Exploring metrics for post-editing effort: and their ability to detect errors in machine translated output

Author: Cristina Cumbreño Díez

Advisors: Nora Aranberri

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua

Language Analysis and Processing

Final Thesis

2019-06-13

Departments: Computer Systems and Languages, Computational Architectures and Technologies, Computational Science and Artificial Intelligence, Basque Language and Communication, Communications Engineer.

Abstract

As more companies integrate machine translation (MT) systems into their translation workflows, it becomes increasingly relevant to accurately measure post-editing (PE) effort. In this paper we explore how different types of errors in the MT output may affect PE effort, and take a closer look at the techniques used to measure it. For our experiment we curated a test suite of 60 EN > ES sentence pairs controlling certain features (sentence length, error frequency, topic, etc.) and had a group of 7 translators post-edit them using the PET tool, which helped collect temporal, technical and cognitive effort metrics. The results seem to challenge some previous error difficulty rankings; they also imply that, once other sentence features are controlled, the type of error to be addressed might not be as influential on effort as previously assumed. The low correlation values between the metrics for the different effort aspects may indicate that they do not reliably account for the full PE effort if not used in combination of one another.

Key words: machine translation, post-editing, post-editing effort, post-editing time, keystrokes, manual scoring, HTER

Index

List of tables

List of figures

1 Motivation

The landscape in the translation industry is changing. Today's interconnectedness has brought along a rapid increase in the content being produced which, in turn, has resulted in higher demand of fast, quality translations. Meanwhile, Machine Translation (MT) systems have become better, more widely available, and the subject of more scientific research. The wide array of studies (Guerberof, 2009; Plitt et al., 2010; Parra Escartín et al., 2015) showing that MT systems increase productivity has encouraged many translation companies to integrate them in their workflow, meaning that professional translators get progressively fewer translation jobs and more post-editing offers (Gaspari et al., 2015).

Post-editing consists in correcting and improving the fluency, accuracy and textual adequacy of an automatically translated text to bring it closer to human standards. Post-editing remuneration sits between the translation rates and the proofreading ones; while post-editing is assumed to be faster than translating from scratch, the quality is often not high enough to reach the standards of human output and thus allow for swift proofreading. This pricing should be a good compromise for both companies and translators; nevertheless, it is common to hear of frustrated translators complaining about post-editing jobs or refusing them altogether. While these translators are often brushed off as being negatively biased against a technological advance that threatens their careers, after conducting a series of interviews with professional translators and editors Guerberof (2013) concluded that their attitude towards working with MT systems was not negative, but that most considered the payment to be unfair when compared to the energy invested.

This energy devoted to completing a post-editing task is commonly referred to as post-editing effort, or PE effort for short. According to Krings (2001), there are three aspects to post-editing effort: temporal (how long it takes to perform an edit), technical (the physical actions taken to modify the text) and cognitive (the type of intellectual processes experienced while post-editing).

One could assume that Krings' three approaches to measure post-editing effort may correlate well: for example, if a big percentage of a sentence is modified, typing the modifications will take time, and finding out what to modify will take mental effort. However, let us imagine another situation: some MT output where a single but very difficult term has been mistranslated. The post-editor will have to read the source segment

(i.e. sentence), locate the error and consider how to approach the edit (cognitive effort) and may spend time looking through Translation Memories or terminological databases for the correct translation (temporal effort); the technical effort, however, will be low, as only one word has been substituted.

Many different techniques are used to measure post-editing effort, but they often focus on just one of the aspects. As we have seen, on some occasions these metrics could be used indistinctly, but on others trusting the measurements of one effort aspect might mean grossly underestimating another one. In order to investigate these discrepancies, our first research question will consist on exploring how techniques for measuring different aspects relate to each other, and whether their results converge or diverge when presented with the same sentence.

The example we posed introduces another interesting issue: the kind of error present in a sentence. Research (Temnikova, 2009; Popovic et al., 2014) has been aimed at trying to determine the influence that an error has in the post-editing effort of correcting the segment; some of it has been aimed at rankings of error difficulty. Nevertheless, it is still not clear how relevant specific errors might be to the total effort required to post-edit the sentence. Our second research question will focus on the effect that different types of errors have on effort, and whether the measuring techniques commonly used in the translation industry to measure PE effort can detect any differences.

This work attempts to explore these two research questions. In order to do so, we carry out an experiment in which a group of translators has to post-edit machine translated output containing specific errors. Their performance is annotated with different measuring techniques and the results are analysed.

The structure of the rest of this work is as follows: the second chapter reviews the literature on the subject; the third explains the experiment design, including the characteristics of the participants, the errors and the techniques used to measure them; the fourth section details the results of the experiment and discusses them; the final chapter includes the conclusions and future work.

2 State of the art

This chapter will review previous research work for the questions in hand. It has been divided into three sections: the first one explains basic concepts about the translation industry and how translators work; the second one examines papers related to our first research question, exploring different measuring techniques; finally, the third section reviews relevant literature for our second research question, concerning errors and their effects on post-editing.

2.1 A word on the translation industry

The translation industry has specific *modus operandi* and terminology with which we should get acquainted before delving any further into the research questions. Currently, when a translator accepts a job offer, they often receive a file that is opened with a Computer Assisted Translation tool, or CAT tool, such as SDL Trados Studio[1] or MemoQ[2].

CAT tools originated in the 1980s as workstations for translators integrating a text processor, dictionaries and a terminology database. The idea behind this software was to facilitate the translators' tasks by grouping various useful tools in a single location. Nowadays CAT tools offer many more features; first of all, a more structured way of visualizing texts (see Figure 1). CAT tools break the texts into segments (i.e. sentences) which are presented as a row of cells. If the source segment (i.e. sentence in the source language which must be translated) matches segments which have been translated in previous projects (which are stored in Translation Memories), the CAT tool will retrieve a potential translation and offer it to the translator so that they correct it instead of translating from scratch. The CAT tool will compute the similarity between the retrieved sentence and the source segment as a percentage; the higher this percentage is, the smaller the payment the translator will receive for it (as it is assumed to be easier); these are called "fuzzy matches". 100% match segments are sometimes considered as proofreading.

Additionally, some CAT tools may display a machine translated version of the source segment; these are often offered when the Translation Memory matches do not

[1] You may find this tool at https://www.sdl.com/es/software-and-services/translation-software/sdl-trados-studio/

[2] You may find this tool at https://www.memoq.com/es/

reach a certain similarity threshold. Post-editing rates may vary from one project to the next, but they are usually fixed for all segments in that project, and do not increase or decrease depending on the quality of the Machine Translation output. Hence the importance of estimating MT quality: it could completely change the pricing system for post-editing tasks.

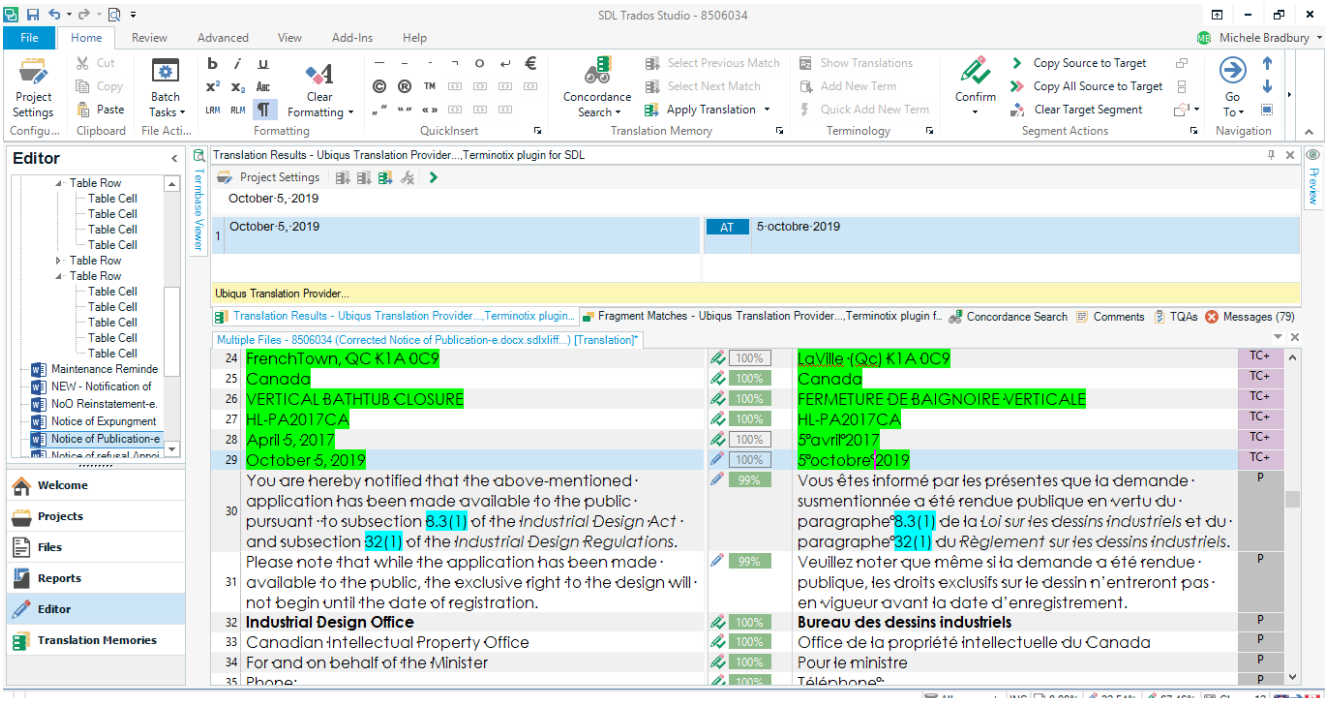


Figure 1 - Screenshot of a project on SDL Trados Studio. From: https://community.sdl.com/

Translators translate segments from a source language into a target language (usually the translator's mother tongue); these language combinations are often presented in this format: EN>ES (i.e. from English into Spanish). When they want to work on a specific segment, translators can open it (i.e. access it) by clicking on the cell. Once the translating or editing is done, the translator will close the segment by clicking outside the cell or moving onto the following segment. Translators will often perform the task in one round, and use one or more subsequent rounds to proofread their work.

2.2 Post-editing effort measuring techniques

The first research question of this paper focuses on comparing the performance of different measuring techniques. Krings' (2001) definitions of the different aspects of post-editing effort have been widely accepted amongst the research community, with research focusing on developing, testing and combining techniques to measure each aspect of effort to find the most complete way of capturing PE effort.

Accurately measuring the effort associated to a post-editing task is incredibly relevant. For example, it directly influences the performance of Confidence Estimation (CE) models, which are used to determine the quality of machine translated output. These results are in turn used to ascertain which segments are good enough to be post-edited, and which ones should be translated from scratch, optimizing the productivity and profit of both translators and translation companies.

Understanding each aspect of effort and what it encompasses is key. Let us begin with the most straightforward aspect: temporal effort, also known as post-editing time. Post-editing time is usually understood as the time frame since the translator opens a segment until they approve and close it. At first, in order to study temporal effort, researchers asked translators to measure themselves with a stopwatch, which was highly unreliable and got in the way of the natural translation flow. Later, as CAT tools were developed, it became easier to develop plugins for these tools that would track the time taken by each segment; such is the case, for example, of the CAT tool SDL Trados and its plugin Studio Time Tracker. Overall, post-editing time is one of the most commonly used metrics for PE effort, inside and outside of the research community, thanks to its simplicity and cost-effectiveness. For example, Plitt et al., (2010) and Parra Escartín et al., (2015) used PE time to measure productivity gains between translation and post-editing.

Regarding technical effort, researchers have taken various approaches to capture it. The most straightforward technique is to measure the number of keystrokes: one stroke corresponds to one physical action taken by the editor. However, keystroke loggers are not integrated in major CAT tools like SDL Trados or MemoQ; they need to be launched alongside and then have the results aligned to each segment, which can be challenging. Moreover, keyloggers such as BlackBox Express[3] are often designed for security purposes and do more than simply tracking which keys are pressed; they can register any other computer activity such as web searches, email client programs, passwords, etc. along with screenshots. This adds much more data to parse and is more intrusive.

A more popular tool in the research world is the automatic metric Translation Edit Rate, or TER. TER originated as part of DARPA's GALE program but gained notoriety after being described by Snover et al. (2006). TER computes the minimum number of editing operations (i.e. insertions, deletions, word substitutions or phrase shifts) to be

[3] You may find this tool at http://www.asmsoftware.com/

performed on a given machine translated output so that it becomes an exact match of its reference human translation, normalized by the number of words in the reference.

Compared to other automatic metrics, TER is cheap and easy to use; it also seems to correlate well with human judgments of translation quality (Snover et al., 2009). Nevertheless, TER has shortcomings; for example, it gives every edit the same score (even though some edits may be more challenging than others), and it does not take into account the edits that a translator could perform and then discard, only the ones that appear on the final version.

Improved versions of TER have been developed to cope with some of the original metric's issues; one such version is the human-targeted TER, most commonly referred to as HTER (Snover et al., 2006). HTER is based on the fact that the human references used to compute the TER score are only some of the potential translations of a given source text, and that one of those other possible translations could have a smaller edit distance to the hypothesis that any of the references. Because TER does not consider the semantical content of a sentence and only tallies exact word matches, sentences with the exact same meaning as the reference but very different wording could be wrongly considered to require a lot of editing. What HTER does to bypass these problems is giving the hypothesis and the reference translation(s) to a human editor, who performs as little editing as possible on the hypothesis to make sure that it is semantically equivalent to the references. Finally, the edited hypothesis is used as a reference to calculate TER, which should now be lower than if it was computed directly on one of the original reference translations.

HTER has been shown to have very high correlations with human judgements of quality. Currently, HTER is not used in its strictly original sense, since its need for human editors made it prohibitively expensive for large tasks, and thus not very frequently used. HTER is now commonly used to refer to the edit-distance between some MT output and its post-edited version, and is usually presented as a value between 0 and 100, which represents the percentage of the sentence that must be edited.

Both keystroke logging and HTER are used to compute technical effort, even though their methods to measure it are quite different: one tracks every literal keystroke, while the other only considers the number of word-level edits between the MT output and final post-edited version. While keystrokes tend to be used for academic research complementing other metrics such as post-editing time due to its technical constraints (some such studies will be mentioned later in this section), HTER's simplicity has

allowed it to be frequently used within the translation industry. For instance, HTER was favoured by Specia et al. (2010) as a base to build confidence estimation models; it was shown to give better results than other commonly used sentence features for CE, such as sentence length.

Finally, the third PE effort aspect is cognitive effort, which focuses on identifying the errors and deciding how to solve them. Cognitive effort is the most difficult aspect to quantify (as it delves into subconscious processes and mental strain), but it is arguably the most important one. Determining how much strain a task imposes on the brain, or how much frustration it sparks, could help create models that predict fatigue and strategize the work accordingly to improve productivity.

Attempts have been made to measure cognitive effort through complex techniques: think-aloud protocols (Krings, 2001) consist in making the translators explain their edits as they happen, but in doing so they affect the natural flow of the translation, fail to tap into the subconscious processes, and do not offer comparable results; choice network analysis (O'Brien, 2006b) explores the different ways a segment can be edited, with the assumption that the more options there are, the more effort it takes to choose among them, but it does not take into account that not all options are available to all post-editors; finally, eye-trackers follow the editor' gaze, assuming that the segments where the gaze stays the longest are more cognitively demanding.

Eye-trackers have gained momentum in recent years, moving from a relatively expensive and unexplored technique to a budding source of reliable cognitive effort measurements. Average fixation time and count have been used to determine the quality of MT output (O'Brien, 2011; Moorkens, 2018), as well as to assess translators' reactions to new CAT tools (Mesa Lao, 2013). Eye-trackers have also been applied to measuring productivity; da Silva et al. (2017) and Carl et al. (2011) noticed a significant increase of cognitive effort in translation from scratch as opposed to post-editing. Similarly, Alves et al. (2016) used eye-trackers to compare Interactive Machine Translation (i.e. where the tool displays suggestions as the translator writes) to non-interactive MT and found that the first one decreased the cognitive effort. Finally, eye-trackers have been employed to determine when and how different types of errors were recognized (Schaeffer et al., 2019) and their impact on cognitive effort (Daems et al., 2017).

As we can see, eye-trackers are very promising and open a new field of research for cognitive effort in post-editing, but they have remained largely confined to the academic fields for now. This may be partly due to the novelty of the technique and the

expertise required to apply it, and partly to the prohibitive cost of using eye-trackers on a large scale.

Another approach to measuring cognitive effort consists in analysing the presence of pauses or "thinking" time (Plitt and Masselot, 2010) within the sentence. It is assumed that the more a translator pauses before an edit, the more cognitively challenging the edit is; because of this, researchers have studied the pause-typing ratios, as well as the duration, frequency and distribution of pauses in the sentence. For example, Lacruz et al. (2012) and Lacruz et al. (2014a) linked the presence of clusters of short pauses with cognitively challenging edits. Similarly, Probst (2017) found differences in the pause length prior to post-editing certain error types; on the other hand, O'Brien (2005, 2006a) examined pauses in segments containing specific source text features believed to increase cognitive effort and segments without them, but found no significant differences.

Another way of investigating cognitive effort is to simply ask the people involved to assess how difficult they considered the task, before or after performing it (Koponen, 2012); this is often referred to as manual evaluation or perceived effort. This method is cheaper than other approaches, but very subjective: inter-annotator agreement tends to be very low. Because of this, even when used to get an overview of translator behaviour or perception, researchers discourage basing Confidence Estimation models that will be decisive on translation workflows solely on human ratings (Moorkens et al., 2015).

Attempts have also been made to compare and combine various metrics, from the same or different effort aspects, in order to accomplish different tasks. For instance, Aziz et al. (2013) used HTER, post-editing time and keystroke logging to create new golden standards for MT system ranking. In addition, Specia (2011) created CE models based off texts annotated with either TER, post-editing time or perceived effort, and obtained the best results with PE time, which they also considered to be the simplest and most objective metric.

Koponen et al. (2012) and Aziz et al. (2014) compared post-editing time and HTER, finding out disparities between both metrics results. They concluded that, by giving the same weight to all edits, HTER fails to fully capture post-editing effort. Koponen et al. (2012) went further by proposing post-editing time as a possible measure for cognitive effort, arguing that most cognitively difficult errors (as per Temnikova's error ranking, which will be discussed at greater length on section 2.3) appeared in the sentences taking the longest time to post-edit.

Lacruz et al. (2014b) found strong correlations between pause-word ratio, HTER and perceived effort. Eye-trackers have also been shown to have good correlations with other metrics of effort; Doherty et al. (2010) and O'Brien (2011) used eye-trackers to obtain average fixation time and count and drew good correlations with HTER and perceived effort. Moorkens (2018) also correlated average fixation duration with technical effort, and average fixation count with temporal effort.

Many of these comparisons would not have been possible without the development of research-focused tools that collected several metrics at once while keeping a CAT tool-like interface. Some examples of this are PET[4] (Aziz et al, 2012), Translog II[5] (Carl, 2012) or Matecat[6] (Federico et al., 2012). PET collects time, keystrokes, perceived effort, edit operations and HTER, and is highly customizable; its main issue is the lack of clear online instructions to learn how to use it. Translog II measures post-editing time and keystrokes, and additionally integrates gaze data tracking; however, this requires the expertise of installing and operating the eye-tracking systems and cameras. Finally, Matecat measures time and HTER, but it is now more focused for commercial use than for research.

2.3 Post-editing of different error types

Another interesting point when looking into PE effort is the kind of errors present in the MT output, their frequency, and whether their presence leads to increased difficulty and effort. It stands to reason that there would be differences between, for example, finding a word that is missing from the source segment and inserting it, and correcting the number agreement between a noun and a verb. In the first case, the translator would have to look at the source and reflect on the translation for that word while, in the second case, the translator would only have to use their knowledge of the language's grammar to add or subtract a few letters.

Error detection and classification has a relevant purpose: when MT system developers analyse the quality of the segments identifying errors, their types and their frequencies is crucial to improving the system. Because of this, research has focused on

[4] You may find this tool at http://wilkeraziz.github.io/dcs-site/pet/index.html

[5] You may find this tool at https://sites.google.com/site/centretranslationinnovation/translog-ii

[6] You may find this tool at https://www.matecat.com

different ways of defining error categories: for example, the European Union's Horizon 2020 research project Quality Translation 21 developed the Multidimensional Quality Metrics[7] (MQM), which encompasses a comprehensive hierarchy of quality issues in translation, including standard naming that has been applied in the translations industry. The main seven categories, according to this classification are: accuracy, fluency, design, locale convention, style, terminology and verity.

On a more academic approach, Vilar et al. (2006) proposed five main categories to analyse errors in MT output: missing words, word order, incorrect words, unknown words and punctuation errors. These categories were then split into subcategories that allowed for finer error classification, and these were used to analyse the distribution of errors within text according to language pairs and directions. Popovic (2011) even developed a method to automatically classify machine translation errors into Vilar et al.'s main categories with a tool called Hjerson.

While error classifications allow us to detect and group errors, they do not provide any information about the effort involved in editing them. In a post-editing context, it is also important to know whether some errors are more difficult than others to address. For this reason, Temnikova (2010) adapted Vilar et al.'s categories and ranked them from 1 to 10 by cognitive effort, based on Harley's cognitive model of reading (2008), Baddeley and Hitch's working memory theory (1974) and Larigauderie's written error detection studies (1998). Going from simply categorizing errors into ranking errors by difficulty has very important applications: MT system developers can focus on eliminating these errors according to their priority, CE models can aim at detecting these errors as indicators of low quality, etc.

According to Teminkova's ranking, the cognitively easiest errors to correct are the ones happening at morphological level (correct word with incorrect form), followed by those at lexical level (incorrect style synonyms, incorrect words, extra words, missing words and mistranslated idioms). The most difficult errors happen at the syntactic level, having to do with punctuation (wrong or missing) and word order (at word or phrase level).

This ranking has been used as a way to test different metrics or check whether their measurements reflect the different difficulty of the errors. Koponen (2012) followed this premise and found that sentences with low perceived effort scores involved changes

[7] You may find this classification at http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

in word order or word class, or mistranslated idioms, while "easier" sentences involved changes in word form or substitutions of words of the same class. Similarly, Koponen et al. (2012) correlated long post-editing times with cognitively challenging errors according to Temnikova's ranking. Popovic et al. (2014) also researched the effects of different types of edit operations on difficulty, finding that lexical and word order edits received worse perceived effort scores, while lexical edits took the longest; removing extra words, however, had little effect on effort. Probst (2017) reached similar conclusions to Popovic et al. (2014) by analysing the length of the pauses right before errors were corrected.

Exploring the effect of errors on post-editing effort is a research area that has been gathering more attention in these last years, but that still merits much more research efforts. So far, Temnikova's effort ranking remain largely unchallenged, with many papers using it as the base for their research. Nevertheless, papers like those of Popovic et al. (2014) or Probst (2017) point that the ranking could use further confirming or tuning. Establishing robust effort ranking that takes into account all aspects of effort could have many potential ramifications, from developing more reliable CE models to revolutionizing the pricing system for post-editing tasks.

3 Experiment design

Our objective in this paper is to explore our two research questions; for this we have designed an experiment which will allow us to compare various measuring techniques and errors. This chapter presents the experimental setup. The first section introduces the characteristics of the participants and the selection process; the second explains the test suite; the third section presents the process to determine the errors we would analyse; the fourth part concerns the aspects we measured and the measuring techniques we used to collect the information. Finally, section six summarizes the task and how it was presented to the participants.

3.1 Participants

Our experiment required of a group of participants to carry out the task, which would consist on post-editing a series of sentences. This allowed us the choice between two profiles: translators or editors. Editors are rarer to find, and lots of translators also carry out post-editing or mixed tasks, which means that they develop additional competences, so we chose to go for the latter profile. Moreover, we selected only professional translators, instead of students, because we wanted to replicate the actual behaviour that translators may have, which is only acquired through experience. We decided that a minimum of 5 participants would be needed to obtain enough variability amongst the results.

The participants were found through a job posting on the professional website ProZ. Around 50 applications were received on the first day, from which 7 participants were chosen after reviewing their CVs. The participants worked in the EN > ES (Spain) language pair and had at least one year of experience in translation and at least 3 months of experience in post-editing. All of them except one had language-related studies (either bachelor's or master's degrees), mostly in translation or specialized translation.

The participants were asked to fill in a short survey before completing the task. The survey consisted of a series of statements that they could give their opinion about, ranging from fully disagree (1) to fully agree (5), and a question concerning the fairness of post-editing remuneration, which ranged from very unfair (1) to very lucrative (5).

The intentions behind these questions were twofold: first, they would help disqualify any translator with extreme opinions about post-editing to avoid them from

introducing bias into the experiment intentionally; second, the answers would help us measure what the general attitudes in the community are, and whether there is agreement.

Table 1 shows the questions, along with an explanation of their relevance and the average results. While not all translators agreed, as evidenced by the standard deviation, there were no outliers that needed to be discarded, and the similar answers in some questions were very revealing. In general, the translators enjoy translating more than post-editing; this might be due to the fact that translating is a more creative and entertaining task than correcting. Regarding PE remuneration, opinions were divided between 3 (fair) and 2 (unfair); none of the translators regarded PE remuneration neither as very lucrative, nor as very unfair. Finally, translators also considered the quality of MT output not to be good enough, yet they said they do not always have a way to check it before accepting the job offer. All this information seems consistent with previous research into translators' opinions like Guerberof's (2013).

Statements	**Why this question was asked**	**Average**	**Standard deviation**
It takes me less time to post-edit a text than to translate it from scratch	Research suggests that PE boosts productivity by saving translators time.	3.36	0.95
I enjoy post-editing	This question is aimed at assessing the translators' general attitude towards PE.	3.48	1.10
I like translating more than post-editing	This question is aimed at assessing the translators' general attitude towards PE and translation.	4.10	0.68
I accept all post-editing jobs proposed to me	Translators often refusing PE jobs may point to bad past experiences or weariness.	3.36	0.81
Post-editing jobs tend to be frustrating	This question is aimed at assessing the translators' potential bias against PE.	3.48	1.06
I cannot assess the difficulty of a post-editing job before accepting it	There are general complaints about the MT output quality, which could be avoided if translators could see a sample of the text in advance.	4.10	1.24
The quality of machine translated text tends not to be good enough so that the job is profitable for me	While research suggests that PE increases productivity, translators sometimes complain that it makes them lose money	3.91	1.01

The retribution for post-editing jobs is...	PE retribution is controversial, since it is not always faster or easier than translation, but it always pays less	2.56	0.48

Table 1 - Survey questions asked to the translators, average and standard deviation of their answers

An additional space was provided for the translators to write comments or clarifications if desired. Several pointed out that post-editing jobs were varied; where some could be enjoyable and profitable, some would be very frustrating, depending on the quality of the MT output. They agreed that MT could help but was not useful in every situation. Often, they commented, they would end up translating segments from scratch, but for a reduced fee. Another added that this situation was dangerous because some translators would try to skim through the text as fast as possible and, as a result, let mistakes and false friends slide. The same translator concluded that a sample fragment of the text should always be provided for post-editing jobs, but that this is not yet common practice in the industry.

3.2 Dataset

The objective of the experiment was to analyse the influence that different errors may have on the PE effort of the sentence; thus, the dataset had to consist of source sentences in English and machine translated output in Spanish, containing specific errors.

Several datasets commonly used for research were considered, such as those used at the different Workshops on Machine Translation from the Conference on Empirical Methods in Natural Language Processing. Upon analysing these datasets, we realized that the sentences contained within were very different from one another, often varying greatly in length and error frequency. We considered that, our objective being to study the effects of errors, we could never be sure that other varying features of the sentence were not interfering with the results. Only if we controlled as many external factors as possible and isolated the errors within the sentences could we be confident, within reasonable doubt, that any potential variation in the results amongst errors was caused by the errors themselves.

A dataset of such characteristics was not available, so we carefully curated our own test suite. Test suites are collections of sentences presenting specific characteristics, which would not usually happen together in the same text. Test suites have been proposed

in past studies, such as Guillou and Hardmeier (2016) or Burchardt et al. (2016), as the best way to analyse a specific aspect of a sentence. Recently, Schaeffer et al. (2019) used a test suite to analyse errors in human translation proofreading, which allowed them to limit total and local error frequency.

Our first task was to decide the aspects over which we wanted to have control. We realized that in several past papers, researchers would acknowledge that their conclusions were tentative because they did not have enough instances of the features they were trying to analyse or compare (Moorkens et al., 2015; Probst, 2017). Thus, the most important thing when designing the dataset was to try to ensure that every single one of the errors we aimed to analyse appeared enough times to make the results levelled and comparable, even if these errors would not happen in the same frequency naturally.

Next, we focused on the number of errors we needed each segment to contain. Having more than one error per sentence would make the sentence selection task much easier; we found more naturally occurring sentences containing several errors than containing just one. This may be due to the fact that the presence of an error will often cause the occurrence of another. Nevertheless, we decided that each sentence would contain just one kind of the chosen errors; otherwise, the final results would be very hard to analyse and compare, not knowing which error had more weight in the difficulty of the sentence. In general, there was only one instance of said error, but in some occasions, such as agreements between words, we allowed for the error to affect more than one word if they were close and clearly related.

We also decided that these errors had to be naturally occurring. That is, we would not alter the MT output, but rather look for sentences where the errors happened spontaneously after passing them through the MT system. On occasion, we modified the source segments superficially so that they would meet our desired specifications, but we never introduced, removed or corrected any part of the MT output. This was crucial because MT systems do not create the same errors as humans may.

The next feature we desired to control was the sentence length. According to Tatsumi (2009), Koponen (2012) and Popovic (2014), among others, sentence length can negatively affect human scores on MT output, because longer reading times make sentences appear more difficult to post-edit, affecting perceived cognitive effort, and have an impact on post-editing time. Establishing maximum and minimum sentence length limits should help reduce the impact of this variable.

Finally, we introduced restrictions on terminology, formality and style by extracting all sentences from the same source, around the same time, and about the same topic. This control on language reduced the impact that the rest of the words in the sentence would have on the results (although of course this variable was impossible to eliminate completely).

Additionally, we had to choose the MT system we wanted to use to extract the errors. We decided to use Google Neural Machine Translation system as it was both free and state-of-the-art technology.

With all these aspects considered, these are the characteristics of our test suites: they are sentences extracted and, occasionally, adapted, from the online International Edition of the newspaper The Guardian. The news articles span from January 23rd to February 15th and all discuss the Venezuelan crisis. The sentence length ranges from 20 to 25 words, both inclusive, and the sentences contain just one error each. There are 10 sentences for each kind of error, amounting to a total of 60 sentences. The sentences were arranged so that they would be narratively cohesive, and it would be always clear what person or situation they were referring to. The full test suite can be found in Annex 1.

3.3 Errors

We considered several popular error classification methods, such as MQM and the ten categories proposed by Temnikova (2009). MQM was eventually discarded because of its heavier focus on human translation editing. MQM's classification is very broad, but many of the areas do not apply to post-editing, while important PE classical errors are not present.

Temnikova's categories are more interesting since, as it has been previously explained, she proposes a ranking by cognitive effort that has been widely used in research. At first, we started choosing sentences for the test suite according to this classification, but we soon realized some categories were almost void, while others had so many instances that allowed for more nuance. Moreover, the categories that had more instances were assumed to be more representative of the most common MT problems. This approach has been used in research such as Schaeffer et al. (2019), who chose a prior classification system and adapted it to the frequency of the errors on their dataset to achieve statistical significance.

Table 2 shows the transition from Temnikova's classification (left) to our final chosen categories (right). There is also a brief description of the categories taken from Temnikova (2009) and an explanation of how these were adapted to our selection.

Original Temnikova Error	**Description**	**Action**	**Final categories / ranking**
Correct word, incorrect form (e.g. number or case)	Error correction requires replacing with a different ending	Split into two categories due to the large number of instances	Agreement of number / gender
			Agreement of time / aspect
Incorrect style synonym	Error correction requires a different style synonym	Discarded; not enough instances found.	
Incorrect word	Error correction requires replacing with a completely different lexical item	Used as is	Mistranslation of 1 word
Extra word	Error correction requires deleting the extra word	Used as is	Extra word
Missing word	Error correction requires adding the missing word	Used as is	Missing word
Idiomatic expression	Error correction requires replacing with the correct translation of the idiomatic expression	Transformed; focus will be on phrasal verbs and other multi-word expressions.	Mistranslation of 2 or more words
Wrong punctuation	Error correction requires replacing with the correct punctuation sign(s).	Discarded; not enough instances of these categories, plus all found instances were almost the same error (no variation).	
Missing punctuation	Error correction requires adding the missing punctuation sing(s).		
Word order at word level	Error correction requires moving single words	Discarded; word order is very difficult to find not co-occurring with other errors.	
Word order at phrase level	Error correction requires moving whole phrases		

Table 2 - Transition from Temnikova's error difficulty ranking into our final ranking

Our categories consist on two different types of agreements (a general number and gender one, and a second one focused on verbal tenses and aspects); missing and extra words; and mistranslation of one or more words. The final categories are presented in Table 3 ranked by cognitive difficulty according to Temnikova's original classification; our assumption is that the results will follow this ranking.

The columns on the right of the table represent the number of times these errors appeared in the total of analysed sentences, and the corresponding frequency. We can see that the most common error for this MT system to make is mistranslation, while the least common is gender or number agreement. This stands to reason, as agreement is easy to infer from the surrounding words while, without context, it may be hard for a MT system to choose the correct sense of a word. Moreover, agreement errors tend to be more common as the sentences become longer and the subjects are separated from their corresponding verbs; since these sentences had a controlled length that was rather short, these errors were not as common.

Code name	Description	Total count	Frequency
Agr N/G	Wrong number or gender of one or more words	20	0.03
Agr T/M	Wrong tense or mode (aspect) of one or more verbs	53	0.08
Mistr 1	Mistranslation of one word	89	0.14
Extra w.	Extra word (not present in source sentence)	51	0.08
Missing w.	Word present in source sentence but missing in machine translated output	32	0.05
Mistr 2+	Mistranslation of two or more words (multi-word expressions)	67	0.11
Others	No errors / other errors / more than one error	288	0.48
Total		600	1

Table 3 - Description of our final error difficulty ranking

We chose 10 instances of each type of error from all the available sentences, trying to combine them in a coherent way so that the final dataset would be telling a story. Our resulting test suite, thus, contained 60 sentences.

3.4 Measured aspects

Post-editing effort, as previously stated, has three main aspects: temporal effort, technical effort and cognitive effort. Each of these aspects can be measured with different metrics that are more or less complex, rare or expensive. We decided to focus on metrics more commonly used in the industry, rather than in academic research.

While some CAT tools have integrated plug-ins that allow measuring some of these aspects, we decided to look for an open-source tool that would combine as many metrics as possible, and that gave us enough raw information to be able to compute other aspects. We decided to use PET (Post-Editing Tool), a graphical user interface for translation and post-editing developed by Wilker Aziz and Lucia Specia which allows researchers to gather effort indicators and is highly customizable to the researchers' needs. The user part of PET has a CAT tool aspect, which displays the source segments on one column and the segments to post-edit on the other (see Figure 2). As you can see on the screenshot on Figure 3, the segments remain blocked until you click on them, and previous or following segments are also unreadable. Since translators sometimes start reading the following segment before closing the current one (thwarting the chronometer results for both segments), PET was customized so that translators could not read segments unless they accessed them; that way any reading or reflection time will be captured in the correct segment.

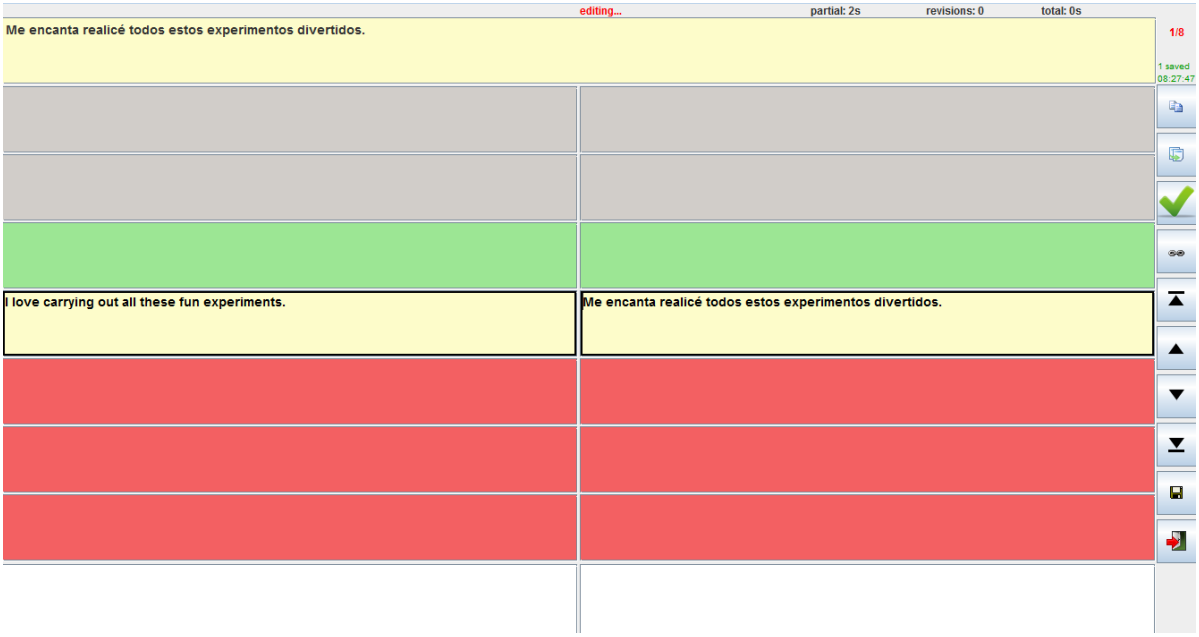


Figure 2 - Screenshot from PET showing a segment before being opened

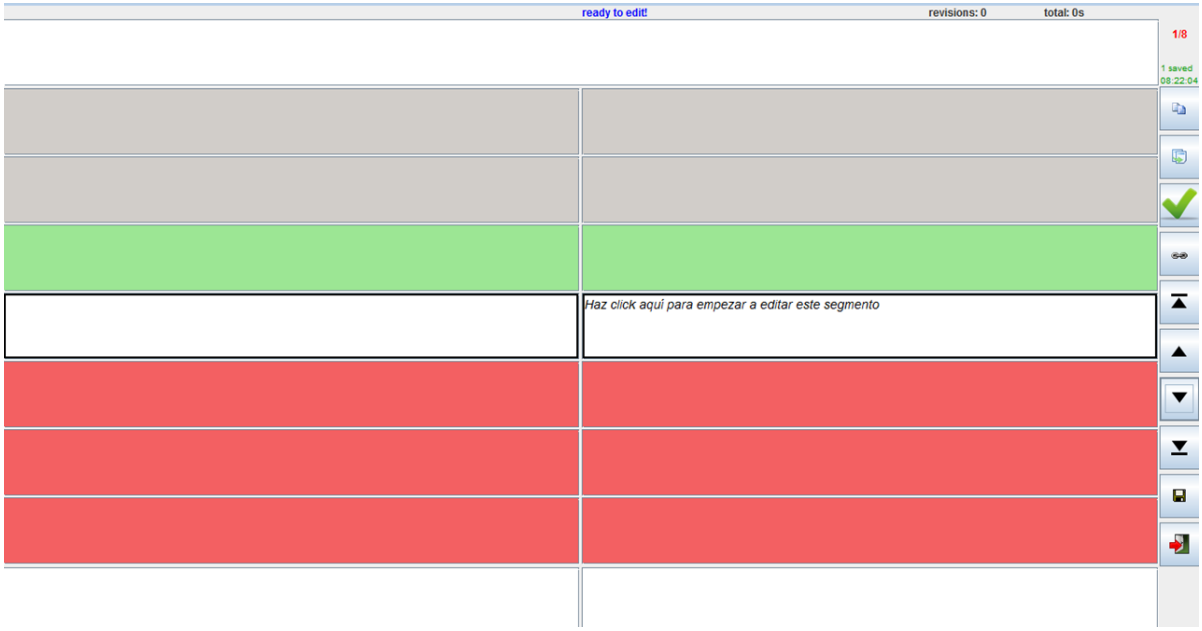


Figure 3 - Screenshot from the test PET task showing an open segment

Figure 4 shows a screen caption of PET's results in their raw form. We have chosen a rather short example, where only a couple of letters were introduced and the segment was only opened once. PET includes a way to automatically parse these files and extract the results, which already include quite a lot of information: post-editing time, keystrokes, perceived effort. HTER is computed during the parsing.

```
<S producer="Cristina">That prompted Maduro to rule as a dictator; the
assembly has been reduced to an impotent NGO, stripped of its constitutional
powers.</S>
<MT producer="Cristina">Eso llevó a Maduro a gobernar como dictador; la
asamblea se ha reducido a una ONG impotente, despojada de sus poderes
constitucionales.</MT>
<annotations revisions="1">
  <annotation r="1">
    <PE producer="user.Cristina">Eso llevó a Maduro a gobernar como un
    dictador; la asamblea se ha reducido a una ONG impotente, despojada de
    sus poderes constitucionales.</PE>
    <assessment id="difficulty (optional)">
      <score>3. Difícil</score>
    </assessment>
    <indicator id="autoaccept" type="flag">false</indicator>
    <indicator id="unchanged" type="flag">false</indicator>
    <indicator elapsed=",0" id="assignment" length="135" offset="0" t0=",0"
    type="change">Eso llevó a Maduro a gobernar como dictador; la asamblea se
    ha reducido a una ONG impotente, despojada de sus poderes
    constitucionales.</indicator>
    <indicator elapsed=",0" id="insertion" length="3" offset="34" t0=",0" typ
    ="change"> un</indicator>
    <indicator id="editing" type="time">1m18s,968</indicator>
    <indicator id="assessing" type="time">23s,184</indicator>
    <indicator id="letter-keys" type="count">2</indicator>
    <indicator id="digit-keys" type="count">0</indicator>
    <indicator id="white-keys" type="count">1</indicator>
    <indicator id="symbol-keys" type="count">0</indicator>
    <indicator id="navigation-keys" type="count">0</indicator>
    <indicator id="erase-keys" type="count">0</indicator>
    <indicator id="copy-keys" type="count">0</indicator>
    <indicator id="cut-keys" type="count">0</indicator>
    <indicator id="paste-keys" type="count">0</indicator>
    <indicator id="do-keys" type="count">0</indicator>
    <events>
      <flow t="0">EDITING_START</flow>
      <change offset="0" t="67">
        <in>Eso llevó a Maduro a gobernar como dictador; la asamblea se ha
        reducido a una ONG impotente, despojada de sus poderes
        constitucionales.</in>
      </change>
      <cursor dot="0" mark="0" t="119"/>
      <cursor dot="34" mark="34" t="68327"/>
      <keystroke offset="34" t="75671"> </keystroke>
      <change offset="34" t="75672">
        <in> </in>
      </change>
      <cursor dot="35" mark="35" t="75672"/>
      <keystroke offset="35" t="76181">u</keystroke>
      <change offset="35" t="76181">
        <in>u</in>
      </change>
      <cursor dot="36" mark="36" t="76182"/>
      <keystroke offset="36" t="76639">n</keystroke>
      <change offset="36" t="76639">
        <in>n</in>
      </change>
      <cursor dot="37" mark="37" t="76640"/>
      <flow t="78968">EDITING_END</flow>
      <flow t="78968">ASSESSING_START</flow>
      <flow t="102152">ASSESSING_END</flow>
    </events>
```

Figure 4 - Screenshot of PET's raw results

Nevertheless, we realized that these files offered a lot of potential for finding out new information; for this reason, we wrote a computer script aimed at extracting more metrics. These were editing time, total pause time, total pause count, length of the initial pause, length of the final pause, length of pauses during editing and number of pauses during editing. Table 4 presents all the metrics grouped by effort aspect, along with brief descriptions of how they were computed and why it was relevant to obtain them.

Effort aspect	**Metric**	**Description**
Temporal	Total time	*Computed as*: The time spent working on a sentence, computed as the time elapsed since the translator opens the segment, until they close it. *Relevance*: Segments that take a long time to correct are assumed to be more difficult; either because there is a lot to correct, or because it takes time to find the right correction to perform.
	Editing time	*Computed as*: The total time minus the pause time. It is considered as the time spent typing and editing. *Relevance*: A high editing time means many things must be corrected; we expect it to correlate well with keystrokes.
Cognitive	Pause time	*Computed as*: Any lapse of time between keystrokes over a certain threshold was considered as pause time. The threshold was established at 0.3 seconds, following Lacruz et al. (2012), who determined this was the shortest possible time elapsed for a pause to be considered as such. *Relevance*: Pause time is assumed to be spent planning corrections or revising; long pause times point to high difficulty.
	Editing pause time	*Computed as*: The length of the pauses that take place between the first and last edits. *Relevance*: Long pauses between edits could point to indecisiveness or trying different options, which means the edit is not straightforward.
	Initial pause	*Computed as*: The length of the pause before the first edit, if there is one. This is assumed to be time spent reading and finding the error. *Relevance*: Difficult edits will take a longer initial pause to figure out how to solve them.
	Final pause	*Computed as*: The length of the pause after the final edit, if there is one. This is assumed to be re-reading, revision time. If no

		editing has been carried out during an annotation, the total time is considered as revision time. *Relevance*: Long revision times could mean the translator is deciding whether or not to keep an edit.
	Pause count	*Computed as*: The number of pauses over 0.3 seconds per segment. *Relevance*: We expect difficult segments to contain more pauses.
	Editing pause count	*Computed as*: The number of pauses that take place between the first and last edits. *Relevance*: A high concentration of pauses between edits may mean the segment is difficult and the translators are reconsidering as they type.
	Perceived effort	*Computed as*: After closing each segment, translators were asked to rate the difficulty of the segment on a 1 to 3 scale, with 1 being easy. The perceived effort time was not considered in the total time. *Relevance*: We assume that if all translators point to one segment being easy or difficult, it will be so.
Technical	Keystrokes	*Computed as*: The number of keyboard keys pressed. These include digit, symbol and letter keys; copy, cut and paste keys; navigation keys; any action keys (Enter, delete, shift, etc.) and the space bar. *Relevance*: From a technical point of view, difficult edits could both contain more total edits and more rewriting.
	HTER	*Computed as*: The edit distance between the machine translated output and the final human post-edited version. *Relevance*: Easier edits should present lower HTER; changing the ending of a word will take fewer total edits than rewriting a full idiom.

Table 4 - Description of all metrics by aspect, how they were computed and their relevance

3.5 Task

The experiment we designed consisted in the following: a dataset of English segments with specific characteristics (length, topic, etc.) were passed through Google's

NMT system and translated to Spanish. From the automatic translations a subset of 60 sentences were selected, each containing one instance of the same type of error.

These were given to a group of 7 participants, who were asked to post-edit each of them. The translators participating in the experiment received an explanatory mail containing several documents. The first one was a general description of the task with instructions on what was expected from them; these included directions such as the fact that each sentence contained only one error, that they must post-edit errors but not style, and that they should close the segments down when taking a break. Additionally, some mock-up examples where presented, showcasing how to post-edit them. The second document was a step-by-step explanation on how to install PET depending of the operating system of the computer they would be working on. The third file was a PowerPoint presentation with a screen captions of short practice test, so that the translators would get familiarized with the working environment by following it on the side. This practice task was meant to lessen the impact of the novelty of using a new tool on the results. Finally, the last document was a timeline of the Venezuelan crisis; it contained a summary of the main events and characters involved, so that the translators would not have to use time looking for information to understand the contents of the text, if they were not familiar with the situation. These documents can be found as annexes 2, 3, 4 and 5.

4 Results

This chapter presents and discusses the results of the experiment. The first section addresses the analysis of the post-edited sentences, and how some had to be discarded to build our final working dataset. The second section consists of a comparison of the measuring techniques and the correlations between them. These results help us explore our first research question: how do different metrics relate to each other? We will analyse the results of comparing all metrics, comparing techniques from different or the same aspects to find relevant relations or differences between their results. The third and final section if meant to address our other research question, which delves into the effect that different types of errors may have on the difficulty of the sentences that contain them. With this objective, we will group all errors by metric and compare them, analysing any patterns that emerge.

4.1 Dataset

The test suite that the translators worked on consisted of 60 sentences, 10 for each type of error, which contained one error each and were between 20 and 25 words long. After carrying out the experiment and analysing the sentences from each translator, a big percentage of them had to be eliminated due to various reasons, as you can see in Table 5.

From the original 420 sentences (60 for each of the 7 participants), 61 had to be discarded because the translators had corrected more items that they were supposed to, usually style, word order or punctuation. For example, Translator 3 changed all the quotation marks from “” to «», which implied that around 20 sentences were no longer valid. While the instructions provided for the task were insistent on the fact that each sentence only contained one error, maybe more emphasis should have been made on the fact that any additional corrections would make the sentence be discarded. Additionally, 34 sentences had to be further discarded because no corrections were performed, while 62 were excluded because the translators had corrected something other than the intended error. Upon analysing these sentences, some seem to repeat among translators, which potentially means that the error was not as clear as we assumed when choosing the sentences or could be construed to be a style error. In total, 158 had to be eliminated, leaving 262 sentences left to analyse. These were automatically annotated with different

metrics thanks to the PET interface, and several other metrics were inferred from the results.

	Too many corrections performed	No corrections performed	Wrong correction performed	Intended correction performed
Translator 1	2	1	12	45
Translator 2	2	3	10	45
Translator 3	29	4	3	24
Translator 4	4	5	11	40
Translator 5	16	6	10	28
Translator 6	8	4	6	41
Translator 7	0	11	8	41
Total	61	34	62	264

Table 5 - Breakdown of how many sentences had to be discarded for each translator and the reasons for it

4.2 Distributions and correlations between metrics, by error

Our first research question concerns the relations and differences between metrics. Since different measuring techniques are aimed at measuring different aspects of effort, we expect to see techniques correlate well within their effort aspect group and have worse correlations with metrics from other aspects.

First, we present a correlation matrix of all the metrics, which can be seen on Figure 5. The correlation matrix displays correlations between measuring techniques, with the darker results meaning low to negative correlations, and the lighter results meaning high correlations. The colour map ranges from -0.25 to 1.

The correlation matrix represents correlations between results for all sentences. There are general patterns which catch the eye instantly; for example, total time and pause time correlate very well to one another and have good correlations with other pause-related metrics such as editing pause time, first pause and last pause. It is interesting to

remark that the correlations with editing time are nevertheless quite low; since all these metrics are essentially fragments of total time, they could be expected to be good.

Editing time does have quite good correlations with keystrokes, pause count and editing pause count. This stands to reason, as editing time essentially represents keystroke pressing time, and pauses only happen in between keystrokes; their total number should be similar. These four metrics correlate very poorly with all other measuring techniques.

HTER displays negative correlation values with all other metrics, and perceived effort does not correlate well with any other effort measuring techniques, either. It is interesting to remark that HTER has such low results; the fact that it correlates so poorly with other metrics, even with keystrokes, could mean that it is not generally representative of post-editing effort, and it should be used with caution. A similar case could be made of perceived effort, whose results do not have good correlations with any other metrics, including those supposed to also be telling of cognitive effort.

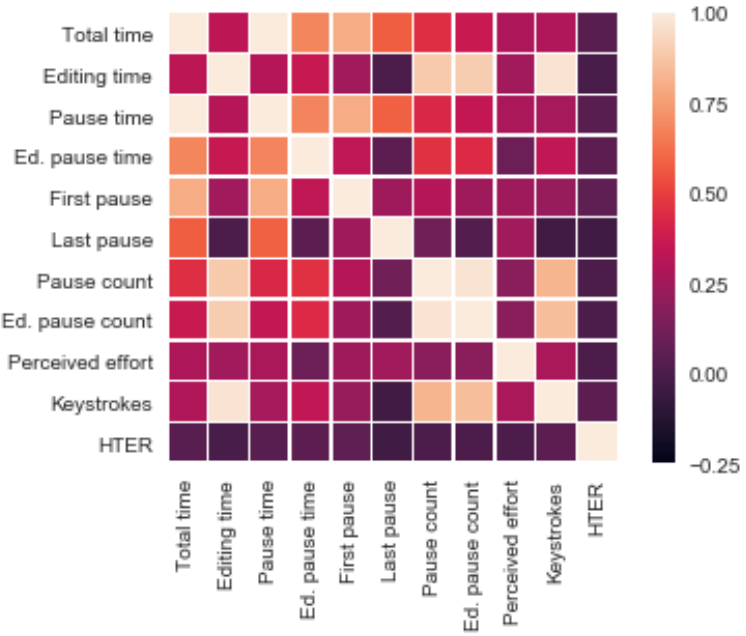


Figure 5 - Correlation matrix between all metrics

Seeing these correlation results, we wondered whether there would be significant differences if the results were split by error. The resulting correlations can be seen on Figures 6 to 11. In general, we can observe similar patterns, with slight differences for some of them. For example, on Figure 8 correlations between total time, editing time and keystrokes are negative; this could be caused by the fact that finding a word to remove may take some time (reading the source and machine translated segments) while actually removing it does not take a lot of typing. In general, N/G agreement (Figure 6) and one-

word mistranslation (Figure 10) present the best correlations. This may mean that the different aspects of effort are more balanced on these instances, taking a similar amount of different types of effort.

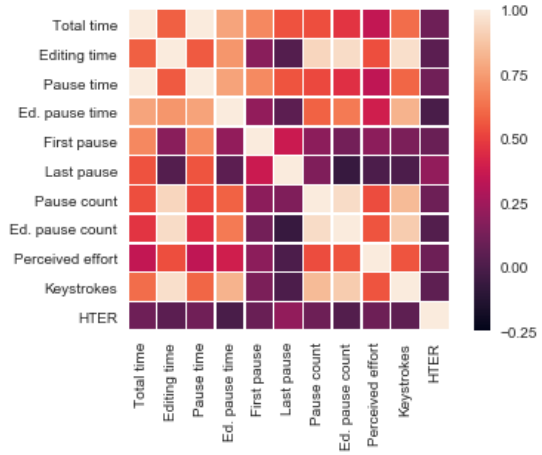


Figure 6 - Corr. matrix for N/G agreement

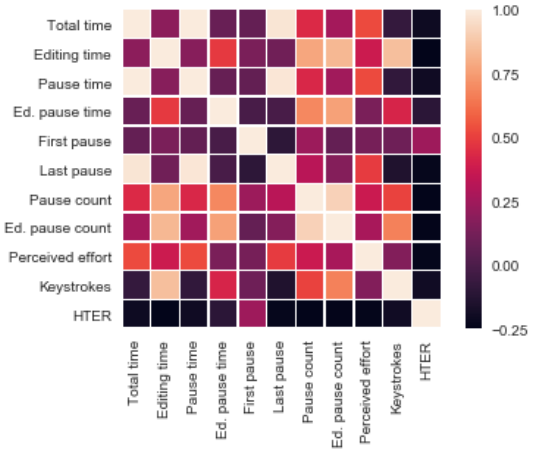


Figure 7 - Corr. matrix for T/A agreement

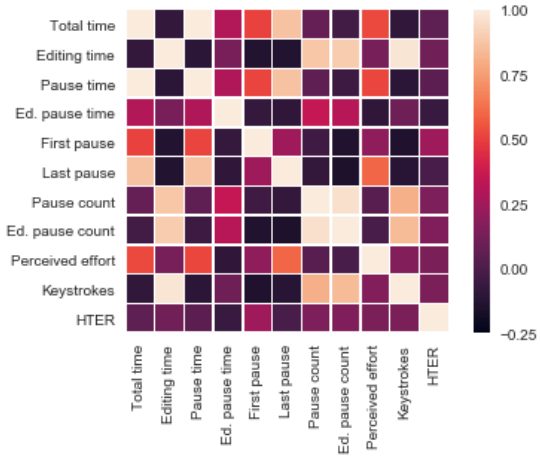


Figure 8 - Corr. matrix for extra word

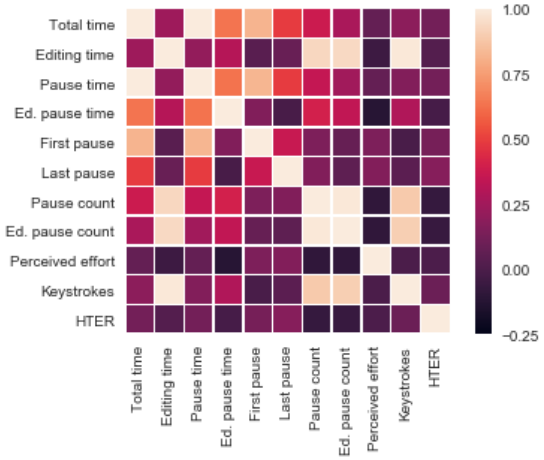


Figure 9 - Corr. matrix for missing word

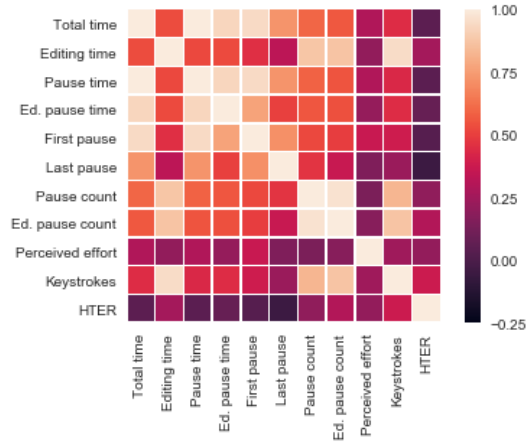


Figure 10 - Corr. matrix for mistranslation 1w

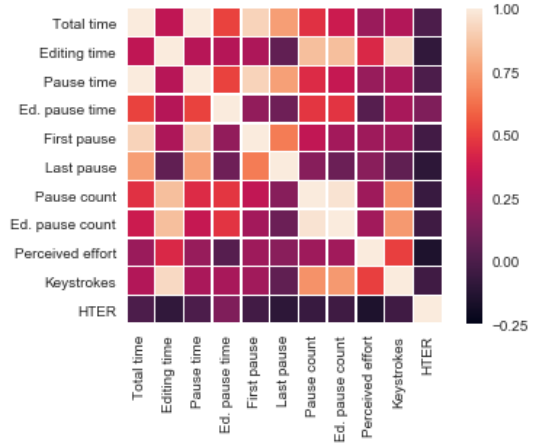


Figure 11 - Corr. matrix for mistranslation 2+w

In order to analyse whether the results of the metrics are similar or different, we have also drawn a box plot (see Figure 12). Box plots show the distribution of the results: the main box represents 50% of the results, while the "whiskers" represent the remaining two quartiles; the darker band within the box marks the median, or limit between the second and third quartiles; the "X" mark corresponds to the average; finally, any individual points outside the main plot are outliers. Narrow box plots mean that all the results are very close together, that is, there is agreement amongst the results; wide box plots mean that the results are scattered and there is disagreement.

In order to be able to compare the results from all techniques, which are not necessarily in the same scale, we have normalized the results to a 0 to 1 range, represented in the y axis of the box plots. This transformation was applied by computing each metric's absolute lowest and highest results amongst all sentences, establishing these as 0 and 1 respectively, and distributing the rest in between.

This box plot shows big differences between the different metrics. HTER obtains the highest (i.e. worst) results on average. The remaining measuring techniques have at least 75% of their results within the 0 to 0.2 range: perceived effort is the worse metric from this group, closely followed by keystrokes and editing time; next are pause count and editing pause count, then total time, pause time and first pause. The worst results are those of editing pause time and last pause.

There are several trends that we can observe in these results. First, pairs of metrics with similar distribution and averages also correlated well on the correlation matrix. Second, technical effort is on average higher than the indicators of other kinds of efforts. It is also interesting to observe that on average the initial pause is longer than the final

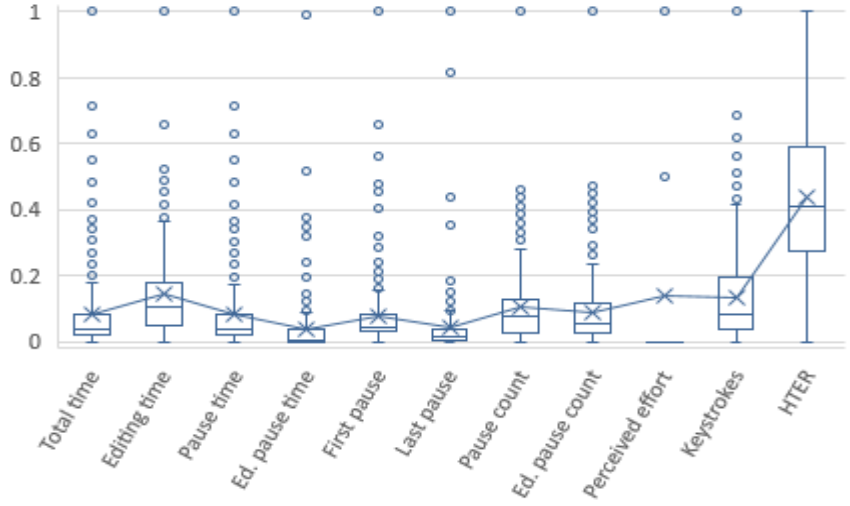


Figure 12 – Distribution graph comparing all measuring techniques

pause, which means that more time is taken to assess the error and decide how to correct it than in revision afterwards.

In general, both ways to compare the metrics lead to similar conclusions: none of these metrics are able to fully capture post-editing effort. Carrying out research, such as CE models (Specia et al., 2010) or productivity studies (Plitt et al., 2010; Parra Escartín et al., 2015), based on just one or two of these metrics, risks reaching results that misrepresent the total post-editing effort. Future work in this area could focus on testing combinations of metrics to best detect increased effort of any kind.

4.3 Distributions of errors by tool

The second research question of this paper aims at finding the effect that different types of errors have on effort, and whether metrics can detect these differences. Since errors are assumed to have a big impact on the difficulty of the sentence, we assume the results will show noticeable differences between them. We also assume Temnikova's (2009) error ranking will be confirmed by the results.

This section presents a series of box plots which compare the results for all errors, as per the measurements of one metric at a time. These have been grouped by effort aspects: the first subsection, for temporal effort, shows total time and editing time; the second one, for cognitive effort, displays the box plots for pause time, editing pause time, first pause, last pause, pause count, editing pause count and perceived effort; finally, the third subsection, for technical effort, presents keystrokes and HTER. A fourth subsection analyses and discusses the general patterns that were found.

The divisions by effort aspect will allow us to compare the metrics which should show the most similar results and check whether there are similar patterns in the ways errors affect them.

4.3.1 Temporal effort

The two box plots in this section represent the total time and editing time (which results from subtracting pause time to the total time). The units in the y axes of these plots are milliseconds, and they have been normalized by the number of words in the post-edited versions of the sentences. We decided to normalize the time metrics because not all sentences have the same number of words, so this transformation was necessary if we were to compare them. Other metrics did not require normalization, as they referred

directly to the correction of the errors and were not affected by the length of the sentence. In these plots, the x axes show the different errors, ordered following Temnikova's (2009) ranking: N/G agreement, T/A agreement, one-word mistranslation, extra word, missing word and mistranslation of several words.

The first thing we observe on Figures 13 and 14 is the distribution of time; most of it is pause time, with very short time dedicated to editing. Aside from this, we remark that the results for different errors remain within a similar range. In general, Mistr1w results are higher than those of extra word, and T/A agreement is higher than N/G agreement. These patterns, which repeat over most metrics, challenge Temnikova's ranking by subverting the difficulty order of errors.

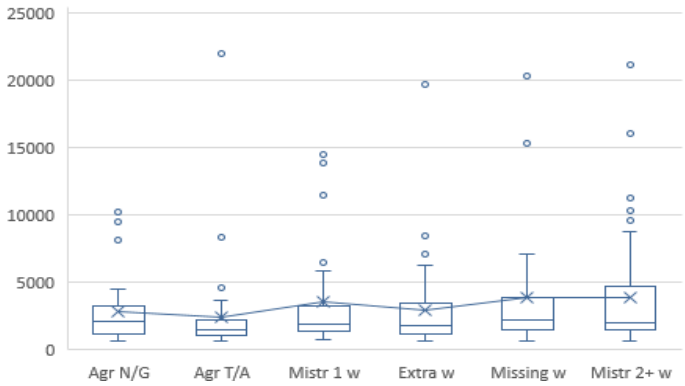


Figure 13 - Distribution graph for total time

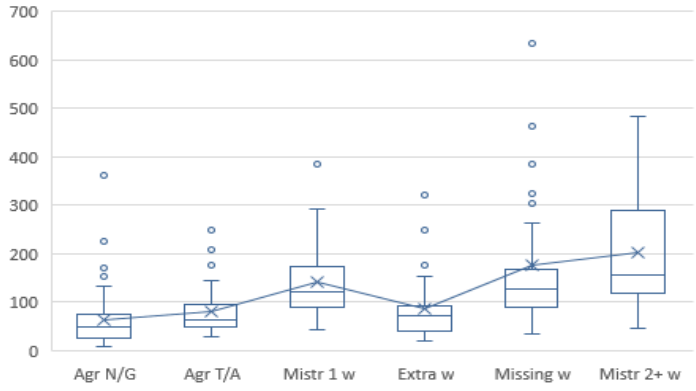


Figure 14 - Distribution graph for editing time

4.3.2 Cognitive effort

This section analyses all the metrics used to measure cognitive effort: pause times, pause counts and perceived effort; the box plots can be seen on Figures 15 to 18.

As we have mentioned previously, pauses are considered to be "thinking" time; that is, time spent reflecting on the correction of an error. A closer look at the distribution of pause time shows that around a fourth of it happens between the first and last edits (Figure 16), which seems to imply that once the translator has decided on an edit, they carry it out without stopping between keystrokes. The rest of the pause time is distributed between the first and last pauses (see Figures 17 and 18), with slightly more on the first pause. It is possible that the first pauses were longer on average, as seen in the previous section, but the fact that revision rounds (where no editing happens) were added to the last pause may have evened the results out.

It is also interesting to note that the first pause is longer for the two mistranslation classes. While it seems plausible that the edits where words must be substituted take the longest thinking-and-deciding time, it again contradicts Temnikova's difficulty ranking. Regarding the last pause, or revision time, N/G agreement gets the worst results.

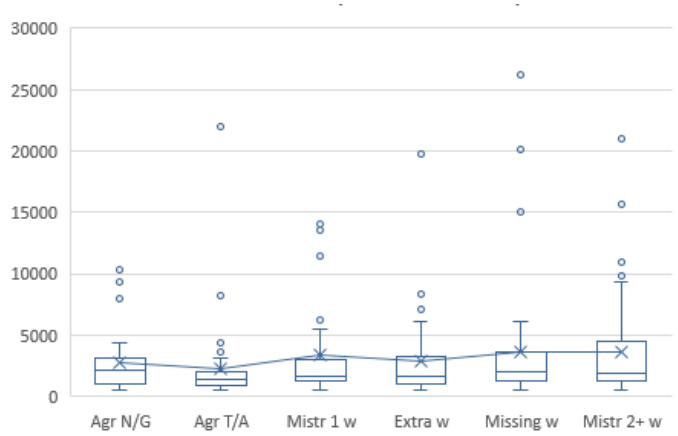


Figure 15 - Distribution graph for pause time

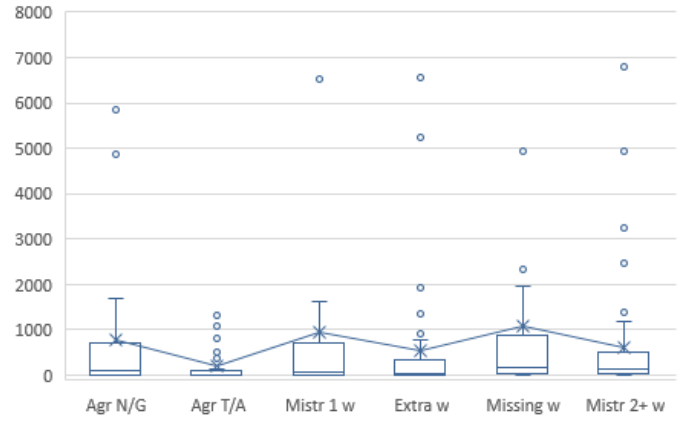


Figure 16 - Distribution graph for editing pause time

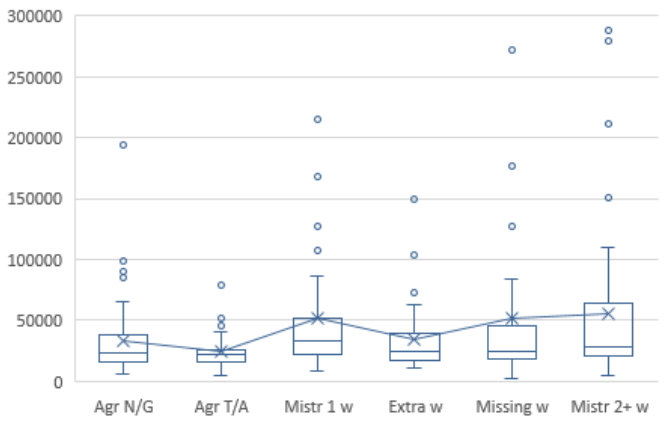


Figure 17 - Distribution graph for first pause

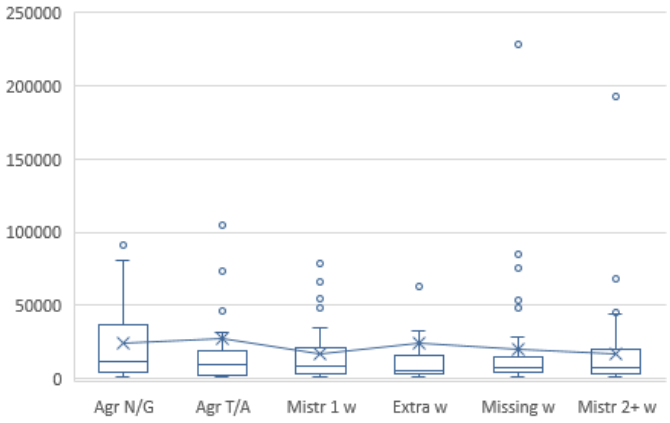


Figure 18 - Distribution graph for last pause

The number of pauses in a sentence may also be indicative of cognitive effort, since stopping or slowing down between keystrokes may denote indecisiveness. The pause counts displayed on Figures 19 and 20 show slightly bigger discrepancies between errors than the pause durations; both mistranslation types and missing word get the worst results, followed by extra word, N/G agreement and T/A agreement. On average, correcting a Mistr 2+w implies 4 more total pauses than T/A agreement (see Figure 19). The number of pauses between the first and last edits (see Figure 20) follow the same pattern, and constitute more than half of the total pauses. This could imply that not all translators have rounds in which they revise without editing.

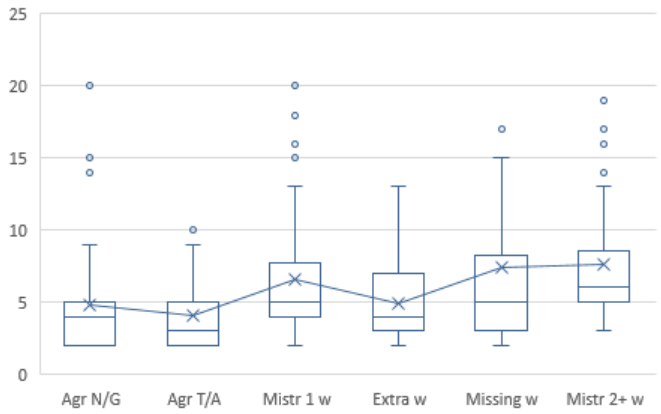


Figure 19 - Distribution graph for pause count

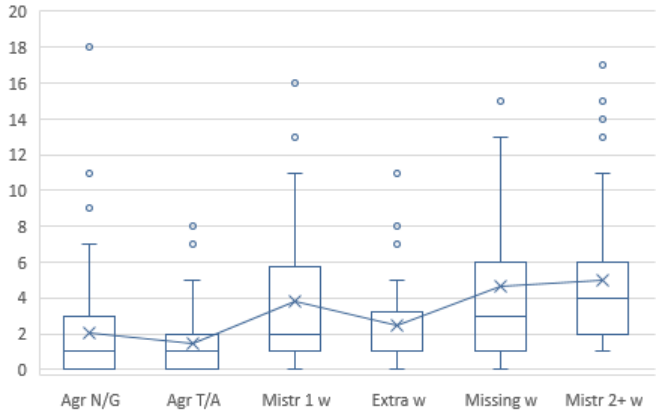


Figure 20 - Distribution graph for editing pause count

Finally, the third method used to measure cognitive effort consisted in asking the translators to rate the difficulty of the sentence, with 1 being the best score and 3 being the worst. The perceived effort scores given by the translators (see Figure 21) displayed similar patterns to those of previous metrics. Average perceived effort of the difficulty ranged between 1.2 and 1.4, meaning that most sentences were considered between easy and medium. It is also interesting to remark that in all cases excepting T/A agreement there were instances of translators choosing all three different possible scores; this points to low inter annotator agreement.

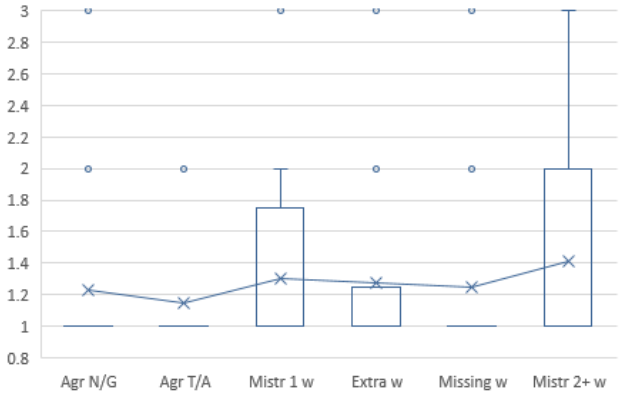


Figure 21 - Distribution graph for perceived effort

4.3.3 Technical effort

Technical effort was measured through two different techniques: keystroke logging and HTER. Keystrokes represent the number of times a translator has pressed a key on the computer, either to type or to move around the text. HTER, on its side, is the minimum edit distance between the MT output and its post-edited version; that is, the

number of edit operations that the translator performed, oblivious to the edits the translator may have tried out and discarded before settling on the final one.

Figures 22 and 23 shows the results for both metrics, which are rather different. Keystrokes displays a similar pattern to the other tools, but the differences are larger: the two mistranslation types and missing word imply, on average, 10 more keystrokes than extra word or agreement issues. Considering that the first implies writing entire words down, while the latter consist on either correcting word endings or deleting, these differences are logical.

On the other hand, HTER's results are quite opposite to the general pattern of the other metrics. N/G agreement gets the worst, highest score, while Mistr2+w receives the lowest. In general, the distributions are quite similar, with averages ranging between 25 and 28. These results seem to show, again, that HTER should be used with caution and full acknowledgement of its shortcomings.

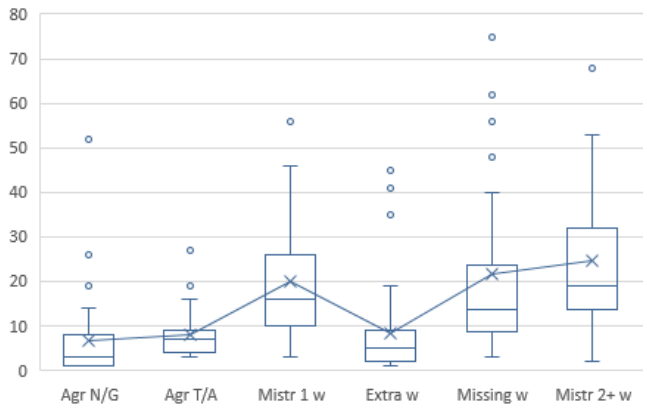


Figure 22 - Distribution graph for keystrokes

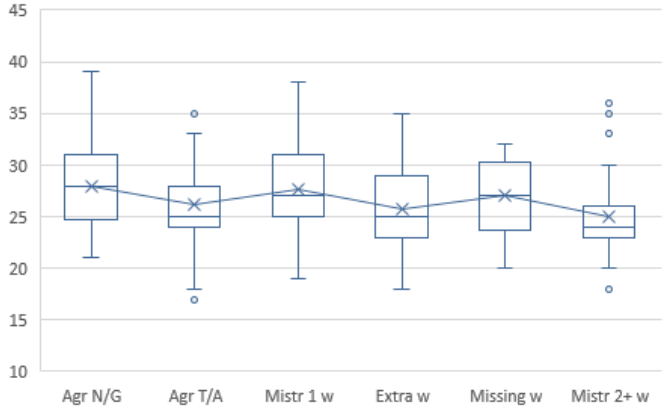


Figure 23 - Distribution graph for HTER

4.3.4 Discussion of results

In most tools, the results for simple mistranslation are worse (i.e. higher) than for extra or missing word, which challenges Temnikova's ranking. Moreover, T/A agreement obtains better (i.e. lower) results than N/G agreement. There appears to be a pattern in our results; repeating this experiment with more participants, a larger test suite and maybe even more error types, could help further confirm these trends and establish a better error ranking.

It is nevertheless important to acknowledge that the differences between errors are very limited. While research has focused on linking the presence of errors to increased difficulty and effort, it seems that once the error has been isolated (that is, controlled other features of the sentence to lessen their repercussions), its effects on effort are not as great

as previously assumed. It is possible that the presence of other words in the sentence, the sentence length or the potential combined effect of different errors within the sentence has been underestimated. These are new research options that could be very interesting for this field.

5 Conclusions and future work

This paper aimed to take a closer look into the real effect that different types of errors have on post-editing effort, and the reliability of the metrics currently used in the translation industry to measure it.

In order to do so, we designed an experiment in which a test suite consistent of 60 sentences (one per type of error) was given to a group of 7 translators to post-edit using the software PET. The features to be controlled were chosen after careful consideration of our objectives and previous research shortcomings. We aimed to isolate the errors as much as possible by controlling every other feature within our possibilities, while still having errors that the MT system had created naturally. Moreover, we made sure for every error to be equally represented within the dataset even if they did not naturally occur with the same frequency.

The post-edited sentences were reviewed and almost half had to be discarded for various reasons, such as the translators correcting too much, not correcting anything at all, or performing the wrong corrections. Going forward with similar experiment designs, it would be advisable to draft even clearer instructions to make sure that translators perfectly understand what is and is not expected from them on the task. While we were trying to measure post-editing effort, this was not a typical post-editing task and it could have been made clearer to translators that leaving sentences untouched or correcting the style of the sentences would result in those sentences being discarded, instead of just telling them that these things should not be done without making it obvious why.

The post-editing tool PET, which the translators used to perform the task, collected data about how long each segment was open for, when and what keys the translators had used, how difficult they perceived the segments to be, and the total amount of edits (HTER). These results were used to infer other metrics, such as pauses times and pause counts. All the metrics were then analysed to see whether they offered answers to our research questions.

The first question consisted in studying the correlations between metrics from different effort aspects and analysing whether their results were or were not similar.

We had assumed that results for different aspects of effort would correlate well with each other but be different from other effort aspects. This was found to be true for some aspects such pause metrics (cognitive effort), which had good correlations on

accounts of being based off each other. Perceived effort, however, did not correlate well with pauses, or any other aspects. This may point out to the fact that asking people to rate the difficulty of sentences is not a good strategy to obtain useful, reliable results. Perceived effort is always relative to the dataset, so before seeing it translators may have biased expectations of what "easy" and "difficult" will look like, which may have affected the results.

Keystrokes and HTER (technical effort) returned very different results which did not correlate at all; again, because keystroke tracking tallies all keys used, while HTER only considers final edits, this was to be expected. HTER is a very common tool in the industry, used as the basis for quality estimation models. The fact that HTER correlates so poorly, even negatively, with all other tools should give pause to anyone who wants to use this metric as the sole source of post-editing effort measurements. Regarding keystrokes, they did correlate quite well with editing time, pause count and editing pause count because all of them rely heavily on the same principle of editing being key-pressing time, and pauses happening between edits.

The second research question explored the effect that different types of errors have on effort, and whether metrics could detect any differences. In this case, we assumed that differences between errors would be quite clear and follow Temnikova's effort ranking. Nevertheless, our results challenge both assumptions. First, the differences between errors are quite faint, with results ranging around similar values. While past research has concluded that the presence of certain errors greatly affects the difficulty of the sentence, they often did not isolate the errors as much as we have done in our experiment, meaning that other sentence features could be blowing the results up. Once these features, like sentence length, error frequency and error combination are controlled, the influence of different types of errors seems to be subtler.

Concerning Temnikova's (2009) ranking, even with the minimal differences in the results some patterns emerge, that repeat themselves over most tools. N/G agreement obtains worse results than T/A agreement, and 1-word mistranslation appears to be more difficult than extra or missing word. According to these results, Temnikova's ranking could need some revising.

Other aspects of our experiment also merit further comments; for example, our test suite. Having used them for this experiment, a word of advice is in order: test suites are difficult and time consuming to obtain; isolating errors proved to be more challenging than previously expected, especially when the MT system is so performant that many

sentences do not contain errors at all. Moreover, test suites have limited reusability potential since their characteristics are very specific and often fitted to a certain task. Finally, it is worth mentioning that when we analysed the post-edited sentences, we realized that many translators were consistently failing on the same segments, which means that they were not as obvious as we might have thought when choosing them. In future research, it would be desirable to do a test run with a person who has never seen the sentences before carrying out the actual experiment to get external feedback about the test suite; this would help avoid using sentences where the error is not evident.

In conclusion, while test suites have proven to be a very useful and generally untapped resource to boost future research, they should be used with caution and full knowledge of the challenges they entail.

This experiment had many limitations due to its humble scope. Future research could focus on drafting similar experiments with more participants, new metrics, or a larger test suite in order to confirm the patterns that have been found. Alternatively, the number of errors could be increased, including categories such as punctuation or word order, considered to be the most cognitively demanding by Temnikova's ranking.

Another direction could be replicating the experiment using eye-trackers; if the results were in line with those obtained through other metrics, it could mean that using such complicated techniques are not compulsory to study most post-editing problems.

This experiment could also be repeated including sentences to be translated from scratch; this would allow to establish productivity thresholds for each metric. It would also be interesting to delve deeper into which aspects each metric measures best, and what errors it is the most sensitive to, and establish a method that combines metrics in an optimal way to capture and predict real post-editing effort.

6 Bibliography

Alves, F., Szpak, K., Luiz Gonçalve, J., Sekino, K., Aquino, M., Castro, R., Koglin, A., Fonseca, N. & Mesa-Lao, B. (2016) Investigating cognitive effort in post-editing: A relevance-theoretical approach. In Hansen-Schirra, S. & Grucza, S. (eds.), *Eye-tracking and Applied Linguistics* (pp. 109–142). Berlin: Language Science Press. doi: 10.17169/langsci.b108.296

Aziz, W., Koponen, M., & Specia, L. (2014). Sub-sentence Level Analysis of Machine Translation Post-Editing Effort. In S. O'Brien, L. Balling, M. Simard & L. Specia (eds). *Post-editing of Machine Translation: Processes and Applications* (pp. 170–199). Newcastle upon Tyne: Cambridge Scholars Publishing.

Aziz, W., Mitkov, R., & Specia, L. (2013). Ranking Machine Translation Systems via Post-Editing. In *Text, Speech, and Dialogue (TSD)* (pp. 410–418). Pilsen, Czech Republic: Springer-Verlag Berlin Heidelberg.

Baddeley, A.D., & Hitch, G. (1974). Working memory. In G.H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (8): 47-89. New York: Academic Press.

Burchardt, A., Harris, K., Rehm, G. & Uszkoreit, H. (2016). Towards a Systematic and Human-Informed Paradigm for High-Quality Machine Translation. In *Proceedings of the LREC 2016 Workshop "Translation Evaluation – From Fragmented Tools and Data Sets to an Integrated Ecosystem"*. Portoroz, Slovenia.

Carl, M., Dragsted, B., Elming, J., Hardt, D. & Jakobsen, A. L. (2011) The Process of Post-Editing: A Pilot Study. *Copenhagen Studies in Languages, 41*, 131-142.

Carl, M. (2012). Translog - II: a Program for Recording User Activity Data for Empirical Reading and Writing Research, In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).

da Silva, I., Alves, F., Schmaltz, M., Pagano, A., Wong, D., Chao, L., Leal, A., Quaresma, P., & Eduardo da Silva, G. (2017) Translation, Post-Editing and Directionality: A Study of Effort in the Chinese-Portuguese Language Pair. In Jakobsen, A. L. & Mesa-Lao, B. (eds). *Translation in Transition: Between cognition and Technology* (pp. 91 - 117). Amsterdam: Benjamins Translation Library.

Daems, J., Vandepitte, S., Hartsuiker, R. & Macken, L. (2017). Identifying the Machine Translation Error Types with the Greatest Impact on Post-editing Effort. *Frontiers in Psychology*, 8, 1282.

Doherty, S., O'Brien, S., and Carl, M. (2010). Eye Tracking as an Automatic MT Evaluation Technique. *Machine Translation*, 24(1):1–13.

Federico, M., Cattelan, A., & Trombetti, M. (2012). Measuring user productivity in machine translation enhanced computer assisted translation. In *Proceedings of Conference of the Association for Machine Translation in the Americas (AMTA)*. San Diego, CA, USA.

Gaspari, F., Almaghout, H. & Doherty, S. (2015). A survey of machine translation competences: Insights for translation technology educators and practitioners. *Perspectives: Studies in Translatology,* 23(3), 1–26.

Guerberof Arenas, A., Depraetere, H. & O'Brien, S. (2013). What we know and what we would like to know about post-editing. *Revista Tradumàtica: tecnologies de la traducció*. 211-218

Guerberof Arenas, A. (2013). What do professional translators think about post-editing? *Journal of Specialised Translation*, *19*.

Guerberof, A. (2009). Productivity and quality in MT post-editing. In *Proceedings of the MT Summit XII-Workshop: Beyond Translation Memories: New Tools for Translators MT*. Ottawa: Association for Machine Translation in the Americas.

Guillou, L. & Hardmeier, C. (2016). PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)* (pp. 636-643). Portoroz, Slovenia.

Harley, T.A., (2008). The Psychology of the Language: from data to theory. Psychology Press: Hove and New York.

Herbig, N., Pal, S., Vela, M., Krüger, A. & van Genabith, J. (2019) Multi-modal indicators for estimating perceived cognitive load in post-editing of machine translation. *Machine Translation*, 1-25.

Koponen, M. (2016a). Is Machine Translation Post-Editing Worth the Effort? A Survey of Research into Post-editing and Effort. *Journal of Specialised Translation*, *25*, 131-148.

Koponen, M. (2016b). Machine Translation Post-editing and Effort: Empirical Studies on the Post-editing Process. Helsinki: University of Helsinki.

Koponen, M. (2013). This translation is not too bad: An analysis of post-editor choices in a machine translation post-editing task. In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice* (pp. 1-9). Nice, France.

Koponen, M. (2012). Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the Seventh Workshop on Statistical Machine Translation* (pp. 181-190). Montreal, Canada.

Koponen, M., & Salmi, L. (2015). On the correctness of machine translation: A machine translation post-editing task. *Journal of Specialised Translation*, *23*, 118-136.

Koponen, M., Aziz, W., Ramos, L., & Specia, L. (2012). Post-editing Time as a Measure of Cognitive Effort. In *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice*. San Diego, United States.

Krings, H. (2001): Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes (Geoffrey S. Koby, ed.). The Kent State University Press, Kent, Ohio & London.

Lacruz, I., Shreve, G., & Angelone, E. (2012). Average Pause Ratio as an Indicator of Cognitive Effort in Post-editing: A Case Study. In *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice*. San Diego, CA, USA.

Lacruz, I. & Shreve, G. (2014a). Pauses and Cognitive Effort in Post-Editing. In O'Brien, S., Winther Balling, L., Carl, M., Simard, M. & Specia, L. (Eds.), *Post-Editing of Machine Translation: Processes and Applications* (pp. 244–272). Cambridge: Cambridge Scholars Publishing

Lacruz, I. & Shreve, G. (2014b). Translation as a higher order cognitive process. In Porter, C. & Bermann, S. (Eds.). *A companion to translation studies* (pp. 107-118). Blackwell Companions to Literature and Culture. Malden, MA, USA: John Wiley & Sons.

Larigauderie P, Gaonac'h D. & Lacroix N. (1998). Working memory and error detection in texts: what are the roles of the central executive and the phonological loop? *Applied Cognitive Psychology*, *12*: 505-527

Mesa-Lao, B. (2013). Eye-Tracking Post-editing Behaviour in an Interactive Translation Prediction Environment. In *Proceedings of the 17th European Conference on Eye Movements*. Lund, Sweden.

Moorkens, J. (2018). Eye tracking as a measure of cognitive effort for post-editing of machine translation. In Walker, C. & Federici, F. (eds). *Eye Tracking and Multidisciplinary Studies on Translation* (pp. 55-70). John Benjamins.

Moorkens, J., O'Brien, S., da Silva, I., de Lima Fonseca, N. & Alves, F. (2015) Correlation of perceived post-editing effort with measurements of actual effort. *Machine Translation, 29*(3-4), 267-284.

O'Brien, S. (2011). Towards predicting post-editing productivity. *Machine Translation*, *25*(3), 197-215. doi: 10.1007/s10590-011-9096-7

O'Brien, S. (2006a). Eye-tracking and translation memory matches. *Perspectives: Studies in Translatology,* 14(3), 185–205.

O'Brien, S. (2006b). Pauses as Indicators of Cognitive Effort in Post-editing Machine Translation Output. *Across Languages and Cultures*, 7, 1–21.

O'Brien, S. (2005). Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Translatability. *Machine Translation*, 19(1), 37–58.

Parra Escartín, C. & Acedillo, M. (2015). Living on the edge: productivity gain thresholds in machine translation evaluation metrics. In *Proceedings of 4th Workshop on Post-Editing Technology and Practice (WPTP4)* (pp. 46-56). Miami, FL, US.

Plitt, M. & Masselot, F. (2010). A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*, 93, 7-16

Popovic, M., Lommel, A., Burchardt, A., Avramidis, E., & Uszkoreit, H. (2014). Relations between different types of post-editing operations, cognitive effort and temporal effort. In *Proceedings of the 17th annual conference of the European Association for Machine Translation, EAMT 2014* (pp. 191–198). Dubrovnik, Croatia.

Popovic, M. (2011) Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 96, 59-68

Schaeffer, M., Nitzke, J., Tardel, A., Oster, K., Gutermuth, S., & Hansen-Schirra, S. (2019). Eye-tracking revision processes of translation students and professional translators. *Perspectives: Studies in Translatology*, 1–15. doi: 10.1080/0907676x.2019.1597138

Snover, M., Madnani, N., Dorr, B., & Schwartz, R. (2009). TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, *23*(2-3), 117-127. doi: 10.1007/s10590-009-9062-9

Snover, M., Dorr, B., Madnani, N., & Schwartz, R. (2009). Exploring Different Human Judgments with a Tunable MT Metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation* (pp. 259–268). Athens, Greece: Association for Computational Linguistics.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas* (pp. 223-231).

Specia, L., & Farzindar, A. (2010). Estimating Machine Translation Post-Editing Effort with HTER. In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEG 10)* (pp. 33-41). Denver, CO.

Specia, L. (2010). Exploiting Objective Annotation for Measuring Translation Post-editing Effort. In *Proceedings of the 15th Conference of the European Association for Machine Translation* (pp. 73-80). Leuven, Belgium

Temnikova, I. (2009). Cognitive Evaluation Approach for a Controlled Language Post--Editing Experiment. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)* (pp. 3485-3490). Valletta, Malta.

Vilar, D., Xu, J., D'Haro, L. & Ney, H. (2006). Error Analysis of Statistical Machine Translation Output. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy.

7 Annexes

7.1 Annex 1: Test suite

Source segment	MT output	Error
Popular discontent at austerity and corruption saw the election in 1998 of Chávez, a charismatic army officer who had led a failed coup.	El descontento popular por la austeridad y la corrupción vio la elección en 1998 de Chávez, un oficial del ejército carismático que había dirigido un golpe de estado fallido.	Mistranslation 1 word
In his first post-election speech, the comandante promised he would not rest while there were still children in the streets and families going hungry.	En su primer discurso posterior a las elecciones, el comandante prometió que no descansaría mientras todavía hubiera niños en las calles y que las familias pasaran hambre.	Extra word
“Chávez would go into the barrios, into any house, and whichever house Chávez went into they’d say: ‘Have a little cup of coffee, my presidente,’”.	"Chávez entraría a los barrios, a cualquier casa, y en cualquier casa a la que entrara, dirían: 'Tome una taza de café, mi presidente'".	Agreement (TENSE, ASPECT)
However, by the time he died in 2013 he had bankrupted his country, running up debt and strangling the private sector with numerous controls.	Sin embargo, para cuando murió en 2013, había llevado a la bancarrota a su país, acumulando deudas y estranguló al sector privado con numerosos controles.	Agreement (TENSE, ASPECT)
Maduro's authoritarian rule, enforced by violence, has exacerbated social divisions, undermined democratic institutions and free media, caused millions to flee abroad and alienated neighbouring countries.	El gobierno autoritario de Maduro, impuesto por la violencia, ha exacerbado las divisiones sociales, ha socavado las instituciones democráticas y los medios de comunicación libres, ha hecho que millones huyan al extranjero y han alejado a los países vecinos.	Agreement (NUMBER, GENDER)
He presides over a broken economy and a country from which around 10% of the population have fled in the past three years.	Preside sobre una economía rota y un país del cual alrededor del 10% de la población ha huído en los últimos tres años.	Extra word
The United Nations estimates 3 million people have fled the country since 2015 to escape chronic food shortages, crumbling healthcare and an economy in freefall.	Las Naciones Unidas estiman que 3 millones de personas han huido del país desde 2015 para escapar de la escasez crónica de alimentos, derrumbando la atención médica y una economía en caída libre.	Mistranslation 1 word
The US, Canada, most Latin American nations and many European states labelled Nicolas Maduro’s second-term election win last May fraudulent.	EE. UU., Canadá, la mayoría de las naciones latinoamericanas y muchos estados europeos calificaron como fraudulentas la victoria electoral de Nicolás Maduro en el segundo mandato en mayo pasado.	Agreement (NUMBER, GENDER)

That prompted Maduro to rule as a dictator; the assembly has been reduced to an impotent NGO, stripped of its constitutional powers.	Eso llevó a Maduro a gobernar como dictador; la asamblea se ha reducido a una ONG impotente, despojada de sus poderes constitucionales.	Missing word
In an interview this week with El País, Chávez's former oil minister, Rafael Ramírez, said that Maduro was "out of time".	En una entrevista esta semana con El País, el ex ministro de petróleo de Chávez, Rafael Ramírez, dijo que Maduro estaba "fuera de tiempo".	Mistranslation 2 + words
Then came the sudden political shake-up that has convinced many Venezuelans the curtains are coming down on Nicolás Maduro's catastrophic six-year rule.	Luego vino la repentina sacudida política que ha convencido a muchos venezolanos de que las cortinas están cayendo sobre el catastrófico gobierno de seis años de Nicolás Maduro.	Mistranslation 2 + words
When Juan Guaidó declared himself Venezuela's interim president last month, he appeared to leapfrog a generation of rival opposition leaders.	Cuando Juan Guaidó se declaró a sí mismo como presidente interino de Venezuela el mes pasado, pareció superar a una generación de líderes opositores rivales.	Mistranslation 1 word
He quickly won the support of the US, the UK, Canada and some Latin American countries, who issued strong public statements recognising his authority.	Rápidamente ganó el apoyo de EE. UU., el Reino Unido, Canadá y algunos países latinoamericanos, quienes emitieron declaraciones públicas firmes en reconocimiento de su autoridad.	Extra word
On Monday, a succession of European governments, including Britain, France, Germany, Portugal and Spain, recognised Guaidó as Venezuela's legitimate leader.	El lunes, una sucesión de gobiernos europeos, entre ellos Gran Bretaña, Francia, Alemania, Portugal y España, reconocieron a Guaidó como el líder legítimo de Venezuela.	Agreement (NUMBER, GENDER)
Tens of thousands of Venezuelan protesters streamed through the capital, Caracas, on Saturday to demand the exit of the president.	Decenas de miles de manifestantes venezolanos viajaron por la capital, Caracas, el sábado para exigir la salida del presidente.	Mistranslation 2 + words
"It's essential we stay … because what is coming is good and it guarantees our future," he said, grinning cheek to cheek.	"Es esencial que nos quedemos ... porque lo que viene es bueno y garantiza nuestro futuro", dijo con una sonrisa de mejilla a mejilla.	Mistranslation 2 + words
"I think these are the final days," said Alberto Paniz-Mondolfi, a doctor who was one of tens of thousands protesting in Barquisimeto, Venezuela's fourth-largest city.	"Creo que estos son los últimos días", dijo Alberto Paniz-Mondolfi, un médico que fue uno de los miles de personas que protestaban en Barquisimeto, la cuarta ciudad más grande de Venezuela.	Agreement (NUMBER, GENDER)
"If it's a question of days or months, who knows? But you can be sure it won't make it through the year".	"Si es una cuestión de días o meses, ¿quién sabe? Pero puedes estar seguro de que no lo hará a través del año".	Mistranslation 2 + words

In an interview with the Guardian, Juan Guaidó insisted his country's march into a new political era was unstoppable and Maduro's "cruel dictatorship" doomed.	En una entrevista con The Guardian, Juan Guaidó insistió en que la marcha de su país hacia una nueva era política era imparable y la cruel dictadura" de Maduro condenada.	Mistranslation 2 + words
He has claimed his economically devastated nation was living through an "almost magical moment" in its newly revived quest for democracy.	Afirmó que su nación económicamente devastada estaba viviendo un "momento casi mágico" en su recién resucitada búsqueda de democracia.	Missing word
He repudiated Maduro's claim this week that the opposition's challenge had collapsed, saying it was a mix of propaganda and delusion.	Repudió la afirmación de Maduro esta semana de que el desafío de la oposición se había derrumbado, diciendo que era una mezcla de propaganda y engaño.	Agreement (TENSE, ASPECT)
Américo de Grazia, an opposition politician from the south-eastern state of Bolívar, said he was convinced Venezuela's military would soon ditch its embattled commander-in-chief.	Américo de Grazia, un político opositor del estado del sudeste de Bolívar, dijo que estaba convencido de que el ejército de Venezuela pronto abandonaría a su asaltado comandante en jefe.	Mistranslation 1 word
The generals fear that Juan Guaidó's offer of an amnesty for the billions they have stolen will not be honoured.	Los generales temen que la oferta de Juan Guaidó de una amnistía por los miles de millones que han robado no será respetada.	Agreement (TENSE, ASPECT)
The primary aims must be to map a consensual, peaceful way forward, promote national reconciliation – and swiftly alleviate the people's grievous suffering.	Los objetivos principales deben ser trazar un camino consensual y pacífico hacia adelante, promover la reconciliación nacional y aliviar rápidamente el grave sufrimiento del pueblo.	Mistranslation 1 word
But the rise of the fresh-faced opposition leader was orchestrated by a Harvard-educated economist with a checkered history in Venezuelan politics: López.	Pero el ascenso del nuevo líder de la oposición fue orquestado por un economista educado en Harvard con una historia a cuadros en la política venezolana: López.	Mistranslation 1 word
"When he was in prison he sent me in his place to speak with other party leaders, and I would relay their messages to him".	"Cuando estuvo en prisión, me envió en su lugar para hablar con otros líderes del partido, y yo le transmitiría sus mensajes".	Agreement (TENSE, ASPECT)
Sources confirmed scores of meetings between US officials and López surrogates – including Tintori – in Washington and around the globe.	Las fuentes confirmaron decenas de reuniones entre funcionarios de EE. UU. y sustitutos de López, incluido Tintori, en Washington y en todo el mundo.	Agreement (NUMBER, GENDER)
It was López who ensured Guaidó would lead the national assembly when Maduro began his second term in early January.	Fue López quien aseguró que Guaidó encabezaría la asamblea nacional cuando Maduro comenzara su segundo mandato a principios de enero.	Agreement (TENSE, ASPECT)
Guaidó's Twitter profile describes him as a civil servant and engineer – as well as interim president of the republic.	El perfil de Guaidó en Twitter lo describe como funcionario e ingeniero, así como a presidente interino de la República.	Extra word

“López has been sewing together an opposition that’s totally united and strong and pushing in the same direction, which is what we are seeing now.”	"López ha estado cosiendo una oposición totalmente unida y fuerte y empujando en la misma dirección, que es lo que estamos viendo ahora".	Mistranslation 2 + words
“Doing politics in Venezuela is a risk that you can pay with your life,” he said, pointing to more than 400 political prisoners.	"Hacer política en Venezuela es un riesgo que puedes pagar con tu vida", dijo, señalando a más de 400 presos políticos.	Missing word
At a specially convened meeting of the UN security council on Saturday, US secretary of state, Mike Pompeo, urged members to rally behind Guaidó.	El sábado, en una reunión especialmente convocada por el consejo de seguridad de la ONU, el secretario de Estado de los EE. UU., Mike Pompeo, instó a los miembros a unirse detrás de Guaidó.	Mistranslation 2 + words
The sanctions represent the US’s toughest economic move against Maduro to date and come five days after Guaidó’s declaration sparked Venezuela’s latest political crisis.	Las sanciones representan el movimiento económico más duro de Estados Unidos contra Maduro hasta la fecha y se producen cinco días después de que la declaración de Guaidó provocó la última crisis política de Venezuela.	Agreement (TENSE, ASPECT)
Bolton said the sanctions were an attempt to alleviate “the poverty and the starvation and the humanitarian crisis” gripping the South American nation.	Bolton dijo que las sanciones eran un intento de aliviar "la pobreza y la inanición y la crisis humanitaria" que afectaba a la nación sudamericana.	Agreement (NUMBER, GENDER)
“The authoritarian regime of Chávez and Maduro has allowed the penetration by adversaries of the United States, not least of which is Cuba.”	"El régimen autoritario de Chávez y Maduro ha permitido la penetración de los adversarios de los Estados Unidos, entre ellos, Cuba".	Extra word
"We think that is a strategic significant threat to the United States and there are others as well, including Iran’s interest in Venezuela’s uranium deposits.”	"Creemos que es una amenaza estratégica importante para los Estados Unidos y hay otros también, incluido el interés de Irán en los depósitos de uranio de Venezuela".	Agreement (NUMBER, GENDER)
His updated “axis of evil”, now Iraq and North Korea have been sorted out, comprises Cuba, Nicaragua and Venezuela (with an eye still on Iran).	Su "eje del mal" actualizado, ahora Irak y Corea del Norte se han resuelto, comprende Cuba, Nicaragua y Venezuela (con un ojo todavía en Irán).	Missing word
In his televised broadcast Maduro accused Bolton and Trump of seeking to destroy his “Bolivarian” administration through a coup that risked plunging Venezuela into conflict.	En su transmisión televisada, Maduro acusó a Bolton y Trump de tratar de destruir a su administración "bolivariana" a través de un golpe de estado que arriesgaba a sumir a Venezuela en un conflicto.	Extra word
Asked if the challenge to his rule meant he was now “against the ropes”, Maduro admitted he was facing a “tough” fight against powerful opponents.	Al preguntarle si el desafío a su regla significaba que ahora estaba "contra las cuerdas", Maduro admitió que se enfrentaba a una "dura" lucha contra oponentes poderosos.	Mistranslation 1 word

“They use sledgehammers instead of boxing gloves,” Maduro said of the US, which he claimed was seeking to topple him to seize Venezuela’s oil.	"Usan martillos en lugar de guantes de boxeo", dijo Maduro sobre los Estados Unidos, que según él buscaba para derrocarlo y apoderarse del petróleo de Venezuela.	Extra word
What is the cause? Is it iron? Is it aluminium? Is it gold, or diamonds? What is the cause?” Maduro asked.	¿Cuál es la causa? ¿Es hierro? ¿Es de aluminio? ¿Es oro, o diamantes? ¿Cuál es la causa?” preguntó Maduro.	Extra word
The whole world waded in after Juan Guaidó declared himself interim president, but the global tug-of-war is dangerous and unhelpful.	El mundo entero se desvaneció después de que Juan Guaidó se declarara a sí mismo presidente interino, pero la lucha global es peligrosa e inútil.	Mistranslation 2 + words
Protesters like Bellorín and González said they were aware of all the potential dangers and hoped Maduro agreed to step down.	Los manifestantes como Bellorín y González dijeron que estaban al tanto de todos los peligros potenciales y que esperaban que Maduro accediera a retirarse.	Extra word
But she also doubted that Maduro – who continues to enjoy the backing of Russia and China, as well as the military– was about to fall.	Pero también dudaba que Maduro, que sigue disfrutando del respaldo de Rusia y China, así como de los militares, estaba a punto de caer.	Agreement (TENSE, ASPECT)
“Venezuela’s oil belongs to the Venezuelan people and the oil money will now go to them through the legitimate government of Guaido.”	"El petróleo de Venezuela pertenece al pueblo venezolano y el dinero del petróleo ahora irá a través del gobierno legítimo de Guaido".	Missing word
Guaidó, the opposition politician leading the push to topple Nicolás Maduro, has urged one of the Venezuelan president’s key international backers, China, to abandon him.	Guaidó, el político de la oposición que lidera el impulso para derrocar a Nicolás Maduro, ha instado a uno de los principales patrocinadores internacionales del presidente venezolano, China, a que lo abandone.	Mistranslation 1 word
China’s trying to secure oil resources, construction contracts and a geopolitical foothold in a region the US has long considered its “backyard”.	China está tratando de asegurar los recursos petroleros, los contratos de construcción y una posición geopolítica en una región que Estados Unidos ha considerado durante mucho tiempo como su "patio trasero".	Extra word
But Beijing has become increasing aware the situation in Venezuela was “unsustainable” and it is unlikely to mourn Maduro’s political passing, if it came.	Pero Pekín se ha vuelto cada vez más consciente de que la situación en Venezuela era "insostenible" y es poco probable que llore el paso político de Maduro, si se produce.	Mistranslation 1 word
During a visit to a navy base in Aragua state on Sunday he tried to fire up troops by citing Hamlet.	Durante una visita a una base naval en el estado de Aragua el domingo, trató de disparar tropas citando a Hamlet.	Mistranslation 2 + words

Maduro paints the rebellion – and Guaidó – as part of an "imperialist" plot to destroy the Bolivarian revolution he inherited from Hugo Chávez.	Maduro pinta la rebelión, y Guaidó, como parte de un plan "imperialista" para destruir la revolución bolivariana que heredó de Hugo Chávez.	Missing word
Gredy Arrieta had travelled 700km from Maracaibo to the pro-Maduro rally, and said he was aware all was not well in his oil-rich nation.	Gredy Arrieta había viajado 700 kilómetros desde Maracaibo hasta el mitin a favor de Maduro, y dijo que sabía que no estaba bien en su nación rica en petróleo.	Missing word
The apparent attempt to overturn it by a Yanqui-picked, middle-class political neophyte has produced a viscerally negative reaction, with little thought for the revolution's failings.	El aparente intento de anularlo por un neófito político de clase media, escogido por los yanqui, ha producido una reacción visceralmente negativa, con poca reflexión sobre las fallas de la revolución.	Agreement (NUMBER, GENDER)
In an interview with Euronews, Maduro boasted that his political foes had "failed totally" in their quest to topple him.	En una entrevista con Euronews, Maduro se jactó de que sus enemigos políticos habían "fallado totalmente" en su búsqueda para derrocarlo.	Mistranslation 1 word
Maduro also sent a message to his opposition challenger Guaidó, sparking what many believe could be a final showdown between the two sides.	Maduro también envió un mensaje a su rival de la oposición, Guaidó, lo que muchos creen que podría ser un enfrentamiento final entre las dos partes.	Missing word
If one day a coup comes to pass, if one day a gringo military intervention comes to pass, your hands will be covered in blood	Si un día se produce un golpe de estado, si un día se produce una intervención militar gringa, se te cubrirán las manos con sangre	Mistranslation 2 + words
Pisani said she felt uneasy about the prospect of foreign military intervention to unseat Maduro, but was adamant he had to step down.	Pisani dijo que se sentía incómoda ante la posibilidad de una intervención militar extranjera para destituir a Maduro, pero estaba convencido de que tenía que renunciar.	Agreement (NUMBER, GENDER)
"The poor areas are where most people live, and since we don't support them any more I guess they want to kill us all"	"Las áreas pobres son donde vive la mayoría de la gente, y como ya no las apoyamos, supongo que quieren matarnos a todos"	Agreement (NUMBER, GENDER)
She and other residents mentioned a series of war-like operations in the days after opposition leader Juan Guaidó sparked the current political crisis.	Ella y otros residentes mencionaron una serie de operaciones bélicas en los días posteriores a que el líder opositor Juan Guaidó provocó la actual crisis política.	Agreement (TENSE, ASPECT)
After the police killed two dozen demonstrators last week, the protests are likely to die down within a couple of days.	Después de que la policía mató a dos docenas de manifestantes la semana pasada, es probable que las protestas terminen en un par de días.	Agreement (TENSE, ASPECT)

Human rights groups report that at least 26 people have been killed since the latest phase of protests began last week.	Grupos de derechos humanos informan que al menos 26 personas han sido asesinadas desde que comenzó la última fase de protestas la semana pasada.	Missing word

7.2 Annex 2: General instructions

Tarea

Este es un experimento en el que vamos a tratar de averiguar cómo la presencia de diferentes tipos de errores afecta a la productividad de los post-editores. Para ello hemos creado un dataset de 60 frases problemáticas que hemos pasado por un traductor automático. **Tu tarea consistirá en corregir los errores** presentes en estas frases.

Errores

Las frases que tendrás que post-editar generalmente contienen **un solo error**. Algunas frases contienen dos errores, pero son del mismo tipo (e.g. tiempo verbal erróneo, concordancia de género errónea, etc.).

Antes de corregir algo, trata de analizar si se trata de un problema real o de una mejora de estilo. Si la traducción es gramaticalmente correcta y conserva el sentido completo del segmento de origen, **no es necesario que mejores el estilo o la naturalidad de la frase**.

Temática

Estas frases tratan sobre la crisis de Venezuela y se han extraído de diferentes noticias del periódico The Guardian durante los meses de enero y febrero. Para familiarizarte con el tema y los personajes principales, hemos preparado **un timeline con eventos importantes** que encontrarás en la misma carpeta que estas instrucciones.

Las frases están ordenadas para tener una cierta coherencia en la temática; por ejemplo, si en un punto se habla sobre una persona sin mencionar su nombre, podrás saber de quién se trata mirando las frases inmediatamente anteriores. Es importante que tengas esto presente para ajustar el género en caso de que sea necesario.

No obstante, **las oraciones no encajan perfectamente**. Es posible que te encuentres con frases que parecen decir lo mismo o que no tienen el mismo tiempo verbal que las anteriores; en caso de duda, asegúrate de adaptar la traducción o post-edición al significado y sintaxis de la frase de origen.

Aquí tienes un ejemplo de tres frases que podrías encontrarte seguidas en la tarea, y que presentan varias de las características que hemos comentado: la segunda frase contiene dos errores, pero ambos son del mismo tipo. Para solucionarlos tienes que haber prestado atención al género de la primera frase. La tercera frase tiene un tiempo verbal diferente a las anteriores, pero se respeta.

Segmentos origen consecutives	**María García** said **she** was considering moving abroad. García was a student leader and dreamt of becoming a lawyer before the situation degenerated. María answers to the questions with a concerned look.

Traducciones automáticas	María García dijo que estaba considerando mudarse al extranjero. García era **un** líder estudiantil y soñaba con convertirse en **abogado** antes de que la situación degenerara. María **responde** a las preguntas con una expresión de preocupación.
Segmentos corregidos	María García dijo que estaba considerando mudarse al extranjero. García era **una** líder estudiantil y soñaba con convertirse en **abogada** antes de que la situación degenerara. María **responde** a las preguntas con una expresión de preocupación.

Tiempo

Tu productividad se medirá de varias maneras; una de ellas es cronometrando cuánto tardas en completar un segmento. Para garantizar la precisión de los resultados, te pedimos que **no realices otras tareas mientras estás trabajando en un segmento**. Puedes cambiar de página para consultar un diccionario o similar, pero si necesitas un momento para hablar con alguien, mirar el móvil, ir al aseo, etc. es preferible que cierres el segmento y lo vuelvas a abrir más tarde.

No es necesario que realices la tarea de una sola sentada, pero si guardas y cierras el programa, asegúrate de seguir trabajando sobre el archivo más reciente (el que aparece más abajo en la lista) para no sobrescribir tus datos.

Encuesta de dificultad

Después de completar cada segmento tendrás que dar una nota a la dificultad de la traducción o sobre el porcentaje de la frase que has tenido que post-editar. Se trata de una percepción personal y por tanto no hay respuestas más o menos correctas; lo único que te pedimos es que seas **consistente con tus notas**. Adicionalmente dispondrás de una casilla para dejar un comentario si hay algo que quieras aclarar.

7.3 Annex 3: Installing PET

Windows

1. Descargar fichero y descomprimirlo
2. Click en run.bat

Linux

1. Descargar fichero y descomprimirlo. En caso de encontrar algún problema al descomprimir, introduce *sudo apt-get install unrar* en tu terminal y prueba a descomprimir el archivo de nuevo
2. Abrir la terminal e introducir los siguientes comandos

   cd Downloads/PET-master

   sudo apt install maven

   mvn compile

   mvn package

   ./run.sh (o *bash ./run.sh*)

Mac OS

1. Descargar fichero y descomprimirlo
2. Abrir la terminal e introducir los siguientes comandos

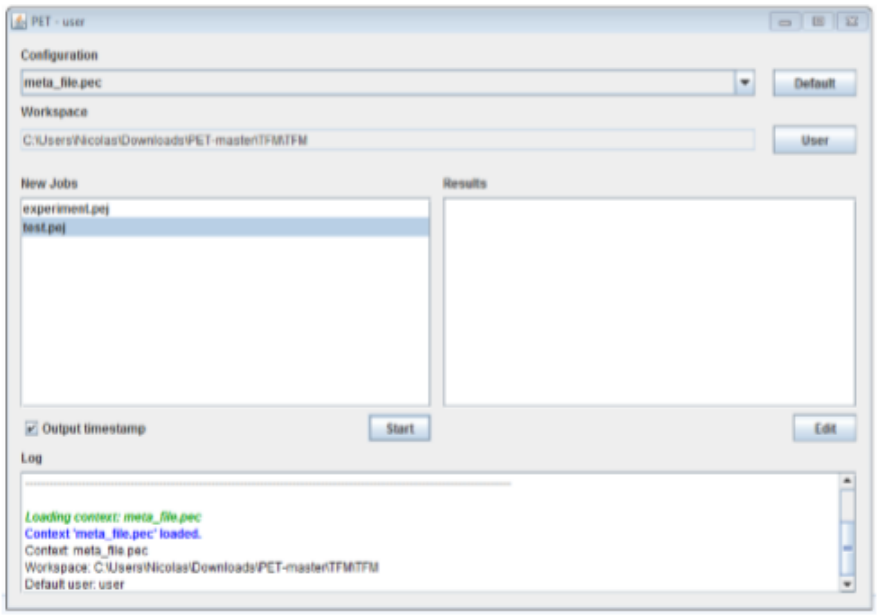
   cd Downloads/PET-master
   bash ./run.sh

7.4 Annex 4: PET test

Si sigues las instrucciones presentadas anteriormente para instalar PET, deberías encontrarte con la siguiente pantalla

1

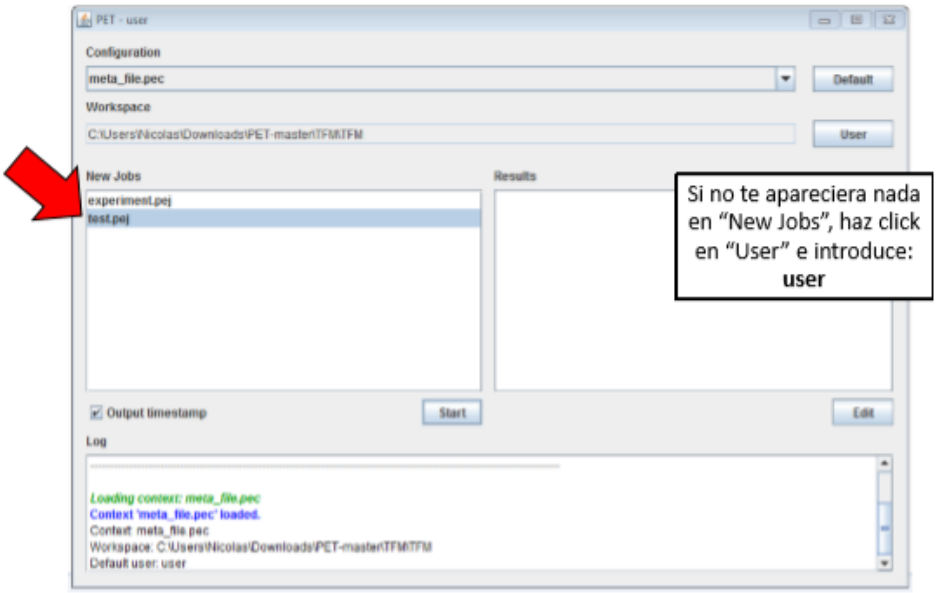
Esta es la pantalla que te encontrarás al abrir el programa



2

Primero, elige la tarea que quieras completar.

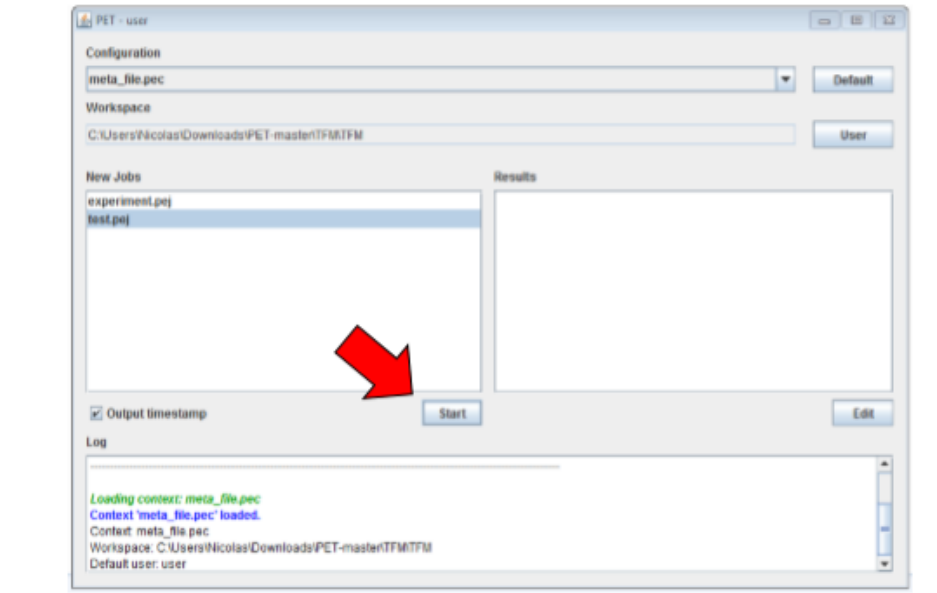
Para esta ronda de prueba elige "test"; para el experimento en sí deberás escoger "experiment"



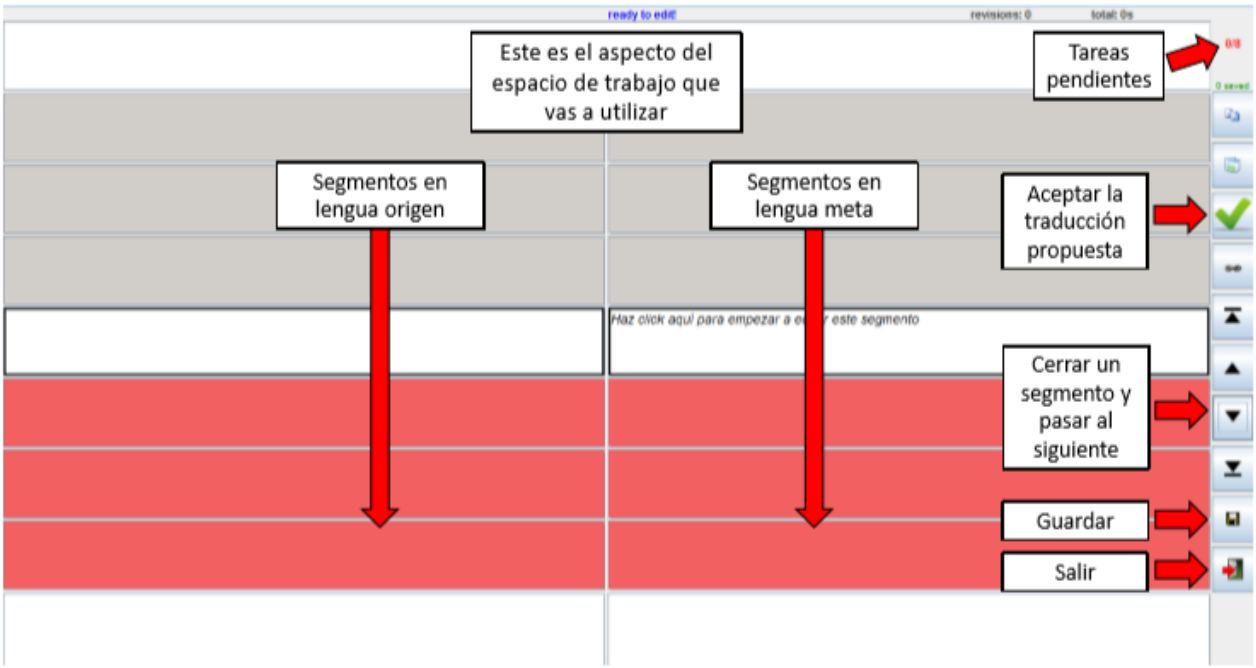
Si no te apareciera nada en "New Jobs", haz click en "User" e introduce: **user**

3

Haz click aquí para comenzar la tarea



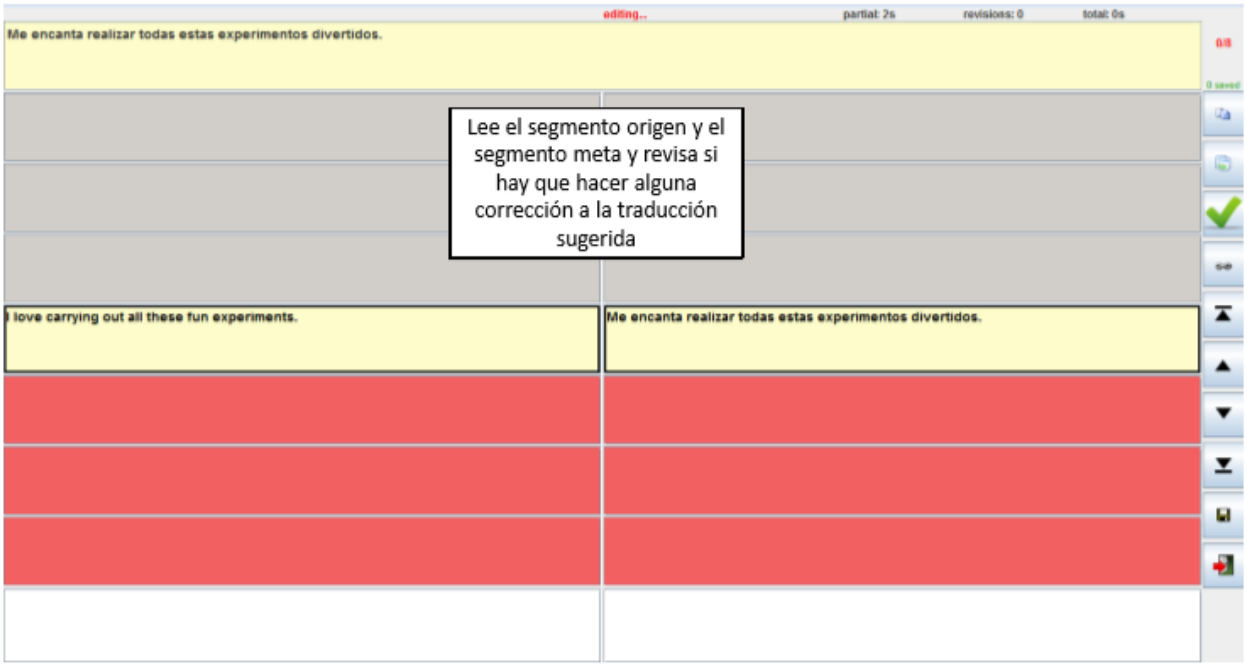
4


Este es el aspecto del espacio de trabajo que vas a utilizar
Tareas pendientes
Segmentos en lengua origen
Segmentos en lengua meta
Aceptar la traducción propuesta
Cerrar un segmento y pasar al siguiente
Guardar
Salir

5


Vamos a empezar la tarea
Haz click aquí para abrir el primer segmento

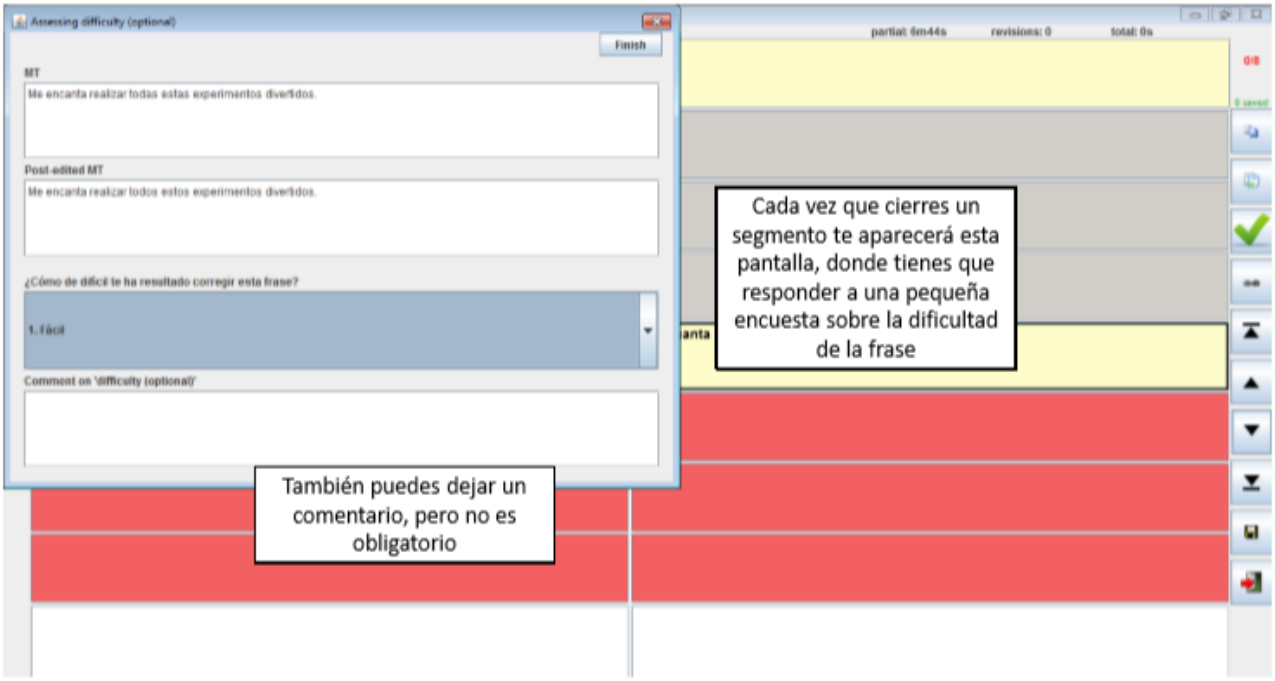
6


editing... partial: 2s revisions: 0 total: 0s
Me encanta realizar todas estas experimentos divertidos.
Lee el segmento origen y el segmento meta y revisa si hay que hacer alguna corrección a la traducción sugerida
I love carrying out all these fun experiments.
Me encanta realizar todas estas experimentos divertidos.

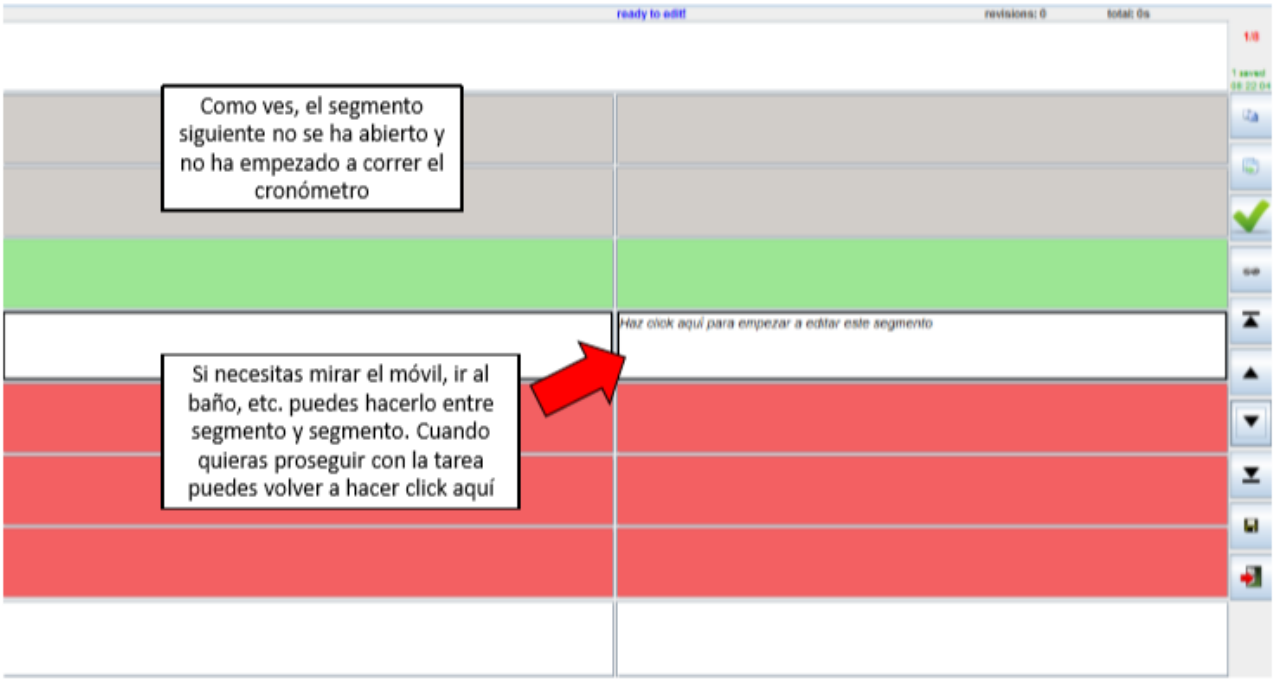
7


editing... partial: 13s revisions: 0 total: 0s
Me encanta realizar todas estas experimentos divertidos.
En este caso hemos corregido la concordancia de género
I love carrying out all these fun experiments.
Me encanta realiza todos estos experimentos divertidos.
Una vez realizados los cambios, haz click aquí para guardar y cerrar el segmento y pasar al siguiente

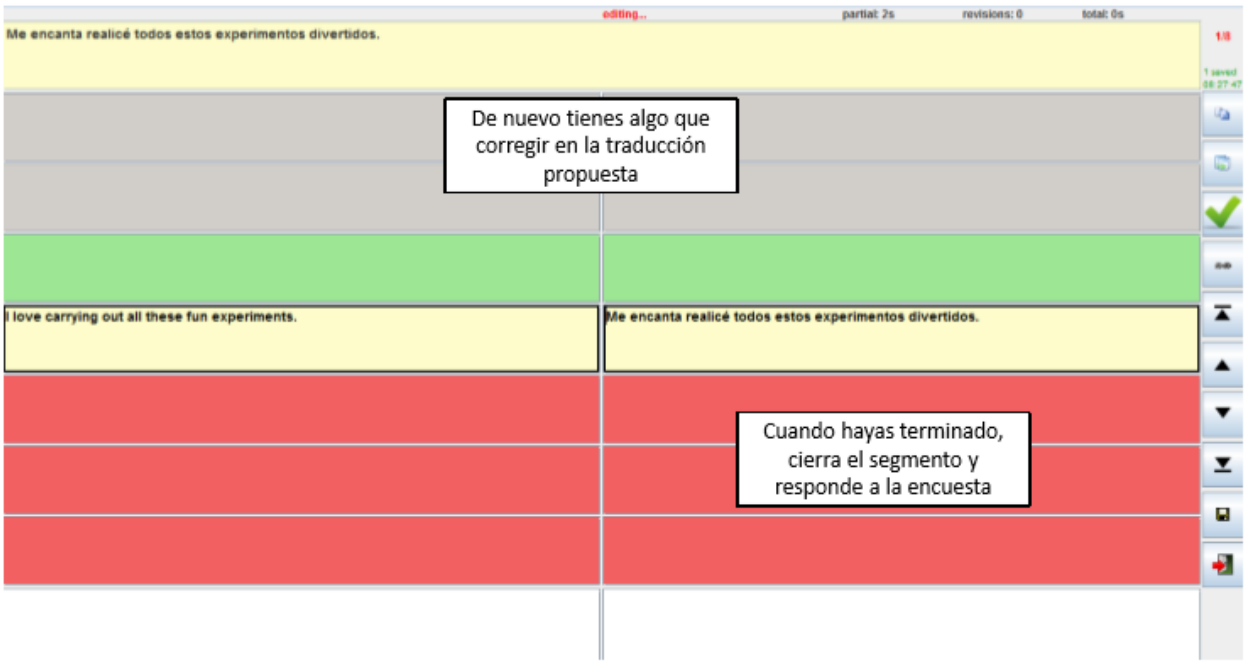
8


Assessing difficulty (optional)
Finish
MT
Me encanta realizar todas estas experimentos divertidos.
Post-edited MT
Me encanta realizar todos estos experimentos divertidos.
¿Cómo de difícil te ha resultado corregir esta frase?
1. Fácil
Comment on 'difficulty (optional)'
partial: 6m44s
revisions: 0
total: 0s
Cada vez que cierres un segmento te aparecerá esta pantalla, donde tienes que responder a una pequeña encuesta sobre la dificultad de la frase
También puedes dejar un comentario, pero no es obligatorio

9


ready to edit!
revisions: 0
total: 0s
Como ves, el segmento siguiente no se ha abierto y no ha empezado a correr el cronómetro
Haz click aquí para empezar a editar este segmento
Si necesitas mirar el móvil, ir al baño, etc. puedes hacerlo entre segmento y segmento. Cuando quieras proseguir con la tarea puedes volver a hacer click aquí

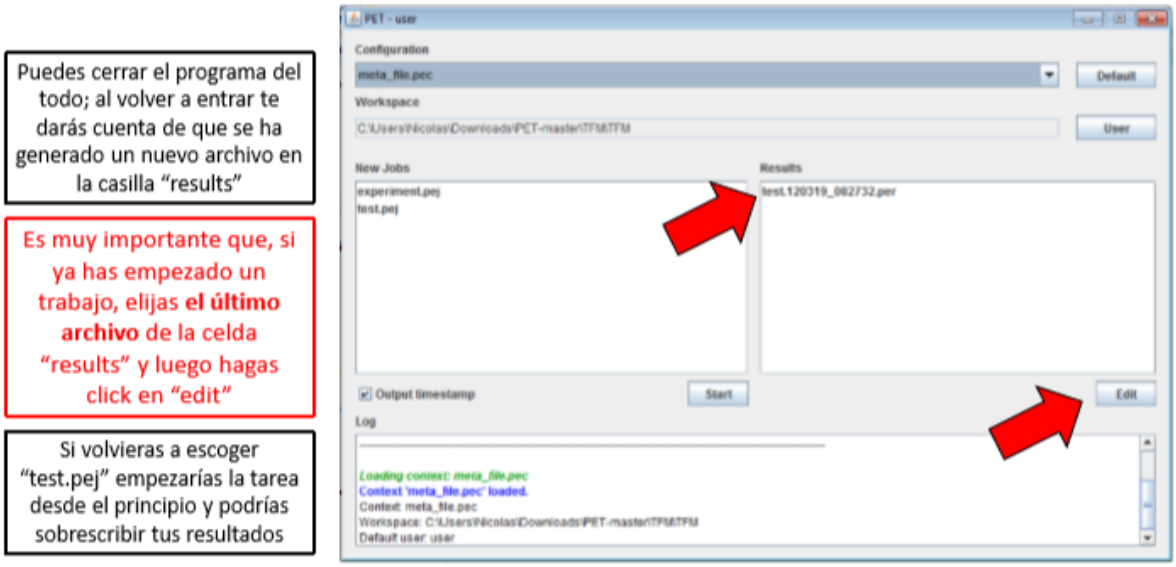
10


editing...
partial: 2s
revisions: 0
total: 0s
Me encanta realicé todos estos experimentos divertidos.
De nuevo tienes algo que corregir en la traducción propuesta
I love carrying out all these fun experiments.
Me encanta realicé todos estos experimentos divertidos.
Cuando hayas terminado, cierra el segmento y responde a la encuesta

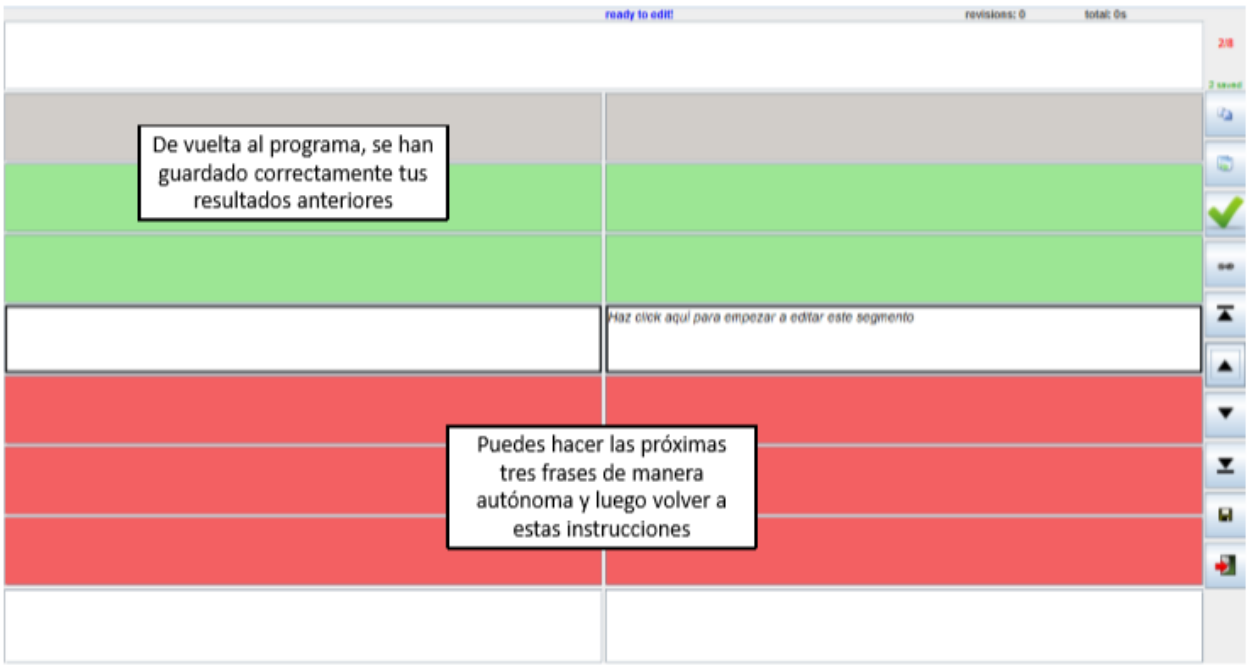
11


ready to edit!
revisions: 0
total: 0s
Llegados a este punto, imagina que quisieras hacer una pausa y cerrar el programa
Haz click aquí para empezar a editar este segmento
Primero guarda y después sal

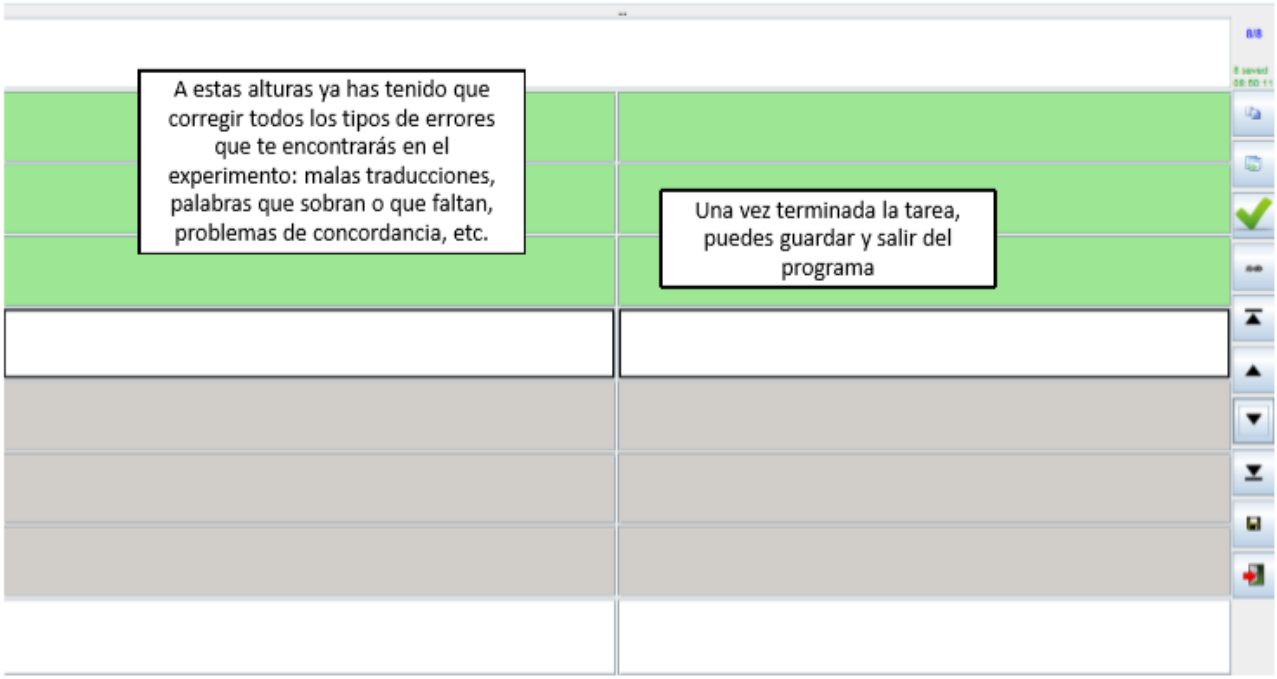
12


Puedes cerrar el programa del todo; al volver a entrar te darás cuenta de que se ha generado un nuevo archivo en la casilla "results"
Es muy importante que, si ya has empezado un trabajo, elijas **el último archivo** de la celda "results" y luego hagas click en "edit"
Si volvieras a escoger "test.pej" empezarías la tarea desde el principio y podrías sobrescribir tus resultados
PET - user
Configuration
meta_file.pec
Default
Workspace
User
New Jobs
Results
experiment.pej
test.pej
test.120319_082732.per
Output timestamp
Start
Edit
Log
Loading context: meta_file.pec
Context 'meta_file.pec' loaded.
Context: meta_file.pec
Workspace: C:\Users\Nicolas\Downloads\PET-master\TFM\TFM
Default user: user

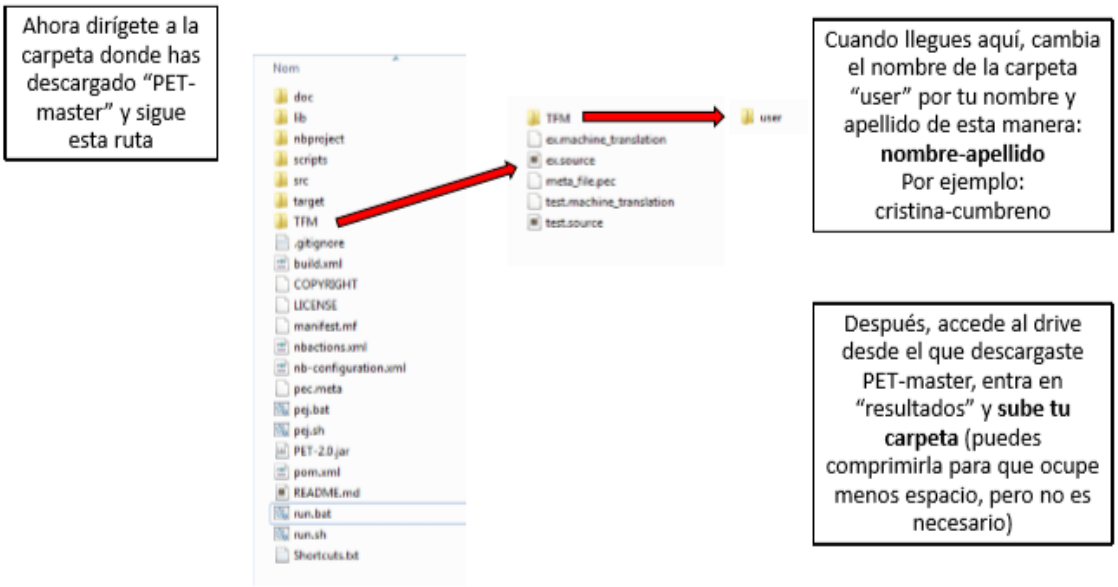
13


ready to edit!
revisions: 0
total: 0s
De vuelta al programa, se han guardado correctamente tus resultados anteriores
Haz click aquí para empezar a editar este segmento
Puedes hacer las próximas tres frases de manera autónoma y luego volver a estas instrucciones

14


A estas alturas ya has tenido que corregir todos los tipos de errores que te encontrarás en el experimento: malas traducciones, palabras que sobran o que faltan, problemas de concordancia, etc.
Una vez terminada la tarea, puedes guardar y salir del programa

15


Ahora dirígete a la carpeta donde has descargado "PET-master" y sigue esta ruta
Nom
doc
lib
nbproject
scripts
src
target
TFM
.gitignore
build.xml
COPYRIGHT
LICENSE
manifest.mf
nbactions.xml
nb-configuration.xml
pec.meta
pej.bat
pej.sh
PET-2.0.jar
pom.xml
README.md
run.bat
run.sh
Shortcuts.txt
TFM
ex.machine_translation
ex.source
meta_file.pec
test.machine_translation
test.source
user
Cuando llegues aquí, cambia el nombre de la carpeta "user" por tu nombre y apellido de esta manera: **nombre-apellido** Por ejemplo: cristina-cumbreno
Después, accede al drive desde el que descargaste PET-master, entra en "resultados" y **sube tu carpeta** (puedes comprimirla para que ocupe menos espacio, pero no es necesario)

16

7.5 Annex 5: Venezuelan timeline

1992	
4 febrero	**Hugo Chávez**, un comandante del ejército venezolano, da un golpe de estado que falla rápidamente. Pide perdón por televisión y es encarcelado.
1994	
27 marzo	Chávez es liberado tras recibir un indulto gracias a su creciente popularidad.
1998	
6 diciembre	Gracias al desencanto generalizado de la población con los partidos políticos establecidos, Chávez gana las elecciones con su proyecto ideológico y social: la "Revolución Bolivariana". Consigue aprobar una nueva constitución y propone medidas económicas y sociales de corte populista, socialista y anti-Estados Unidos.
2001	
	Chávez promulga 49 leyes sobre la administración y redistribución de tierras. Crece la preocupación de que Chávez esté tratando de concentrar poder político y económico en el estado.
2002	
11 abril	Tras varias protestas y manifestaciones masivas, se da un golpe de estado. El presidente de la cámara de empresarios se autoproclama presidente con apoyo de la Confederación de Trabajadores de Venezuela y políticos de la derecha. Inicialmente Chávez es encarcelado, pero esa misma noche vuelve al poder. La oposición organiza nuevas protestas y solicita un referéndum revocatorio.
2004	
	Se convoca un referéndum para decidir si Chávez debe seguir gobernando los dos años y medio que le quedan a su legislatura, y sale victorioso.
2005	
12 enero	Chávez decreta una reforma agraria para beneficiar a las clases menos favorecidas de las áreas rurales, atacando la propiedad privada. También impone nuevas regulaciones a los medios de comunicación, con fuertes multas y hasta cárcel en caso de difamación de figuras públicas.
4 diciembre	Los partidos chavistas ganan en la Asamblea Nacional. La oposición no acude a las elecciones alegando "falta de garantías".
2006	
Diciembre	Chávez gana las elecciones presidenciales con el 63% de los votos. **Leopoldo López** se convierte en líder de la oposición y lucha para lograr reformas en el sistema judicial.
2007	
Enero	Chávez anuncia que las principales compañías energéticas y de comunicaciones serán nacionalizadas.
2008	
Abril	Leopoldo López anuncia su candidatura a las elecciones para la alcaldía de Caracas. El gobierno le acusa de corrupción y le sanciona impidiéndole presentarse a cargos públicos durante 7 años. La Corte Interamericana de Derechos Humanos revisa su caso y emite un fallo por unanimidad a favor de López. En respuesta, el Gobierno venezolano informa de que el fallo está lleno de "contradicciones y hechos inexactos" y el Tribunal Supremo de Justicia ratifica la inhabilitación.
Noviembre	Venezuela y Rusia firman un acuerdo de cooperación en las áreas de gas y petróleo.

2009	
Febrero	Los venezolanos aprueban en referéndum una enmienda a la constitución que elimina el límite al periodo que una persona puede ostentar un cargo de gobierno.
2012	
Noviembre	Chávez gana su tercer mandato consecutivo con el 54% de los votos y una participación del 81%.
2013	
5 marzo	Chávez fallece de cáncer a los 58 años por complicaciones de un cáncer de colon. Se convocan nuevas elecciones.
14 abril	**Nicolas Maduro** (entonces vicepresidente) gana por un estrecho margen (50,61%) frente a **Henrique Capriles**, otro líder de la oposición. El comando de campaña de Capriles presenta una impugnación del proceso electoral ante el Tribunal Supremo de Justicia; este declara la solicitud "inadmisible".
Mayo	Los resultados de la elección presidencial desencadenan grandes manifestaciones en las que mueren 28 personas. Esto se aúna a un deterioro marcado de la economía, un aumento de los índices de criminalidad a nivel nacional y denuncias de corrupción en organismos públicos.
2014	
Febrero	Una serie de manifestaciones lideradas por Leopoldo López se saldan con 43 muertos, por lo que es acusado de instigar actos de violencia y se entrega a las autoridades. Es encarcelado en la prisión de Remo Verde.
2015	
10 septiembre	López es condenado a casi 14 años de prisión por incitación pública a la violencia.
6 diciembre	La coalición opositora Unidad Democrática gana las dos terceras partes de la Asamblea Nacional, poniendo fin a 16 años de control del Partido Socialista sobre el parlamento.
2016	
5 enero	A causa de las presiones del Tribunal Supremo de Justicia, tres diputados de Unidad Democrática renuncian a la Asamblea Nacional, dejando a la coalición sin la mayoría necesaria para bloquear la legislación propuesta por Maduro. Además, la mayoría de las leyes ya aprobadas por la Asamblea Nacional son declaradas inconstitucionales por el TSJ.
Septiembre	Cientos de miles de personas protestan en Caracas exigiendo la renuncia de Maduro.
2017	
Marzo	El Tribunal Superior de Justicia anuncia que asumirá las funciones de la Asamblea Nacional y prohíbe a Henrique Capriles ejercer cargos públicos durante 15 años.
Abril	A pesar de que el TSJ se retracta de sus decisiones a causa de la presión internacional, se inicia una nueva ola de protestas a nivel Nacional que para mediados de julio se han cobrado la vida a más de 90 personas.
8 julio	Leopoldo López sale de la cárcel y pasa a arresto domiciliario por problemas de salud. En este momento comienza a reunirse más activamente con miembros de la oposición y a enviar a su esposa, **Lilian Tintori**, a reuniones con embajadores. Comienzan las reuniones con el gobierno de Trump.
16 julio	Se convoca un referéndum para crear un nuevo cuerpo legislativo: la Asamblea Constituyente. Aunque su tarea principal es reescribir la Constitución, pronto empieza a apropiarse de otras tareas legislativas, como despedir a la fiscal general que estaba investigando el fraude electoral en las últimas elecciones.

2018	
Mayo	Maduro vuelve a salir victorioso en las urnas, en unas elecciones con poca participación y sospechas de compra de votos. Un gran número de países, incluyendo a Estados Unidos y al Grupo de Lima, se niegan a reconocer los resultados.
Agosto	Durante un desfile en el que Maduro estaba dando un discurso, dos drones cargados con explosivos son detonados cerca del presidente. Maduro acusa a Colombia y a Estados Unidos de urdir un plan para asesinarlo, pero no proporciona pruebas.
2019	
5 enero	**Juan Guaidó**, un miembro de la oposición poco conocido, ocupa el cargo de presidente de la Asamblea Nacional tras haber sido nombrado en diciembre de 2018.
10 enero	Maduro inaugura su segunda legislatura. Un gran número de países se niega a reconocerlo como presidente. Maduro confirma que el órgano legislativo será el único poder legítimo de Venezuela, quitando a la Asamblea Nacional todo su poder.
23 enero	Guaidó se autoproclama presidente interino de Venezuela siguiendo el artículo 233 de la Constitución, que dice que ante un vacío de poder o usurpación el presidente de la Asamblea Nacional deberá asumir la presidencia interina. A los pocos minutos del anuncio, Estados Unidos reconoce oficialmente a Guaidó.
24 enero	Un gran número de países reconoce a Guaidó como presidente interino. China, Cuba, Bolivia, Turquía y Rusia se posicionan a favor de Maduro.
25 enero	EEUU, Canadá y otros países anuncian que enviarán ayuda humanitaria a Venezuela.
28 enero	Guaidó promete a las tropas Venezolanas amnistía y protección si desertan. El gobierno de EEUU amenaza con represalias si se ejerciera violencia contra Guaidó.
29 enero	EEUU impone sanciones a la empresa Petróleos de Venezuela S.A. (PDVSA) y bloquea sus activos en suelo americano.
4 febrero	Los gobiernos de España, Francia, Reino Unido y Alemania reconocen a Guaidó, tras haber dado un plazo de ocho días a Maduro para convocar elecciones. La mayoría de los países occidentales reconocen a Guaidó, a excepción de Italia y Nueva Zelanda.
6 febrero	El gobierno venezolano bloquea el puente Tienditas en la frontera entre Venezuela y Colombia
7 febrero	Los primeros convoyes con ayuda humanitaria llegan a la frontera colombiana. Se anuncia que cruzarán la frontera el 23 de febrero.