noted that the language of instruction of all participants is Basque.

Regarding the foreign language, English in this case, the reported level of competence is more widespread even when almost half classify themselves within the B2 level, and almost 40% within the B1 level. As expected, over 75% of the participants report using English less than 25% of the time. All in all, given their reported level of competence in their main and foreign languages, this group of participants proved adequate to study the impact of using MT to produce texts in a foreign language where their competence is low, starting from their language of instruction. Therefore, the foreign language is at the independent user level according to the CEFR, whereas their main language of instruction is at the proficient user level.

### 2.2 Tasks

This experiment aims to recreate a real scenario where a user avails of MT due to his/her lack of full competence in the foreign language. Considering that each user has a different language competence and style (even in their main languages), we decided to ask each of them to produce their own *source* texts. Also, as they would in a real context, we allowed them to use any language resource except MT to complete the tasks. This mainly involved online bilingual dictionaries and grammar-related sites.

Letting participants completely freely choose the text to write would have biased the results. Therefore, in order to make it possible to compare the results and draw conclusions from the work

[1]Common European Framework of Reference for Languages – Self-assessment Grid available at https://europass.cedefop.europa.eu/sites/default/files/cefr-en.pdf

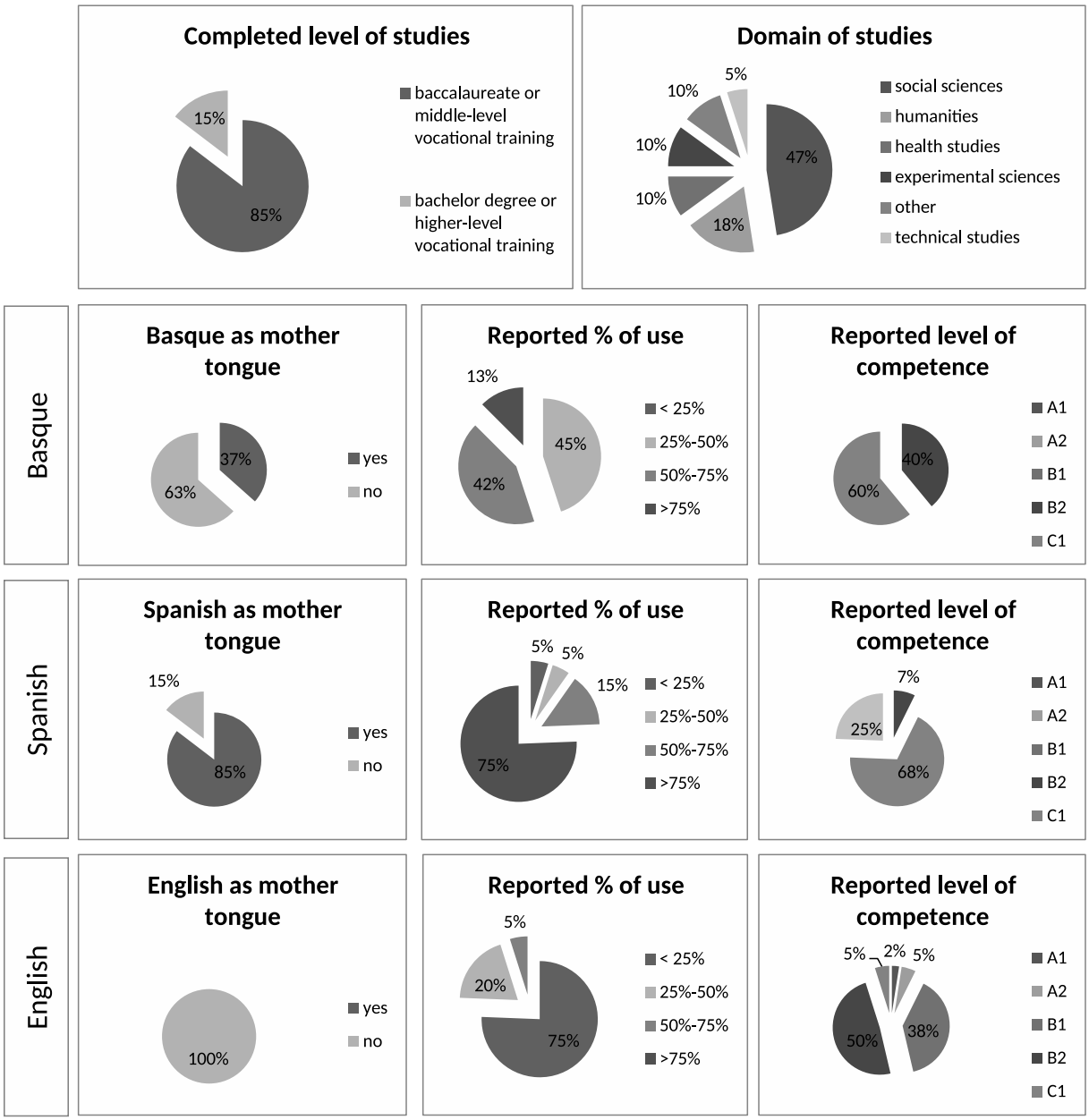


**Figure 1:** Studies- and language-related information of participants. Notice that legends for all possible answers have been displayed for the individual diagrams for easy comparison.

they performed, we set a guided task that aimed to somewhat define the genre, the domain and the length of the text to be produced, while still providing ample room for free contribution. Specifically, participants were asked to write a piece of flash fiction, that is, a short narrative with a full plot, a tool used successfully to promote writing within young adults (Batchelor and King, 2014). Aware of the effort of dealing with long texts in a foreign language, we asked participants to write texts of around 150 words. Also, the stories should be based on a storyboard.

The use of storyboards in linguistic research is widely accepted (Bochnak and Matthewson, 2015). Contrary to targeted storyboards which aim to elicit specific language, we opted for non-targeted storyboards, which aim to elicit language in general, and mainly, narratives (Burton and Matthewson, 2015). Storyboards would also help participants avoid the *blank page effect* when we asked them to come up with creative stories on the spot. We opted for persona-based scenarios (Cooper et al., 2003), which, according to Grudin and Pruitt (2002), are more effective, as the user may feel more represented in the storyline. We provided participants with hints about the setting,

the actors and the goal or ending, and asked them to invent the actions and write a complete story (Kantola and Jokela, 2007; Rosson et al., 2002). The storyboards created for this experiment consisted of three vignettes: (1) the initial situation where the setting and the characters were presented, one of the characters being the participant, (2) a blank vignette representing the events and actions the participants would have to create, and (3) the final situation showing the setting and characters at the end of the story.

Given that the goal of the experiment was to compare the difference between writing texts in the foreign language from scratch and using MT, participants were asked to write two stories based on two storyboards. First, they wrote a piece of flash fiction in English. Secondly, they wrote another piece in Basque and edited its English MT output until they were satisfied with the final result. The participants worked on a customised web site where the different tasks were presented to them, with their respective storyboards and the MT version for their second text, which was obtained through the Google Translator API for the Basque–English pair. They were also asked to fill in an initial questionnaire about their studies and language competence and use, and to answer a final question to assess the level of help provided by the MT system. We also collected the users' permission to use their contribution through this site.

## 3 Analysis of results

The analysis presented here focuses on using a series of metrics to compare texts produced in both set-ups, (1) when participants write directly in the foreign language, English, and (2) when they start with their main working language, Basque, and edit the English MT output provided by Google's Translator. Many researchers in the area of language acquisition have long considered complexity, accuracy and fluency as the three main aspects that capture the foreign language competence (see Housen and Kuiken (2009) for a discussion). In reference to the experiment that concerns us, this means that texts can be classified as better or worse depending on how natural and native-like they sound, the grammatical and lexical inaccuracies they contain, and the complexity of the structures included. Therefore, for this study, we concentrate on the surface closeness, syntactic and lexical complexity, and edit types, while also considering participants' view on usefulness. We report count averages together with the standard deviations, and when possible, calculate statistical significance of the differences (with a 95% confidence interval) using the unpaired Student's t-test to compare the two set-ups (significance is marked with a †).

### 3.1 Does the text produced using MT look more like English?

Let us focus on accuracy and fluency first. Not fully competent speakers tend to make grammatical mistakes and awkward lexical choices to a higher or lower degree. However, if the source text presented to a machine translation system is written by a fully competent speaker, that is, it includes no errors and it is natural, given the features of current neural MT systems, the system is expected to produce a fluent output with no (or few) grammatical mistakes and a (relatively) sound lexical choice. Whereas meaning issues might be present, that is, the output does not express exactly what the user intended to, the machine translated texts tend to comply with the target language features to a considerable degree.

To observe whether participants produced a text that reads more like English with or without using MT, we measured textual closeness through perplexity. In machine translation research, perplexity is used to measure how much a translation fits a language model. In other words, a low perplexity indicates that a text is similar to the language model used as reference. For that reason, we compared the perplexities of the texts produced by the participants in both set-ups in order to find out which of the two displayed sequences that were closer to the language model. We calculated the perplexity at word- and POS-level to account for the surface form but also for a structure-level form, albeit shallow.

To train the language models, we first compiled the corpus for English. Whereas languages tend to comply with overall linguistic features that are intrinsic to them, it is also true that each textual genre brings its own linguistic features and distributions with it. Therefore, if we are to measure surface closeness, it is only fair that the language model is trained using texts that belong to the same genre as the one produced by the participants. As, to our best knowledge, no purposely-build corpus of flash fiction is readily available for NLP testing, we

| | Original English | | Post-edited English | | |
|---|---|---|---|---|---|
| **Level** | **Average** | **Std. dev.** | **Average** | **Std. dev.** | **t-test** |
| token-based | 177.480 | 135.268 | **111.380** | 35.002 | t(78) = 2.992, p = 0.0037$^{\dagger}$ |
| POS-based | 10.582 | 5.813 | **7.765** | 1.080 | t(78) = 3.013, p = 0.0035$^{\dagger}$ |

**Table 1:** Results for the perplexity metric

opted for a main news corpus and complemented it with a number of popular classic literature works that recount stories, tales and adventures. Specifically, we used the first 3 million lines (74.7 million words) of the News Crawl corpus 2019, shuffled and deduplicated,[2] and a 0.5-million-word corpus of stories obtained through the Gutenberg Project.[3]

We built the language models with Modified Kneser-Ney smoothing and no pruning. For the word-level model, we considered n-grams up to $n$ = 5. For the POS-level model, we used *ixa-pipes* (Agerri et al., 2014) to annotate the English corpus at POS level first. The tool uses the Penn Treebank POS tagset, which consist of 36 classes. For the language model, we considered n-grams up to $n$ = 6 and had to assign default parameters to singletons, even when they are not present in the PoS-annotated corpus.

The results show that, on average, perplexities are lower for the post-edited texts both at word-level and at POS-level, the difference being statistically very significant (see Table 1). This indicates that participants obtain surface sequences that are more similar to English when using MT than when they produce the texts directly in that language. Even when MT systems have been reported to produce output that has interference from the source language (Toral, 2019), it seems that participants' competence in the foreign language (independent users according to the CEFR) is not sufficient to outperform the MT system. Participants might be producing either word sequences that are closer to their main languages or word sequences that are incorrect in the foreign language and therefore, using MT seems to help them produce texts that read more like English.

Let us now turn to the complexity aspect. Leaving aside the correctness and appropriateness of the language, a feature that displays the language competence of a person is his or her ability to exploit the linguistic resources available in a language. In line with this, we would expect that texts written directly in the main language of a person display more diversity, precision and information density as the person has the ability and resources necessary for it. Machine translation could prove beneficial in overcoming the more limited access to resources in the foreign language by allowing users to produce the text in their main language, for which their linguistic ability is high, and obtain a foreign language text that mirrors that complexity. Whereas MT is not designed to help with other discourse or textual factors such as adequacy, coherence or cohesion, which are properties linked to cross-linguistic communicative strategies, it does provide the opportunity to assist with the selection and sentence-level arrangement of linguistic elements.

To observe whether differences emerged in the texts produced by the participants in terms of complexity, we looked at a number of lexical and syntactic features. We considered that lexical complexity could be accounted for in terms of frequency, diversity and density. To obtain those measures, we used the information provided by the *ixa-pipes* through the *Analhitza* application (Otegi et al., 2017) to obtain the relevant counts for types, tokens and POS.

We first considered POS frequency. This analysis was intended to observe whether certain grammatical categories were more or less present when writing in one of the two set-ups. For example, we can argue that nouns and verbs are more basic and central categories than adjectives and adverbs, which are used to modify the former. Similarly, pronouns, prepositions and conjunctions are considered to be more complex categories and a higher level of competence is required to use them.

By considering the POS proportions in both set-ups (see Table 2), we observe that some differences emerge. Whereas not significant differences were noticed for the more basic categories, it was interesting to see that the use of prepositions or subordinate conjunctions and pronouns was significantly higher when using MT.

We next considered lexical diversity, that is, the

[2] http://data.statmt.org/news-crawl/en/

[3] https://www.gutenberg.org/ - We used 9 books covering some of the works by Arthur Conan Doyle, Agatha Christie, the Grimm brothers, Mark Twain, and H.G. Wells.

| | Original English | | Post-edited English | | |
|---|---|---|---|---|---|
| **POS** | **Average** | **Std. dev.** | **Average** | **Std. dev.** | **t-test** |
| nouns | 18.809 | 3.630 | 20.010 | 2.247 | t(78) = 1.7792, p = 0.0791 |
| adjectives | 4.408 | 1.986 | 4.341 | 1.666 | t(78) = 0.1635, p = 0.8705 |
| verbs | **22.896** | 2.450 | 21.143 | 2.019 | t(78) = 3.4758, p = 0.0008† |
| adverbs | 5.222 | 1.898 | 5.726 | 1.713 | t(78) = 1.2451, p = 0.2168 |
| determiners | 10.444 | 1.867 | 11.243 | 2.921 | t(78) = 1.4583, p = 0.1488 |
| prep. or sub. conj. | 3.682 | 1.287 | **4.949** | 1.299 | t(78) = 4.3803, p = 0.0001† |
| pronouns | 12.209 | 2.422 | **15.061** | 2.192 | t(78) = 5.5214, p = 0.0001† |

**Table 2:** Results for the lexical proportion metric

| | Original English | | Post-edited English | | |
|---|---|---|---|---|---|
| **POS** | **Average** | **Std. dev.** | **Average** | **Std. dev.** | **t-test** |
| nouns | 0.722 | 0.088 | 0.697 | 0.091 | t(78) = 1.2825, p = 0.2035 |
| adjectives | 0.910 | 0.137 | 0.942 | 0.0821 | t(78) = 1.2592, p = 0.2117 |
| verbs | 0.580 | 0.068 | 0.601 | 0.0694 | t(78) = 1.3732, p = 0.1736 |
| adverbs | **0.752** | 0.164 | 0.601 | 0.069 | t(78) = 5.3829, p = 0.0001† |
| determiners | 0.235 | 0.086 | 0.295 | 0.394 | t(78)= 0.9399, p = 0.3502 |
| prep. or sub. conj. | **0.300** | 0.109 | 0.213 | 0.077 | t(78) = 4.0763, p = 0.0001† |
| pronouns | 0.470 | 0.098 | 0.444 | 0.099 | t(78) = 1.1933, p = 0.2364 |
| overall | 0.530 | 0.039 | 0.553 | 0.054 | t(78) = 2.1830, p = 0.0320† |

**Table 3:** Results for the lexical variety metric

| Original English | | Post-edited English | | |
|---|---|---|---|---|
| **Average** | **Std. dev.** | **Average** | **Std. dev.** | **t-test** |
| 0.499 | 0.029 | 0.512 | 0.023 | t(78) = 2.2566, p = 0.0268 |

**Table 4:** Results for the lexical density metric

variation in the words used to produce the text. We would expect that a lower competence would result in lower diversity, as the lexical resources available would be more limited. This should result in the use of more generic words and absence of synonyms and hyponyms.

However, lexical diversity as measured by the type/token ratio does not exhibit differences between the set-ups (see Table 3). In fact, it seems that the diversity for adverbs and prepositions or subordinate conjunctions is very significantly higher in the text written directly in English. There may be several reasons why this is the case. Firstly, we must remember that research has shown that MT output results in a lower lexical variety as compared with manual translation (Toral, 2019), which indicates a tendency to reduce the vocabulary produced. Secondly, we must also bear in mind that the task carried out by the participants involved writing a short piece of fiction. It is possible that, given the limited size of the text, lexical diversity is not the optimum metric to account for complexity. A more qualitative analysis that considers the exact words used and their respective difficulty could shed light into these questions. It might be the case that the diversity is similar in both set-ups, but that the precision and difficulty of the words produced is greater in one over the other.

Finally, we considered the lexical density of the texts. It is possible that a higher competence in a language allows for condensing more details within the texts. In this case, the MT system would allow this condensation to be transferred to the final English text. A comparison between the average lexical density, measured as the ratio of the number of content words and the total number of words, displayed no significant differences (see Table 4). Again, a qualitative analysis would be necessary to pinpoint the reasons for this trend, which could be related to MT weakness or to the limited communicative competence of the participants.

| | Original English | | Post-edited English | | |
|---|---|---|---|---|---|
| **Semantic functions** | **Average** | **Std. dev.** | **Average** | **Std. dev.** | **t-test** |
| coordinating conjuctions | 6.78 | 2.87 | **10.03** | 3.69 | t(78) = 4.3978, p = 0.0001† |
| subordinating conjuctions | 4.50 | 1.93 | 4.43 | 2.79 | t(78) = 0.1397, p = 0.8893 |
| manner | 0.83 | 0.90 | 0.75 | 1.01 | t(78) = 0.3509, p = 0.7266 |
| purpose or reason | 2.10 | 1.24 | 1.55 | 1.36 | t(78) = 1.8944, p = 0.0619 |
| temporal | 8.13 | 3.34 | 8.73 | 3.30 | t(78) = 0.8089, p = 0.4210 |
| object | **21.30** | 6.43 | 16.35 | 6.36 | t(78) = 3.4607, p = 0.0009† |
| object complement | 4.03 | 2.73 | 4.75 | 2.58 | t(78) = 1.2205, p = 0.2260 |
| predicative complement | **7.48** | 2.79 | 5.63 | 2.74 | t(78) = 2.9896, p = 0.0037† |
| noun modifier | 38.38 | 10.57 | **43.43** | 10.72 | t(78) = 2.1222, p = 0.0370† |
| adjectival or adverbial modifier | 3.38 | 2.00 | 4.10 | 2.35 | t(78) = 1.4867, p = 0.1411 |
| prepositional modifier | 14.15 | 5.45 | **20.48** | 6.48 | t(78) = 4.7236, p = 0.0001† |
| apposition | 1.48 | 1.57 | 1.10 | 1.06 | t(78) = 1.2537, p = 0.2137 |
| n. of sentences | 14.13 | 5.09 | 14.13 | 4.88 | t(78) = 0.0000, p = 1.0000 |
| sentence length | 14.08 | 2.70 | 14.95 | 2.43 | t(78) = 1.5234, p = 0.1317 |

**Table 5:** Results for the syntactic complexity metric

The study of the syntactic complexity was carried out focusing on the presence of certain structures in the text produced by the participants. As the language competence of a learner increases, the basic subject and predicate sentence structure gains intricacy, and additional elements, constituents and semantic roles start to be present.

In order to check whether differences existed in the texts produced in the set-ups, we examined the occurrence of a number of syntactic-semantic characteristics of the texts. Specifically, we focused on semantic dependency relations, which represent the grammatical function in terms of the role that each dependent element plays with respect to its head.

We automatically analysed the texts produced by participants using *ixa-pipes*, which provides a wrapper for the English dependency parser and semantic role labeller based on mate-tools (Björkelund et al., 2009; Vossen and others, 2016) and it is trained on the dependency structures as defined for the CoNLL-2008 Shared Task (Johansson, 2008). We selected 12 dependency relations (see Table 5) that signal complexity, such as the presence of coordinating and subordinating conjunctions, elements that indicate manner, purpose, reason or temporal modifiers, prepositional modifier or adjectival and adverbial modifiers. It is expected that the number of these complex relations will be higher in the texts written using MT because participants were able to express themselves more competently in the language of instruction.

The results in Table 5 show the average occurrence of each type of relation in both set-ups. Whereas the rates for most relations do not seem to vary, several differences surface. The post-edited texts display a significantly higher presence of coordinating conjunctions, nouns modifiers and prepositional modifiers. Also, the presence of objects and predicative complements is higher when writing directly in the foreign language. However, we must concede that the latter are often compulsory elements required by transitive verbs, whereas modifiers and conjunction can be freely used to produce more elaborate text. As a result, we could argue that writing in their language of instruction and using MT to translate it into the foreign language is allowing participants to produce more complex structures to a certain degree.

### 3.2 How do users approach the MT version?

The fact that MT might prove useful in obtaining a more fluent and complex text in the foreign language does not guarantee that the produced text will be error-free and absolutely natural-sounding, or that it will express exactly what the user intended. MT is still imperfect and users still have to perform an additional step before they can consider the text finished: post-editing. In order to fully identify the effects MT has in foreign language text production, it is necessary to analyse what users do with the MT output. Are they able to identify errors and awkward sequences introduced by the system? Can they measure to what extent the system is expressing what they originally intended? Are they aware of the impact the nuances introduced by the system may produce on readers?

As a first step toward identifying user editing

| Metric | Average | Std. dev. |
|---|---|---|
| TER | 9.69 | 8.43 |
| number of edits | 24.02 | 23.59 |
| insertions | 4.88 | 7.19 |
| deletions | 6.25 | 6.82 |
| substitutions | 11.40 | 11.60 |
| shifts | 1.53 | 1.92 |

**Table 6:** Edit information calculated by the TER metric

behaviour, we used edit distance measurements as calculated by TER. Given the shared foreign language competence of the participants, and the characteristics imposed on the text by the task description (text genre, initial and final settings and characters, length considerations), we assumed that the quality of the source texts was rather similar, which should, in turn, result in MT output of rather similar quality, allowing some room for comparison.

As we can observe in Table 6, the average TER value is close to 10, which is a rather good score for the metric, indicating that participants did not consider that a high number of changes were necessary to improve the MT output. The reasons for this can vary. On the one hand, it is possible that the MT quality was very good, and therefore, no changes were necessary. However, it is also possible that the MT output was imperfect but the participants were not sufficiently competent to improve the output, or even identify mistakes.

Nevertheless, it is interesting to consider the standard deviation, which indicates a rather dissimilar behaviour among participants. A closer inspection showed that 12.5% did not introduce any change in the MT output, whereas 20% modified more than one in every five words. Therefore, we can argue that the approach followed to edit the MT output was diverse. The total edits performed and its standard deviation also reflect this trend. Whereas we see the average at 24.02 edits, the standard deviation is extremely high at 23.59.

It is worth noting that changes introduced by the users may originate from diverse needs and also lead to different outcomes. Just to mention a few, we identified cases where editing was performed to adjust the meaning expressed by the MT output to the originally intended (see Example 1), to make stylistic changes – with various results (see Example 2), or even with the intention of improving the MT output but introducing errors (see Example 3).

**Example 1:** Required meaning adjustments.
**Basque source:** *Plater bat jan eta beste bat ateratzen zuen.*
**MT output:** He ate one dish and took another.
**Post-editing:** When one dish was finished she served another.

**Example 2:** Stylistic changes.
**Basque source:** *Udako oporrak ziren.*
**MT output:** It was a summer vacation.
**Post-editing:** This story happened in a summer holiday.

**Example 3:** Introduction of errors.
**Basque source:** *Zer esango diot?*
**MT output:** What will I say?
**Post-editing:** What will I told her?

Even when we must remember that the optimisation logic used by the TER metric does not always match the linguistic intuition used by users when editing text, it is worth considering the edit types calculated by the metric. We see that shifts were, by far, the less frequent, which indicates that the MT system output the information in an acceptable order for the participants. Insertions and deletions remained at around 5-6 on average, and substitutions were twice as frequent at 11.40 on average.

The observed results reveal the complexity of the editing behaviour in this type of set-up and, albeit out of the scope of the present analysis, call for a comprehensive manual analysis of the edits to shed light into behavioural patterns.

### 3.3 How do participants view machine translation?

Let us finally address participants' perception of MT usefulness. After performing both tasks, participants were asked to assess how much the MT system made the task easier for them. In a scale of 1-5, where 1 is not at all and 5 is completely, participants rated the usefulness of the MT system to produce short fiction narratives at 3.95 on average, with a standard deviation of 0.95. This clearly shows the positive attitude towards the technology.

Participants reporting a very positive attitude towards machine translation emphasised that they greatly valued that the MT output provided them with the translation of words that were unknown to them and that the system dealt with verb tenses and forms properly for them. They also claimed that the MT system showed them translations they

would have never considered, as they differed considerably from the original structure or use of words, allowing them to learn alternative ways to express their ideas. While they acknowledged the difference between Basque and English in terms of *how things are said*, they conceded that they produce foreign language texts that follow their main language's patterns. These participants noted that they had to make few changes, which involved either correcting errors or adjusting the meaning.

Participants who were more critical towards MT tended to acknowledged its value and then added the negative aspects encountered during the task. Among their complains, worries and regrets were the fact that the MT service was not interactive, that they could keep parts of the output but had to modify others, that the meaning was sometimes distorted, and that the system was unable to handle irony or identify specific intents. It was interesting to read a comment conceding the lack of competence in the foreign language to properly assess the quality of the MT output.

## 4 Conclusions

Given the increase in the translation quality provided by automatic systems, the option of using online freely available systems to produce text in a language in which we are not fully competent by exploiting our main language is more and more appealing. With this in mind, this work analyses the effects of using MT when writing flash fiction in the foreign language.

To examine this, we asked participants, who were advanced users of Basque (language of instruction) and independent users of English (foreign language), to write two pieces of flash fiction of around 150 words each, with and without using MT. We compared features of the stories produced in each set-up with the aim to examine the effect of starting the writing process in a language in which the participants were competent and having an MT system provide them with a preliminary translation. Specifically, we aimed to observe whether MT can help to produce a text that sounds more English, and whether it can increase the complexity of the text. To that end, we compared word- and POS-level perplexities, lexical proportions, diversity and density, and the frequency of semantic relations that involve complex structures.

Results suggest that using MT participants produced final foreign language texts that followed English word- and POS-sequences more closely, indicating a higher fluency. We also observed that the proportion of pronouns and prepositions or subordinate conjunctions was higher in this set-up, even when no significant difference was observed in lexical variety and density. Dependency relations, in turn, revealed that the frequency of noun and prepositional modifiers, as well as coordinating conjunctions was also significantly higher. Overall, we can conclude that the texts produced using MT display certain traits that are typical of better quality texts.

We also considered the post-editing work of the participants. By examining TER scores and edits counts, we discovered that the participants approach the MT output differently. While it is true that the edit-distance is rather low in general, some make no changes to the output, whereas others change over 20% of the words, with most staying somewhere in between. Finally, it was encouraging to learn that participants perceived that the MT system was useful for the task (it obtained a score of 3.9 on average in a 1–5 scale), which shows the advance of MT quality for Basque and the positive attitude towards the technology.

While this research has revealed a number of interesting features from a quantitative perspective, further research into the actual lexical choice in each of the set-ups is now necessary to highlight differences in terms of lexical precision and difficulty between set-ups. Also, what remains to properly account for is the level of proficiency participants show in addressing MT output. Interesting results would be provided by research reporting on the elements that prompt users to introduce changes and on the impact these have at a linguistic level but also from the reader's perspective. Complementary research on the linguistic characteristics of the texts and user performance could also shed light into second language acquisition processes and teaching opportunities, as well as guide MT development.

## Acknowledgements

The research leading to this work was partially funded by the Spanish MEIC and MCIU (UnsupNMT TIN2017-91692-EXP and DOMINO PGC2018-102041-B-I00, co-funded by EU FEDER), and the BigKnowledge project (BBVA foundation grant 2018).

## References


Agerri, Rodrigo, Josu Bermudez, and German Rigau. 2014. Ixa pipeline: Efficient and ready to use multilingual nlp tools. In *Ninth International Conference on Language Resources and Evaluation, May 26-31, Reykjavik, Iceland*, pages 3823–3828.

Aranberri, Nora, Gorka Labaka, A Diaz de Ilarraza, and Kepa Sarasola. 2014. Comparison of post-editing productivity between professional translators and lay users. In *Proceeding of AMTA Third Workshop on Post-editing Technology and Practice, October 22-26,Vancouver, Canada*, pages 20–33.

Batchelor, Katherine E. and April King. 2014. Freshmen and five hundred words. *Journal of Adolescent & Adult Literacy*, 58(2):111–121.

Björkelund, Anders, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task, June 4-5, Boulder, Colorado*, pages 43–48.

Bochnak, M Ryan and Lisa Matthewson. 2015. *Methodologies in semantic fieldwork*. Oxford University Press, USA.

Bowker, Lynne and Jairo Buitrago Ciro. 2015. Investigating the usefulness of machine translation for newcomers at the public library. *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association*, 10(2):165–186.

Bowker, Lynne and Jairo Buitrago Ciro. 2019. *Machine translation and global research: Towards improved machine translation literacy in the scholarly community*. Emerald Group Publishing.

Bowker, Lynne. 2009. Can machine translation meet the needs of official language minority communities in Canada? A recipient evaluation. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, (8):123–155.

Burton, Strang and Lisa Matthewson. 2015. Targeted construction storyboards in semantic fieldwork. In M. Ryan Bochnak, Lisa Matthewson, editor, *Methodologies in semantic fieldwork*, chapter 5, pages 135–156. Oxford University Press, USA.

Cooper, Alan, Robert Reimann, and Hugh Dubberly. 2003. *About face 2.0: The essentials of interaction design*. John Wiley & Sons, Inc.

Grudin, Jonathan and John Pruitt. 2002. Personas, participatory design and product development: An infrastructure for engagement. In *Proceedings of the Participatory Design Conference, June 23-25, Malmo, Sweden*, pages 144–161.

Housen, Alex and Folkert Kuiken. 2009. Complexity, accuracy, and fluency in second language acquisition. *Applied linguistics*, 30(4):461–473.

Johansson, Richard. 2008. Dependency syntax in the conll shared task 2008.

Jordan-Nuñez, Kenneth, Mikel L. Forcada and Steve Clua. 2017. Usefulness of MT output for comprehension – analysis from the point of view of linguistic intercomprehension. In *Proceedings of MT Summit XVI, Sep. 18-22, Nagoya, Japan*, pages 241–253.

Kantola, Niina and Timo Jokela. 2007. SVSb: simple and visual storyboards: developing a visualisation method for depicting user scenarios. In *Proceedings of the 19th Australasian conference on Computer-Human Interaction: Entertaining User Interfaces, November 28-30, Adelaide, Australia*, pages 49–56.

Martín Peris, Ernesto. 2011. La intercomprensión: concepto y procedimientos para su desarrollo en las lenguas románicas. *Y. Ruiz de Zarobe y L. Ruiz de Zarobe, La lectura en lengua extranjera, London: Portal Editions*, pages 246–270.

Martín, Ernesto. 2005. *EuroComRom-los siete tamices: un fácil aprendizaje de la lectura en todas las lenguas románicas;[con CD-ROM; español-català-français-italiano-português-românã-galego-occitan]*. Shaker.

Nurminen, Mary and Niko Papula. 2018. Gist MT users: A snapshot of the use and users of one online MT tool. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation, May 28-30, Alacant*, pages 199–208.

Otegi, Arantxa, Oier Imaz, Arantza Díaz de Ilarraza, Mikel Iruskieta, and Larraitz Uria. 2017. Analhitza: a tool to extract linguistic information from large corpora in humanities research. *Procesamiento del lenguaje natural*, 58:77–84.

O'Brien, Sharon, Michel Simard, and Marie-Josée Goulet. 2018. Machine translation and self-post-editing for academic writing support: Quality explorations. In Moorkens, J. et al., editor, *Translation Quality Assessment*, pages 237–262. Springer.

Parra Escartín, Carla, Sharon O'Brien, Marie-Josée Goulet, and Michel Simard. 2017. Machine translation as an academic writing aid for medical practitioners. In *MT Summit XV, Sep. 18-22, Nagoya, Japan*, pages 254–267.

Rosson, Mary Beth, John M Carroll, and Natalie Hill. 2002. *Usability engineering: scenario-based development of human-computer interaction*. Morgan Kaufmann.

Toral, Antonio. 2019. Post-editese: an exacerbated translationese. In *Proceedings of MT Summit XVII, August 19-23, Dublin, Ireland*, pages 273–281.

Vossen, Piek et al. 2016. Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems*, 110:60–85.