eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Analysis of user post-edited texts and a proposal for assistance through iSTS

Author: Tatiana González Ferrero

Advisors: Nora Aranberri Monasterio

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua

Language Analysis and Processing

Final Thesis

2021-06-14

Acknowledgements

First of all, I would like to thank you, Nora Aranberri, for your infinite patience, support and guidance up to the very end.

I would also like to thank each and every one of the participants for their time and effort, as well as their altruism. Without you, this project would not have been possible.

Thank you also, German Rigau, for orienting Nora and me when we were losing our way.

I would like to acknowledge my fellow classmates, especially Cecilia, Margot, Kevin, and Nayla for their encouragement during these two challenging years.

Last, but not least, I want to thank my family and especially you, Piotr, for restraining me every single time I was about to give up (and there were plenty).

Departments: Computer Systems and Languages, Computational Architectures

and Technologies, Computational Science and Artificial Intelligence,

Basque Language and Communication, Communications Engineer.

Abstract

In the present investigation, the way lay users employ machine translation (MT) systems has been studied. The benefits of applying these systems for the dissemination have been examined to limit the scope, and the Polish-Spanish language pair has been selected. Prior to analyzing the similarities and differences in the compositions that these users have created, with and without the assistance of MT, various state-of-the-art systems, along with their strengths and weaknesses, have been described. Moreover, some of the available MT evaluation methods that address those weak points have been presented. However, it is considered that there is still a gap when it comes to assisting lay users in post-editing. It is believed that interpretable semantic textual similarity (iSTS) could fill this void. Nevertheless, some refinements of its annotation guidelines might be necessary.

Key words: machine translation, post-editing, iSTS, lay users, quality estimation, automatic evaluation metrics

Index

1 Introduction 5
2 State of the Art 7
2.1 Uses of Machine Translation 7
2.1.1 MT for assimilation 7
2.1.2 MT for dissemination 10
2.1.3 MT as a pedagogical tool 14
2.2 Strengths and weaknesses of MT systems 19
2.3 Methods for MT evaluation 23
2.3.1 Quality estimation 23
2.3.2 Automatic evaluation metrics 25
2.4 Interpretable Semantic Textual Similarity 32
3 General objectives 34
4 Analysis of lay users writings with and without MT assistance 36
4.1 Methodology 36
4.1.1 Participants 36
4.1.2 Design of the experiment 38
4.2 Analysis 45
4.2.1 Text length (sentences, words, characters per text) 47
4.2.2 Sentence length (words per sentence) 49
4.2.3 Lexical proportion 50
4.2.4 Lexical variety 52
4.2.5 Lexical density 54
4.2.6 Readability 55
4.2.7 Most frequent words (CREA) 59
4.2.8 SRL (rfunc) 62
4.2.9 Perplexity 65
4.3 Conclusions 68
5 Potential of iSTS for assisting users in post-editing 72
5.1 Introduction and objectives 72
5.2 Methodology 74
5.2.1 Corpus 74
5.2.2 Attempt of applying iSTS annotation scheme 77
5.3 Proposal of new guidelines for iSTS annotation 81
5.4 Conclusions 91
6 General conclusions 93
7 Future work 96
8 References 97
9 Appendix 102

1 Introduction

The translation industry has experienced massive growth in the last few years since the world is becoming more globalized. The rapid increase of the need for translated content has led professional translators to side with technology in order to facilitate the task in terms of speed and productivity (*Specia et al.*, 2018), (*Esplà-Gomis et al.,* 2018). This has resulted in the development of various machine translation (MT) systems. The use that the professional translators make of the tools, as well as their effort when post-editing the obtained output or their general attitude towards MT have been a recurrent subject of study, mainly due to the willingness of research scientists to improve the performance of the MT systems to the maximum extent possible (*Rei et al*, 2020).

A significant evolution can be observed in the quality of the current MT engines, especially those based on neural networks, in contrast to the first models, mostly statistical and rule-based (*Bentivogli et al.,* 2016), (*Toral and Sánchez-Cartagena,* 2017), (*Koponen et al.,* 2019). In fact, the advent of neural machine translation (NMT) has resulted in a ubiquitous use of MT, meaning that no longer just professional translators are taking advantage of the existing tools, but also lay users are benefiting from some of the many available free online systems (*Specia et al.*, 2018).

First, the texts that these users produce with and without the assistance of MT will be analyzed in the present investigation. In other words, the existing similarities and differences between the compositions that users create when writing directly in the foreign language (FL) supported by dictionaries, grammars or conjugators, and the texts that these same users produce with the help of MT systems starting from their first language (L1). To know whether users are benefiting from these tools or not, this analysis will be carried out in terms of complexity. For this purpose, different aspects will be studied, namely text length, sentence length, lexical proportion, lexical variety, lexical density, readability, basic vocabulary, syntactic structures, and perplexity. While the results in isolation may not be conclusive, the combination of the findings of all these metrics will indicate whether users produce more complex texts when assisted by MT. However, depending on the level of proficiency of the users in the FL, the differences between the two setups can be either enormous or barely remarkable.

Again, it should be noted that the users in this study are neither professional nor trained translators. Therefore, they may not be benefiting the most from the output of the MT systems. That is why, in this research, based on the aforementioned compositions that users have written in their L1 and have been subsequently translated into the FL by means of MT, the potential of interpretable semantic textual similarity (iSTS) for assisting them when post-editing will be explored. It is believed that this technique, which measures the semantic equivalence between sentences, could provide useful indications to users. To meet the objective, the annotation guidelines of iSTS will be applied to the described scenarios and further refined whenever the situation requires it.

Both objectives together with the respective findings will be discussed in more detail in the following chapters. The study is therefore organized as follows. The first chapter corresponds to the state of the art. The second chapter will present the general objectives in more depth. The third chapter covers the analysis of lay users compositions with and without the help of MT. The fourth chapter will discuss the potential of iSTS for assisting users in post-editing. The fifth chapter will provide the general conclusions. And the sixth chapter will make a short reflection on future work.

2 State of the Art

This chapter will provide an overview of the context in which the project is embedded. To this end, first, the possible uses of MT will be discussed. This first section has been subdivided into three parts, namely MT for assimilation, MT for dissemination, and MT as a pedagogical tool. The present study has been designed taking into account both the use of MT for dissemination and the use of MT as a pedagogical tool. Once the potential uses of automatic translation systems have been reviewed, their strengths and weaknesses will be analyzed. For this purpose, reference will be made to different investigations that have examined the strong and weak points of current systems, such as neural (NMT), statistical (SMT), or rule-based (RBMT) machine translation systems. Afterward, widespread MT quality evaluation methods will be described. These methods could be of great help for end-users to address the weaknesses and to make the most out of the strengths mentioned in the previous section. Quality estimation and automatic metrics such as BLEU, (H)TER, METEOR, Hjerson, and COMET will be presented here. In the last section of this chapter, a brief description of iSTS, a technique used to assess the similarity between word sequences, will be included. It is believed that, even if iSTS is not used within the MT field, it could contribute with relevant information in assisting lay users in making optimal use of the output.

2.1 Uses of Machine Translation

As mentioned in the introduction, lay users are nowadays those who make greater use of MT (*Specia et al.*, 2018). However, very little research has been carried out so far on them. Little as it might be, the studies performed can be classified in three different areas, namely MT for assimilation, MT for dissemination, and MT as a pedagogical tool. They all will be described in the next paragraphs.

2.1.1 MT for assimilation

One of the most widespread uses of MT is related to gisting or assimilation. It should be noted that the level of knowledge of the target language is not a limiting factor for using MT for this purpose. While it is true that users with little or no command of the language are more likely to make use of these tools to be able to comprehend the message, this does not mean that users with medium or high command will not do so as well. Perhaps the goal in the latter case is not understanding but verifying what has been understood. Moreover, it is worth mentioning that nowadays, it is increasingly common for people to live abroad. While it does not necessarily mean that the official language(s) of their new home country differs from their mother tongue, this is often the case. It is also likely that there are linguistic minorities within the borders of a nation or even a region. However, even though their language is in most cases one of the official languages, there may be situations in which they are kept in the background. In either case, the need for translated content is evident. *Bowker* (2009) and *Bowker and Buitrago Ciro* (2015) set Canada up as the perfect environment to study the use of MT for assimilation purposes since it is a bilingual and multicultural country. Both researches aimed at testing the users' level of satisfaction on the quality of not only human and machine-generated translations but also of post-edited ones and whether they could be sufficient to meet their demands. The methodology followed in both studies was largely similar. It was first required to evaluate the quality of different MT engines (statistical, rule-based, and hybrid) to determine which one to use in the next steps. A survey was then carried out to know the reasons why the participants would want to have the contents translated. Afterward, they were given four different translations of the same source text, namely a professional human translation, a maximally post-edited MT (both the content and the style were corrected), a rapidly post-edited MT (only content errors were corrected), and a raw (unedited) MT, and asked about their satisfaction. The respondents were later provided with the typology of the texts, the costs, and the time invested in producing them. Before knowing any of the details, some participants (specifically those who were not language professionals) chose the MT system as their preferred option instead of selecting human translation. Once they received all the information, some changes took place, leading fewer people to opt for machine-translated texts. In both studies, it was concluded that the favourite translations were not only the maximally but also the rapidly post-edited ones.

It must also be taken into account that sometimes a high-quality translation for a specific language pair cannot be obtained, which may endanger understanding certain types of texts. A way to overcome the problem could be to use MT to translate the content that needs to be assimilated into another language that shares several features with the requested one. This subject of inter-comprehension within the context of MT has been of interest to *Jordan-Núñez et al.* (2017), who attempted to give an answer to different research questions, such as the preference of the readers in terms of usefulness for an automatic translation into their L1 or a human translation into a language belonging to the same family; the level of comprehension of machine-translated texts whose source language is of a different language family from readers' L1, and the target language belongs to the same one; and the usefulness for a reader of having a text in their L1 that has been either automatically translated or translated by non-native speakers. To address the aforementioned research questions, a cloze test methodology was applied. The participants, whose L1 was Spanish, were given texts with different degrees of specialization (natural sciences, human and social sciences, journalistic content, and general topics) and were then tested in four situations in which they were required to fill in the gaps in professionally translated texts. As a hint, they were provided with either a machine or a human translation based on the previously described characteristics (the languages used in the study were Italian, French, and English). It was concluded that highly specialized texts written in a language from the reader's L1 language family are more useful than the MT output produced in their own mother tongue and that there is a relevant preference for an automatic translation with a source language from a different language family over a manual translation into the L1 produced by a non-native speaker.

As mentioned at the beginning of this section, the studies presented above refer only to a couple of situations in which MT might be of great help for assimilation. However, there are many more cases that are currently on the rise. This is the case of, for example, e-commerce or social media (*Specia et al.*, 2018). As discussed throughout this section, the raw MT output may not perfectly meet the objective of assimilation, and the professional human translation of certain content might not be available. Therefore, concerning the last two cases mentioned, the ideal situation would be that the users, regardless of their competence in the source language, could have at their disposal some kind of a tool. That tool would indicate whether the information they are reading on the Internet in a language they understand, and which has been provided by an MT system, is the same as that of the source language. In other words, if the original meaning is kept, if the MT system has added or omitted information, or if the used terminology is correct, among other things.

2.1.2 MT for dissemination

As the need for producing content in a foreign language is becoming increasingly common, the use of MT is progressively being seen as a beneficial tool for meeting that objective. This may either be because the command of the target language is not sufficient to satisfy specific needs or because it is essential to produce content in that language in a fast and low-cost manner. The research studies devoted to this topic cover different scenarios, which will be reviewed in the following paragraphs: academic writing, medical context, stories, and crisis situations.

2.1.2.1 Academic writing

As is well known, English has become the dominant language within the scholarly communication system. Since not all researchers possess the adequate level to produce their investigation reports directly in that language, the use of MT is increasingly being taken into consideration. In recent years, and in view of the significant development of the MT systems, several attempts have been made to ascertain whether the needs of the scholars are successfully fulfilled, or there are still some aspects that need to be improved. The study conducted by *O'Brien et al.* (2018) is intended to see the effect of MT as an aid for writing academic texts in English as a Foreign Language (EFL). To this end, a group of researchers who desired to publish and disseminate their work were told to write a short academic abstract (+/- 500 words) in their field of expertise. Two groups were established: while the first subgroup had to produce the first half of the abstract in English and the second half in their L1 (which differed from one participant to another), the second subgroup was required to carry out the task in the other way around. Afterward, the part written in their L1 was translated into English with the help of an MT system, and they were asked to check and produce the final text.

Moreover, the researchers had to write down the time invested for every part (translation and post-editing) as well as the resources employed. Finally, a professional reviewer was in charge of judging the final version of the manuscript without knowing any details either of the text or of the author. An analysis of the edits and the quality was then performed. As could be expected, the participants alleged that drafting in their L1 was easier than in EFL. However, they also admitted that the revision of the abstract written in EFL was less challenging than self-post-editing the text in L1, which was highly time-consuming for the authors who were unfamiliar with the task. Furthermore, it has to be noted that, based on the edits made by the professional reviewer, the quality was equal in both versions.

2.1.2.2 Medical context

Although this section is closely related to the preceding one, it has been decided to deal with the field of medicine independently since it usually presents some peculiarities that make it differ from any other academic environment. As it is desirable to provide global access to medical research findings, it has been analysed whether MT is able to fulfil the needs. An interesting approach in this matter is that of *Parra Escartín et al.* (2017), in which a group of medical practitioners whose L1 was Spanish served as study subjects. Prior to starting the experiment, they were asked to self-assess their English level and later to take a language test. Afterward, a publication initially written by each of them in their L1 was automatically translated into English with Google Translate, and the participants were requested to review it by using the "track changes" functionality in MS Word. Once the post-editing was made, a professional translator was hired to proofread the resulting texts. The amount and typology of edits, namely essential edits, preferential edits, essential edits not implemented, and introduced errors, were analyzed. Even though it was found out that many medical practitioners are making use of MT as a writing aid to disseminate their research, not all of them are satisfied with the output. Some of the complaints involved its literalness, the incorrect use of synonyms and grammar, and the lack of terminology. Among the other conclusions, it was interesting to see that, when self-post-editing, essential and preferential edits were the most frequent changes performed by the participants, while in the case of the proof-reader, the rate of preferential modifications was considerably higher.

2.1.2.3 Stories

The scenario proposed by *Aranberri* (2020) was designed to study the way users typically make use of the existing tools to produce texts in a language they do not master. The research was conducted in the Basque Country, a special location due to the coexistence of two official languages. The 40 selected participants reported a similar level of competence in both Spanish and Basque, consequently. However, since Basque was the language of instruction, it was decided to set it as the source language for the investigation. As one of the goals was to see how to proceed when trying to produce content in a foreign language as realistically as possible, the students were asked to write their own source text. They were allowed to use any language resource except for MT. To not affect the results, it was decided to establish a specific genre, domain, and minimum length, which was a piece of flash fiction of around 150 words taking a storyboard as a guideline. In order to compare the differences between writing in the foreign language from scratch and using MT, the participants were required to create two stories: one written directly in English and one in Basque to later post-edit the English MT output. According to some quantitative measurements, the texts produced by the students using MT seemed to be more complex, and their surface form was more similar to original English texts. This could be observed in the use of prepositions or subordinate conjunctions and pronouns, which was bigger in the post-edited version. However, the manual translations were characterized by a wider lexical variety. It is worth mentioning that Translation Error Rate (TER) revealed that not many changes were deemed necessary to improve the MT output. This may be due to either the fact that it was very good or to the lack of competence of the participant, which is not sufficient to correct it. In any case, the questionnaires filled out before and after the experiment showed a positive attitude towards the use of MT, also pointing out that it was very beneficial for learning unknown words and considering new structures for their translations.

2.1.2.4 Crisis situations

As described by *O'Brien and Federici* (2019), there is a particular context in which language translation is of vital importance, and it is that of crisis situations. For the researchers, the concept of 'crisis situation' encompasses natural hazards, human-driven disasters (including terrorism), and conflicts in multilingual and multicultural societies. Any of these events can take place within a city, a region, a nation, or even across borders between multiple countries. Therefore, the need for not only a linguistic but also a cultural transfer from one language to another through written, oral, signing, or multimodal channels is obvious. Since it is not always possible to count on trained professionals, the use of MT to help people act as cultural mediators is increasingly being taken into account. However, some studies show that still much needs to be done to raise awareness of the existing tools. As an example, *Cadwell and O'Brien* (2016) sought to analyse the role of Information and Communication Technology (ITC) in disasters, such as the earthquake that originated a tsunami, which in turn set off a nuclear accident in Japan in March 11, 2011. To this end, 28 individuals with ages ranging from 20 to 50 belonging to 12 nationalities, who were living in Japan at that time, were subjected to face-to-face, in-depth, semi-structured, individual interviews in 2013. They all had different skills of Japanese – from beginner to almost native-level – and worked in various fields. The responses to the interviews, together with official reports of the disaster, other grey literature, and an illustrative corpus of actual communication, contributed to creating a special model of how language, culture, and translation impact ICT use during a disaster. When taking a look at the results, the authors were surprised about the fact that almost none of the participants alluded to the use of OMT tools for providing linguistic mediation for other of their nationals. Among the possible explanations, the technical (insufficient electric power supply), demographic (a small number of foreign residents compared to the overall affected population), socio-cultural (the implementation of specialised translation was of low priority in a resource-poor disaster setting), and natural ones (the destruction of necessary network infrastructure) stood out.

Without going any further, the current Covid-19 pandemic has confirmed the potential use of MT systems to produce content in several different languages during a crisis. In fact, many people have lived this situation in places where they do not know the language, so it has been vital to use these tools to provide essential information such as the measures applied, how to proceed when getting sick, or what to do to get vaccinated.

Although there have been several cases mentioned in which the use of MT for dissemination has been studied, these are not the only situations where these tools have a place for producing content in a foreign language. In fact, the conclusions drawn from the above-described research studies may be applicable to other scenarios. While MT can be useful to provide a first and broad response to the needs of users who want to create this type of content, it is not always sufficient. In order to optimally meet the objective, it is necessary to review the output provided by the system. However, end-users do not always have the required knowledge or the adequate tools to do so. For example, if the system itself could highlight what parts of the sentences may be incorrect or have a different meaning than that of the source text, users could make better use of the output.

2.1.3 MT as a pedagogical tool

Although it is still a topic that raises discrepancies among the teaching staff, a growing trend towards MT's inclusion as a supplementary tool for language learning can be observed. *Niño* (2020), for example, advocates the use of online machine translation (OMT) for independent language learning (ILL). To prove its usefulness, she analyzed how advanced language students of Spanish (specifically C1 level according to the CEFR) dealt with OMT systems to translate a text of their own choice. They were free to choose the topic and the OMT engine with which to proceed as well as the manner of using the output (while some decided to post-edit it, others opted for comparing the translation obtained from several different systems). Although at first, the participants were under the impression that they could do better without the OMT output, they admitted later that it actually helped them to produce better quality translations, since not only were they presented with new structures and vocabulary, but it also allowed them to check the grammar of their own produced content. Furthermore, as the students had an advanced command of the target language, they could guess quite accurately, which were the strengths and weaknesses of these (for them) newly discovered online language reference tools.

However, this is not the first attempt to demonstrate the benefits of implementing MT to teach a foreign language. *Niño* (2004) presents another experiment to show how advantageous the correction of MT output is for making L2 learners acquire linguistic accuracy and therefore be more confident when editing their own writing since they will be able to learn from the correct translation and from correcting errors. To this end, a group of advanced Spanish students was divided into two subgroups: one subgroup was asked to post-edit the machine translation of an advice column and rewrite it into a more correct version, while the other was required to translate the exact same text from scratch. Afterward, an analysis of the errors committed by the MT system and the students was performed. A comparison was made between the manual and the post-edited translations. Among the conclusions of the study, it is highlighted that, even if the predominant error class is the grammatical one in all types of translations, the lowest percentage is found in the post-edited content. This situation may be explained by the fact that the students of the post-editing subgroup were mostly focusing on the grammar. Spelling errors, meanwhile, were less often present on the manual translations since the participants made greater use of dictionaries. A few years later, *Niño* (2008) follows the same methodology with a slight modification. In this case, the group in charge of post-editing the MT output was given a 10-days training course in which the participants did not only learn what machine translation was about and how it worked but also practiced post-editing with several types of texts with different features. Although the post-editing was performed in groups during the course, the final task had to be carried out individually within a week. In addition, in contrast to the previous study, the students were required to work with 8 different text types during both the training session and the experiment. Prior to performing the error analysis, which covered 50 error categories within four domains, the tasks were manually corrected using a colour-code system (green for a correct post-editing and red for a wrong MT translation and post-editing). Although the conclusions drawn from this experiment are rather similar to those of the preceding research, it is interesting to note that there was a significant difference in scores regarding spelling errors, which were more frequent in manual translations. Those were justified as human errors. On the other hand, this subgroup dealt better with discursive errors than the one in charge of post-editing.

The research presented by *García and Pena* (2011) introduces a slight change in regards to the profile of the participants of the study by no longer focusing on advanced learners but on beginners and early intermediates. The aim was to test the convenience of using MT to help develop the students' writing skills in the foreign language. The research was conducted with sixteen university students of Spanish divided into two subgroups, depending on their level. All of them were asked to perform two tasks. The first one was the same for both groups, and the second one level specific, consisting of producing texts (50 words each for beginners (A1.2) and 100 words each for early-intermediates (A2.3)). Half of the texts had to be directly written into the L2 and the other half in L1 to be subsequently translated with Tradukka. Afterward, the students were first required to pre-edit the source and then to post-edit the output. The resulting data were then analyzed by taking two perspectives into account, namely, writing as a product and writing as a process. The first one considers aspects such as the number of written words or the quality. In contrast, the second one is focused on the pauses, proofreading, and editing intervention, for which it was required to record the screens and catch the cursor movements as well as the keyboard log. For assessing the quality of the translations, the acquired knowledge of both groups was taken into account, concluding that all of them seemed to benefit from the MT system in more or less the same proportion. It should also be noted that beginners only performed edits in the source text (pre-editing), while early intermediates dared to modify the target text as well. However, this does not necessarily mean that all interventions were successful or, in other words, improved the end text. What was most striking was that, in contrast to beginners, who produced the content in L1 based on what they knew in L2, some of the early intermediates did not succeed in producing error-free source texts.

Along this line, *Lee* (2019) introduces a similar, but at the same time, a different approach to evidence the helpfulness of using MT as a computer-assisted language learning tool (CALL). In this case, a group of Korean students of English was first played a video and, after the visualization, was told to produce a one-page text on the topic. The following task was to translate it manually from scratch by using dictionaries or grammar books. The next step was to enter the original text into an MT system of their choice. Instead of post-editing the obtained output, as is the usual procedure, the students were given a chance to revise their manual translation aided by the machine translation. Finally, after a short interview of the participants and a collection of reflections, a quantitative as well as a qualitative analysis of both (pre-MT and post-MT) versions were carried out. According to the quantitative data, even if its amount decreased considerably after the revision of the manual translation, the most frequent type of errors was the grammatical one, distantly followed by the lexical type. Statistically speaking, there were no significant lexical or sentence complexity differences between the pre- and post-edited versions. The results of the qualitative analysis showed that, despite the initial scepticism about MT accuracy, the tool was especially useful when the participants wanted to use a certain word in particular, since dictionaries usually offer many possibilities, and it is sometimes difficult to choose the best option for a specific context. It was also pointed out that this task helped the students improve their lexico-grammatical awareness.

A small variant of this methodology has been proposed by *Lee and Briggs* (2020), the outline being almost the same except for a few changes. In this case, the students were asked to write an essay as a source text, and they were not given a chance to choose the MT system, as the use of Google Translate had already been prescribed by the researchers. Moreover, the analysis performed was more focused on the error types and their categorization than the previously described one. Among the results, it was observed that while the number of words increased in the revised texts, errors were reduced, as was the case in the formerly described study. In addition, the MT output helped the students to correct errors in articles, prepositions, noun plurals, and substitutions, which were also the most frequent error types. It is worth mentioning that the participants who committed fewer mistakes when writing directly in L2 were the ones that made more changes during the revision phase. Most of the changes were positive; in other words, the errors were corrected. However, there were also neutral changes, implying no change at all, and negative changes, which introduced new errors.

It is important to note that, although it is not the standard way of proceeding when it comes to employing machine translation as a pedagogical tool, some researchers have seen the potential of pre-editing to teach a foreign language. *Shei* (2002) presents three case studies to prove the pros and cons of modifying the input in order to obtain the desired translation. While two of the studies were based on the pre-editing of a text written in the participants L1 (for the first one, it was Chinese, and for the second one - English), the third study involved pre-editing a text in their L2, which posed the biggest challenge. The students had to follow the same pattern in all three cases: enter the text into one of the two proposed MT systems; observe the output; write down the limitations of MT (structural, lexico-semantic, idiomatical, cultural, and operational); modify the input; and repeat the process until getting satisfactory results or until the output could not be further improved. Once they had finished the task, the participants also provided a few pre-editing strategies for improving the performance of MT engines, relying on their experience (reorganisation, simplification, addition, replacement, pre-translation, punctuation). The main conclusion drawn from the experiment was that the MT system became a grammar checker for the language learners, helping them flag a word or phrase, call attention to aspects of punctuation, draw awareness to polysemy, experiment with structures, and raise consciousness about their interlanguage. The matter of grammaticality had already been studied by

Richmond (1994), who also defends the process of pre-editing for teaching and learning a foreign language. For him, the aim was not to produce texts in L2 but to understand the processes by which meaning is expressed in a specific way in the L2. To this end, a group of students from the first and second year of French, who had problems with the language, were asked to pre-edit a text using French Assistant in interactive mode (allowing them to choose a word, expression, or form within a concrete context). The software presented some limitations, as it was only possible to enter the text sentence by sentence. Unlike the previously seen study, the participants were given the correct final translation so that they knew how the output had to look like after pre-editing. It was pointed out that backward translation increases the students' awareness of the differences between L1 and L2 because it puts the emphasis on linguistic processes and linguistic input. The participants were able to retain more structures of the target language in a more playful environment, which helped reduce the language class stress since the participants were not forced to produce a text in L2, even though they were continually working with the target language[1].

The procedure used by *Briggs* (2018) differs entirely from all the ones seen so far since the sought goal, in this case, is not to make the teaching staff aware of the benefits of using MT in their classroom but the students. In order to make it all the more dynamic, the project was designed in the form of a contest. Eighty students from different fields of knowledge were divided into four groups and were given three types of surveys written in their L1 (Korean) - for whose answers they were granted a score. The first survey was aimed at getting more information about the L2 (English) level of the participants as well as of the frequency of use of web-based machine translation (WBMT) engines; the second one contained questions in Likert style about their attitude towards WBMT; and the third questionnaire was meant to evaluate a series of machine-translated texts (KO to EN). The vast majority of the participants reported frequent use of WBMT tools, both inside and outside the class, to look up a new vocabulary. Most students also claimed that they felt insecure with their writing skills and that, consequently, WBMT could help them support their language learning efforts. What was particularly worrying was that some automatic translations that contained obvious

[1] It should be remembered that in the 90s MT systems were rule-based. This means that they included transfer rules to transform the structures and vocabulary of the source language into the structures and vocabulary of the target language and functioned in a deterministic manner. However, with corpus-based approaches, this is not the case anymore. Therefore, the effectiveness of using MT for this purpose with current systems remains to be considered.

errors were given a positive score and that some students were not able to improve the output – instead of restructuring the sentences, they either omitted them or changed the order of some words.

As could be observed in this section, the potential application of MT to language learning can be carried out from many different perspectives. While some propose to work with the same type of text, others suggest that it is better to deal with different sorts of content; while some recommend that all students perform the same tasks, others claim that it is more interesting to divide them into two groups and give them diverse assignments; while some state that it is only suitable for advanced learners, others confirm its benefits for beginners and early intermediates; while some limit themselves to post-editing the output, others consider also pre-editing the source text, etc. Despite the dissimilar approaches to the same topic, they all conclude that, with specific teaching goals and embedded within a well-structured didactic sequence, MT can help learners to improve their writing skills, to learn new structures and vocabulary, and to raise their awareness of the similarities and differences between languages.

2.2 Strengths and weaknesses of MT systems

As observed in many of the studies described in the previous chapter, more and more people are starting to discover the benefits of MT and are willing to use it in their everyday lives, even if, at first, the mere mention of the tool often generated inevitable rejection. It is surprising how the attitude towards MT has changed within the last decade – between the 'do not use it, it is terrible' perspective to the 'use it, it is amazing' one, there is only a few years difference. However, as also mentioned earlier, MT is not perfect, and one should be aware of its limitations. This section will cover the strengths and weaknesses of the different available MT systems based on various research works. This will reveal that there remains much to be done, even though the improvements have been enormous. Moreover, it is important to know their weak and strong points, not only to be careful when using them but also to try to make the most out of them. The aim of doing this analysis is to be able to better assist users in taking maximum advantage of the MT output. Once the strengths and weaknesses of the different existing systems are identified, more specific indications on how to proceed with the automatically translated content can be given. Some of the research studies that have been paving the way for improving such indications will be described below.

The approach followed by *Bentivogli et al.* (2016) employs the data available for the English- German task of IWLST Evaluation 2015. It makes a comparison of the first four top-ranked participants, which are one neural MT (NMT) system and three statistical MT systems, namely a standard phrase-based MT (PBMT), a hierarchical PBMT, and a combination of phrase-based and syntax-based MT. The decision to operate with that language pair was prompted by the challenging morphology and the word-order differences. It is worth mentioning that the nature of the data also plays an important role in the performance of the tools. In this case, the study was conducted on TED talks, leading the systems to deal with oral language, which is structurally less complex, formal, and fluent than a written discourse. Moreover, the talks covered a wide selection of topics and were carried out by speakers with very different styles. This lexical and thematic variety is a factor that works in favour of NMT since it has proven to be better with diversity than any other SMT system. Furthermore, NMT seemed to produce morphologically more correct translations and with fewer lexical errors than PBMT. However, whereas NMT is more accurate with word reordering and works well with verbs and nouns, its performance with prepositions, negative particles, and articles is as poor as that of PBMT. One of the biggest deficiencies of NMT must also be mentioned, which is its difficulty in operating with long sentences. Even though the best results were obtained with that system, the decrease in quality when sentences get longer (specifically, from 35 words onwards) is more dramatic compared to any other PBMT.

Toral and Sánchez-Cartagena (2017) employed a similar methodology but introduced a couple of modifications. Instead of working with a single language pair, the research was carried out on nine language directions belonging to four language families, enabling the authors to obtain more general conclusions. It was also decided to work with news stories with the best PBMT and NMT systems submitted to the WMT16 translation task since not only were they state-of-the-art, but in addition, all outputs were publicly available. The type of texts selected for this study is made up of longer sentences than those in transcribed speeches. This confirmed the aforementioned investigation results: NMT outperformed PBMT up to sentences of length 36-40 units, while PBMT outperformed NMT for longer sentences. When analysing the output similarity of all translations out of English, it was interesting to see that NMT led to considerably different outputs compared to PBMT, even though the systems belonged to the same paradigm. This, however, can be explained by the fact that concepts are represented by numeric vectors. Although NMT output was more fluent due to its lower perplexity and performed better in terms of inflection and reordering, the differences with PBMT regarding lexical errors were much smaller and inconsistent.

The investigation conducted by *Koponen et al.* (2019) evaluated the strengths and weaknesses of the different MT systems from a different perspective, namely that of a post-editor. 33 translation students, from now on post-editors, were asked to correct the output of an NMT, a statistical MT (SMT), and a rule-based MT (RBMT) for the language pair English-Finnish. The selected source text (ST) was obtained from the WMT16 news task dataset, which in turn came from the BBC website. This ST, which was later subdivided into 165 subsegments, contained 27 sentences and 385 words. The analysis of the resulting post-edited texts was performed taking two approaches into consideration: a product-based and a process-based approach. With respect to the product-based approach, in other words, the differences in the distribution of edit types between the examined MT systems, some of the results differed slightly from those of the previously described studies. For example, in the case of word form changes such as verb forms, while the output of NMT required fewer edits than that of SMT, the amount of these edits was still greater than in the case of RBMT. This could be due to the selected language pair since morphologically rich languages have proven to be challenging for NMT systems. Moreover, mistranslations, lexical errors, and omissions were rather common in the sentences produced by NMT, although the number of insertions needed was greater in the output of SMT. Regarding the post-editing of RBMT, this system was the one that presented the highest amount of word order changes. Furthermore, the number of deletions when post-editing the translation of this system was more than double of the output of the other MT engines combined. It is worth mentioning that NMT and SMT posed some problems in terms of ambiguity, which were, however, handled correctly by RBMT in most cases. In regard to the process-based approach, the technical, cognitive, and temporal PE efforts were measured. It was observed that the number of keystrokes, and consequently the technical effort, was higher when post-editing the NMT output. Nevertheless, the length of the pauses was shorter compared to the case of the other systems, which involved a significantly bigger cognitive effort. It should be noted that the aspects of effort did not necessarily correlate since some errors were easy to spot but needed much editing, while some other errors were quickly corrected but were challenging to identify.

While it is true that the above-reported investigations were focused on the study of the strong and weak points of the existing MT systems, the conclusions extracted were slightly different since they did not follow the same approaches. It has therefore been deemed necessary to summarize these results. As could be deduced, NMT systems have the highest amount of strengths of all the examined MT systems: they deal better with diversity than any other SMT system; are more accurate with word reordering; work well with verbs and pronouns; are more fluent; perform better in terms of inflection; provide morphologically more correct translations and with fewer lexical errors. The last couple of features depend on the language and text typology, though. The main strength of SMT systems (either PBMT or SBMT) is their performance with long sentences, which is significantly better than that of NMT. When it comes to RBMT, one of its strong points is that it deals well with ambiguity. However, as discussed earlier, no MT system is perfect. Among the weaknesses of NMT systems, it should be noted that their performance with prepositions, negative particles, and articles is as poor as that of SMT; they have difficulties in operating with long sentences; have more problems with word forms than RBMT; present a higher number of mistranslations, lexical errors and omissions than RBMT; and the post-editing of their output involves more technical effort than that of SMT and RBMT. With regard to SMT, these systems exhibit more lexical errors than NMT; omit more information; and its post-editing implies a more cognitive effort. Lastly, RBMT systems deal badly with word reordering; and their output contains many extra words in comparison with the source text.

As pointed out at the beginning of this section, the results obtained from the previously described studies can be of great help in developing tools that address more specifically the strong and weak points of the MT systems. End-users would be the most likely to benefit from these tools since they would be provided with precise indications to make the most optimal use possible of the output.

2.3 Methods for MT evaluation

The previous section revealed that, although current MT systems of all kinds present several strengths, their weaknesses, and therefore their imperfect performance is no less relevant. That is why a revision of the output is needed, in particular, when such output is used for dissemination purposes. While this task is more accessible for professional translators, it poses a great problem for lay users. This is, first, due to the fact that the latter do not necessarily possess the required translation skills, and second because they may not have sufficient linguistic competence. To enable end-users to make the most out of the good qualities of MT systems and to overcome the bad ones, it is believed necessary to provide them with post-editing guidance. For example, this could be done by identifying potential errors in the translation. Already existing tools developed for evaluating the quality of MT systems might be an option to fulfill this purpose. To know to what extent these tools can be useful for lay users considering the kind of supplied information, they will be explored in more detail in the paragraphs below. The quality estimation method will be described first, followed by the traditional automatic evaluation metrics. However, it is worth noting that the analysis of these metrics will be done from a particular perspective. As most of them require reference translations in order to work, it is already assumed that they would not be useful for lay users. Nevertheless, if these metrics were based on quality estimation, what information could they provide to end-users to assist them in using the MT output in the most profitable way?

2.3.1 Quality estimation

An area of growing interest for NLP applications, especially within the context of those that produce natural language as output (e.g., MT), is quality estimation (QE). Its goal is to provide an estimate of the quality or reliability of the results returned by any of these applications without the need for gold-standards (*Specia et al.*, 2018). This derives from the willingness to adapt NLP applications to real-world settings, where the demand for information about the output quality is continuously increasing, and the access to reference outputs is difficult or almost impossible. It is worth mentioning that QE can be made at sentence-level, but also at word-level, phrase-level, and document-level. However, while it is essential to have different levels for different applications, sentences are the most natural unit for QE. In fact, readers and many NLP applications, such as MT, tend to focus on one sentence at a time when dealing with translations. Since most research in QE has been made around applications that are directly targeted at end-users, QE models have been built in such a way that the needs can be met. Such applications range from estimating the post-editing effort to support the work of translators, to gisting information from social media or e-commerce platforms. QE models need to be composed of at least a few thousand examples, which are in turn annotated using different types of labels and described through a number of features. The annotating labels can either be binary (0,1; good, bad ...), range from 1-4, or be based on a Likert scale. When it comes to the features, four main groups are established: complexity features (extracted only from the source), fluency features (extracted only from the target), confidence features (extracted from the MT system), and adequacy features (extracted from both the source and target sentences). It should be pointed out that QE uses machine learning methods to assign quality scores. The nature of the previously mentioned labels serves to choose the most suitable algorithm. In other words, labels represented as continuous scores (BLEU, HTER, post-editing time...) lead to choosing regression algorithms such as linear regression, random forests, or single- and multi-layer perceptron, among many others. However, discrete labels (binary, 1-5 point scale ...) require the election of classification algorithms, such as naive Bayes or SVM.

It is not surprising that QE has become a very popular method to evaluate the quality of MT. Actually, one of its main strengths is the fact that not only MT developers or professional translators can profit from the results provided by the method, but also end-users of any NLP applications, at whom QE is especially aimed. It is not only valuable that there is no need for having reference translations in order to obtain an estimation score of the output quality of MT systems, but also that it can work at any level. In fact, word- and phrase-level QE can be particularly useful for lay-users with any level of knowledge of the target language since words or phrases that are not reliable and hence require attention or revision appear highlighted within the MT output.

2.3.2 Automatic evaluation metrics

Although human evaluation has many advantages, such as the fact of being extensive and providing reliable and very detailed information, it also presents a large amount of disadvantages, especially in terms of cost (Papineni et al., 2002). Not only does it take long to annotate data, but it is also rather expensive to hire professionals to perform those annotations. Moreover, it should not be forgotten that human evaluation is marked by a high level of subjectivity (Snover et al., 2006). This is why it has been determined to create automatic evaluation methods. The following sections will describe in more detail some of the state-of-the-art automatic metrics most commonly used in the field of MT evaluation, ranging from the traditional BLEU, (H)TER, METEOR, or Hjerson, to the recently developed COMET. Besides providing details on their functioning, a few reflections on their strengths and weaknesses will be made. This will be done taking their usefulness for lay-users into consideration, who are the object of the present study. As pointed out in the introduction to this section, the advantages and disadvantages of these metrics will be analyzed under the assumption that they are based on quality estimation. In other words, it will be discussed what information each of them could provide to end-users if there were no need for reference translations for their execution. It is worth noting that the description of the metrics will be made based on their true nature. The hypothetical case of absence of references will only be addressed when discussing their strengths and weaknesses.

2.3.2.1 BLEU

BLEU (*Papineni et al.*, 2002) is one of the most widely used metrics for automatic MT evaluation since it is not only quick and language-independent, but it also correlates well with human evaluation and has a little cost per run. It works at a sentence-level by measuring the closeness of a machine-translated sentence to at least one reference human translation relying on a numerical metric. This metric ranges from 0 to 1, being 1 the highest possible score, which means that the hypothesis is identical to the reference. It should be noted that the amount of reference translations plays a key role when calculating the results, in that the more they are, the higher the score. BLEU evaluates the previously mentioned translation closeness by comparing n-grams of the hypothesis with the n-grams of the reference, and counting the matches, which are position independent. As can be deduced from the maximum feasible result, the higher the number of matches, the better the hypothesis is. This precision-based metric first focuses on computing the unigram matches, and then moves to longer n-gram matches. It is interesting to point out that these matches can also be used to assess adequacy when taking the unigrams into account, and fluency when observing the length of the n-grams. By using the n-gram precision score, BLEU is able to differentiate between two or more machine translations with a similar quality and distinguish between two or more human translations with a different quality. The length of the hypotheses is also of great importance for calculating the final score. Although the n-gram precision penalizes all words that appear in the hypothesis but not in the references, it is not completely reliable for ensuring that the candidate is either too long or too short. This is solved by adding a brevity penalty, which is computed over the entire corpus and rewards the hypotheses that match the references in length.

Although it is widely used for the development of MT systems, the results that BLEU provides can be very difficult to interpret by non-experts. Furthermore, the fact that this metric is highly sensitive to the number of given references makes it more costly than it seems at first since it requires the prior work of professional translators. Lastly, it should be noted that the approach to recall made through the brevity penalty might result in extremely low scores for short sentences, which are not necessarily incorrect nor have missing information.

Now, if, as said at the beginning of this section, this metric did not require reference translations (that is, it can be learnt to predict BLEU through QE models) nor provided exclusively numerical information, it could supply some useful indications to the end-users. For example, it would highlight the words that appear in the target and that are also present in the source text. In addition, some special font could be used to indicate whether a sequence of words in the target is aligned with the source without alterations of order, insertions, or omissions. However, the fact that it would only display indications in the output would prevent from knowing if the information is missing. Perhaps all the words in the target would be in the source, but not all the words in the source would be in the target. Moreover, attempting to solve this by penalizing sentence length may not be entirely effective. It might be the case that either synonyms have been used or that the target language requires slightly fewer words to express the same meaning as the source language (and vice versa).

2.3.2.2 (H)TER

Translation Edit Rate (TER) is an automatic measure that evaluates the output of an MT system by computing the minimum amount of editing that a human would have to perform in order to change a machine-translated sentence so that it exactly matches a reference translation (*Snover et al.,* 2006). That explains why, as opposed to other automatic evaluation metrics, the lower the score, the better the quality. It should be pointed out that, in the case of more than one reference, only the number of edits of the closest one will be calculated. The possible edits include insertion, deletion, and substitution of a single word, as well as shifts of word sequences. Interestingly, all of these edits have equal cost. Moreover, punctuation tokens are considered normal words, and the incorrect capitalization counts as an edit. However, since the scores provided by this metric do not always reflect the acceptability of the hypothesis, a slightly new version of TER has been created. Human-targeted TER (HTER) consists of finding the minimum number of edits to be performed in a new targeted reference. In other words, a fluent speaker of the target language creates a new reference translation targeted for the system output by editing the machine-translated sentence until it has the same meaning as the other reference(s) and is fluent. It is noteworthy that HTER with a single targeted reference reduces the edit rate by 33% compared to TER with 4 untargeted references. Furthermore, it has been found that HTER correlates better with human judgements than individual human judgements correlate with each other.

As one can conjecture, one of the weaknesses of HTER is its cost compared to other automatic metrics since it takes some time for a human to annotate the sentences. However, the high cost in terms of time does not necessarily correlate with the economic cost, since the annotator does not have to be bilingual, which is notably cheaper. Moreover, this is also supported by the fact that this metric is less sensitive to the number of references compared to other automatic metrics. An important strength of HTER that should be pointed out is its high correlation with human judgements and its low subjectivity in contrast to them. Finally, as is the case of most automatic metrics, the results provided by HTER might not be very meaningful for lay-users. However, if HTER, besides not needing references, would provide users with more information than merely numeric scores, it would be very useful for them in order to use the MT output efficiently. By highlighting the insertions, deletions, substitutions of single words, as well as shifts of word sequences, users would know whether the MT system has entered extra information, missed some information, or re-ordered the existing one. These users, however, would be in charge of judging whether, for example, the shifts lead to a correct or an incorrect translation since the metric would not provide such feedback. We must remember, however, that HTER, similarly to the other metrics described here, works on a word-form level. This means that the metrics do not compare or understand the meaning of the words but rather focus on the characters each consists of. This, we believe, is highly limiting.

An approach to test the usefulness of HTER if it was based on QE has been made by *Esplà-Gomis et al.* (2018). Their research made a valuable contribution to future systems whose creation may be of great help to professional translators. However, not only this target audience could take advantage of this potential development since the proposal might well be suitable for all types of users. Most investigations in terms of QE have been made at a sentence level, and in many instances, the information obtained from the method is a score for the whole machine translated sentence that serves to know whether it is worth post-editing that sentence or not. This research, on the other hand, focuses on the word level. In this case, the words of the machine translation that need to be edited, that is to say, replaced or deleted, are automatically identified. Although this already eases the work of post-editors to a great extent, the researchers felt the need to include a new factor, namely the insertion positions. In summary, the main goal of the study was not to know the total amount of words to be inserted but where they had to be entered. This was successfully accomplished by using sources of bilingual information (SBI), specifically three MT (Apertium, Lucy, and Google Translate) and a bilingual concordancer (Reverso Context), to extract features that were later used by neural networks (NNs) for making predictions of words and insertions. They evaluated several different feature sets and NNs on two publicly available datasets (WMT15 for the language pair English-Spanish, and WMT16 for the language pair English-German). The newly created method, compared to those that do not identify the gaps, gave very competitive results using considerably fewer features. Actually, among their experiments, the simultaneous identification of word deletions and word positions achieved better results than just detecting word deletions. The next goal for the investigators is to adapt the method to the sentence level in order to be able to predict the total post-editing effort required for a sentence, which up to now is mainly done with HTER (*Snover et al.*, 2006).

Although there is still much to be done, the previously described approach is shaping up to be of great help for users, regardless of their expertise. It does not only provide very useful information for post-editing the output of MT systems, but it also presents a great advantage compared to other evaluation metrics since it does not need a reference translation to perform its task. This makes it less-costly and faster than other methods. Moreover, it is also intended to assist users with no knowledge of the source language when using MT for assimilation since it would provide information about the reliability of the translation.

2.3.2.3 METEOR

The METEOR (*Lavie and Denkowski*, 2009) automatic metric for MT evaluation has also become very popular within the field of MT development, especially because of the on-going efforts to keep improving it. This metric measures the lexical similarity between a machine-translated sentence and one or more reference translations by creating word alignments between them. These alignments do not consist only of exact words, i.e., words that have identical surface forms, but also of stem words (two words with identical stems) and synonymy words (two words considered synonymous when sharing synonyms sets according to an external database, such as WordNet). Once the word-to-word matching has been performed, METEOR provides a score between 0 and 1 to each sentence. The score has proven to correlate well with human judgements of translation quality since this metric, besides precision, relies on recall. Furthermore, METEOR features a fragmentation penalty that accounts for the preservation of word order and three free parameters, namely controlling the relative weights of precision and recall in the Fmean score (initially set at 0.9), controlling the shape of penalty as function fragmentation (initially set at 3.0), and the relative weight assigned to the fragmentation penalty (initially set at 0.5). As these parameters can be tuned, METEOR is suitable for several languages other than English. Actually, the stemmers used by the automatic metric already include support for other European languages.

METEOR attempts to address some of the weaknesses found in BLEU and proves to be successful. The most remarkable strength of this automatic metric is the fact that it takes stemming and synonymy into account, which makes the scores more reliable. However, as is the case with other automatic metrics, the results provided by METEOR, while helpful for developing MT systems, are not very clarifying for lay-users. Therefore, if it did not only provide a numerical score nor needed reference translations, this metric could be very useful for these users. First of all, it would display the alignments between the words, allowing them to know what information is present and what is missing within the translations. Moreover, it would also indicate whether the word order has been altered or not. While it is true that order modifications can change the overall meaning of a sentence resulting in a bad translation, this is not always the case. End-users would be the ones to judge whether these re-orderings are relevant or not. In addition, the fact of giving a score to each sentence might help to get an idea of whether it is worth post-editing or translating it from scratch.

2.3.2.4 Hjerson

Hjerson (*Popović*, 2011) is a tool for the automatic classification of errors in MT output. This tool detects five word level error classes, namely morphological errors, reordering errors, missing words, extra words, and lexical errors. The choice of these error classes was based on the work of *Vilar et al.* (2006). Hjerson implements a method based on the Word Error Rate (WER), which in turn derives from the Levenshtein edit distance algorithm, combined with the precision and recall error rates. In order to run the tool, it is essential to have at least a reference translation (in the case of having more than one, the tool will make use of the reference translation with the lowest WER score), a hypothesis translation, and the base forms of both reference and hypothesis translations. While Hjerson is a language-independent tool that has been tested in various language pairs and tasks, the base forms are indispensable. Otherwise, the lack of them, especially in the case of morphologically rich languages, may result in undetected errors and noisy output. Additionally, it is possible to call the tool with further parameters, such as part of speech (POS) tags, to get more detailed results. The default output is a file containing the overall raw counts and error rates; however it is also possible to get those computations for each sentence. The most user-friendly output option, though, is an HTML file that includes the original sentences with visualized error categories, using different colors and font styles, such as pink and italic for inflectional errors, green and underlined for reordering errors, blue and bold for missing and extra words, and red, bold and italic for lexical errors. Moreover, a text file containing all the original words tagged with their corresponding error category can be obtained as well.

As stated above, one of the main strengths of Hjerson is its speed when classifying and analyzing errors, which is much faster than what could be done by humans. In addition, it is worth noting that the results of this tool have a high correlation with those obtained by human evaluators. Moreover, although the default output may not be very simple to interpret by non-specialist users, Hjerson provides the opportunity to obtain more visual results where the end-users can get a better idea of the existing errors and their location. When it comes to weaknesses, the fact that the tool requires at least one reference translation limits its use a bit since it is not possible for a great number of MT users to comply with that parameter. This applies also to the need for base forms, given that not all users know what they are and/or how to get them. However, if, as discussed in the previous metrics, neither references nor the base forms were necessary, the output of Hjerson would indeed be particularly useful for end-users. This way, they would know whether there are changes in the inflection of terms, word re-ordering, missing or extra words, or if the lexical choice is not entirely correct by comparing the source and the target text. Post-editing would therefore be much easier since they would know exactly where to make the modifications if necessary.

2.3.2.5 COMET

COMET (*Rei et al.*, 2020) is a PyTorch-based neural framework that aims to train with different types of human judgements highly multilingual and adaptable MT evaluation models, which in turn can function as metrics. The quality of the MT output is more accurately predicted compared to other automatic metrics due to the fact that the models make use of information from both the source input and the target-language reference translation. In fact, inspired by QE models and distancing itself from traditional evaluation metrics, attempts have been made to achieve high levels of correlation with human judgements without even making use of a reference translation. COMET supports two distinct architectures, namely the estimator model, and the translation ranking model. The first one is trained to regress directly on a quality score, while the second one is trained to minimize the distance between a good hypothesis and both its reference and original source. To demonstrate their effectiveness, two versions of the estimator model and one of the ranking model have been trained with data from three different corpora. COMET-HTER is the first version of the estimator model that regresses on HTER and has been trained with the QT21 corpus. This corpus, which is made of tuples with the source sentence, the human-generated reference, the MT hypothesis, and the post-edited MT, is a publicly available dataset containing industry-generated sentences from an information technology and life sciences domain. The second version of the estimator model, COMET-MQM, regresses on the proprietary implementation of multidimensional quality metrics (MQM) corpus, which is a proprietary internal database of MT-generated translations of customer support chat messages. In this corpus, English appears only as of the source and never as the target language. The last MT evaluation model, COMET-RANK, is a version of the translation ranking model that has been trained with the WMT DARR corpus from 2017 and 2018. This corpus is a collection of human judgements in the form of adequacy direct assessments (DA) and is available in both high and low-resource language pairs. All previously described models have proven to be successful not only when working with translations with English as the source or as the target language, but also with language pairs that do not involve English at all.

Although with the growing interest in neural networks, efforts have mainly been made to improve MT systems, there have been contributions of enormous value in the field of MT evaluation. A case in point is COMET. The main strength of this framework is the fact that it is not essential to have a reference translation in order to evaluate the quality of the MT output. It is fast, low-cost, and very reliable. Moreover, it has a high correlation with human judgements. The only drawback is perhaps, as is also the case of some of the metrics described in the previous sections, that the results may not be easily interpretable by non-specialized users. However, it is true that the metrics could give them an accurate idea of the quality of the MT output, and then it is up to the users to decide what to do with it.

2.4 Interpretable Semantic Textual Similarity

If, as discussed above, the evaluation metrics described so far were based on QE and did not require any reference translations to operate, they could provide quite useful information to users so that they could have an idea of the quality of the MT output. However, it should be remembered that these metrics mainly focus on the form of the words that make up the automatic translations but not on their meaning. How could users be indicated that the source text is semantically equivalent to the target text?

Could they be warned that, although it seems to be a perfect translation, both texts are the complete opposite with respect to meaning?

Having information about semantics, in addition to word form, could be of great help for users when facing the post-editing task. Semantic Textual Similarity (STS) could contribute to provide this information. Although it has not yet been applied to the field of MT with the aim of giving feedback to end-users, it is thought that this technique measures the level of semantic equivalence between two sentences (*Agirre et al.*, 2015) through a score range from 0 to 5 could have potential to assist users in taking the most out of the MT output. Nevertheless, its interpretable version (iSTS) is believed to be even more useful for them (*Agirre et al.*, 2016). This is because it will no longer provide an overall numeric score on the degree of similarity between whole sentences, but instead, it will analyze the sentences by small parts, namely chunks. In addition to giving those chunks a score, they will also be assigned an informative label. These tags may indicate either that the chunks are equivalent (*EQUI*), or that their meanings are in opposition (*OPPO*), or that they are semantically similar (*SIMI*), or that one is more specific than the other (*SPE1/SPE2*), or that their meanings are related (*REL*), or that there is no semantic equivalent in the other sentence being compared (*NOALI*). Furthermore, although both STS and iSTS have been designed primarily for English, several attempts have been made to apply them to a multilingual context (*Cer et al.*, 2017). This is a further reason to consider the application of these techniques to the field of MT.

In the present research, the potential of iSTS for assisting lay users in post-editing will be explored. This measure, which could serve to indicate the existing differences between the source and the target text, will be described in more detail in the following chapters.

3 General objectives

The present investigation is divided into two parts. The aim is to give answers to two objectives that, although different, are complementary to each other.

As discussed in the previous chapter, the uses of MT systems are many and diverse. Also, the types of users who can benefit from the output of these tools are very different. To delimit the scope of the research, it has been decided to address the use of MT for dissemination, also closely linking it to the use of MT as a pedagogical tool. This is because the study subjects have the status of both lay users and language learners. The aim of this first part is to analyze the texts that lay users create when writing directly in the foreign language (FL) and the compositions they produce when post-editing the output of an MT system where the source is a text produced by themselves in their L1. Moreover, while some studies have focused on either users with an advanced command of the FL or on basic users, this investigation will attempt to cover different levels of proficiency. It has also been decided to conduct the study with a single language pair, specifically Polish-Spanish. The choice was motivated by the desire of not working with languages belonging to the same linguistic family in order to avoid interferences.

To examine the similarities and differences between these texts, the focus will be set on their complexity. To this end, several aspects will be investigated. Are sentences longer when users write directly in the FL or when they do it first in their L1? Is there more diversity of parts of speech (POS) in the post-edited texts or in those written directly in the FL? Which compositions have a higher proportion of content words and are therefore more informative? Are the texts written without the aid of MT easier to read than the post-edited ones? Which of the writings contains more basic vocabulary? Do both texts have the same syntactic structures? Do the post-edited texts read more like the FL or the other way around? These and other questions will be addressed in the following chapter. Although it is presumed that the users will have a better command of the L1 compared to the FL, the previously mentioned weak points of the MT systems must also be taken into account. Could users be provided with some indications on how to overcome these weaknesses when post-editing the MT output?

This is where the second objective of this research comes into play. In the state-of-the-art chapter, a measure that calculates the degree of semantic equivalence between two sentences was presented. Although this could be applied to the field of MT, its interpretable version is believed to have the greatest potential for creating a special tool that could assist users in the post-editing task. Instead of working with whole sentences, iSTS operates with chunks. Moreover, those chunks are assigned a score and a label that indicate not only the type of differences between two sentences but also how big those differences are. This could be really useful for giving feedback to users on the quality of the automatic translations. To explore its potential in this area, it would first be necessary to investigate whether the original design of iSTS would fulfill this goal or whether modifications would be required. How would users benefit from it? What information would iSTS provide them compared to the previously described metrics? This will be addressed in more depth in the second part of this investigation.

4 Analysis of lay users writings with and without MT assistance

4.1 Methodology

As stated in the objectives chapter, in this first part, an analysis of the similarities and differences in the compositions that lay users create when writing directly in the FL and when post-editing the automatic translation of a text they had produced in their L1 will be performed. To this end, a single language pair (Polish-Spanish) has been studied and different levels of proficiency have been taken into account. In the following subsections, it will be described in more detail who the participants were, what type of texts they had to produce to conduct such an analysis, and what their attitude was towards online language tools and MT systems before and after completing the experiment.

4.1.1 Participants

The study was carried out in Poznań, Poland, between December 2020 and February 2021. However, due to the unprecedented pandemic situation, the vast majority of participants performed it online. On the one hand, this way of proceeding presented some challenges, which will be described in the following sections. On the other hand, it enabled the involvement of some participants who were neither in the city nor the country at this time.

Although the profiles of the participants varied, they all had something in common, namely their L1 (Polish) and the fact of being learners of Spanish as a foreign language. The first attempt to recruit participants for this research study was made at a local language academy. It was considered to be the ideal place in order to make a proper division among the proficiency levels. However, the number of students who were willing to take part in the experiment was not large enough to extract reliable conclusions. Therefore some of them were asked whether any of their acquaintances were learning Spanish. In the end, the number of participants was 21. Although more

Poles were interested in joining the research, some of them dropped out of the process before completing all of the tasks.

As commented above, there were not many commonalities among the participants. That is why it was decided to conduct a survey prior to performing the experiment, with the aim of gathering some data that could influence, the final results on a certain level,. Knowing the level of the learners, for example, was essential for the subsequent analysis, as it would allow drawing more reliable conclusions. In addition, it was interesting to find out whether the participants knew other foreign languages and how they used them. The aim was to investigate if the knowledge of additional languages could have any influence on the users' writing. Moreover, having basic information about their academic background and employment status was also of great importance since the goal of this study was to work with lay users. In fact, the presence of translators or language professionals could affect the results.

The age of the Spanish learners ranged between 20 and 55 years. According to CEFR, three of them (14.29%) were basic users (A1-A2); 10 (47.62%) were independent users (B1-B2), and 8 (38.10%) were proficient users (C1-C2). Aside from Polish and Spanish, all of them had some knowledge of another foreign language. Specifically, English was the primary L2 (90.48%), followed by German (52.38%) and French (38.10%). Other mentioned languages were Russian, Portuguese, Ukrainian, Italian, Basque, and Chinese. It was also interesting to see what all these languages were used for. While many agreed on the utilization of Spanish for leisure (watching films/series, listening to music, reading) and touristic purposes, communication and work purposes were the most mentioned causes for using their other L2s. Last but not least, the participants were asked about their academic background and employment status. It is worth mentioning that the great majority had studied either a degree in the field of Arts and Humanities (42.86%) or Social and Legal Sciences (42.86%). The remaining participants had a background in the domains of Engineering and Architecture (19.05%), Health Sciences (9.52%), or Science (4.76%). Those numbers add up to more than 100%, because some participants were qualified in more than one field. At the time of the experiment, a large number of participants were unemployed (28.57%). Concerning the active ones, Education (19.05%) and IT (19.05%) were the most often mentioned sectors of work, followed by Human Resources (9.52%), Economics and Finance (9.52%), Culture and Creative Industries (4.76%), Healthcare (4.76%), Law (4.76%), and Tourism and Leisure (4.76%). As was the case with educational background, the total of percentage exceeds 100%, since a couple of the participants belonged to more than a single working sector.

It is worth noting that the number of participants with a high command of Spanish was substantially higher than the number of basic users. Therefore, although some conclusions could be drawn, the results would not be as reliable as those of the groups with a larger number of learners. As regards the background, it was very diverse, which meant that the participants were perfectly adequate for conducting the experiment.

4.1.2 Design of the experiment

The study was divided into three parts. As commented above, the first one was a survey in which participants were asked several questions about their command of Spanish as well as of other languages, their academic background, their employment status, and their attitude towards MT and other online language tools such as dictionaries, grammars or conjugators. Some of the data extracted from this first survey served to establish a profile of the participants and were therefore described in the previous section. However, the outcomes related to the users' attitudes towards MT and other online language tools will be described in more detail in the upcoming sections. The second part was, itself, subdivided into two parts. The first subpart consisted of a single step, which was that the participants had to write a text directly in Spanish. From now on, reference will be made to this text as *esDIR*. The second subpart involved three steps. For the first step, the users were required to write a composition in their L1 (from now on *plDIR*). For the second step, the *plDIR* text was pasted into an MT system and automatically translated from Polish into Spanish (from now on *esMT*). And finally, for the third step, the learners were told to correct the MT output (from now on *esPE*). Once they were done with all the writing tasks, the participants had to fill out another survey where they were asked about their experience when performing the previous part. This was the third and last part of the experiment. The next sections will cover in more detail how the above-mentioned parts were structured.

4.1.2.1 Writing tasks

The aim of the present investigation was to simulate, to the maximum extent possible, a scenario in which participants could make use of MT systems coupled with post-editing for producing written content in the foreign language. To this end, it was necessary to determine the number of parameters. Among them, the size of the texts to be written was an important aspect to consider. Therefore, the experiment was designed based on the guidelines set out by Instituto Cervantes for the Diplomas de Español como Lengua Extranjera (DELE), since they are official, internationally recognized certificates accrediting the level of competence of the Spanish language. In order to be suitable to all proficiency levels, it was determined that the ideal length of the texts to be written should be 200 to 250 words. However, as will be discussed in the upcoming chapters, more than half of the participants produced greater length content. Regarding the text type, it was opted for a travel blog. Regardless of how it may seem, the decision was not random. First, it should be taken into account that the targets of our study were lay users, so it was necessary to reflect on what situations they would need to make use of MT engines. Among the considered options were also product reviews and comments on social media (*Specia et al*, 2018). However, they were discarded for being, on average, too short. Second, it was desired to have creative texts in order to pose some challenges to the machine translator. Third, it should not be forgotten that the participants were learners of Spanish, meaning that the topic had to be appealing to them since they had to make an effort to write in a foreign language. Moreover, not only were blog formats included in the guidelines of DELE, but traveling is one of the few things learners of foreign languages have always had in common. The participants, who received the instructions in both Spanish and Polish to facilitate their complete understanding of the study, were not informed about the text type until just before starting to write it to prevent them from being biased.

The research was carried out as a series of virtual meetings via Zoom. These meetings were adapted to the availability of the participants. The number of learners per encounter ranged between 1 and 3. The researcher was always present to guide them and assist them in the event of issues. Depending on the request of the learner, the instructions were provided either in Spanish, Polish, or English. As mentioned above, the requirement of proceeding online led to some challenges. The main problem to be faced was connected to the part in which the participants had to produce the *esDIR* texts. Even though they could use any online language tool (monolingual, bilingual, or contextual dictionaries, grammars, conjugators), it was essential that none of them made use of MT engines. To ensure that this condition was fulfilled, they were asked to record their screens while performing the experiment. The screen recorder chosen for this purpose was OBS Studio. The choice was based on the fact that the software was free and open-source, and its interface was user-friendly. Since it was only needed to have proof of their screens, the participants were told to disable both desktop and microphone audio. Once solved this issue, it was time to select a word processor where the learners could type their texts in either of the subparts. Google Docs was considered the most suitable tool, given that the files could be easily accessed by both the participants and the researchers. A further advantage was the fact that all participants were familiar with the text editor. Although some were more aware of the wide variety of options offered by the tool than others, since they were using it on a daily basis, none of them had technical difficulties in completing the task. To prevent the participants from modifying the texts some time after the experiment, the documents were downloaded as soon as it was reported that the assignment had been completed.

As commented above, they were informed about the text type to be produced just prior to writing. A short brainstorming exercise was carried out beforehand to spare them the blank-page syndrome. Just like it happened with the instructions, this activity was carried out in all three aforementioned languages. The initial purpose was to avoid, to the maximum extent, any possible use of Spanish to not activate the vocabulary in their minds in advance. However, it turned out to be more complicated than what was thought. It is worth noting that, while they were asked to write texts on the same subject in the first and second subpart, the participants were requested to create entirely new content when writing in Spanish and in Polish. To prevent them from performing translations between the languages, they were encouraged to start by drafting in Spanish the introduction to their blog, that is their first post, and continue by producing in Polish their second post. To ensure that the compositions were not similar, they were told to describe one trip when creating the blog post in Spanish and report a totally different trip when writing in Polish. However, they were not forbidden to establish connections between the narrated trips. It would have been ideal to keep in secret all the steps and reveal them little by little to prevent them from being biased, but since their screens had to be recorded, some of them were distrustful and wished to know every single detail in advance. Eventually, it was decided to provide all participants with some general information so that they would all be in the same conditions. Not everyone was entirely satisfied with it, though, and decided not to participate in the end.

Due to time constraints, the tasks had to be done one right after the other. There were small breaks in between to allow the researchers to download the documents and explain, in more detail, how to proceed in the following steps. It should be mentioned that, even though there was no time limit, the first task was the most costly. On average, the participants needed 40 minutes to write the *esDIR* texts. Furthermore, the step related to post-editing was, on average, the next one in terms of time-consuming. Around 17 minutes took the users to complete this task in contrast to the average of 15 minutes that cost them to produce the *esPL* blog posts. This was somehow surprising since it was initially predicted that writing a text from scratch, even in the L1, would take longer than revising the output of an MT system.

The selected MT engine to perform the translations from Polish to Spanish was Google Translate. As occurred in previously reported cases, the choice was not random. Since it was desired to emulate the way users proceed within the private sphere to the greatest extent possible, several Poles were asked in a loose, informal conversation external to the research about their experience with MT. Curiously, the only MT service mentioned was the one developed by Google. The decision was also based on the fact that it is nowadays a state-of-the-art neural MT tool2 whose performance and quality have improved over the years, becoming a reference service.

It was difficult to simulate with high accuracy the way to proceed when using online MT services given the many different contexts where the translations can be performed. That is why, at the time of performing the post-editing, it was opted for creating two columns in the text file where the participants had typed their blog post in Polish. The column on the left contained the *plDIR* text, and its machine-translated version (*esMT*) was pasted in the column on the right. This gave the participants the opportunity to easily compare the texts since they had both versions at the same height. The idea of placing the texts on different pages was discarded because having to scroll up and down, besides not being practical, could have led to ignoring some of the errors made by the MT system. The participants were then asked to revise the MT output and correct it until they considered it could be posted on their imaginary blog. They were required to carefully verify that all the information they had produced in their L1 was present in the translation. To post-edit the automatic translation, the learners did not have to indicate the changes in a different color or format nor add comments. They simply had to remove or add to the automatically generated text anything deemed necessary to create a flawless translation. To that end, similarly to the way they proceeded when writing directly in Spanish, the participants could take advantage of any language tool, including the spell checker of the word processor.

4.1.2.2 Attitude towards MT and online language tools

As discussed earlier, the learners had to fill in a survey prior to beginning with the writing tasks and another one after finishing them. Both surveys could be carried out in both Spanish and Polish. In the previous sections, a reference was made to one of the goals of the first survey, which was basically gathering more information about the participants in order to set different profiles that could be useful for the analysis. However, this was not the only aim of this initial survey. It was deemed interesting to study whether the attitude of the users towards MT systems and other online language tools remained the same before and after the experiment.

To this end, the Poles involved in this investigation were asked, prior to performing the writing tasks, whether they make translations to their L1 from any of the languages they know (and vice versa) on a daily basis. A total of 9.5% responded that ‘never’, while 9.5% said that ‘always’. Almost half of the participants (47.6%) indicated that they do it from time to time, 23.8% that they do it almost always, and 9.5% that they hardly ever do it. In addition, they were also asked whether they use any tool to help them with this task, such as monolingual dictionaries (WSPJ, PWN, RAE...), bilingual dictionaries (Pons, Reverso...), contextual dictionaries (Linguee, Context...), grammars (Diccionario panhispánico de dudas, Nueva Gramática de la Lengua Española...), or conjugators (Reverso, Pons...). 19% of the participants indicated that they always use language tools to make translations, 33.3% that almost always, 42.9% that from time to time, and only a 4.8% stated that they never do it. The vast majority

[2] https://cloud.google.com/translate/docs/languages?hl=en

(85.7%) mentioned that they use bilingual dictionaries. This was closely followed by contextual dictionaries since a total of 71.4% confirmed to make use of this tool. Conjugators are used by 52.4% of the participants, monolingual dictionaries by 38.1%, and grammars by 28.6%. The Spanish learners were also given a chance to indicate other tools of which they made use. Those mentioned included: Wiktionary, WordReference, pl.bab.la, Glosbe, and Wikipedia. They were later told to rate on a scale from 1 to 5 the degree of usefulness of these language tools, *1* being 'not useful' and *5* 'very useful'. None of them considered them to be not useful. More than half (57.1%) claimed that they are, in fact, very useful. 23.8% of the users scored their usefulness with a 4, and 19% did it with a 3. The next questions were about their attitude towards MT. That is why they were requested to grade on a scale from 1 to 5 how frequently they used any MT service, *1* being 'never' and *5* 'always'. It was surprising to see that only 4.8% always make use of MT systems as opposed to 19% who stated that they never do it. 14.3% claimed that they do it almost always, while 23.8% reported that they hardly ever do it. 38.1% of the participants admitted that they take advantage of automatic translation from time to time. As was the case with the previously mentioned language tools, the learners were also requested to evaluate on the same scale from 1 to 5 the degree of usefulness of machine translators in general. In contrast to the outcomes obtained for the same question regarding the language tools, there were users that stated that MT systems are not useful at all (9.5%) or very little useful (9.5%). Only 19% indicated that these engines are very useful. 38.1% of the participants remained impartial, while 23.8% leaned towards its usefulness. A further question was added in order to know whether they were satisfied with the quality of the MT tools they had previously used or not. To this end, another scale from 1 to 5 was selected, 1 being in this case 'not satisfied at all' and 5 'very satisfied'. It was striking to observe that, although 23.8% of the participants said that they are quite satisfied with the quality of MT systems, none of them selected the highest degree of satisfaction. In fact, 9.5% of the users stated that they were not satisfied at all, and 28.6% that they were not very satisfied. The remaining 38.1% expressed a fairly neutral judgment. The results of the first survey confirmed the hypothesis about the users' attitude towards machine translation. While it is true that the results were not entirely negative, there is a certain reluctant attitude towards its use. Would these opinions remain the same after the experiment?

The second survey, which also corresponded with the last part of the study, aimed to learn more about how the participants perceived the experiment once they had performed all the aforementioned tasks. The participants were first asked whether they had used any specific tool as an aid for producing the text directly in Spanish. A total of 85.7% of the learners indicated that they did use language tools, while 14.3% stated that they did not do it. Afterward, they were required to specify these tools and evaluate their level of usefulness on a scale from 1 to 5, *1* being ‘not useful’ and *5* ‘very useful’. More than half of the participants (71.4%) pointed out that they had used bilingual dictionaries. Other tools selected were conjugators (28.6%), contextual dictionaries (23.8%), grammars (14.3%), and monolingual dictionaries (4.8%). As was the case with the first survey, an empty field was provided so that they could indicate the specific tools they used if they wished to. Some of the ones mentioned were: Wiktionary, Wikipedia, Glosbe, pl.bab.la, Thesaurus, the Google Docs spell checker, and even their own notebook. Regarding the level of usefulness of the previously detailed tools, almost half of the participants (42.9%) stated that they were quite useful, and 28.6% said that they were very useful. A neutral response was given by 19% of the users. The percentages obtained from indicating the non-usefulness (4.8%) or low usefulness (4.8%) of the tools were very low. The same questions applied to the part related to post-editing. Similar to the first survey, the learners were later asked to grade on a scale from 1 to 5 how useful the selected MT service was for them, 1 being ‘not useful’ and 5 ‘very useful’. In this case, almost half of the participants (47.6%) stated that it had been quite useful, and 19% even selected the highest degree of usefulness. For 28.6% of the users, having the output of an MT system did not seem to play an important role. Only 4.8% indicated that it was not useful at all. The learners of Spanish were later asked whether they were satisfied with its quality or not, 1 being in this case ‘not satisfied at all’ and 5 ‘very satisfied’. The outcomes were very different from those obtained by posing the same question before performing the experiment. Almost half of the participants (42.9%) marked the highest score with regard to their degree of satisfaction with the quality of the selected MT system. 33.3% were quite satisfied, and 19% remained neutral. Only 4.8% indicated that they were not very satisfied. It should be noted that none of the participants expressed absolute dissatisfaction with the quality of the automatic translations. It is important to emphasize that these were not general questions but related to the tasks they had just completed. The participants were then requested to indicate the deficiencies they had observed in the machine translations and to briefly describe what and what not did the tool help them most for. The answers to these questions can be found in the Appendix. The next question was whether they would use any MT engine for producing content in any language different from their L1 again. To answer this, they were presented with a scale from 1 to 5, 1 being ‘never’ and 5 ‘always’. The outcomes were rather evenly distributed. 9.5% of the participants claimed that they would never use any MT system again; 19% said that they would not use it in most cases; 19% indicated that they would use it always; 23.8% stated that they would use it often, and the remaining 28.6% had a neutral opinion. Lastly, the participants were provided with an empty box so they could make comments or observations about the study. These comments and observations can also be found in the Appendix.

4.2 Analysis

Once the dataset containing the blogs produced by the Spanish learners was obtained, it was time to analyze it. A total of four texts per participant were available, i.e., the one written directly in Spanish (*esDIR*), the one written directly in Polish (*plDIR*), the automatic translation of the Polish text (*esMT*), and the post-edited version of the MT output (*esPE*). However, the following subsections will be based on examining the behavior of the users when writing texts in Spanish with and without the support of an MT system. That is, the analysis will be performed on *esPE* and *esDIR* texts. This behavior will be examined from the perspective of linguistic complexity. The goal was to study whether participants produced more complex compositions when working directly with the foreign language or when operating first in their L1 and then automatically translating the content into Spanish. The analysis was performed both at a general level, i.e., using the totality of the texts, and at competency level. In other words, the complexity of the blog posts written by basic users (A1-A2), independent users (B1-B2), and proficient users (C1-C2) was evaluated.

The first aspect in terms of complexity to be analyzed was the number of sentences, words, and characters per text. Although the participants were required to produce compositions with a minimum amount of terms, the aim was to see whether they limited themselves to that minimum or exceeded it and, in case they exceeded it in both set-ups, to examine if there were substantial differences between them. Furthermore, while the number of these elements alone already provided some information, analyzing them as a whole gave an initial idea of the length of the sentences and words in each text. The metric selected for this purpose was text length. The next subsection would serve to extend the information obtained from the previous measure. Although the situation could already be inferred, it was determined to study the sentence length to observe precisely how big the differences were. After knowing the number of words per sentence, it was desired to examine which terms composed them. First, the presence of some basic grammatical categories in the texts was analyzed since the use of some POS would require a good command of the language. This was done with the lexical proportion metric. Then, the goal was to find out how diverse these words were. Even if there were great amount of some of these grammatical categories considered more complex present in the texts, there might also have been a relevant number of repetitions. This would not necessarily mean a higher degree of complexity. For this purpose, the lexical variety metric was used. Subsequently, the degree of informativeness supplied by the texts was evaluated. To this end, it was necessary to discard functional words (prepositions, conjunctions…) and focus on the proportions of words that had meaning on their own, i.e., nouns, adjectives, verbs, and adverbs. A high ratio of these types of words would indicate that the writer has produced texts in a more explicit way. This was measured with the lexical density metric. The next thing to be observed was the number of complex words in each blog post, in other words, how many terms with more than three syllables were present. A high number of those words would suggest that the text was more complex compared to another one composed of shorter words. The selected metric was readability. The next step was to analyze the specific vocabulary contained in the blog posts: were the terms used within the basic vocabulary that a learner of Spanish should know, or did they go beyond? To do this, a comparison of the words found in these texts was made against three lists containing the 1000, 5000, and 10000 most frequent words in *Corpus de Referencia del Español Actual* (CREA). Then, it was studied whether there were more complex syntactic structures in one of the set-ups compared to the other. For this purpose, a semantic role labeler was used to provide the type and frequency of occurrence of all the syntactic structures present in the texts. Not all of them were analyzed. In fact, only 7 of them were selected as indicative of complexity. Finally, efforts were made to evaluate how far the compositions written by the participants were from a language model. To find out which of the set-ups read more like Spanish, the metric used was perplexity. Since the goal was not only to analyze the texts at the surface level but also at the structure level, two language models were created, namely a token-based model and a POS-tag-based model.

In the following subsections, the objectives of the aforementioned metrics, along with the results obtained, will be discussed in more detail.

4.2.1 Text length (sentences, words, characters per text)

Although the experiment was designed in a way that the participants had to create blog posts with a minimum amount of words, it was deemed interesting to observe whether they had limited themselves to that minimum or they had exceeded it. Moreover, it was desired to make a comparison between the text length of the *esDIR* and *esPE* texts. The purpose of examining this aspect was none other than seeing the potential usefulness of MT systems for producing longer and consequently more complex texts. However, the focus was not only on the number of terms included in those texts, but also on the number of characters, which provided an initial idea of the length of the words, and on the total number of sentences.

As discussed earlier, the participants were asked to produce texts with a minimum amount of words (200-250). Although this was also a requirement that had to be fulfilled in other experiments, not all succeeded due to the fact that there was a time limit (*Chon et al.*, 2021). However, in the present study, the learners could take as long as they needed to write their blog posts. As can be observed in Table 1, this is why not only was the minimum met, but it was greatly exceeded.

The number of sentences, words, and characters per text was obtained using UNIX Shell commands. First, the average and the standard deviation were calculated for all participants, and then the language proficiency levels were considered. Even though each one of the elements was important for the investigation, they had to be analysed together in order to get more reliable conclusions. In fact, the number of sentences combined with the number of words and characters provided a first impression of how long the sentences were. This, however, will be commented on in detail in the upcoming sections.

Interestingly, the number of words and characters was noticeably higher in the post-edited texts than in the ones directly written in Spanish. This could be due to several reasons. First, although participants were required to produce blog posts with a minimum of words, no maximum was set. When writing the texts in their L1, the learners of Spanish appeared to feel more confident and exceeded the specified minimum by far. However, this could also be explained by the fact that the MT system might have included new functional items that were not present in the original Polish text, such as prepositions or articles. Nevertheless, the number of sentences was slightly larger in the first blog posts, except in the case of the advanced learners. This may imply that MT was of great help for the participants to produce more elaborated texts since a decrease in sentences and an increase in words lead to longer and consequently more complex sentences. Expectedly, the quantity of words and characters increased in accordance with the proficiency levels. The basic users (A1-A2) created texts with a length more limited to what they were asked to produce. As the competency increased, the learners were more confident and wrote the tasks disregarding the minimum of words. Such a phenomenon was particularly evident in the post-edited texts.

		esDIR			***esPE***		
		sentences /text	**words/text**	**characters/text**	**sentences /text**	**words/text**	**characters/text**
General	**Average**	19.905	257.762	1544.667	19.048	288.286	1733.905
	Std. dev.	5.243	40.843	264.692	4.307	58.071	326.701
A1-A2	**Average**	22	231	1436.333	16.333	258.667	1558.667
	Std. dev.	7.937	22.338	303.638	0.577	20.133	157.246
B1-B2	**Average**	21.1	259.4	1556.5	19.5	279.9	1690.6
	Std. dev.	4.408	53.596	346.623	4.994	60.537	350.893
C1-C2	**Average**	17.625	265.75	1570.5	19.5	309.875	1853.75
	Std. dev.	5.069	23.383	115.616	4.140	61.629	327.610

Table 1: Results of the count of the sentences, words and characters per text

4.2.2 Sentence length (words per sentence)

Based on the premise that the bigger the amount of words per sentence, the more complex the texts are, another aspect that was desired to analyze was the sentence length of the participants' compositions. (*Jagaiah,* 2017) suggested that this measure serves to give an idea of the syntactic complexity of the texts. In fact, according to (*Yan and Xu*, 2017), the longer the length, the better the writer's command of syntactic structures is, along with the vocabulary. Both studies agreed that sentence length is a reliable metric to determine the quality of a written text. In the present investigation, it was attempted to make a comparison between the *esDIR* and *esPE* blog posts in order to see, even if only superficially, whether the post-edited ones were more complex than those written directly in Spanish or the other way around.

In Table 2, it can be observed that the sentence length was greater in the post-edited texts in all cases, regardless of the language proficiency level. While the general median length of the texts directly written in Spanish was 11.5, and their average length was 13.782 with a standard deviation of 8.947, the median length of the post-edited texts was 14, and their average length was 15.684 with a standard deviation of 9.705. This pattern already started to be apparent in the previous section.

The number of words per sentence was calculated with UNIX Shell commands. The median, the average, and the standard deviation were computed in Excel after having entered all the data of the participants.

Although indeed, there were not many differences in the case of advanced learners, the contrast between the *esDIR* and *esPE* texts produced by the basic users was substantial. When looking at the average length of the *esDIR* texts, the results differed a lot among the levels. However, the *esPE* texts had a similar length in all instances. This could suggest that the use of MT was of great help for the participants, especially for those who had a lower command of Spanish. Knowing that they had the possibility to first write the texts in their L1 and later to translate them automatically encouraged them to produce more extensive sentences.

		Median	***Average***	***Standard deviation***
General	***esDIR***	11.5	13.782	8.947
	esPE	14	15.684	9.705
A1-A2	***esDIR***	9	10.500	7.015
	esPE	15	15.837	8.719
B1-B2	***esDIR***	11	12.294	7.665
	esPE	14	14.428	8.819
C1-C2	***esDIR***	13	15.078	10.928
	esPE	14	15.891	10.960

Table 2: Results of the word count per sentence

4.2.3 Lexical proportion

Which grammatical categories were more present in the *esDIR* texts compared to *esPE* texts? Which less? To provide a proper response to either of these questions, it was required to compute the lexical proportion in both set-ups. (*Aranberri*, 2020) used this metric to make an initial approach to measure lexical complexity. It is worth noting that studying the frequency of appearance of nouns and verbs, as they are the most basic categories, would give a less reliable view of the complexity of the texts than examining the proportion of articles, adjectives, or adverbs, which are modifiers of those basic categories. However, it should be borne in mind that the use of this metric alone was not enough to extract reliable conclusions. A high proportion of a grammatical category that required a high command of the language, such as prepositions or conjunctions, did not necessarily mean that the texts were more complex. In fact, there could have been many repetitions, and this is not exactly an indicator of complexity.

With this in mind, the proportion of the grammatical categories was computed by dividing the total amount of every POS by the sum of all tokens that made up the texts. However, it should be noted that the percentages did not have to be calculated manually since the tool Analhitza (*Otegi et al.*, 2017) provided all the numbers. The grammatical categories listed on the application were the so-called basic categories: nouns, adjectives, verbs, adverbs, determiners, conjunctions, and prepositions. The first four are content words, i.e., words that have meaning on their own, while the remaining three are functional words, i.e., words that have some grammatical function in the sentence but whose meaning is difficult to define. Content words are framed within the open class, that is to say, a class to which new items are constantly added. On the other hand, functional words are part of the closed class, which is a class that does not admit such frequent incorporation of new members.

Although a couple of guesses were made beforehand, it is worth mentioning that there were some unexpected results, as can be seen in Tables 3 and 4. It was first predicted that, since nouns and verbs were the most basic categories, their presence would be higher in the *esDIR* texts. This was, however, only entirely fulfilled in the texts produced by the A1-A2 learners. On the other side, these basic users altered the predictions in the case of adverbs, which were considered to be a more complex category and therefore more present in the post-edited blog posts. However, what was more surprising was the large occurrence of conjunctions in the texts that were directly written in Spanish since the use of this POS usually requires a good command of the language.

The lack of articles in Polish may lead to some difficulties for learners of Spanish as a foreign language. This might explain why the frequency of this grammatical category was higher in the *esPE* texts. Also, the fact that this Slavic language is so highly inflected could justify a general lower appearance of prepositions in the *esDIR* blog posts since it is not that simple to understand when to use this POS.

		nouns	**adjectives**	**verbs**	**adverbs**	**determiners**	**conjunctions**	**prepositions**
General	**Average**	24.40%	7.79%	17.33%	6.22%	14.30%	7.06%	13.98%
	Std. dev.	0.050	0.026	0.026	0.016	0.024	0.019	0.027
A1-A2	**Average**	30.02%	10.48%	16.06%	7.74%	11.96%	5.98%	11.12%
	Std. dev.	0.055	0.058	0.030	0.014	0.046	0.018	0.049
B1-B2	**Average**	23.67%	7.38%	18.20%	5.75%	14.42%	7.29%	14.10%
	Std. dev.	0.047	0.017	0.020	0.017	0.016	0.019	0.021
C1-C2	**Average**	23.21%	7.30%	16.73%	6.23%	15.02%	7.18%	14.91%
	Std. dev.	0.043	0.017	0.032	0.012	0.022	0.019	0.021

Table 3: Results for the lexical proportion metric on *esDIR* texts

		nouns	adjectives	verbs	adverbs	determiners	conjunctions	prepositions
General	Average	24.83%	8.02%	15.54%	6.94%	15.30%	6.36%	14.72%
	Std. dev.	0.035	0.023	0.028	0.019	0.019	0.017	0.024
A1-A2	Average	25.33%	10.88%	14.75%	7.00%	14.10%	5.47%	15.90%
	Std. dev.	0.016	0.019	0.016	0.014	0.015	0.030	0.019
B1-B2	Average	24.51%	7.32%	15.65%	7.12%	15.35%	6.71%	15.24%
	Std. dev.	0.035	0.017	0.028	0.022	0.023	0.017	0.021
C1-C2	Average	25.04%	7.84%	15.70%	6.71%	15.68%	6.25%	13.63%
	Std. dev.	0.043	0.024	0.033	0.019	0.016	0.010	0.027

Table 4: Results for the lexical proportion metric on *esPE* texts

4.2.4 Lexical variety

The degree of diversity in the terms used to produce a text could also provide an idea of the lexical complexity of that text. If a text has a high level of lexical variety, this means that the writer has used several different words with little repetition (*Johansson*, 2008). It is therefore assumed that participants with greater language proficiency will have more lexical resources and hence will write texts with higher lexical diversity. Some of these resources would be synonyms and hyponyms. The goal of using this metric was to examine whether the blog posts produced by the participants in their L1 and subsequently translated with the selected MT system were more diverse lexically than *esDIR* texts.

As in other research studies (*Aranberri*, 2020), the measure selected for this purpose was the type-token ratio (TTR). This metric is computed by dividing the total number of unique words (types) by the whole amount of words (tokens). Similar to the previous section, it was still required to operate with POS. Therefore, Analhitza (*Otegi et al.*, 2017) was the tool chosen to carry out this task. The application developed by the IXA group did not only provide not only the POS, the types, and the tokens but also the TTRs. The results can be seen in Tables 5 and 6.

Previously, it was commented that the frequency of occurrence of conjunctions was higher in *esDIR* texts. However, the level of diversity of this grammatical category was significantly larger in the post-edited blog posts. Likewise, it was striking to observe that there was a wider variety of determiners and prepositions in the texts that were directly written in Spanish than in *esPE*. In fact, the difficulties that these categories pose to Polish learners of the Romance language must be remembered. The case of nouns was quite special. The differences between both set-ups were not as overwhelming as with the other POS. This could be explained by the typology of the texts the participants had to produce. Travel blog posts are usually characterized by a significant presence of nouns, in particular proper nouns, which may be repeated more than once within a text. In general, except for the more advanced learners of Spanish, the remaining categories seemed to be more diverse in *esPE* texts. As commented before, the use of adjectives and adverbs responds to more complexity since they are not basic categories but serve to modify them. This may account for the wider variety of these grammatical categories in the texts produced with the aid of MT.

		nouns	**adjectives**	**verbs**	**adverbs**	**determiners**	**conjunctions**	**prepositions**
General	**Average**	78.24%	87.83%	61.21%	67.12%	24.75%	31.33%	25.04%
	Std. dev.	0.068	0.094	0.114	0.095	0.080	0.101	0.071
A1-A2	**Average**	80.22%	85.32%	55.67%	57.97%	28.26%	27.37%	32.11%
	Std. dev.	0.068	0.165	0.149	0.049	0.167	0.034	0.116
B1-B2	**Average**	76.24%	87.88%	56.51%	65.76%	23.37%	32.97%	22.01%
	Std. dev.	0.079	0.088	0.081	0.082	0.062	0.108	0.060
C1-C2	**Average**	80.01%	88.72%	69.16%	72.26%	25.15%	30.76%	26.16%
	Std. dev.	0.052	0.085	0.103	0.098	0.068	0.113	0.048

Table 5: Results for the lexical variety metric on *esDIR* texts

		nouns	adjectives	verbs	adverbs	determiners	conjunctions	prepositions
General	**Average**	77.13%	90.69%	63.88%	67.18%	21.95%	36.39%	23.65%
	Std. dev.	0.080	0.059	0.097	0.097	0.056	0.088	0.055
A1-A2	**Average**	78.20%	85.67%	65.98%	66.55%	22.92%	41.07%	22.59%
	Std. dev.	0.079	0.076	0.153	0.105	0.074	0.078	0.006
B1-B2	**Average**	77.23%	92.29%	63.47%	69.98%	23.29%	34.74%	23.24%
	Std. dev.	0.094	0.059	0.094	0.087	0.063	0.072	0.063
C1-C2	**Average**	76.60%	90.56%	63.61%	63.92%	19.91%	36.69%	24.58%
	Std. dev.	0.070	0.048	0.094	0.107	0.040	0.111	0.058

Table 6: Results for the lexical variety metric on *esPE* texts

4.2.5 Lexical density

Lexical complexity can also be measured in terms of density. Lexical density serves to assess the degree of informativeness of a text. In other words, the higher the number of lexical words, the more specific and detailed the content will be and, therefore, more complex. Taking the example of the research of (*Aranberri*, 2020), the lexical density of both *esDIR and* esPE texts was computed by dividing the sum of content words, namely nouns, adjectives, verbs, and adverbs, by the total number of tokens.

Again, as occurred with the previously described metrics used to evaluate the lexical complexity of the blog posts, Analhitza (*Otegi et al.*, 2017) was the selected application for obtaining the proportions of the different grammatical categories. In this case, the percentages were manually computed since the tool did not provide that information.

Although when looking at the general results (Table 7), the difference between the proportions of content words in both *esDIR* and *esPE* texts was not substantial, it should be pointed out that the post-edited texts had a fewer amount of these words. In fact, it was the opposite of what was expected. Nevertheless, the most surprising situation came when taking the language proficiency levels into consideration. The most significant contrast was found in the texts produced by the basic users. However, this did not necessarily mean that they were more complex. As commented above, blog posts feature a large number of nouns. Therefore, to reach more reliable conclusions, it was necessary to perform a small manual analysis. The previously mentioned tool, Analhitza, provide not only the proportions of the existing content words but also a list of the specific words that were within each of the grammatical categories along with their frequency. This confirmed that the participants with a lower command of Spanish had a tendency to use a substantial amount of proper nouns when writing directly in that language than when producing texts in their L1. Actually, some *esDIR* texts include a very large sequence of names of cities and countries. This might be explained by the fact that they were required to comply with a minimum length.

In summary, this metric alone did not help much to assess the lexical complexity of the texts.

		esDIR	***esPE***
General	**Average**	55.75%	55.34%
	Std. dev.	0.048	0.026
A1-A2	**Average**	64.30%	57.96%
	Std. dev.	0.078	0.015
B1-B2	**Average**	55.00%	54.60%
	Std. dev.	0.023	0.026
C1-C2	**Average**	53.47%	55.28%
	Std. dev.	0.022	0.026

Table 7: Results for the lexical density metric

4.2.6 Readability

Readability has been used in some investigations as a further element to assess the level of complexity of a text. However, it should be noted that some research studies have not applied the metrics for measuring the readability in itself, but only one of the components that serve to calculate it. For example, the Gunning FOG index, which is a well-known readability formula, is based on the word length for giving a score. Therefore, (*Chon et al.*, 2021) resolved to compute the percentages of complex words, namely the words with more than three syllables, to determine the rarity of the texts.

The assumption was that the texts produced with the help of MT would contain a larger amount of longer words than those written directly in the foreign language.

In the present study, an attempt was made to emulate the way to proceed as proposed in the previously mentioned research. The same criteria were followed to detect complex words in the texts produced in Spanish. The number of words per syllable length was supplied by the online tool *Legible*[3]. The proportions were calculated by dividing the sum of these words by the total number of tokens that made up the texts. As can be observed in Table 8, the number of complex words was, in general, slightly larger in *esPE* texts. The only exception was found in the case of the advanced learners, who seemed to have knowledge of a non-trivial vocabulary.

		esDIR	***esPE***
General	**Average**	9.04%	9.70%
	Std. dev.	0.029	0.016
A1-A2	**Average**	9.50%	9.76%
	Std. dev.	0.046	0.012
B1-B2	**Average**	8.11%	9.78%
	Std. dev.	0.024	0.015
C1-C2	**Average**	10.04%	9.59%
	Std. dev.	0.028	0.020

Table 8: Amount of complex words (more than 3 syllables)

While calculating the proportions of complex words helped to get an initial idea of the degree of rarity within the blog posts, it was desired to analyze the readability scores by using some of the available metrics. It is worth mentioning that the formulas

[3] https://legible.es/

are language-dependent, which means that the ones used in the studies that served as a model for the current research are of no use for working with Spanish texts.

Efforts were made to apply metrics that were as similar as possible to those of English. After an extensive search, it was decided to assess the readability of the Spanish texts by using two formulas: first, the Flesch-Szigriszt Index (*Barrio-Cantalejo*, 2008), which is an interpretation of the formula conceived by Szigriszt-Pazos (Szigriszt-Pazos, 1992) and is calculated as follows:

$$I = 206.835 - \frac{62.3S}{P} - \frac{P}{F}$$

I being the INFLESZ index; *S* - the number of syllables; *P* - the total amount of words; *F* -the number of sentences.

And second, the Fernández-Huerta Index (Fernández-Huerta, 1959), which is calculated as follows:

$$L = 206.84 - 0.60P - 1.02F$$

L being readability; *P* - the average of syllables per word; *F* - the average of words per sentence.

Both of them were computed by the program INFLESZ (*Barrio-Cantalejo*, 2015). The average and the standard deviation of the results can be found in Table 9. In order to get valuable insights from data, it is important to know how to interpret the scales (Table 10). In general, participants have produced slightly more complex texts with the help of the selected MT service. However, both *esDIR* and *esPE* texts are considered to be (slightly) easy, meaning that there were no big differences in terms of readability. The most noteworthy thing, though, is the fact that, according to these formulas, basic users have produced more complex texts when writing directly in Spanish than when post-editing. This phenomenon could be explained by the presence of an outlier. While the results of this metric observed for the *esPE* text of this concrete participant did not contrast much with those of the other participants, the readability score obtained from the *esDIR* text attracted all the attention. According to that outcome, the blog post produced by this participant, who had a low command of Spanish, was substantially more complex than the blog posts written by very advanced users. Therefore, it has been decided to recalculate the readability of the basic users disregarding the mentioned participant. The average score for the INFLESZ readability metric of the *esDIR* texts was 82.305, with a standard deviation of 8.690, and 86.545, with a standard deviation of 8.195 for the Fernández-Huerta metric. In the case of the *esPE* texts, the average score for the INFLESZ metric was 77.55, with a standard deviation of 2.645, and an average of 81.865, with a standard deviation of 2.609 for the Fernández-Huerta metric. These new results show the same trend as those of the other participants; in other words, the texts produced with the aid of an MT system tend to be more complex in terms of readability than those written directly in the foreign language.

		esDIR		***esPE***	
		INFLESZ	**F/H**	**INFLESZ**	**F/H**
General	**Average**	76.267	80.691	74.117	78.565
	Std. dev.	8.505	8.224	4.668	4.531
A1-A2	**Average**	72.913	77.483	73.553	77.980
	Std. dev.	17.389	16.731	7.171	6.977
B1-B2	**Average**	77.339	81.747	74.003	78.482
	Std. dev.	6.799	6.568	3.605	3.510
C1-C2	**Average**	76.184	80.574	74.471	78.889
	Std. dev.	7.374	7.198	5.529	5.351

Table 9: Results of the readability metrics *Flesch-Szigriszt Index* (INFLESZ) and *Fernández-Huerta* (F/H)

Flesch-Szigriszt Index (INFLESZ)		**Fernández-Huerta Index**	
< 40	Very difficult	**90-100**	Very easy
40-55	Slightly difficult	**80-90**	Easy
55-65	Average	**70-80**	Slightly easy
65-80	Slightly easy	**60-70**	Average
> 80	Very easy	**50-60**	Slightly difficult
		30-50	Difficult
		0-30	Very difficult

Table 10: Scale of difficulty according to *Flesch-Szigriszt Index* and *Fernández-Huerta Index*

4.2.7 Most frequent words (CREA)

In this subsection, efforts will again be made to evaluate the lexical complexity along with the lexical richness of the texts produced by the participants. It should be noted that this will be done from a substantially different perspective compared to the previously described attempts. Initially, the idea was to look for lists of words that learners at level X should know. However, although these types of lists were easily accessible for languages such as English or German and were even provided by reference institutions, no such possibility was found for Spanish. Therefore, it was determined to modify the approach slightly. Instead of searching for lists delimited by levels, it was attempted to find general lists that contained the most frequent words in Spanish. This was motivated by the idea that frequent terms are also basic words. The main goal of this metric was to see whether *esDIR* texts had more or less of these frequent words than *esPE* texts. This, in turn, would give an idea of whether the lexical richness of the participants was superior to that of the MT system or just the opposite. The *Real Academia Española* (RAE), which is the cultural institution of reference devoted to the linguistic regularization within the Spanish-speaking world, has made several lists publicly available 4 of the most frequently occurring words in the *Corpus de Referencia del Español Actual* (CREA). Specifically, these lists contained 1000, 5000, and 10000 most common words in current Spanish.

Before anything else, the texts produced by the participants were lowercased, cleaned from punctuation, and tokenized using UNIX shell commands. Afterward, a new document including a list of all the tokens was created for each text. Then, with the help of a Python script, the three lists of CREA were compared against each of the lists created for the texts of the participants. As a result, two new files were created: one that included the terms that appeared in both the blog posts and the reference corpus (from now on *found words*), and another one that contained the words that were present in the creations of the learners but not in CREA (from now on *not-found words*). Each token was added only once to either of these new files. In other words, if a token had already been included in any of the documents, it was discarded to avoid unnecessary repetitions. The percentage of found and not found words of every single text was

4 http://corpus.rae.es/lfrecuencias.html

subsequently calculated by dividing the number of these terms by the sum of unique words that constituted every blog post. The analysis began with the texts directly written in Spanish and continued with the post-edited ones. Once the ratios had been computed for each participant, the average and the standard deviation were calculated for all of them first regardless of their language proficiency level and then taking it into account.

First, the focus will be set on the overall percentages. As can be observed in Table 11, and just as expected, the amount of not-found terms decreases significantly when dealing with bigger word lists. However, it was interesting to see that, in general, almost half of the unique tokens present in the texts produced by the participants were not included within the 1000 most frequently occurring words in CREA. It is worth noting that the percentages obtained for the blog posts directly written in Spanish were substantially higher than those of the post-edited ones. While this may be explained by the tendency of MT services to display unvarying translations for some terms instead of using synonyms, this might also be a result of misspellings or foreign words present in texts of the learners that are not listed in the Spanish corpus. To come to a more reliable conclusion, this had to be combined with some manual analysis. To this end, the files containing the not-found words were consulted, specifically those obtained from the comparison with the list of the 10000 most frequent words in CREA. Efforts were made to detect the differences between the *esDIR* and *esPE* texts of each participant. The phenomena found were noted, and it was examined whether there were similarities among the texts. The analysis, although not exhaustive, gave an overview of the general situation.

Although, as predicted, the texts directly written in Spanish contained a large number of errors of various types (*acompanar, artistos, quiereis, pegueno, cominda, chocolada...*), this was not the only phenomenon appearing in the not found lists. Colloquialisms (*finde, uni, chavales, culito, chulas...*) as well as proper nouns and foreign words (*poznań, rusałka, cytadela, dron, hobby, blog, posts...*) had also a strong presence. In contrast to the post-edited ones, these first blog posts included several expressions of laughter (*jaja...*) and interjections (*hmm...*).

As occurred with the previously described texts, the post-edited blog posts were characterized by a high number of proper nouns and foreign words (*nerds, croissants,* *potemkin, lviv, scooters, fans, autostop, roadtrip...*). Moreover, the terms seemed to be more complex than the ones used in the texts directly written in Spanish, particularly in the case of adjectives, such as *'prestigioso'* or *'perseverantes'*.

However, the lists obtained from both texts shared quite a few features. It seems that some verb tenses such as future or conditional are less likely to appear in CREA, as well as words that are relatively common in the singular but appeared in the texts in their plural form (*castillos, cementerios...*). Lastly, it should be noted that many participants typed terms related to the pandemic situation (*coronavirus, pandemia, confinamiento...*). Although they have become very frequently used and have already been collected in the reference corpus, there are still not among the 10000 most common words.

When taking the proficiency levels into account, the most surprising thing was that the texts created by basic users had a larger amount of not found words. This, however, can be explained by what has been mentioned before, that is, a big number of misspellings, proper nouns, and loanwords. Colloquialisms and complex verb tenses were rather present in the creations of independent and proficient users. Therefore, despite the fact that the results suggest a different conclusion, the manual evaluation plays in favor of the post-editing.

		esDIR			*esPE*		
		1000	**5000**	**10000**	**1000**	**5000**	**10000**
General	**Average**	46.30%	26.19%	17.49%	45.57%	23.30%	14.90%
	Std. dev.	0.060	0.071	0.067	0.062	0.050	0.037
A1-A2	**Average**	52.16%	32.53%	21.82%	50.12%	24.68%	15.36%
	Std. dev.	0.114	0.129	0.138	0.029	0.010	0.022
B1-B2	**Average**	45.18%	24.92%	16.44%	44.40%	23.06%	14.86%
	Std. dev.	0.045	0.058	0.061	0.053	0.032	0.032
C1-C2	**Average**	45.51%	25.40%	17.18%	45.33%	23.08%	14.76%
	Std. dev.	0.049	0.057	0.041	0.077	0.075	0.050

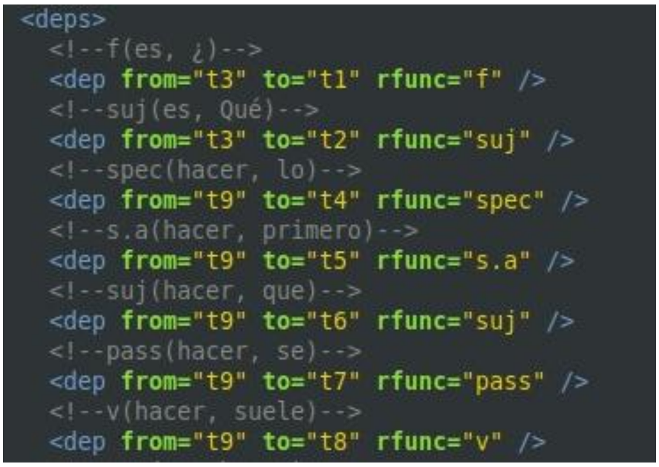
Table 11: Amount of not found words, namely the terms that were present in the texts produced by the participants but not in the list of most frequent words in CREA

4.2.8 SRL (rfunc)

In this subsection, complexity will be analyzed from a slightly different perspective, concretely, one in which both syntax and semantics come into play. As one can predict, the better the command of a language, the more complex the sentences produced by the learners will be. In other words, the participants with a higher level will not be limited to using basic structures such as subject, verb, and direct/indirect object or attribute, but they will tend to add more and more elements in order to express their ideas better. This principle also applies to the context of MT since the users first have to produce a text in their L1. In theory, the structures that make up that text should be more complex than if they had written it directly in any foreign language of which they had a lower command.

To examine the complexity of the sentences produced by the participants in both esDIR and esPE texts, it was decided to make use of a semantic role labeler (SRL). This tool will be responsible for assigning labels to basic elements in a sentence, such as words and phrases, which will indicate their semantic role in that sentence. Specifically, the SRL selected for this study was ixa-pipe-srl. This module, which has been developed by IXA group and forms part of the multilingual NLP IXA-Pipeline (*Agerri et al.*, 2014), provides a wrapper for the Spanish dependency parser and SRL based on Mate tools (*Björkelund* et al., 2009). Prior to using the module, it is required to tokenize and POS-tag the texts in NAF format. It is worth noting that the models used in this tool have been trained with PropBank, NomBank, and AnCora corpus in CoNLL 2009 Shared Task format (*Hajič et al.*, 2009).

Although both semantic role labelling and dependency parsing were performed on the texts, only the latter was taken into account for this research study, in particular *rfunc* values. This decision was made by considering what had been done in previous investigations (*Aranberri*, 2020).



```
<deps>
  <!--f(es, ¿)-->
  <dep from="t3" to="t1" rfunc="f" />
  <!--suj(es, Qué)-->
  <dep from="t3" to="t2" rfunc="suj" />
  <!--spec(hacer, lo)-->
  <dep from="t9" to="t4" rfunc="spec" />
  <!--s.a(hacer, primero)-->
  <dep from="t9" to="t5" rfunc="s.a" />
  <!--suj(hacer, que)-->
  <dep from="t9" to="t6" rfunc="suj" />
  <!--pass(hacer, se)-->
  <dep from="t9" to="t7" rfunc="pass" />
  <!--v(hacer, suele)-->
  <dep from="t9" to="t8" rfunc="v" />
```

Figure 1: Example of the dependency parser output in NAF format

As can be seen in Figure 1, the output was obtained in NAF format. All the values for *rfunc* were extracted with the aid of a Python script, and a new file was created for each text. Apart from the labels, the newly created documents also contained the frequency of occurrence. Specifically, 49 different tags were found in the texts. The tags assigned to these values were collected from the AnCora5 annotation guidelines. They provided information about POS, syntax, syntax-semantics, and named entities. However, the focus was set on the syntactical constituents and functions.

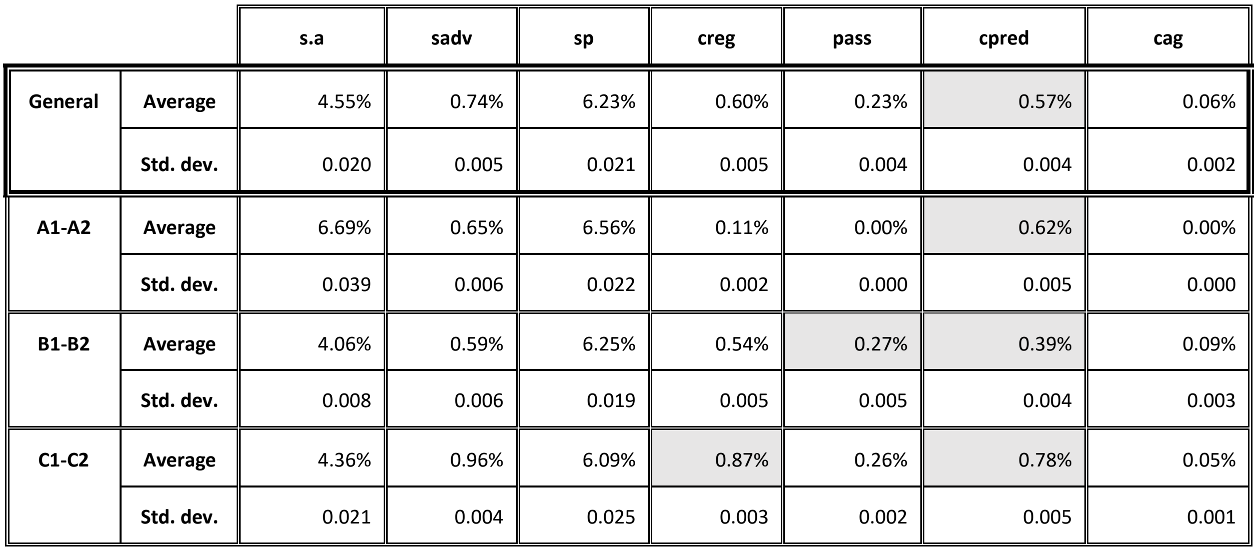
While it is true that the selection of the tags responded to subjective criteria, they served to give an insight into the differences in terms of complexity between *esDIR* and *esPE* texts. Regarding syntactical constituents, the labels chosen to be analyzed were *s.a* (adjective clause, e.g. *aventuras locas; platos extraños; fotos chulas*), *sadv* (adverbial phrase, e.g. *desde entonces; ciudades alrededor; hasta pronto*), and *sp* (prepositional phrase, e.g. *región de; posibilidad de; ganas de*). As for syntactical functions, the studied tags were *creg* (prepositional object, e.g. *hablar sobre; luchar por; sirven para*), *pass* (passive marker, e.g. *se dice; se puede; se construyera*), *cpred* (predicative complement, e.g. *llamado Mongolia; dejará boquiabierto; empezó suave*), and *cag* (agent complement, e.g. *recibidos por; visitada por; conocida por*). The choice of *rfuncs* based on adjectives and adverbs was motivated by the fact that the use of these categories requires a high command of the foreign language since they serve to complement nouns and verbs in order to provide more detailed information. Also, as pointed out in previous subsections, prepositions may pose a problem to speakers of

[5] http://clic.ub.edu/corpus/en/documentation

highly-inflected languages such as Polish. This is why it was determined to keep studying the behavior of the participants towards this category by selecting a couple of *rfuncs* constituted by prepositions. Last but not least, the use of the passive voice usually responds to a greater knowledge of the foreign language. Therefore, it was interesting to examine whether the differences between the texts written directly in Spanish and the post-edited ones when focusing on the number of passive sentences were big or not.

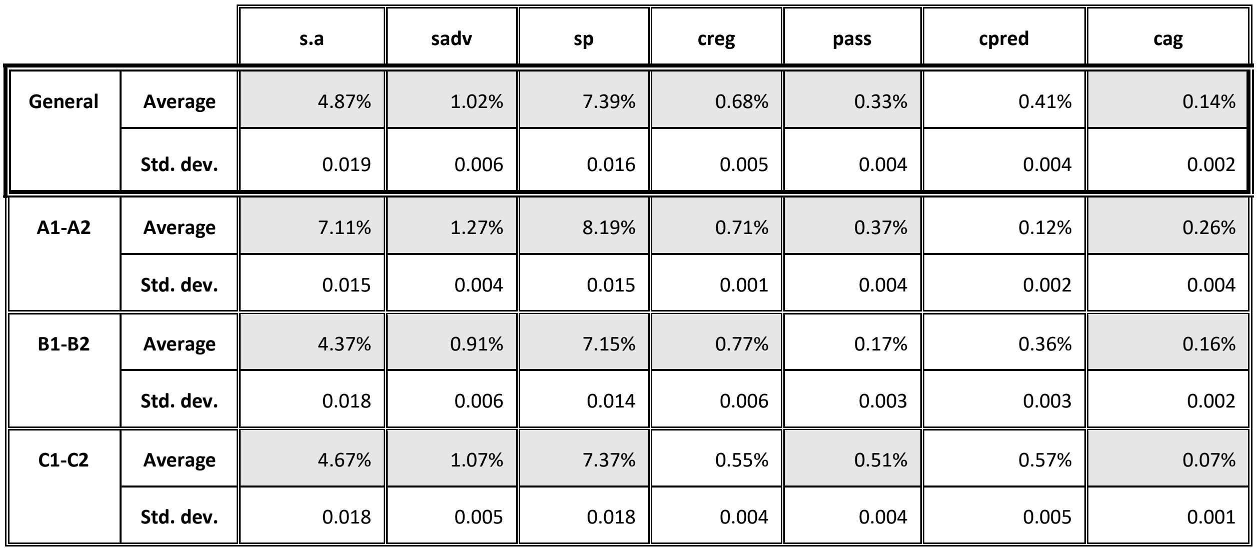
To perform the analysis, the proportion of the selected *rfuncs* was computed by dividing each one of them by the total amount of *rfuncs* within every text. As commented above, to observe more clearly whether MT helped the users produce more elaborate texts, it was determined to discard basic values such as subject, verb, direct/indirect object, or attribute since they were very likely to appear in almost every sentence. It should be remembered that the choice of some of the labels was proficiency-level-oriented, namely the ones related to the passive voice. As expected, these tags were not present in the *esDIR* texts written by basic users (Table 12). However, they did appear in the blog posts produced with the aid of MT (Table 13). At the same time, it was anticipated that proficient users were the ones to use more adverbs when writing directly in Spanish.

Nevertheless, when taking a look at the results of the *esPE* texts, there was a slightly higher occurrence of adverbial sentences in the texts of A1-A2 learners. The selected tags that more often appeared in the blog posts regardless of the competency level were those linked to adjectives and prepositions. It is worth mentioning that, with the exception of predicative complements, the results play, in general, in favor of post-editing, with the most significant difference being found in the texts of basic users.



		s.a	sadv	sp	creg	pass	cpred	cag
General	**Average**	4.55%	0.74%	6.23%	0.60%	0.23%	0.57%	0.06%
	Std. dev.	0.020	0.005	0.021	0.005	0.004	0.004	0.002
A1-A2	**Average**	6.69%	0.65%	6.56%	0.11%	0.00%	0.62%	0.00%
	Std. dev.	0.039	0.006	0.022	0.002	0.000	0.005	0.000
B1-B2	**Average**	4.06%	0.59%	6.25%	0.54%	0.27%	0.39%	0.09%
	Std. dev.	0.008	0.006	0.019	0.005	0.005	0.004	0.003
C1-C2	**Average**	4.36%	0.96%	6.09%	0.87%	0.26%	0.78%	0.05%
	Std. dev.	0.021	0.004	0.025	0.003	0.002	0.005	0.001

Table 12: Amount of syntactical constituents and functions in *esDIR*



		s.a	sadv	sp	creg	pass	cpred	cag
General	**Average**	4.87%	1.02%	7.39%	0.68%	0.33%	0.41%	0.14%
	Std. dev.	0.019	0.006	0.016	0.005	0.004	0.004	0.002
A1-A2	**Average**	7.11%	1.27%	8.19%	0.71%	0.37%	0.12%	0.26%
	Std. dev.	0.015	0.004	0.015	0.001	0.004	0.002	0.004
B1-B2	**Average**	4.37%	0.91%	7.15%	0.77%	0.17%	0.36%	0.16%
	Std. dev.	0.018	0.006	0.014	0.006	0.003	0.003	0.002
C1-C2	**Average**	4.67%	1.07%	7.37%	0.55%	0.51%	0.57%	0.07%
	Std. dev.	0.018	0.005	0.018	0.004	0.004	0.005	0.001

Table 13: Amount of syntactical constituents and functions in *esPE*

4.2.9 Perplexity

A further aspect of being measured was textual closeness. In other words, to see whether participants had produced texts (both with and without the help of MT) that read like Spanish. Taking as a model previous studies (*Aranberri*, 2020), the metric chosen for this purpose was perplexity. Furthermore, this metric is widely used in the field of MT in order to measure at what level an automatic translation suits a language model, which is nothing but a statistical model that assigns probabilities to words and sentences. A high probability would be equal to low perplexity, which means that a text is rather similar to the reference language model.

To create a language model, it is first essential to have a big corpus. It would have been ideal working with datasets containing blog posts, but nothing was found specifically for travel blogs in Spanish. Therefore, it was decided to create a more generic corpus that covered all types of topics and vocabulary, and that could give a general idea of how the texts produced by the participants were. CommonCrawl6, NewsCrawl7, and NewsCommentary8 were the datasets selected for this purpose. The number of 1.8 M lines was the sample size of CommonCrawl, which is a corpus that contains raw web page data, metadata extracts and text extracts. NewsCrawl and NewsCommentary are two datasets composed of news articles from news sites. The sample size of the former was 13.3 M lines, and 0.2 M lines of the latter. Since it was desired to calculate the perplexity at both token-level and POS-level, two language models were created. For the first model, it was first required to tokenize the datasets and then apply the truecaser9 of moses. This was done to keep words in their natural case instead of lowercasing all of them. A 5-gram language model was then created by using unpruned KenLM10 with modified Kneser-Ney smoothing. The second model was a 6-gram language model created based on POS information. The corpus was first tokenized and then POS-tagged with the aid of ixa pipes (*Agerri et al.*, 2014). Again, KenLM with modified Kneser-Ney smoothing, and no pruning was applied. The texts produced by the participants were pre-processed following the same steps, that is, they were tokenized and true-cased for calculating the perplexity with the first model, and tokenized and POS-tagged for measuring the textual closeness with the second model. The computation of the perplexity was done with the moses decoder11, and only the last four lines of the output were taken into account.

The results of the perplexity measured with the 6-gram model based on POS information will be analyzed first (Table 14). When observing the average of all participants, the perplexity obtained from the esDIR texts (5.949) is slightly higher than that of the esPE ones (5.533). Moreover, the standard deviation shows that there were not many discrepancies between the structures of all blog posts. As could be expected,

6 https://commoncrawl.org/
7 https://commoncrawl.org/2016/10/news-dataset-available/
8 https://www.statmt.org/wmt17/translation-task.html
9 https://www.statmt.org/moses/?n=Moses.SupportTools
10 https://github.com/kpu/kenlm
11 https://github.com/moses-smt/mosesdecoder

those who had a basic command of Spanish were the ones that produced texts that differed more from the language model than those written by participants with an intermediate or advanced level. However, the analysis of the text structure was not enough to extract reliable conclusions. Therefore, it was necessary also to make use of the model made of tokens.

The differences between esDIR and esPE when calculating their perplexity using the token-based 5-gram model were overwhelming (Table 15). The post-edited texts read more like Spanish than those written in that language without the help of MT. It is worth noting that the higher the level of Spanish, the smaller the differences between the results. Nevertheless, when looking at the general results, the most striking thing was the standard deviation. This was particularly the case of esDIR texts that included OOVs, that is, unknown words that were also scored by the language model. As pointed out in previous subsections, the presence of an outlier could explain the outstandingly high value for this measure. If the results of that concrete participant were excluded, the average perplexity of the esDIR texts would be 259.810, with the standard deviation of 185.071, when including OOVs; and the average of 224.659, with the standard deviation of 116.357, when not counting OOVs. The average of the esPE texts with OOVs would be 170.557, with the standard deviation of 61.266, while without OOVs the average would be 155.738, with the standard deviation of 48.429. Although the fact of not taking the results of the outlier into account does not result in large differences in the case of the post-edited texts, the average perplexity in the esDIR texts is reduced to almost half. Despite this, esPE texts are still more similar to the language model than those written directly in Spanish.

		esDIR	*esPE*
General	**Average**	5.949	5.533
	Std. dev.	0.751	0.438
A1-A2	**Average**	6.594	5.590
	Std. dev.	2.006	0.772
B1-B2	**Average**	5.884	5.514
	Std. dev.	0.260	0.478
C1-C2	**Average**	5.788	5.535
	Std. dev.	0.402	0.285

Table 14: Results of the perplexity metric with the 6-gram language model based on POS information

		esDIR		*esPE*	
		Including OOVs	**Excluding OOVs**	**Including OOVs**	**Excluding OOVs**
General	**Average**	436.258	307.949	172.046	157.934
	Std. dev.	828.463	398.178	60.104	48.263
A1-A2	**Average**	1415.917	740.921	173.345	173.345
	Std. dev.	2208.047	1067.836	38.206	38.206
B1-B2	**Average**	315.205	260.582	167.138	152.545
	Std. dev.	246.586	147.184	63.634	54.427
C1-C2	**Average**	220.203	204.793	177.695	158.890
	Std. dev.	65.087	60.278	68.019	47.803

Table 15: Results of the perplexity metric with the 5-gram language model based on tokens

4.3 Conclusions

Although the results have been discussed throughout the different subsections, a summary of the findings obtained from the analysis carried out in this first part of the research will be done within the following paragraphs.

It should be remembered that the goal was to examine the level of complexity of the texts produced by the participants when writing directly in the FL and of the compositions written first in their L1 and subsequently automatically translated into the FL. To this end, several aspects were analyzed.

The first thing studied to assess the complexity of participants' productions was the number of sentences, words, and characters per text. While the participants were required to adhere to a minimum length when writing their compositions, they were not provided with a maximum. It was interesting to observe that the minimum text length was by far exceeded in the post-edited samples (with an average of 288.29 words/text). This phenomenon was particularly visible as proficiency level increased (258.67 words/text in the case of basic users; 279.9 words/text in the case of intermediate users; and 309.86 words/text in the case of advanced users). In contrast, the compositions produced directly in the FL had a larger amount of sentences than those produced with the aid of the MT system (19.9 vs. 19.05). An increase in the number of words and a reduction in the number of sentences indicated that, indeed, the sentences of the post-edited texts were substantially longer than those of the writings produced directly in the FL (median of 11.5 in the case of *esDIR* vs. 14 in the case of *esPE*; and an average of 13.78 in the case of *esDIR* vs. 15.68 in the case of *esPE*). The data thus seem to indicate that the participants produced more complex compositions with the help of the selected MT system, at least in terms of text and sentence length.

The nature of the words that appear in the blog posts was the next aspect to be analyzed. For this purpose, the proportions of the primary POS tags were studied. It was initially predicted that the texts written directly in the FL would have a bigger proportion of nouns and verbs than post-edited ones since they are considered the most basic categories. However, this hypothesis was only fulfilled in the case of verbs (17.33% vs. 15.54%). The remaining categories regarded as more complex were found in higher proportions in the compositions produced with the help of MT, with the sole exception of conjunctions (7.06% vs. 6.36%). Although this was initially striking, since the use of this POS generally requires a good command of the language, when analyzing how varied the conjunctions were, it was concluded that the diversity of this category was notably larger in the post-edited texts (31.33% vs. 36.39%).

As observed in the case of conjunctions, a higher proportion of concrete POS does not provide enough information for determining whether a text is complex or not. A further way to support the findings is by measuring how varied these grammatical categories are. In contrast to what was commented in relation to the proportions of the POS, the degree of diversity was more extensive in the case of nouns (78.24% vs. 77.13%) than in the case of verbs (61.21% vs. 63.88%) in the texts written without the assistance of MT. However, it is worth noting that the differences between both setups were not overwhelming.

To examine how informative the compositions were, the proportion of content words, i.e., nouns, verbs, adjectives, and adverbs, was calculated. Thus, the lexical density of the participants' productions was measured. It was interesting to see that the proportion of these words was slightly larger in the texts written directly in the FL (55.75%) than in the post-edited ones (55.34%). This could be explained by the tendency of some participants to write long sequences of city and country names, which was not exactly indicative of a greater level of informativeness.

The degree of readability was also measured in order to determine how complex the texts were. To this end, the compositions were first analyzed according to the proportion of complex words they contained, i.e., words with more than three syllables. Except in the case of advanced users (10.04% vs. 9.59%), the general tendency was to make greater use of complex words when using the MT system (9.04% vs. 9.70%). Readability was also calculated using two well-known metrics, namely Flesch-Szigriszt Index (*Barrio-Cantalejo,* 2008), and Fernández-Huerta Index (Fernández-Huerta, 1959), specifically developed for that purpose. Again, the trend was to produce compositions that were more difficult to read with the aid of MT.

Both setups were also analyzed by measuring which of the two contained a greater amount of complex vocabulary and syntactic structures. With respect to the evaluation of the lexicon employed by the participants, it was decided to compare the terms included within the compositions against three lists containing the 1000, 5000, and 10000 most frequent words in CREA. The percentages suggested that the post-edited texts had a greater number of these terms than those written directly in the FL. A small manual analysis was conducted to see what types of words were in the texts but not in the lists. The *esDIR* texts had a large number of misspellings, colloquialisms, proper nouns, foreign words, expressions of laughter, and interjections, as opposed to the *esPE* texts, which were rather characterized by a big presence of proper nouns and foreign words. When studying the syntactic structures, only some of those considered to be more complex (according to subjective criteria) were examined. Those that had a greater presence in both setups were the ones linked to adjectives and prepositions. In general, with the exception of predicative complements (0.57% vs. 0.41%), the selected syntactic structures tended to appear more in the post-edited texts. The differences were particularly noticeable in the compositions of the basic users.

Finally, the contrast between the setups when evaluating which of the texts read more like the FL was overwhelming. While this phenomenon was not visible when comparing them against a POS-based model (5.949 vs. 5.533), the results were striking with the token-based model (307.949 vs. 157.934). However, the biggest difference was found in the texts created by basic users (740.921 vs. 173.345). The data show that, by far, the compositions produced with the help of MT resemble the language model more closely than the texts written directly in the FL.

While it is true that the higher the level of proficiency in the FL, the smaller the discrepancies between the setups, the several analyses carried out in this study indicate that MT systems enable users to produce more complex texts in the FL. Nevertheless, some of the results obtained from the post-edited texts, especially those from participants who had a lower command of the FL, suggest that users may not be benefiting to the fullest from the MT output. Ideally, they would have a tool to guide them in the process of post-editing, which would provide them with indications on where to perform the modifications, if needed.

5 Potential of iSTS for assisting users in post-editing

5.1 Introduction and objectives

As described in the state-of-the-art chapter, Semantic Textual Similarity (STS) is a measure of the degree of semantic equivalence between a pair of sentences (*Agirre et al.*, 2015). Several studies have explored its applicability to diverse Natural Language Processing (NLP) tasks, such as information extraction, question answering, summarization, and MT evaluation. In fact, although this metric was initially designed to operate with sentence pairs in English, several attempts have been made to transform it into a multilingual and even cross-lingual measure (*Cer et al.*, 2017). This fact has increased substantially its potential application to MT. So far, most efforts in this matter have been concentrated on finding a correlation, albeit a moderate one, between this metric and QE. However, to the best of our knowledge, nothing has been done yet, having the end-user in mind.

The most significant approach to this audience, although unrelated to the field of MT, has occurred with the development of the interpretable version of this metric, the so-called iSTS (*Agirre et al.,* 2016). Its goal is not only to provide a score indicating the level of semantic equivalence between two sentences but also to give an explanation of that particular score. Furthermore, the score in the case of this interpretable measure is no longer global. Every single sentence is split into chunks, and these chunks have to be aligned within the two sentences to be compared. Each alignment, which can be formed by more than one chunk, is given a score and assigned a label that explains the type of semantic relation the aligned chunks have.

The scores range from 0 to 5:

- *5* indicates that the meaning of the aligned chunks is equivalent,
- [*4*, *3*] indicate that the meaning of the aligned chunks is very similar or closely related,

- [*2, 1*] indicate that the meaning of the aligned chunks is slightly similar or somehow related,
- *0* (or *NIL*) indicates that the meaning of the chunk is entirely unrelated. This score is given to a chunk that remains unaligned.

As for the tags:

- *EQUI* is assigned to alignments formed by chunks that are semantically equivalent in the context. It should be noted that there is an interdependence between this label and the score *5*;
- *OPPO* is assigned to alignments formed by chunks that are semantically in opposition to each other in the context;
- *SPE1* is assigned to alignments formed by chunks that are semantically similar, but those of the first sentence are more specific than those of the second one;
- *SPE2* is assigned to alignments formed by chunks that are semantically similar, but those of the second sentence are more specific than those of the first one;
- *SIMI* is assigned to alignments formed by chunks that are semantically similar, but do not fulfill the previously mentioned conditions;
- *REL* is assigned to alignments formed by chunks that have related meanings but are not assigned any of the described tags yet;
- *NOALI* is reserved to the chunks with the *0* or *NIL* score, i.e., those that have remained unaligned.

It is worth noting that each one of the mentioned labels is exclusive, which means that only one can be used simultaneously. However, there are two additional tags that may be assigned (or not) to the alignments regardless of the compulsory ones:

- *FACT* is assigned to alignments formed by chunks in which the factuality is different,
- *POL* is assigned to alignments formed by chunks in which the polarity is different.

As stated above, these measures, both interpretable and non-interpretable, have great potential regarding their application with MT. However, iSTS appears to be ideal for giving feedback to end-users about the quality of the automatic translation and even providing them with indications so that they can post-edit the MT output more effectively. Not only would they be told whether there are differences between the source and the target, but also how big these differences are and where they are located. Nevertheless, a few questions arise about it. For example, would the existing labels be enough to accomplish the goal of assisting users for post-editing? In case that new tags need to be added, how specific should they be? What information would iSTS provide to the users in comparison to the metrics described in the state-of-the-art chapter? Would it make sense to operate with chunks within the field of MT? In the following subsections, some attempts will be made to answer these and other questions. As a result, an annotation proposal will be made in order to more efficiently apply iSTS to MT by taking end-users into account, specifically lay-users who have not been trained to translate professionally.

5.2 Methodology

5.2.1 Corpus

To test the potential of iSTS as a tool for assisting users in making the most out of the output of MT systems, it was decided to keep working with the writing tasks described in the first part of the present study. The compositions that the participants produced directly in Spanish (*esDIR*) were discarded in this section since they were not deemed useful for the subsequent analysis. A gold standard translation (*esGS*) of each of the texts that the participants wrote in their L1 (*plDIR*) was created instead. While no reference translations would have been necessary, this was done for various reasons: first, because the original design of iSTS was to operate with sentence pairs of the same language, and second, it appeared to be easier to see its application for helping users if only one language was used when presenting the annotation proposal. It is worth noting that other equally valid gold standards could have been used. The person in charge of creating the ones employed in this investigation was a native Polish speaker with a very high command of Spanish. This person was required to produce a translation as faithful to the original and as fluent in the target language (in this case, Spanish) as possible, without omitting details or adding extra information. It is assumed that if sentence pairs composed of *plDIR* and *esGS* were annotated according to the original iSTS guidelines, all chunks would have the *EQUI* label and the score *5*. Moreover, no chunks would be unaligned.

The nature of this corpus would allow applying iSTS within different set-ups. It should be remembered that it is not only made up of texts written in Polish (*plDIR*) and their respective human gold-standard translation (*esGS*), but it is also composed by their automatic translation (*esMT*) and the post-edited version of that MT output (*esPE*). The comparison of sentence pairs belonging to any of these scenarios would lead to different but interesting findings. For example, comparing sentences of the *esGS* and *esMT* compositions would serve to indicate to users the differences, if applicable, between the source text and the automatic translation. Furthermore, the comparison of sentences of *esGS* and *esPE* would reveal whether the participants have managed to overcome the weaknesses of the selected MT system. Last but not least, comparing sentences of the *esMT* and *esPE* texts would provide an idea of whether (and where) the participants have performed modifications in the MT output. As stated previously, the sentences to be compared do not need to belong to the same language. Therefore, a comparison between *plDIR* and *esMT, plDIR* and *esPE,* and *plDIR* and *esGS* could also be made. In the latter case, and as mentioned above, all alignments would be formed by semantically equivalent chunks. Thus, it is assumed that the outcomes of the other two possible scenarios would be quite similar to those of the previously described cases. To limit the scope of this study, the focus of the analysis will be set on the comparison between sentences included in the gold-standard translation and the automatic translation. It is believed that the indications provided by iSTS could be of great help for users to better post-edit the output of the MT systems, and consequently, get the most out of it.

In addition to the different setups covered by the corpus, its potential also lies in its length. The average number of sentences per participants' text is 19 with a standard deviation of 4,3. If only sentences in a single language (in this case, Spanish) were compared, a total of approximately 57 sentence pairs could be analyzed for each of the 21 participants. However, if an interlingual analysis were to be carried out, the number of sentence pairs to be examined would amount to approximately 114 per participant. In other words, when operating in a monolingual environment, the present corpus would allow annotating a total of approximately 1197 sentence pairs, and if, additionally, iSTS were to be applied to a multilingual context, it would result in the annotation of approximately 2394 sentence pairs.

As previously pointed out, STS works with sentences as a whole. However, for its interpretable version, it is required to operate with chunks. Although, as demonstrated so far, the corpus could have a lot of potential for applying iSTS, it would first be necessary to split every single text into sentences and subsequently into chunks. Since the present study will be carried out in a monolingual environment, only the chunks of the texts in Spanish have been defined. Ideally, the same pre-processing task should be performed on Polish compositions so that, if desired, any of the previously described scenarios could be analyzed.

The tool chosen to fulfill this purpose was TreeTagger[12] (*Schmid*, 1995). Although it is a multilingual tool that was especially developed for annotating texts with POS and lemma information, it can also be used as a chunker in four languages, among which Spanish is included. This chunker was trained on the IULA Spanish Treebank (*Marimon et al.*, 2012), and the parameter file on the Spanish CRATER corpus (*McEnery et al.*, 1997) together with the Spanish lexicon of the CALLHOME corpus of the LDC[13]. The use of another tool may have resulted in different chunks since it is often not easy to define them. The choice of TreeTagger was motivated not only by its simplicity but also by its consistency in determining the chunks.

The final step with respect to preparing the corpus for the subsequent annotation with the iSTS characteristic labels and scores would be to align the chunks. That is, to link one or more chunks that compose the two sentences to be compared, taking their meaning in the context into account. First, the chunks that have the same (or more or less the same) meaning considering both the context and the interpretation of the sentence should be aligned. In general, these chunks also tend to have similar roles within the sentences. In case the roles were different, but the chunks were related, they should also be aligned. Chunks could also be aligned, regardless of their role, even if the sentences would make reference to different events. If there were unaligned chunks, there would be three options: integrate the chunk into an existing alignment, create a new alignment, or leave the chunk unaligned. The latter would be the least favorable

[12] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

situation. The alignments of this corpus have been done manually. As it is a costly step in terms of time and effort, only chunks belonging to the *esGS* vs. *esMT* setup have been aligned for this investigation. The optimal situation, however, would be to keep on aligning pairs of sentences from all the previously described scenarios.

5.2.2 Attempt of applying iSTS annotation scheme

As stated in the preceding subsection, the corpus can be further improved in order to analyze different contexts. For the present research, however, only sentence pairs belonging to the *esGS* and *esMT* will be considered, as they are the only ones that have been completely pre-processed so far. The evaluation of this particular combination will serve to explore the potential of iSTS to provide indications to users to post-edit the output of the MT system. The goal would be that they could create a translation as faithful as possible to the source text. It is desired to examine the type of information that the original design of iSTS could give them and, in case it is not enough, or it is too much, to modify the guidelines.

To this end, the followed approach was iterative. That is, first, the annotation system had to be designed. This step had already been done since it was determined that the starting point would be the original iSTS guidelines. Afterward, the annotation system was applied to the existing corpus. The following stage was to analyze whether the objectives set were met or not. And finally, if this was not the case, the annotation system was further refined. Thus, the annotation proposal that will be presented in the following paragraphs is the result of several testing and refinement cycles. It is worth noting that, since the pre-processed part of the corpus was still comprised of a large number of sentence pairs, it was decided to carry out the iterations on the compositions of a single participant. Specifically, 30 pairs of sentences have been evaluated.

Although efforts were made to maintain the original iSTS design to the highest possible extent, some modifications had to be made. In the first of the aforementioned cycles, it was concluded that the *REL* tag was no longer necessary. Besides not finding any situation in which this label could be assigned to an alignment instead of another of the existing ones, it was considered that it could also lead users to confusion. This is because it is not a particularly easy task to distinguish between similarity and

[13] https://www.ldc.upenn.edu/

relatedness. Furthermore, the additional tags, i.e., *POL* and *FACT*, are not deemed to provide relevant indications to users when assisting them in performing the post-editing of the MT output. Therefore, both labels were also discarded. However, even though the scoring range was kept unchanged and the remaining tags were believed to be useful, this did not mean that they all were sufficient to accomplish the intended goal. In fact, it was observed that, in many instances, despite the terminology is accurate, there were discrepancies in the word forms that altered the semantic equivalence in a different way, for example, by changing the gender, number, person, or verb tense, among others. How could users be indicated that, although the term selected by the MT is indeed correct, this alteration in the word form prevents it from meaning the same as in the source text? How could they be guided to only modify the word form without changing the whole term?

The first approach to answering these questions was to create new fine-grained labels which specifically indicated the differences between the aligned chunks. However, as more sentence pairs were analyzed, the possible scenarios increased, and, most likely, not all of them were covered by the corpus. Moreover, further refinement of this very specific labeling was dismissed because it is not always better to provide extremely detailed information. In fact, it should be remembered that the lay users are the target audience of this research. Telling them to replace the *preterite* with the *imperfect* may not be very enlightening to them, usually, it is quite the contrary. It is also important to bear in mind that the simpler the labeling, the easier it is to automate the process. For all these reasons, it was determined to discard the inclusion of fine-grained labels. However, a couple of broad tags were created to overcome the aforementioned challenges.

As discussed earlier, all alignments within the sentence pairs would ideally be assigned the *EQUI* label and the score of 5. Furthermore, no chunks would be unaligned. This would indicate that the MT has preserved the meaning of the source text successfully. Nevertheless, as can be assumed, this is not always the case. While the use of the already existing labels and scores can serve to get an idea of how different the aligned chunks are, they are not pointing users to where exactly they should focus their attention. Therefore, the *GRAM* label has been created to address the previously described instances in which the selected terminology is not an issue, but the word form is. The score assigned to that specific alignment would indicate the number of grammatical differences between the chunks. However, it can also be the case that the chosen terminology is not the proper one. The *LEX* tag would then inform the users that it is the word itself that should be modified. Grammatical differences and lexical differences might sometimes coexist. For such cases, a further label has been created, namely *LEXGRAM.* It arises from the combination of the preceding two. The presence of spelling differences should not be overlooked. The original iSTS design stands indifferent to them. However, they are of great relevance when assessing the quality of a translation. In the event of these differences, the aligned chunks, whether or not they were equivalent, will be penalized with an additional tag, namely *MIS*. The use of each of these labels will be illustrated below with some examples. The format adopted for providing the examples is that proposed by (*Agirre et al.*, 2016): token-id-sent1 ⬄token-id-sent2 // label // score. The additional labels will be added right after the score. It is worth noting that the first sentence of each of the examples belongs to the gold standard translation (*esGS*) and the second one to the machine translation (*esMT*).

Sentence 1	[La última vez]1 [quise]2 [contaros]3 [brevemente]4 [acerca de cómo visitar]5 [una ciudad]6 [durante el confinamiento]7, ¡[resultó]8 [bastante]9 [posible]10!
Sentence 2	[La última vez]1 [quise]2 [contarles]3 [brevemente]4 [acerca de visitar]5 [la ciudad]6 [durante el cierre]7, ¡[resultó]8 [que]9 [era]10 [muy posible]11!
Annotation	1⬄1 (EQUI 5); 2⬄2 (EQUI 5); 3⬄3 (SIMI 4 GRAM); 4⬄4 (EQUI 5); 5⬄5 (SPE1 4 LEX); 6⬄6 (SPE2 4 GRAM); 7⬄7 (SIMI 4 LEX); 8⬄8,9,10 (SPE2 4 LEX); 9,10⬄11 (SIMI 4 LEX).

Table 15: Example of the iSTS annotation to illustrate the use of the *GRAM* and *LEX* tags

In this first example (Table 15), several of the aforementioned phenomena can be observed. For instance, in the alignment of chunks 3⬄3, the term is the same in both setups. However, the chunks are not equivalent. The *GRAM* label, together with the score 4 and the *SIMI* tag, indicates that there is a grammatical difference between the chunks. In this case, this difference lies in formality (*tú vs. usted*). This newly created label appears again in the alignment 6⬄6. Here, although the score provided is the same as in the earlier commented example, it is combined with the *SPE2* tag. This will be explained in more detail in the guidelines, but if a chunk in a sentence has a definite article and the chunk with which it is aligned in the other sentence is formed by an indefinite one, the label *SPE1/2* is given to that alignment, depending on where the definite article is located. In the present example, many alignments have been assigned the *LEX* tag. Either the chunks are similar but not equivalent, as in the case of 9,10⇔11, or one of the chunks contains more information than the other, such as 5⇔5.

Sentence 1	[Y]1 ¿[qué]2 [hay]3 [de la comida]4?
Sentence 2	¿[Que]1 [hay]2 [de la comida]3?
Annotation	1⇔Ø (NOALI); 2⇔1 (EQUI 5 MIS); 3⇔2 (EQUI 5); 4⇔3 (EQUI 5).

Table 16: Example of the iSTS annotation to illustrate the use of the *MIS* tag

This second example (Table 16) illustrates the case of chunks that have spelling differences, even though they have been assigned the maximum score and are considered semantically equivalent. The additional label *MIS* would then serve to penalize them. This way, users will know that, although the word proposed by the MT system is correct, they should pay attention to its spelling. While chunk *1* of the first sentence has been left unaligned, the remaining alignments have been assigned the *EQUI* tag and score *5*. However, the alignment 2⇔1 deserves special mention. As stated previously, the newly created *MIS* tag indicates that chunk *1* of the second sentence requires user revision. Specifically, the missing accent mark should be added.

Sentence 1	[Siempre que]1 [sea]2 [posible]3 [escuchar]4 [mutuamente]5 [vuestras divagaciones]6 [nerds]7.
Sentence 2	[Siempre que]1 [puedas]2 [escuchar]3 [tus argumentos]4 [nerds]5.
Annotation	1⇔1 (EQUI 5); 2,3⇔2 (SIMI 4 GRAM); 4⇔3 (EQUI 5); 7⇔5 (EQUI 5); 5⇔Ø (NOALI); 6⇔ 4 (SIMI 3 LEXGRAM).

Table 17: Example of the iSTS annotation to illustrate the use of the *LEXGRAM* tag

While a number of these alignments would deserve comment, this third sentence pair (Table 17) has been selected to exemplify the case where the *LEXGRAM* tag might be of particular use. The concrete alignment that requires the use of this additional label is 6⇔4. In the second sentence, even though the second person has been used, the chunk differs from that of the first sentence in the number, being singular in this case (*vuestras* vs. *tus*). Moreover, the word choice was not the right one. Therefore, the user, in addition to paying attention to the grammar, should modify the vocabulary. The score given to this alignment (*3*) indicates the presence of two differences, one grammatical and one lexical.

The examples above illustrate only a few cases where the newly created additional labels could be used. Although it has been worked with the compositions of a single participant, it is believed that the tags are broad enough to be applicable to the rest of the corpus but also informative enough to be helpful to the users. A first annotation proposal based on the original iSTS guidelines, but with some modifications, will be presented below. This proposal is certainly not definitive. It is not ruled out that, as more sentence pairs from the corpus are annotated, further refinement may be necessary. Furthermore, it should be remembered that it has been implemented in a monolingual environment. iSTS was indeed developed based on English, and Spanish has been the language used in the present investigation. This would suggest its potential of being a multilingual measure, as already predicted in previous research studies (*Cer et al.*, 2017). However, the new labels have not yet been tested in an interlingual setting. It would therefore be ideal to apply these new guidelines to all the above-mentioned setups.

5.3 Proposal of new guidelines for iSTS annotation

This annotation proposal for iSTS with the aim of helping end-users to efficiently post-edit the MT output is based on the SemEval-2016 guidelines[14] (*Agirre et al.*, 2015). As discussed above, the chunks have been created following different criteria since the original guidelines were explicitly designed to operate in English. Depending on the language (or languages) to be worked with, the toolset will vary in most cases. However, it is worth noting that the way of proceeding to align the chunks has remained unchanged. Reference was made to this alignment-making process in the previous sections. On the one hand, it was also commented that not all labels presented in the guidelines of *Agirre et al.* (2015) had been adopted in this proposal. On the other hand, new labels have been created to complement the already existing ones in an attempt to provide more accurate information to users. A great novelty with respect to the first guidelines is that, except for *EQUI* and *NOALI*, the *SIMI*, *SPE1*, *SPE2*, and *OPPO* tags will have to be followed by another label. These other labels will be *GRAM*, *LEX*, and *LEXGRAM*. There will be a fourth tag, *MIS*, which will serve to indicate differences in spelling. This last label can also appear together with *EQUI*. The use of the new labels in conjunction with the existing ones will be explained below. Furthermore, a few examples will be given to illustrate some of the possible scenarios.

First, the use of the *EQUI* tag and the possible application of the new *MIS* tag will be discussed. Subsequently, the use of the newly created labels and their combination with the already existing ones will be described. *GRAM* will be the first, followed by *LEX* and ending with *LEXGRAM*. It should be pointed out that, although the range of scores from *0* to *5* is maintained, each of the label combinations will be scored in a unique way. This will also be addressed in the following paragraphs.

As for the *NOALI* label, the original guidelines have been maintained. That is, those chunks that cannot be aligned will be assigned the *NOALI* tag. Some of the sentence pairs which will be presented below also have unaligned chunks. In fact, the presence of *NOALI* labels can indicate either that the message has not been fully conveyed in one of the two sentences or that information has been left out but may be present in another chunk or another sentence. It is therefore considered that the mere presence of this label could be of help to the user. With this fact in mind, the guidelines proposal will now be described.

As previously mentioned, when the aligned chunks are semantically equivalent, they are assigned the ***EQUI*** label and a score *5* (Table 18).

Sentence 1	[Y]1... ¿[qué]2 [hicimos]3 [realmente]4 [durante todos estos paseos]5?
Sentence 2	[Y]1... ¿[qué]2 [hicimos]3 [realmente]4 [durante todos estos paseos]5?
Annotation	1⇔1 (**EQUI 5**); 2⇔2 (**EQUI 5**); 3⇔3 (**EQUI 5**); 4⇔4 (**EQUI 5**); 5⇔5 (**EQUI 5**).

Table 18: Example of the iSTS annotation to illustrate the use of the *EQUI* label

No additional tags would be needed, unless there are differences regarding the spelling. In that case, the ***EQUI*** tag and the score *5* would be kept, but they would be followed by the ***MIS*** tag (Table 19).

[14] https://alt.qcri.org/semeval2016/task2/data/uploads/annotationguidelinesinterpretablests2016v2.2.pdf

Sentence 1	[Y]1 ¿[qué]2 [hay]3 [de la comida]4?
Sentence 2	¿[Que]1 [hay]2 [de la comida]3?
Annotation	1⇔Ø (NOALI); 2⇔1 (**EQUI 5 MIS**); 3⇔2 (EQUI 5); 4⇔3 (EQUI 5).

Table 19: Example of the iSTS annotation to illustrate the use of the *EQUI* label together with the *MIS* tag

The uses of the ***GRAM*** label would be very diverse, as briefly described below.

In conjunction with ***SIMI***

In case the aligned chunks have grammatical dissimilarities, such as different number, gender, person, formality, verb tense, or even a missing article, among others, the *GRAM* tag would complement the *SIMI* label. The score would be reduced as the grammatical differences increase. That is, if the aligned chunks have only one grammatical difference, the annotation would look as follows *SIMI 4 GRAM*. If they have two differences, the annotation would be *SIMI 3 GRAM*. And so on. The following sentence pair (Table 20) exemplifies this phenomenon. The alignment 8⇔8 has been annotated with *SIMI 4 GRAM* as there is only one difference with respect to the number. However, alignments 3⇔3 and 4⇔4 have two differences, specifically regarding number and formality. Therefore, they have been annotated as *SIMI 3 GRAM*.

Sentence 1	[Si]1 [no]2 [queréis]3 [quedaros]4 [en casa]5, [preparad]6 [una ruta]7 [por las librerías]8.
Sentence 2	[Si]1 [no]2 [quiere]3 [quedarse]4 [en casa]5, [haga]6 [un recorrido]7 [por la librería]8.
Annotation	1⇔1 (EQUI 5); 2⇔2 (EQUI 5); 3⇔3 (**SIMI 3 GRAM**); 4⇔4 (**SIMI 3 GRAM**); 5⇔5 (EQUI 5); 6⇔6 (SIMI 3 LEXGRAM); 7⇔7 (EQUI 5); 8⇔8 (**SIMI 4 GRAM**).

Table 20: First example of the iSTS annotation to illustrate the use of the *SIMI* label together with the *GRAM* tag

While it is true that it is possible to align chunks with the same meaning but different constructions, they might not be entirely equivalent. For example, it could be the case that one of the sentences is personal and the other impersonal. Although this may not seem to be a relevant difference, it is an important one, especially when taking the way of conjugating the verb of the personal sentence into account. This is one of the problems encountered with iSTS. Operating only within the context of the sentence does not allow knowing, for example, the number (singular or plural) or the formality (*tú* or *usted*) the user has opted for throughout the text. In the following sentence pair

(Table 21), it can be seen in the alignment 3⇔3,4 that the participant has chosen the impersonal form (*hay que*), but the MT system has used the verb in the second person singular instead (*tienes que*). Again, as iSTS only works with a sentence at a time, it is not known whether this same pattern has been followed in the rest of the text. It would thus be ideal to draw the user's attention to that specific alignment, so that the difference does not go unnoticed, even though the chunks may appear to be equivalent. Therefore, when encountering this type of personal vs. impersonal phenomenon, the *GRAM* tag will follow the *SIMI* label. A score *4* will be the starting point. However, as observed in other cases, if there are more differences, the score will be reduced.

Sentence 1	[Durante los paseos]1, [ojo]2, [hay que]3 [hablar]4.
Sentence 2	[Durante los paseos]1, [ojo]2, [tienes]3 [que]4 [hablar]5.
Annotation	1⇔1 (EQUI 5); 2⇔2 (EQUI 5); 3⇔3,4 (**SIMI 4 GRAM**); 4⇔5 (EQUI 5).

Table 21: Second example of the iSTS annotation to illustrate the use of the *SIMI* label together with the *GRAM* tag

In conjunction with ***SPE1*** and ***SPE2***

In case the grammatical differences are linked to the article type, for example, one being definite and the other indefinite, the annotation would be slightly different. In such an instance, *GRAM* would not follow *SIMI* but *SPE1/2*. If the chunk in the first sentence consists of a definite article and in the second sentence of an indefinite article, the annotation would be *SPE1 4 GRAM*. If, on the other hand, the indefinite article is in the first sentence and the definite in the second one, the annotation would be *SPE2 4 GRAM*. In addition to differences with respect to the article, there could be another of the above-mentioned grammatical differences. The annotation would be *SPE1 3 GRAM* or *SPE2 3 GRAM*, depending on the location of the definite article. Again, as the number of grammatical differences increases, the score will decrease. The following example (Table 22) illustrates the coexistence of a difference between articles along with another grammatical difference, namely gender.

Sentence 1	[Y]1 ¿[en una más pequeña]2?
Sentence 2	¿[En]1 [el]2 [más]3 [pequeño]4?
Annotation	1⇔ Ø (NOALI); 2⇔1,2,3,4 (**SPE2 3 GRAM**).

Table 22: Example of the iSTS annotation to illustrate the use of the *SPE1/2* labels together with the *GRAM* tag

In conjunction with ***OPPO***

In the analyzed corpus, there were no instances in which it proved necessary to use the *GRAM* tag together with *OPPO*. The situations where grammatical differences were observed could be solved with the previously described labels. Perhaps, in case there would be very big differences such as the presence of a verb in the past tense in one sentence and a verb in the future tense in the other, the *SIMI* tag might not give the user enough information to post-edit the MT output correctly. The *OPPO* label could therefore be introduced here. However, as no sentence pair with this type of difference has been encountered, it is not certain whether this would be the optimal way to approach it. This label is thus left aside, for the moment, in conjunction with *GRAM*.

The ***LEX*** label can also be used in several contexts, as depicted below.

In conjunction with ***SIMI***

The most frequent situation in which the newly created *LEX* tag follows the already existing *SIMI* label is that in which the aligned chunks are formed by words that, although similar, are not semantically equivalent. As pointed out previously, the context should be taken into account. The score would indicate how different the terms are. The starting point would be *4*, which indicates that the chunks are very similar. Then, *3* would mean that they are quite similar, and so on down to *1*, which would indicate a very low degree of similarity. In the following example (Table 23), it can be seen that, although the difference in the alignment 9⇔12 is not very large, the chunks do not mean exactly the same. In fact, a city can be *big* but not *important*, just as an *important* city does not have to be *big*. Such a nuance can also be observed in the alignment 5⇔5. Stating “*durante un rato*” involves a shorter period of time than “*durante un tiempo*”. However, in this context, they have been considered equivalent.

Sentence 1	[Si]1 [lográis]2 [no]3 [hablar]4 [durante un rato]5, [seguramente]6 [encontraréis]7 [en todas las ciudades]8 [grandes]9 [bocadillos]10 [o]11 [pizzas]12.
Sentence 2	[Si]1 [logras]2 [no]3 [hablar]4 [durante un tiempo]5, [seguramente]6 [encontrarás]7 [bocadillos]8 [o]9 [pizzas]10 [en todas las ciudades]11 [importantes]12.
Annotation	1⬄1 (EQUI 5); 2⬄2 (SIMI 4 GRAM); 3⬄3 (EQUI 5); 4⬄4 (EQUI 5); 5⬄5 (EQUI 5); 6⬄6 (EQUI 5); 7⬄7 (SIMI 4 GRAM); 8⬄11 (EQUI 5); 9⬄12 (**SIMI 4 LEX**): 10⬄8 (EQUI 5); 11⬄9 (EQUI 5); 12⬄10 (EQUI 5).

Table 23: First example of the iSTS annotation to illustrate the use of the *SIMI* label together with the *LEX* tag

It may also be the case that there is no semantic equivalence between the aligned chunks due to a word modification through derivational affixes. In particular, in the examined corpus, a large presence of diminutives has been observed. However, this approach is proposed to be followed with any type of affix, i.e. suffixes, prefixes or infixes. The following pair of sentences (Table 24), specifically the alignment 4⬄4, illustrates one of those possible cases.

Sentence 1	¡[Abasteceos]1 [de un termo]2 [y]3 [una cestita]4!
Sentence 2	¡[Abastécete]1 [de un termo]2 [y]3 [una canasta]4!
Annotation	1⬄1 (SIMI 4 GRAM); 2⬄2 (EQUI 5); 3⬄3 (EQUI 5); 4⬄4 (**SIMI 4 LEX**).

Table 24: Second example of the iSTS annotation to illustrate the use of the *SIMI* label together with the *LEX* tag

The *LEX* tag, together with the *SIMI* label, can also be assigned to those chunks that consist of a translation of a proper name. However, the translation differs from the one considered to be ideal. The alignment 5⬄5 of the sentence pair below (Table 25) is an example of this phenomenon.

Sentence 1	[El vino]1 [caliente]2 [y]3 [los croissants]4 [de San Martín]5 [fueron]6 [nuestro kit]7 [básico]8.
Sentence 2	[El vino]1 [caliente]2 [y]3 [los croissants]4 [de Martinica]5 [eran]6 [nuestro set]7 [básico]8.
Annotation	1⬄1 (EQUI 5); 2⬄2 (EQUI 5); 3⬄3 (EQUI 5); 4⬄4 (EQUI 5); 5⬄5 (**SIMI 4 LEX**); 6⬄6 (SIMI 4 GRAM); 7⬄7 (EQUI 5); 8⬄8 (EQUI 5).

Table 25: Third example of the iSTS annotation to illustrate the use of the *SIMI* label together with the *LEX* tag

In the present annotation proposal, it has also been determined that regional variants, even if they mean exactly the same thing, should not be considered equivalent. Although there may be situations in which the meaning of the chunks is easily understood, regardless of the linguistic area of origin, this is not always the case. Therefore, users should be warned that the words proposed by the MT system do not belong to the standard variant of the language into which they are translating. A clear example of this would be the alignment 11⇔11 of the following pair of sentences (Table 26). The Diccionario de la Real Academia Española[15] (DRAE) states that both terms are semantically equivalent in some Latin American countries, but they have very different meanings in Spain.

Sentence 1	[Estaba]1 [en un vaso]2 [cerrado]3, ¿[quizá]4 [era]5 [café]6, [y]7 [solo]8 [os]9 [estoy]10 [vacilando]11?
Sentence 2	[Estaba]1 [en una taza]2 [cerrada]3, [tal vez]4 [era]5 [café]6, ¿[y]7 [solo]8 [te]9 [estoy]10 [pajeando]11?
Annotation	1⇔1 (EQUI 5); 2⇔2 (SIMI 4 LEX); 3⇔3 (EQUI 5); 4⇔4 (EQUI 5); 5⇔5 (EQUI 5); 6⇔6 (EQUI 5); 7⇔7 (EQUI 5); 8⇔8 (EQUI 5); 9⇔9 (SIMI 4 GRAM); 10⇔10 (EQUI 5); 11⇔11 (**SIMI 4 LEX**).

Table 26: Fourth example of the iSTS annotation to illustrate the use of the *SIMI* label together with the *LEX* tag

In conjunction with ***SPE1*** and ***SPE2***

It is proposed that *LEX* complements the *SPE1* and *SPE2* labels in such cases where the chunks of one of the sentences provide more information than the others with which they have been aligned. This scenario is particularly visible in alignments made up of more than one or two chunks. This could be one of the possible ways to indicate users that either the MT system has omitted information or has added it. The example below (Table 27) was presented previously, but it has been brought up again to exemplify the present guidelines proposal. When analyzing the 5⇔5 alignment, it can be seen that the gold standard translation is more specific than the MT output. Just the opposite of what happens with the alignment 8⇔8,9,10. Although it has been decided to maintain the scoring system in order to be consistent with the other possible scenarios, the combination of these two labels, namely *SPE1/2* and *LEX*, will always be assigned the score *4*.

[15] https://dle.rae.es/

Sentence 1	[La última vez]1 [quise]2 [contaros]3 [brevemente]4 [acerca de cómo visitar]5 [una ciudad]6 [durante el confinamiento]7, ¡[resultó]8 [bastante]9 [posible]10!
Sentence 2	[La última vez]1 [quise]2 [contarles]3 [brevemente]4 [acerca de visitar]5 [la ciudad]6 [durante el cierre]7, ¡[resultó]8 [que]9 [era]10 [muy posible]11!
Annotation	1⇔1 (EQUI 5); 2⇔2 (EQUI 5); 3⇔3 (SIMI 4 GRAM); 4⇔4 (EQUI 5); 5⇔5 (**SPE1 4 LEX**); 6⇔6 (SPE2 4 GRAM); 7⇔7 (SIMI 4 LEX); 8⇔8,9,10 (**SPE2 4 LEX**); 9,10⇔11 (SIMI 4 LEX).

Table 27: Example of the iSTS annotation to illustrate the use of the *SPE1/2* labels together with the *LEX* tag

In conjunction with ***OPPO***

Whenever an alignment is formed by chunks whose meaning is the complete opposite, it will be assigned the already existing *OPPO* label followed by the newly created *LEX* tag. The alignment 11,12,13⇔11 of the following sentence pair (Table 28) is a very illustrative example of this phenomenon. Again, just as in the previous case, the score to be given to the alignments with these two labels will be *4*.

Sentence 1	[Sospecho]1 [que]2 [os]3 [habéis]4 [estado]5 [preguntando]6 [sobre los detalles]7 [de tal iniciativa]8: ¿[cuánto]9 [tiempo]10 [puede]11 [pasar]12 [uno deambulando]13 [por la ciudad]14?
Sentence 2	[Sospecho]1 [que]2 [se]3 [ha]4 [estado]5 [preguntando]6 [acerca de los detalles]7 [de tal empresa]8: ¿[cuánto]9 [tiempo]10 [puede pasar eludiendo]11 [la ciudad]12?
Annotation	1⇔1 (EQUI 5); 2⇔2 (EQUI 5); 3⇔3 (SIMI 3 GRAM); 4⇔4 (SIMI 3 GRAM); 5⇔5 (EQUI 5); 6⇔6 (EQUI 5); 7⇔7 (EQUI 5); 8⇔8 (EQUI 5); 9⇔9 (EQUI 5); 10⇔10 (EQUI 5); 11,12,13⇔11 (**OPPO 4 LEX**); 14⇔12 (EQUI 5).

Table 28: Example of the iSTS annotation to illustrate the use of the *OPPO* label together with the *LEX* tag

Finally, the ***LEXGRAM*** tag can also appear in many situations:

In conjunction with ***SIMI***

It may occur that one of the aforementioned situations, where the use of the *SIMI* tag followed by the *GRAM* label is necessary, coexists with another of those that require the use of the *SIMI* tag followed by the *LEX* label. In other words, a case where the aligned chunks differ in grammar but also their meaning is not equivalent. The *LEXGRAM* tag would then be used to complement the *SIMI* label. As for other instances described above, the score given to the alignment will be reduced as the number of differences increases. The starting point, however, will not be *4* but *3*, since it is assumed that there are already two differences between the chunks (Table 29).

Sentence 1	[Si]1 [no]2 [queréis]3 [quedaros]4 [en casa]5, [preparad]6 [una ruta]7 [por las librerías]8.
Sentence 2	[Si]1 [no]2 [quiere]3 [quedarse]4 [en casa]5, [haga]6 [un recorrido]7 [por la librería]8.
Annotation	1⇔1 (EQUI 5); 2⇔2 (EQUI 5); 3⇔3 (SIMI 3 GRAM); 4⇔4 (SIMI 3 GRAM); 5⇔5 (EQUI 5); 6⇔6 (**SIMI 3 LEXGRAM**); 7⇔7 (EQUI 5); 8⇔8 (SIMI 4 GRAM).

Table 29: Example of the iSTS annotation to illustrate the use of the *SIMI* label together with the *LEXGRAM* tag

In conjunction with ***SPE1*** and ***SPE2***

As occurred with the combination of the *OPPO* and *GRAM* labels, there were no instances in the analyzed corpus in which the *SPE1* and/or *SPE2* tags could be applied together with the *LEXGRAM* label. For this to happen, there should have been alignments composed of chunks that, besides being more informative in one sentence than in the other, had at least one of the previously described grammatical differences. Also, as in the preceding case, the score *3* would be the starting point.

In conjunction with ***OPPO***

Earlier it was noted that no instance was found in the corpus in which the GRAM tag followed the OPPO label. However, some situations have been identified in which, in addition to the meanings of the aligned chunks being in opposition, they also account for at least one grammatical difference. This can be seen in the alignment 6⇔7 of the pair of sentences below (Table 30). Thus, those chunks will be annotated with the *OPPO* label together with the *LEXGRAM* tag. Again, the starting score will be *3*, and it will be reduced as the amount of differences increases.

Sentence 1	[Si]1 [es]2 [fin]3 [de semana]4, [hay que]5 [recorrerlas]6 [antes de las 14:00]7.
Sentence 2	[Si]1 [es]2 [un fin]3 [de semana]4, [tienes]5 [que]6 [rodearlos]7 [antes de las 2 pm]8
Annotation	1⇔1 (EQUI 5); 2⇔2 (EQUI 5); 3⇔3 (SIMI 4 GRAM); 4⇔4 (EQUI 5); 5⇔5,6 (SIMI 4 GRAM); 6⇔7 (**OPPO 3 LEXGRAM**); 7⇔8 (SIMI 4 LEX).

Table 30: Example of the iSTS annotation to illustrate the use of the *OPPO* label together with the *LEXGRAM* tag

Something that has not been mentioned so far, but which is believed to be also useful for users, is to observe the numbers of the chunks that form the alignments. Although, in many instances, the alteration of the order does not influence on the meaning of the sentence, this is not always the case. It would therefore be advisable to pay special attention to sentences embodying this phenomenon. The following pair of sentences (Table 31) is an example of this. However, in this case, the meaning has remained the same.

Sentence 1	[En Cytadela]1 [os]2 [encontraréis]3 [ardillas]4, [y]5 [en Rusałka]6, [castores]7.
Sentence 2	[Te]1 [encontrarás]2 [con ardillas]3 [en la Ciudadela]4 [y]5 [castores]6 [en Rusalka]7.
Annotation	**1⇔4** (EQUI 5); 2⇔1 (SIMI 4 GRAM); 3⇔2 (SIMI 4 GRAM); 4⇔3 (EQUI 5); 5⇔5 (EQUI 5); 6⇔7 (EQUI 5); 7⇔6 (EQUI 5).

Table 31: Example of the iSTS annotation to illustrate the numbers of the aligned chunks

As it has been mentioned in the previous subsections, this annotation proposal has been designed based on 30 sentence pairs belonging to one of the many possible setups offered by the corpus. Therefore, although some efforts have been made to keep it as general as possible, and it is believed that it could be applied to the rest of the corpus, it is not ruled out that the guidelines would have to be further refined. Moreover, it should be remembered that the solution has been operated in a monolingual environment. Perhaps there would be labels that would be useless or new tags should be added in an interlingual context.

The next step would be to keep pre-processing the corpus and try to apply the annotation proposal to the other texts that integrate it. The task, however, is far from simple. While there are sentence pairs that are easy to annotate, others involve considerable cognitive effort. In fact, it is possible that, although the tags created are pretty broad in order to reduce subjectivity as much as possible, the presence of more than one annotator could lead to some disagreement.

Before concluding, it should be noted that, even if the users were not specifically told what the differences were between the sentences, the fact that the iSTS works with chunks is already seen as a great help. Regardless of whether the labels are more or less informative, the fact that the users are pointed to the chunk they should check could be valuable.

5.4 Conclusions

The objective of this second part of the investigation was to explore the potential that iSTS could have for assisting users in the post-editing task.

The first step to approach this subject was to analyze the original design of the mentioned technique and observe whether it could satisfy this goal or not. In order to do this, it was necessary to collect a specific corpus that enabled making comparisons between pairs of sentences. The corpus used in the first part of the research was also deemed to be of great help in this one, although with some modifications. It consisted of a total of 21 tuples with a text written in the participants' L1, namely Polish (*plDIR*), its respective human gold-standard translation (*esGS*), and machine translation (*esMT*) into Spanish, and the post-edited version of the MT output (*esPE*). Since all of the compositions were based on the same text (*plDIR*), they were considered very suitable for testing iSTS.

This technique could have been applied to many possible scenarios, but in order to reduce the scope of the study due to time constraints, it was decided to work with the combination of *esGS* vs. *esMT*. First, the intention was to operate in a monolingual environment since the original design was conceived that way. In addition, it was believed that this particular configuration could provide a clearer idea of how iSTS could help users benefit the most from the MT output. To further reduce the scope, as the corpus was still large, the original iSTS annotation system was applied to the compositions of a single participant. Specifically, 30 sentence pairs were examined. These sentences were split into chunks with the aid of TreeTagger (*Schmid*, 1995). Subsequently, the chunks were aligned according to the criteria established in the original guidelines. After that, the available labels and scores were assigned to each of the alignments.

First, it was observed that some of the existing tags, namely *REL*, *POL*, and *FACT*, did not seem to be useful within the context of MT, more specifically for the purpose of giving feedback on translation quality to users. It was also noted that the existing labels, although they covered all the possible casuistries, might not be informative enough for users. It should be highlighted that, in the field of MT and as discussed in other metrics, the form of the words is of great importance. However, iSTS only gives indications regarding semantics. Therefore, it was decided to make an annotation proposal that would include other labels.

Although these new tags were not fine-grained, they would serve to inform the users of the type of differences between the sentences in the two setups. These labels should appear together with the already existing tags. They would also be assigned a score, but, in this case, this score would serve to indicate the number of differences there are between the sentences. If, on the one hand, the differences refer to semantics, the existing tags (*SIMI*, *SPE1*, *SPE2*, and *OPPO*) would be followed by the newly created *LEX* label. If, on the other hand, the differences are related to the word form, these labels would be complemented by the *GRAM* tag. In case the lexical and grammatical differences coexist in the same alignment, the additional label to be used would be *LEXGRAM*. For the equivalent chunks, the annotation system proposed in the original guidelines will be preserved, except in case they contain spelling differences. In such an instance, the *EQUI* label will be followed by the *MIS* tag. There could be alignments with up to three tags since *MIS* could also complement *GRAM*, *LEX*, and *LEXGRAM*. For chunks that do not have a semantic equivalent, i.e., cannot be aligned, the original annotation system would also be maintained.

It should be remembered that this is just a guidelines proposal. It is thought that this new scoring and tagging system could be applied to the rest of the corpus. Moreover, it is believed that it could be very informative for users. However, until it has been implemented, no reliable conclusions can yet be drawn. In fact, the task of applying this annotation system to the other compositions is considered to be quite complex. It is not always evident what type of labels to use and how to align the chunks.

Regardless of whether there are tags or not, what is certain is that the fact that iSTS does not work with whole sentences but with chunks already provides useful information to the users. They will not have to think about where they should check the sentences, since they will be indicated which particular part requires their attention, no matter whether the reason is explained to them or not.

6 General conclusions

The present investigation has consisted of two parts with also two different, although related objectives.

In the first part, an analysis of the similarities and differences between two setups was made, specially focusing on complexity. The first of these setups was made up of texts that learners of a FL created directly in that FL (*esDIR*) with the assistance of online language tools, except for MT systems. The second setup was formed by compositions that those same users first wrote in their L1 (*plDIR*), then translated into the FL using an MT system (*esMT*) and, finally, post-edited (*esPE*).

The degree of complexity of the texts was analyzed considering different aspects. As for text length, post-edited texts contained a greater number of words than those written directly in the FL. The number of sentences, however, was larger in the *esDIR* compositions. This indicated that the sentences of the *esPE* texts were substantially longer than those of the other setup and, therefore, more complex. The nature of the words included in the texts was also analyzed. Verbs, defined as a basic category, appeared in greater proportion in the texts written directly in the FL. The rest of the categories considered complex had a bigger presence in the post-edited texts, with the sole exception of conjunctions. Nevertheless, it was observed that the latter category was more varied in the *esPE* texts than in the *esDIR* ones. The diversity of POS also served to denote the level of complexity of the texts. Although nouns were more varied in the compositions written without MT assistance, there were no major differences between the setups when studying this concrete aspect. The degree of informativeness of the writings was measured by taking the proportion of content words into account. In this case, the results were in favor of the *esDIR* texts. A further aspect examined was the level of readability. It was noted that there was a tendency to produce more complex words, i.e., with more than three syllables, when using MT. While this is one of the elements used to measure readability, a couple of metrics specifically developed for this purpose also revealed that post-edited texts were more complex. The next thing to analyze was complexity with respect to vocabulary and syntactic structures. In the first case, both automatic and manual analyses were carried out. The results indicated that *esPE* compositions contained a higher number of basic terms. However, this could be explained by the big amount of foreign words, colloquialisms, and misspellings that characterized the texts written directly in the FL. In the second case, the analyzed complex syntactic structures appeared more frequently in the post-edited writings. The greatest presence in both setups was found in complements related to adjectives and prepositions. Finally, the texts were compared against two language models, one based on tokens and the other on POS. While in the latter case, the differences between the two setups were not overwhelming, in the former the compositions written directly in the FL diverged greatly from the model.

In general, the texts produced with the help of MT were more complex than those written directly in the FL. As the level of proficiency in that language increased, the differences between the setups became smaller. This would be the link between the first and the second part of the research. To ensure that the differences between the setups are also reduced in the case of users with a lower command of the FL, a special tool should ideally be available for them to give them indications of where and how they should post-edit the MT output. This way, they would be able to create a translation as faithful as possible to the source text.

That is why in the second part the objective was to explore the potential of iSTS to assist users in the post-editing task. To this end, the texts that the participants had created in the first part of this research were used again. However, those written directly in the FL were discarded. A human gold-standard translation for the texts that these participants had produced in their L1 was created instead. The original guidelines were applied to these compositions to see whether they could meet the intended purpose. As the corpus was too big for the scope of this study, only 30 sentence pairs belonging to the newly created gold standard translation (*esGS*) and the MT output (*esMT*) were analyzed. These sentence pairs were divided into chunks and aligned according to the annotation system established for the application of this technique. Each alignment was then assigned a label and a score. However, some of the tags were discarded, specifically *REL*, *POL*, and *FACT*. It was decided to create new labels instead (*GRAM*, *LEX*, *LEXGRAM*, and *MIS*). The reason for this was that, while the innovation of iSTS over the traditional MT evaluation metrics is that it addresses semantic differences, the presence of differences in the word forms is also of great importance in the field of MT. Therefore, although the other labels were kept, the new ones were designed to inform the users of the type of differences on which they would have to focus their attention: grammatical, lexical or spelling.

This annotation proposal was not at all definitive. In fact, it should be applied to the rest of the texts included in the corpus to test if they are really useful. There were cases in which it was difficult to select the labels to be assigned and to create the alignments. Thus, it is possible that it will have to be further refined by following, again, an iterative approach. Although it would then need to be implemented in a real setting, it is assumed that it will be of great help to users. Simply dealing with chunks, regardless of the informativeness of the tags, will ease the task to the users, as they will be told which specific part of the translated sentences to concentrate on.

7 Future work

Due to the limited scope of the study, decisions had to be made in order to narrow the focus and try to address every objective proposed.

Although the analysis carried out in the first part was fairly extensive, further work could still be done. For example, it would be ideal to test a new methodology, so that the participants could complete the writing tasks in person instead of having to proceed online, and observe whether the findings hold. In addition, it would be interesting to perform the same analysis with a larger number of participants, so that conclusions would be more reliable. In fact, it would be optimal that the quantity of participants for each level of proficiency would be more balanced, since in this research the number of basic users was substantially lower than the number of users with an intermediate or advanced command of the FL. Other than analyzing the compositions in terms of complexity, it would also be desired to perform an error evaluation of both *esDIR* and *esPE* texts. Additionally, another language pair could be tested to see whether the same phenomena would be reported. In the present experiment, the studied languages did not belong to the same language family to avoid interference, but it would be interesting to see what could happen if this were the case.

As for the second part, and as already pointed out throughout the project, the next step would be to fully annotate the corpus using the annotation proposal. In case it does not work, the idea would be to further refine it. In fact, to reflect better the situation in the context of MT, the new guidelines should be applied to an interlingual environment. The available corpus would also enable doing so. Once it has been annotated, besides making it publicly accessible for future research, it would be optimal to automate the whole process and even create a tool that could be implemented in the MT systems. This would allow carrying out the subsequent idea, which would be to test the tool with the end-users to learn about their experience and verify whether it is useful for them. If this were not the case, to know how it could be improved.

8 References

Agerri, R., Bermudez, J., Rigau, G. (2014). IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, 26-31.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L., Wiebe, J. (2015). SemEval-2015 task 2: semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceeding of the 9th International Workshop on Semantic Evaluation (SemEval-2015).*

Agirre, E., Gonzalez-Agirre, A., Lopez-Gazpio, I., Maritxalar, M., Rigau, G., Uria, L. (2016). SemEval-2016 task 2: interpretable semantic textual similarity. In *Proceedings of SemEval-2016.*

Agirre, E., Maritxalar, M., Rigau, G., Uria, L. (2015). SemEval-2016: Interpretable STS annotation guidelines.

Aranberri, N. (2020). With or without you? Effects of using machine translation to write flash fiction in the foreign language.

Barrio-Cantalejo, I.M. (2015). *El programa Inflesz*. Legibilidad.com Una web sobre el análisis de la legibilidad de textos escritos en español. https://legibilidad.blogspot.com/2015/01/el-programa-inflesz.html

Barrio-Cantalejo, I.M., Simón-Lorda, P., Melguizo, M., Escalona, I., Marijuán, M.I., Hernando, P. (2008). Validation of the INFLESZ scale to evaluate readability of texts aimed at the patient. In Anales Sis San Navarra, 31(2).

Bentivogli, L., Bisazza, A., Cettolo, M., Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. *Association for Computational Linguists.*

Björkelund, A., Hafdell, L., Nugues, P. (2009). Multilingual semantic role labeling. In *Proceedings of The Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009),* 43-48.

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

Bowker, L., Buitrago Ciro, J. (2015). Investigating the usefulness of machine translation for newcomers at the public library. *Translation and Interpreting Studies*, 10(2):165-186.

Briggs, N. (2018). Neural machine translation tools in the language learning classroom: students' use, perceptions, and analyses. *The JALT CALL Journal*, vol. 14: 3-24.

Cadwell, P., O'Brien, S. (2016). Language, culture, and translation in disaster ICT: an ecosystem model of understanding.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L. (2017). SemEval-2017 task 1: semantic textual similarity multilingual and cross-lingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017).*

Esplà-Gomis, M., Sánchez-Martínez, F., Forcada, M. L. (2018). Predicting insertion positions in word-level machine translation quality estimation.

Fernández Huerta, J. (1959). Medidas sencillas de lecturabilidad. In *Consigna (Revista pedagógica de la sección femenina de Falange ET y de las JONS)*, (214): 29-32.

García, I., Pena, M.I. (2011). Machine translation-assisted language learning: writing for beginners.

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M.A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., Straňák, P., Surdeanu, M., Xue, N., Zhang, Y. (2009). The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In Pr*oceedings of the Thirteenth*

Conference on Computational Natural Language Learning (CoNLL 2009). Shared Task, 1-18.

Jagaiah, T. (2017). Analysis of syntactic complexity and its relationship to writing quality in argumentative essays.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective.

Jordan-Núñez, K., Forcada, M.L., Clua, E. (2017). Usefulness of MT output for comprehension – an analysis from the point of view of linguistic intercomprehension. In *Proceedings of XVI Machine Translation Summit*, vol. 1: Research Track, 241-253.

Koponen, M., Salmi, L., Nikulin, M. (2019). A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. *Machine Translation*, vol. 33: 61-90.

Lavie, A., Denkowski, M. J. (2009). The METEOR metric for automatic evaluation of machine translation.

Lee, S.M. (2019). The impact of using machine translation on EFL students' writing. *Computer Assisted Language Learning*, vol. 33 (3): 1-19.

Lee, S.M., Briggs, N. (2020). Effects of using machine translation to mediate the revision process of Korean university students' academic writing. *ReCALL*: 1-16.

Marimon, M., Fisas, B., Bel, N., Arias, B., Vázquez, S., Vivaldi, J., Torner, S., Villegas, M., Lorente, M. (2012). The IULA Treebank. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).*

McEnery, A.M., Sanchez-Leon, F., Gaultier, E., Oakes, M. (1997). CRATER corpus. *European Language Resources Association (ELRA).*

Niño, A. (2004). Recycling MT: A course on foreign language writing via MT post-editing.

Niño, A. (2008). Evaluating the use of machine translation post-editing in the foreign language class.

Niño, A. (2020). Exploring the use of online machine translation for independent language learning. In *Research in learning technology*, vol. 28: 1-32.

O'Brien, S., Federici, F.M. (2019). Crisis translation: considering language needs in multilingual disaster settings. *Disaster Prevention and Management: An International Journal*.

O'Brien, S., Simard, M., Goulet, M.J. (2018). Machine translation and self-post-editing for academic writing support: quality explorations.

Otegi, A., Imaz, O., Díaz de Ilarraza, A., Iruskieta, M., Uria, L. (2017). ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research. In *Procesamiento del Lenguaje Natural* 58: 77-84.

Papineni, K., Roukos, S., Ward, T., Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation.

Parra Escartín, C., O'Brien, S., Goulet, M.J., Simard, M. (2017). Machine translation as an academic writing aid for medical practitioners.

Popović, M. (2011). Hjerson: an open source tool for automatic error classification of machine translation output. In *The Prague Bulletin of Mathematical Linguistics*, vol. 96 (1):59-68.

Rei, R., Stewart, C., Farinha, A. C., Lavie, A. (2020). COMET: a neural framework for MT evaluation.

Richmond, I.M. (1994). Doing it backwards: using translation software to teach target-language grammaticality. In *Computer Assisted Language Learning*, vol. 7 (1): 65-78.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*

Shei, C. (2002). Teaching MT through pre-editing: three case studies.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, vol. 200.

Specia, L., Scarton, C., Paetzold, G. H. (2018). Quality estimation for machine translation.

Szigriszt Pazos, F. Sistemas predictivos de legilibilidad del mensaje escrito: fórmula de perspicuidad. (1992) *Tesis Doctoral inédita*. Universidad Complutense de Madrid.

Toral, A., Sánchez-Cartagena, V.M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions.

Vilar, D., Xu, J., D’Haro, L. F., Ney, H. (2006). Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 06).*

Yan, J., Xu, X.Y. (2017). The relationship between syntactic complexity and writing quality of Chinese EFL learners.

9 Appendix

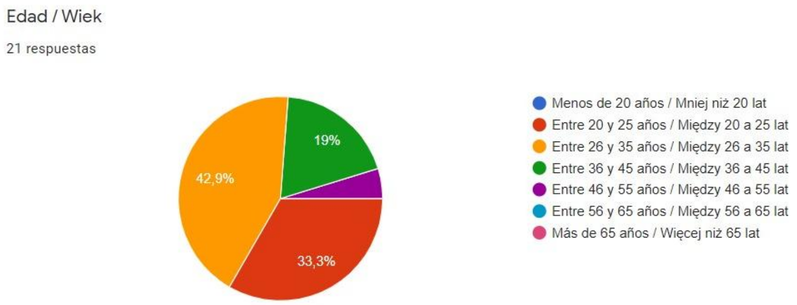


Figure 2: Age of the participants

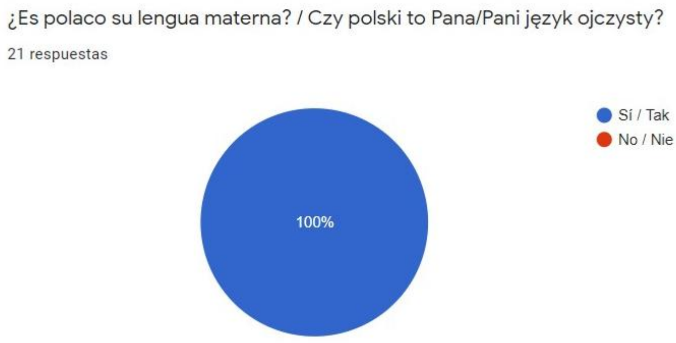
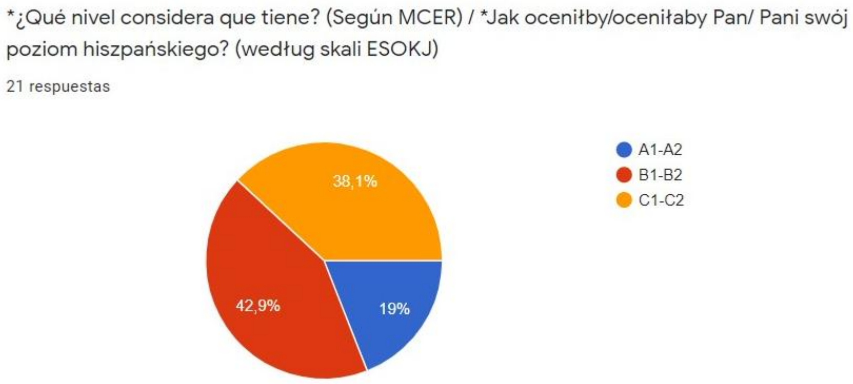


Figure 3: L1 of the participants – Polish

Para hablar con amigos, leer textos, ver las series en Netflix/ do romawiania z przyjaciółmi, czytania tekstów, oglądania seriali
para comunicar con mis amiagas y mejorar mis competencias en espanol para estudiar en Barcelona
para hablar con mi amiga, Tatiana <3 para comunicar con la gente de España, y además para leer artículos, entender las canciones, las series españolas, ver TED X etc.
Uso el español para hablar con mis amigos y para cuando hago turismo en paises hispanofalantes.
Durante el curso y para leer.
Para leer libros y peridicos, para ver peliculas y series.
Para comunicarse cuando viajo e para entender la Petra de las canciones cuando bawiło salsa.
Para hablar con españoles, ver La Resistencia y escuchar La Vida Moderna
Uso el español con mayor frecuencia durante mis clases universitarias o cuando escucho música y veo series de televisión
Para estudiar, trabajar
Para comunicarme con mis amigos, para ver peliculas y leer cosas interesantes, para ensenyarlo a los ninyos

Para comunicarse con la gente en España, donde actualmente vivo
podroze
Para el placer, con los fines turisticos, en el trabajo
Para viajar/ver unas series en espanol
para leer, para hablar con mis amigos
El trabajo en la universidad
estudio la filología española
Para los estudios. Es que estudio la etnolingüística (inglés +español)
Słuchanie piosenek oglądanie filmów. Hobby
para viajar y hablar con unas amigas espanolas

Figure 4: Use of Spanish

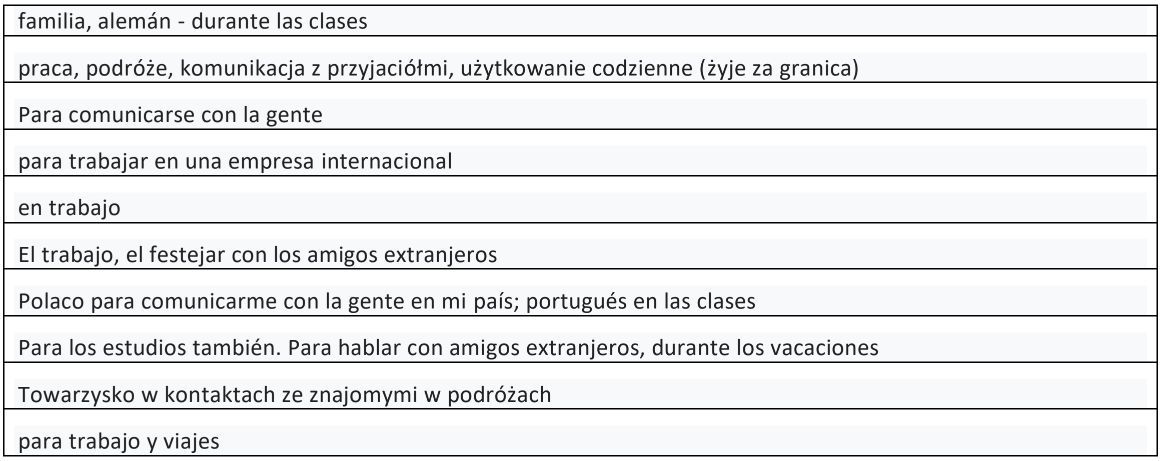


*¿Qué nivel considera que tiene? (Según MCER) / *Jak oceniłby/oceniłaby Pan/ Pani swój poziom hiszpańskiego? (według skali ESOKJ)

21 respuestas

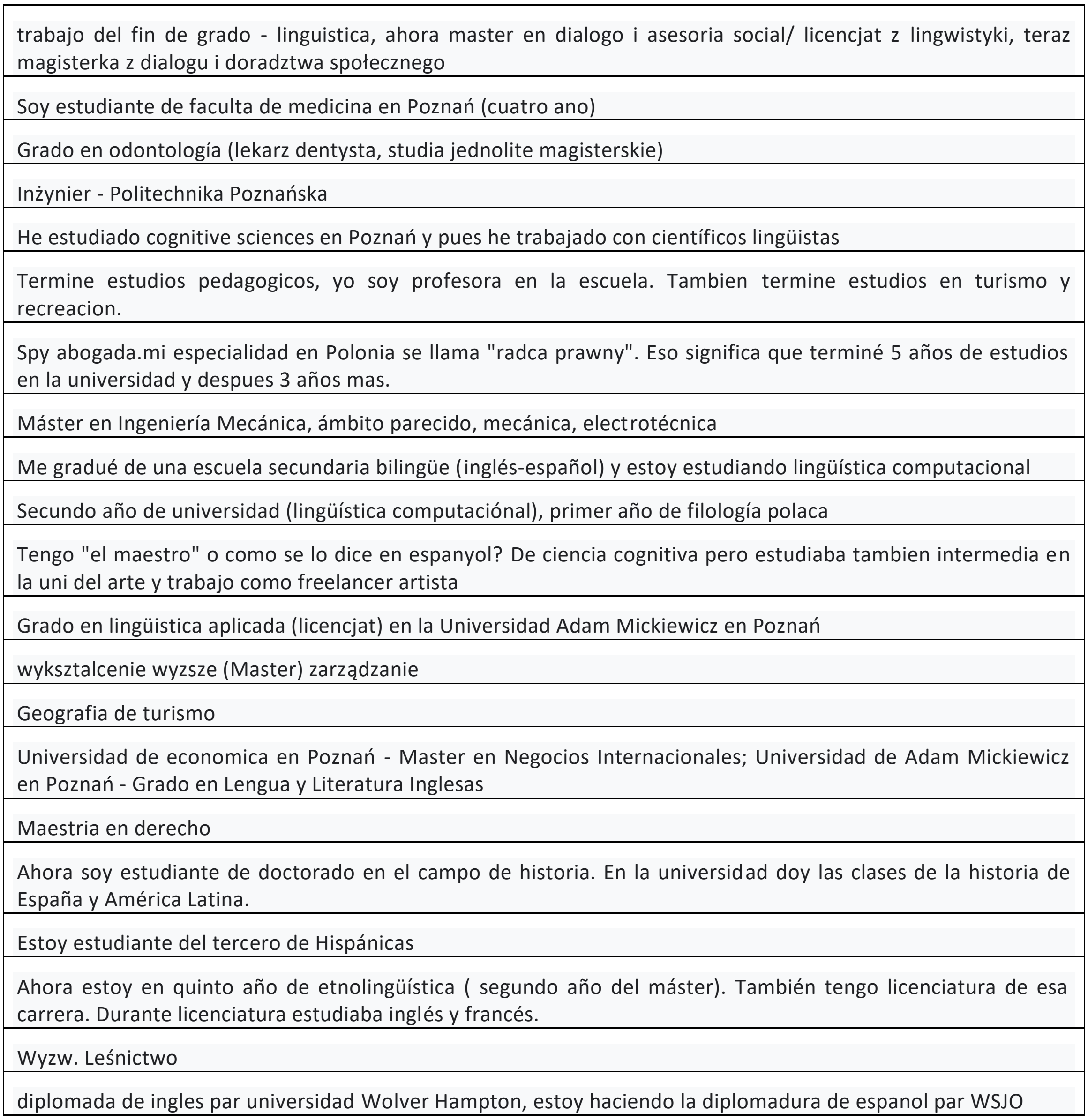
Figure 5: Participants' level of Spanish

Para hablar con amigos, leer textos, ver las series en Netflix/ do rozmawiania z przyjaciółmi, czytania tekstów, oglądania seriali
leer las textos de mi universidad (medicina), comunicar con mis amigos de extranjero, participar en conferencias internacionales
para comunicar con la gente de los países extranejros, que no son hispanohablantes, para ver las series y las películas en original
Para hablar con familia y amigos en sus lenguas maternas. Para poder hacer turismo y aprender culturas diferentes. Para poder acceder a las informaciones en estas lenguas para no depender de traductores.
Para leer, ver películas. Trabajo en inglés, en una empresa internacional. En francés hablo durante el curso y de veces cuando estoy viajando.
Mas a menudo para conversar. Tambien yo soy guia de la ciudad y necesito ingles para guiar los grupos.
Para conocer lan gente nueva, para leer los articulos, ver las peliculas. A veces tambien lo uso en mi trabajo.
para ver y leer cosas en Internet
para chatear con amigos, en la universidad y en el uso diario, como navegar por Internet
Para trabajar, estudiar, leer, escuchar
Polaco es mi idioma materna, ingles para muchas cosas, aleman y frances casi no uso
polaco - comunicación con mi familia y amigos; inglés - comunicación con mi novio y a veces sus amigos y



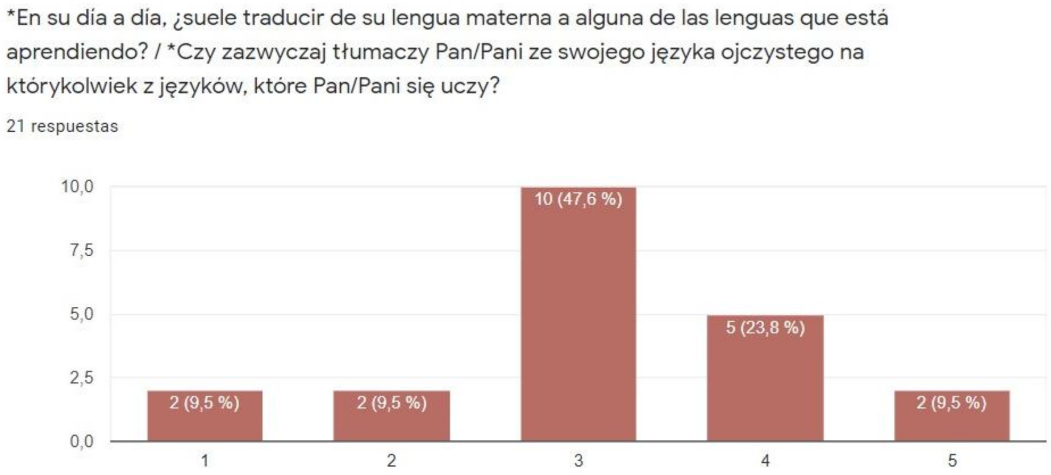
familia, alemán - durante las clases
praca, podróże, komunikacja z przyjaciółmi, użytkowanie codzienne (żyje za granica)
Para comunicarse con la gente
para trabajar en una empresa internacional
en trabajo
El trabajo, el festejar con los amigos extranjeros
Polaco para comunicarme con la gente en mi país; portugués en las clases
Para los estudios también. Para hablar con amigos extranjeros, durante los vacaciones
Towarzysko w kontaktach ze znajomymi w podróżach
para trabajo y viajes

Figure 6: Use of the foreign languages known by the participants (other than Spanish)



trabajo del fin de grado - linguistica, ahora master en dialogo i asesoria social/ licencjat z lingwistyki, teraz magisterka z dialogu i doradztwa społecznego
Soy estudiante de faculta de medicina en Poznań (cuatro ano)
Grado en odontología (lekarz dentysta, studia jednolite magisterskie)
Inżynier - Politechnika Poznańska
He estudiado cognitive sciences en Poznań y pues he trabajado con científicos lingüistas
Termine estudios pedagogicos, yo soy profesora en la escuela. Tambien termine estudios en turismo y recreacion.
Spy abogada.mi especialidad en Polonia se llama "radca prawny". Eso significa que terminé 5 años de estudios en la universidad y despues 3 años mas.
Máster en Ingeniería Mecánica, ámbito parecido, mecánica, electrotécnica
Me gradué de una escuela secundaria bilingüe (inglés-español) y estoy estudiando lingüística computacional
Secundo año de universidad (lingüística computacional), primer año de filología polaca
Tengo "el maestro" o como se lo dice en espanyol? De ciencia cognitiva pero estudiaba tambien intermedia en la uni del arte y trabajo como freelancer artista
Grado en lingüistica aplicada (licencjat) en la Universidad Adam Mickiewicz en Poznań
wyksztalcenie wyzsze (Master) zarządzanie
Geografia de turismo
Universidad de economica en Poznań - Master en Negocios Internacionales; Universidad de Adam Mickiewicz en Poznań - Grado en Lengua y Literatura Inglesas
Maestria en derecho
Ahora soy estudiante de doctorado en el campo de historia. En la universidad doy las clases de la historia de España y América Latina.
Estoy estudiante del tercero de Hispánicas
Ahora estoy en quinto año de etnolingüística (segundo año del máster). También tengo licenciatura de esa carrera. Durante licenciatura estudiaba inglés y francés.
Wyzw. Leśnictwo
diplomada de ingles par universidad Wolver Hampton, estoy haciendo la diplomadura de espanol par WSJO

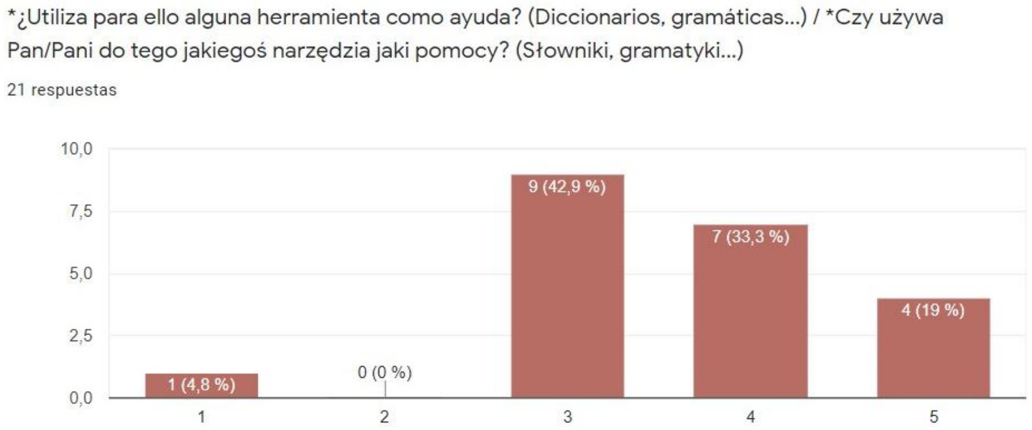
Figure 7: Academic background of the participants



*En su día a día, ¿suele traducir de su lengua materna a alguna de las lenguas que está aprendiendo? / *Czy zazwyczaj tłumaczy Pan/Pani ze swojego języka ojczystego na którykolwiek z języków, które Pan/Pani się uczy?

21 respuestas

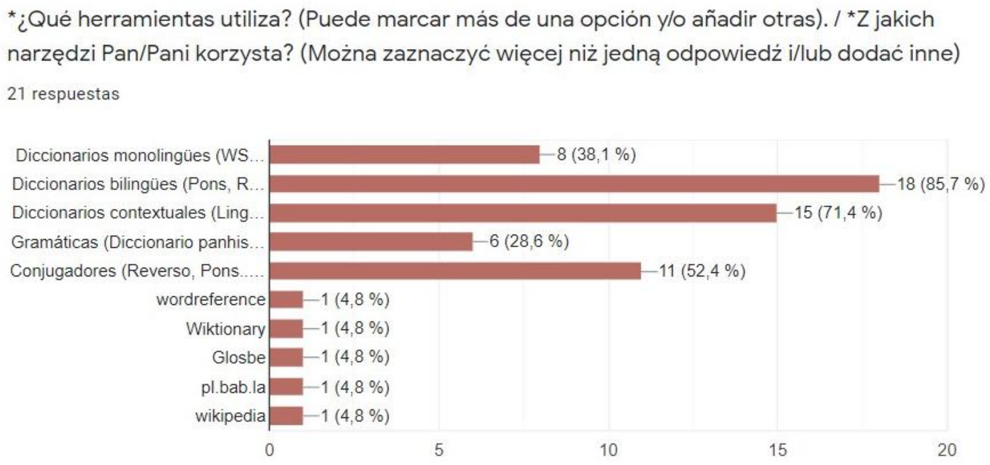
Figure 8: Frequency in which participants translate from or to their L1



*¿Utiliza para ello alguna herramienta como ayuda? (Diccionarios, gramáticas...) / *Czy używa Pan/Pani do tego jakiegoś narzędzia jaki pomocy? (Słowniki, gramatyki...)

21 respuestas

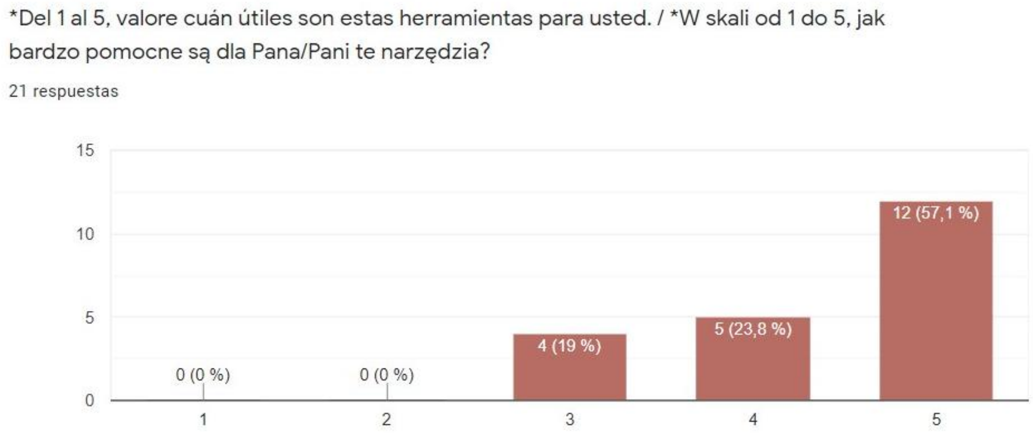
Figure 9: Use of language tools for making translations – general context



*¿Qué herramientas utiliza? (Puede marcar más de una opción y/o añadir otras). / *Z jakich narzędzi Pan/Pani korzysta? (Można zaznaczyć więcej niż jedną odpowiedź i/lub dodać inne)

21 respuestas

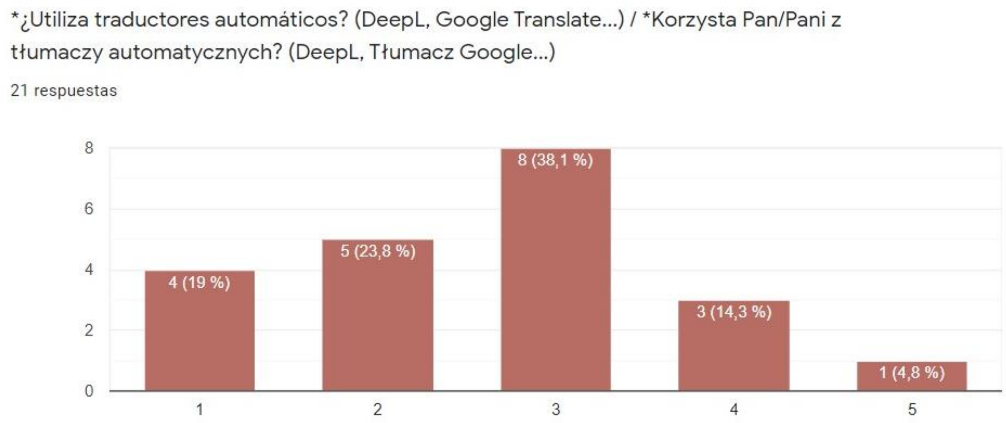
Figure 10: Types of the language tools used – general context



*Del 1 al 5, valore cuán útiles son estas herramientas para usted. / *W skali od 1 do 5, jak bardzo pomocne są dla Pana/Pani te narzędzia?

21 respuestas

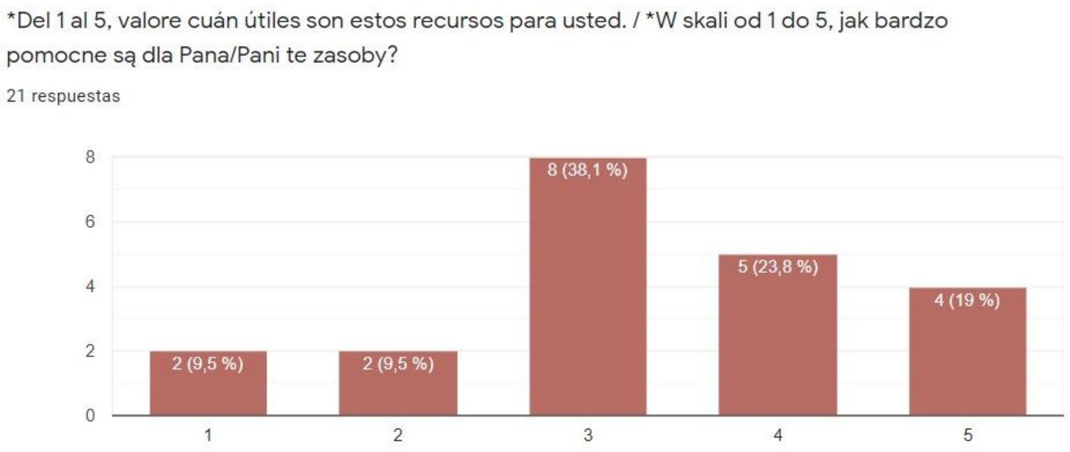
Figure 11: Degree of usefulness of language tools – general context



*¿Utiliza traductores automáticos? (DeepL, Google Translate...) / *Korzysta Pan/Pani z tłumaczy automatycznych? (DeepL, Tłumacz Google...)

21 respuestas

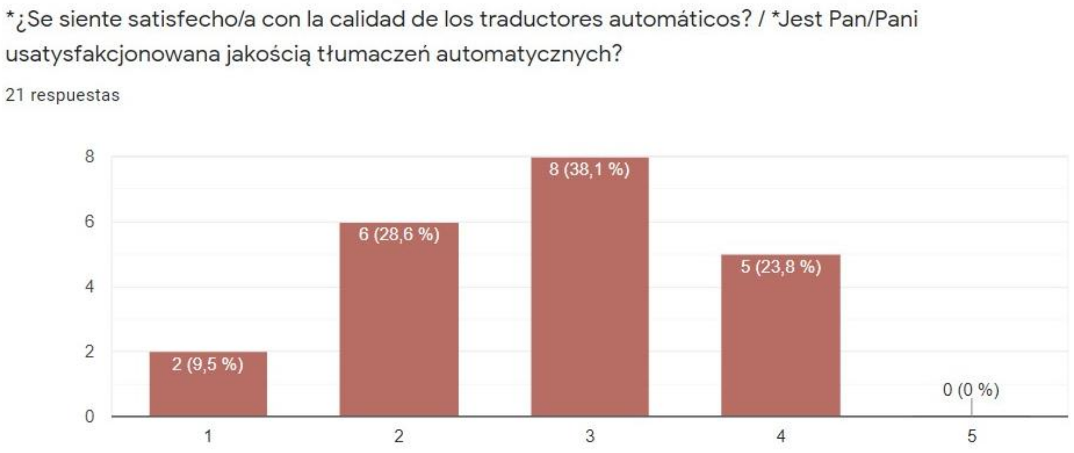
Figure 12: Use of MT systems – general context



*Del 1 al 5, valore cuán útiles son estos recursos para usted. / *W skali od 1 do 5, jak bardzo pomocne są dla Pana/Pani te zasoby?

21 respuestas

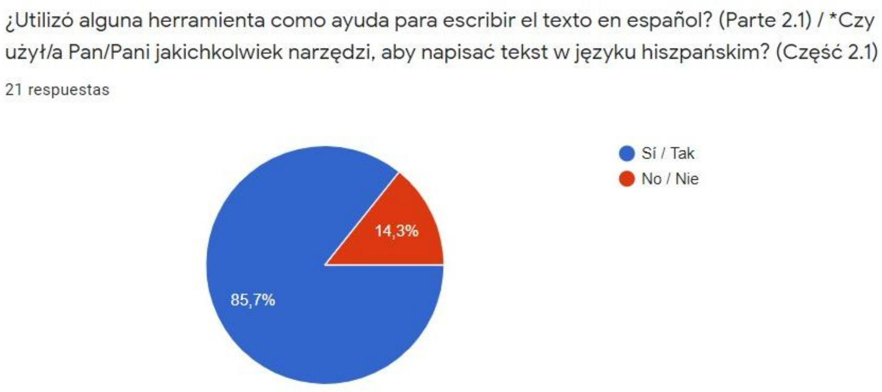
Figure 13: Degree of usefulness of MT systems – general context



*¿Se siente satisfecho/a con la calidad de los traductores automáticos? / *Jest Pan/Pani usatysfakcjonowana jakością tłumaczeń automatycznych?

21 respuestas

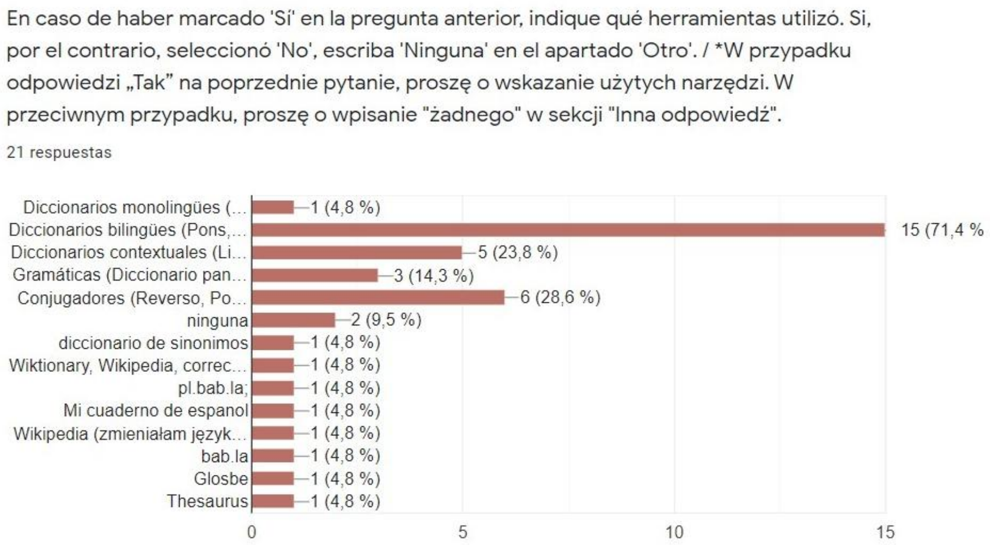
Figure 14: Degree of satisfaction with the quality of MT systems – general context



¿Utilizó alguna herramienta como ayuda para escribir el texto en español? (Parte 2.1) / *Czy użył/a Pan/Pani jakichkolwiek narzędzi, aby napisać tekst w języku hiszpańskim? (Część 2.1)

21 respuestas

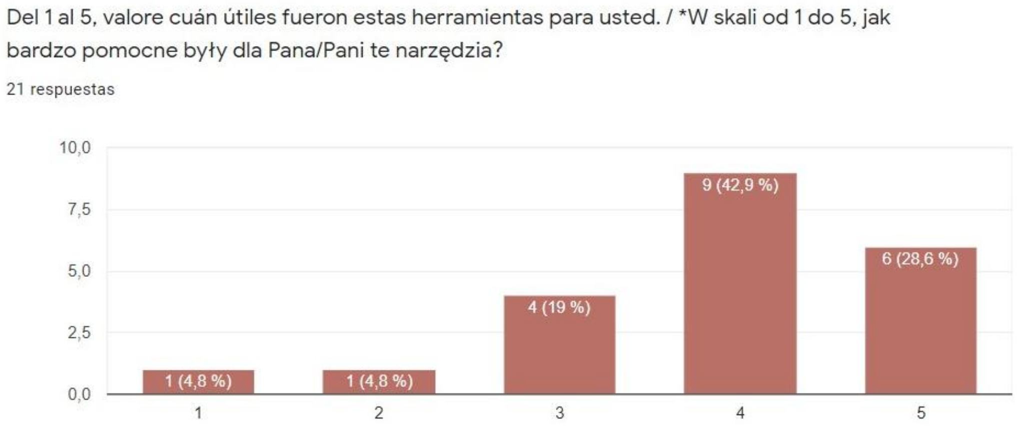
Figure 15: Use of language tools for writing *esDIR* texts



En caso de haber marcado 'Sí' en la pregunta anterior, indique qué herramientas utilizó. Si, por el contrario, seleccionó 'No', escriba 'Ninguna' en el apartado 'Otro'. / *W przypadku odpowiedzi „Tak" na poprzednie pytanie, proszę o wskazanie użytych narzędzi. W przeciwnym przypadku, proszę o wpisanie "żadnego" w sekcji "Inna odpowiedź".

21 respuestas

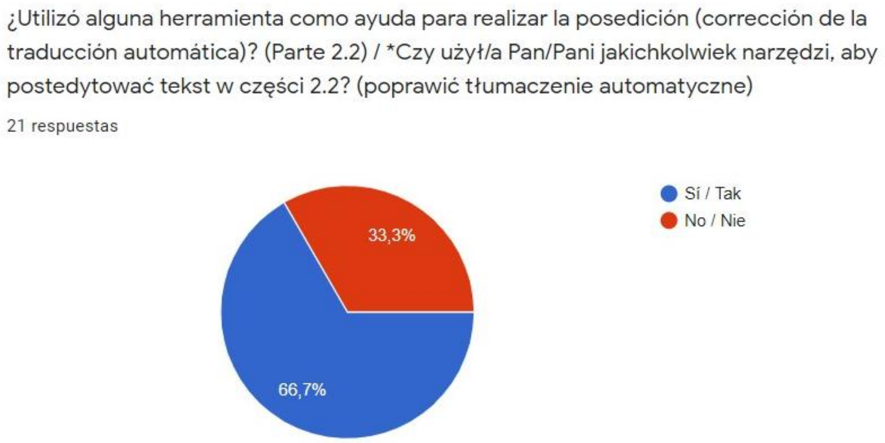
Figure 16: Types of the language tools used – *esDIR*



Del 1 al 5, valore cuán útiles fueron estas herramientas para usted. / *W skali od 1 do 5, jak bardzo pomocne były dla Pana/Pani te narzędzia?

21 respuestas

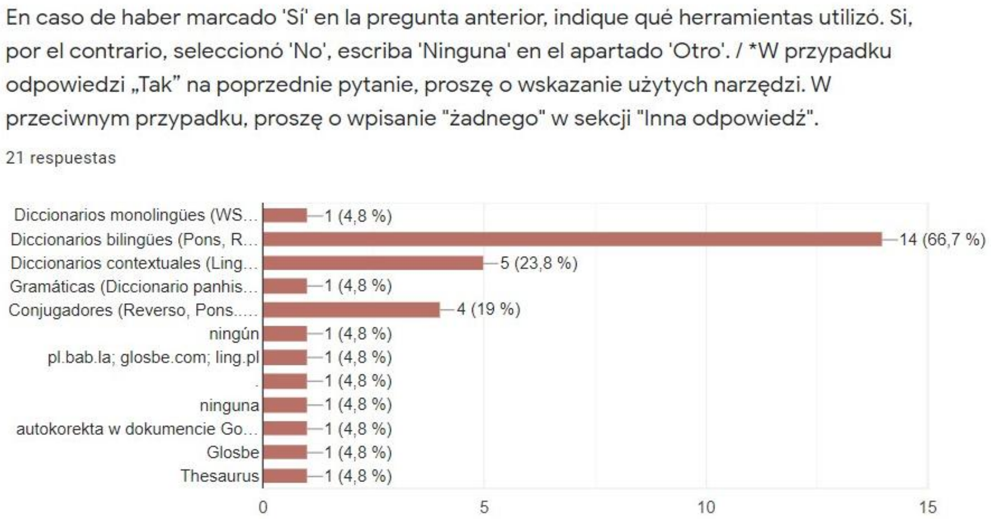
Figure 17: Degree of usefulness of language tools – *esDIR*



¿Utilizó alguna herramienta como ayuda para realizar la posedición (corrección de la traducción automática)? (Parte 2.2) / *Czy użył/a Pan/Pani jakichkolwiek narzędzi, aby postedytować tekst w części 2.2? (poprawić tłumaczenie automatyczne)

21 respuestas

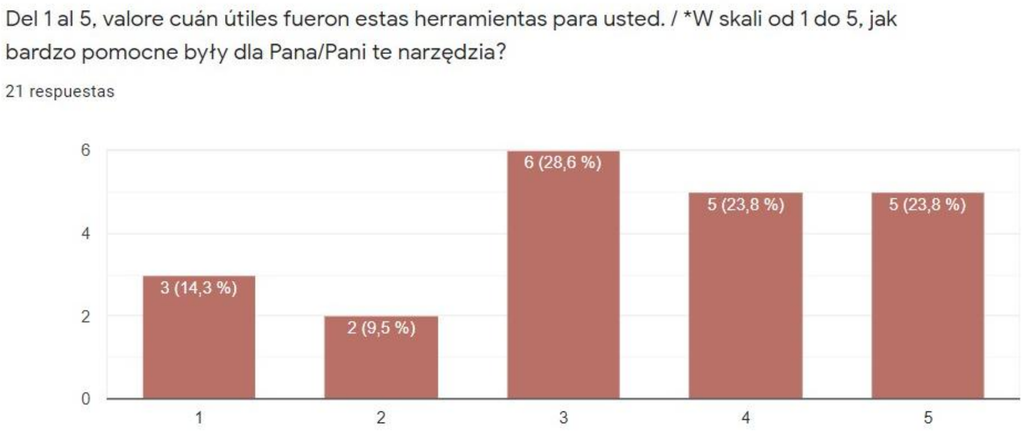
Figure 18: Use of language tools for post-editing (*esPE*)



En caso de haber marcado 'Sí' en la pregunta anterior, indique qué herramientas utilizó. Si, por el contrario, seleccionó 'No', escriba 'Ninguna' en el apartado 'Otro'. / *W przypadku odpowiedzi „Tak" na poprzednie pytanie, proszę o wskazanie użytych narzędzi. W przeciwnym przypadku, proszę o wpisanie "żadnego" w sekcji "Inna odpowiedź".

21 respuestas

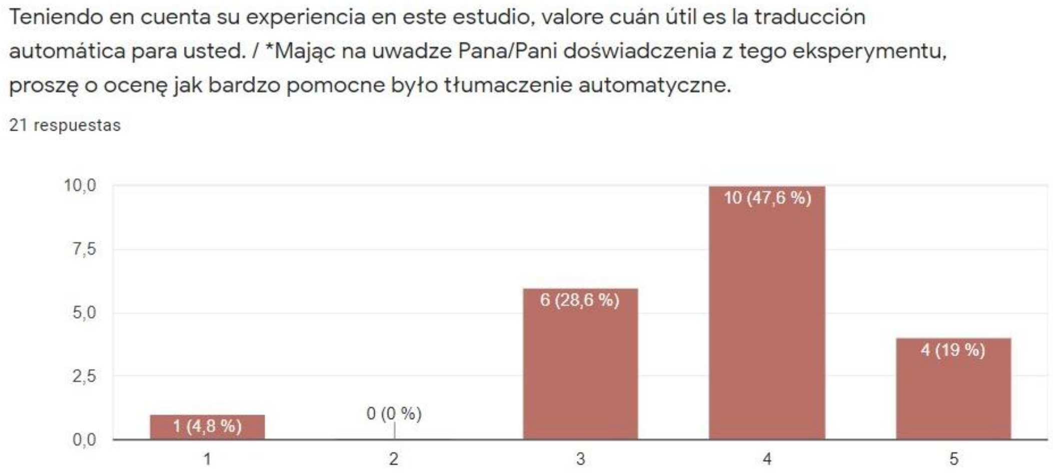
Figure 19: Types of the language tools used – *esPE*



Del 1 al 5, valore cuán útiles fueron estas herramientas para usted. / *W skali od 1 do 5, jak bardzo pomocne były dla Pana/Pani te narzędzia?

21 respuestas

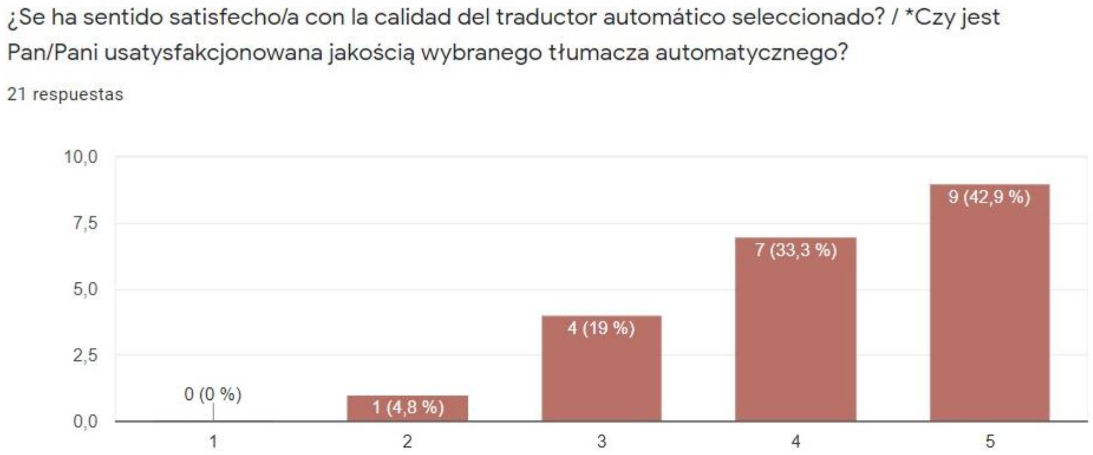
Figure 20: Degree of usefulness of language tools – *esPE*



Teniendo en cuenta su experiencia en este estudio, valore cuán útil es la traducción automática para usted. / *Mając na uwadze Pana/Pani doświadczenia z tego eksperymentu, proszę o ocenę jak bardzo pomocne było tłumaczenie automatyczne.

21 respuestas

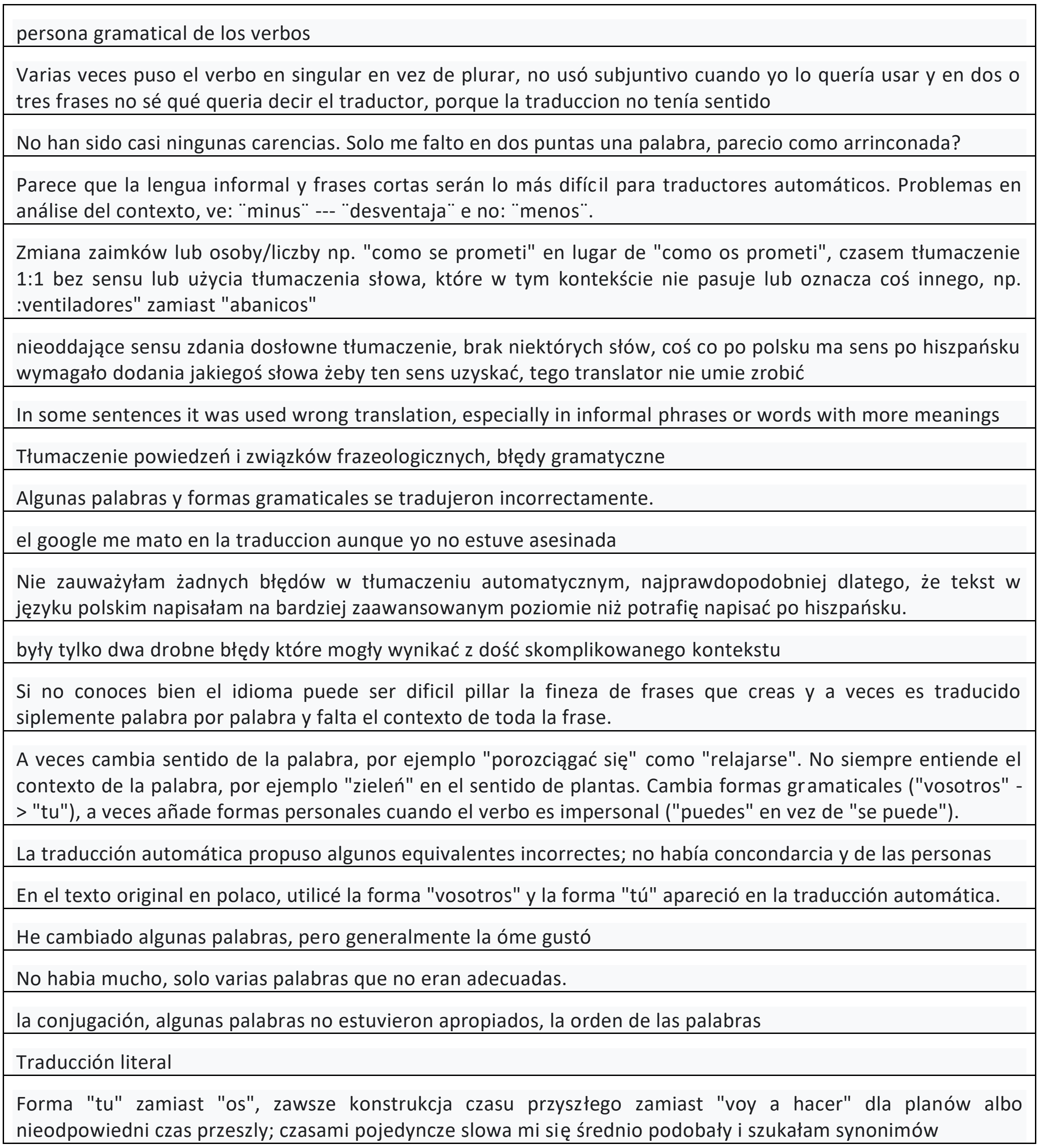
Figure 21: Degree of usefulness of MT systems – ***esPE***



¿Se ha sentido satisfecho/a con la calidad del traductor automático seleccionado? / *Czy jest Pan/Pani usatysfakcjonowana jakością wybranego tłumacza automatycznego?

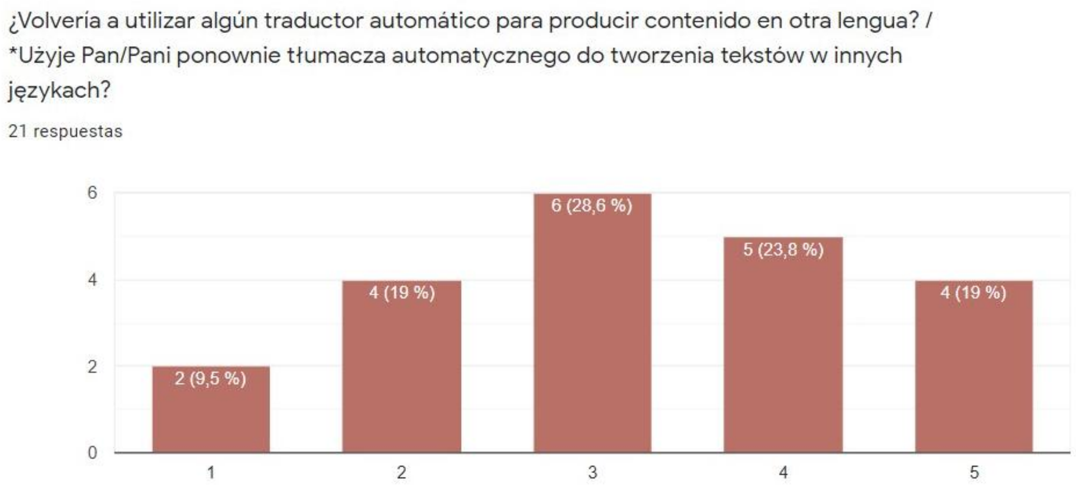
21 respuestas

Figure 22: Degree of satisfaction with the quality of MT systems – ***esPE***



persona gramatical de los verbos
Varias veces puso el verbo en singular en vez de plurar, no usó subjuntivo cuando yo lo quería usar y en dos o tres frases no sé qué queria decir el traductor, porque la traduccion no tenía sentido
No han sido casi ningunas carencias. Solo me falto en dos puntas una palabra, parecio como arrinconada?
Parece que la lengua informal y frases cortas serán lo más difícil para traductores automáticos. Problemas en análise del contexto, ve: ¨minus¨ --- ¨desventaja¨ e no: ¨menos¨.
Zmiana zaimków lub osoby/liczby np. "como se prometi" en lugar de "como os prometi", czasem tłumaczenie 1:1 bez sensu lub użycia tłumaczenia słowa, które w tym kontekście nie pasuje lub oznacza coś innego, np. :ventiladores" zamiast "abanicos"
nieoddające sensu zdania dosłowne tłumaczenie, brak niektórych słów, coś co po polsku ma sens po hiszpańsku wymagało dodania jakiegoś słowa żeby ten sens uzyskać, tego translator nie umie zrobić
In some sentences it was used wrong translation, especially in informal phrases or words with more meanings
Tłumaczenie powiedzeń i związków frazeologicznych, błędy gramatyczne
Algunas palabras y formas gramaticales se tradujeron incorrectamente.
el google me mato en la traduccion aunque yo no estuve asesinada
Nie zauważyłam żadnych błędów w tłumaczeniu automatycznym, najprawdopodobniej dlatego, że tekst w języku polskim napisałam na bardziej zaawansowanym poziomie niż potrafię napisać po hiszpańsku.
były tylko dwa drobne błędy które mogły wynikać z dość skomplikowanego kontekstu
Si no conoces bien el idioma puede ser dificil pillar la fineza de frases que creas y a veces es traducido siplemente palabra por palabra y falta el contexto de toda la frase.
A veces cambia sentido de la palabra, por ejemplo "porozciągać się" como "relajarse". No siempre entiende el contexto de la palabra, por ejemplo "zieleń" en el sentido de plantas. Cambia formas gramaticales ("vosotros" -> "tu"), a veces añade formas personales cuando el verbo es impersonal ("puedes" en vez de "se puede").
La traducción automática propuso algunos equivalentes incorrectes; no había concondarcia y de las personas
En el texto original en polaco, utilicé la forma "vosotros" y la forma "tú" apareció en la traducción automática.
He cambiado algunas palabras, pero generalmente la óme gustó
No habia mucho, solo varias palabras que no eran adecuadas.
la conjugación, algunas palabras no estuvieron apropiados, la orden de las palabras
Traducción literal
Forma "tu" zamiast "os", zawsze konstrukcja czasu przyszłego zamiast "voy a hacer" dla planów albo nieodpowiedni czas przeszly; czasami pojedyncze slowa mi się średnio podobały i szukałam synonimów

Figure 23: Deficiencies observed in the machine translations – ***esPE***



¿Volvería a utilizar algún traductor automático para producir contenido en otra lengua? / *Użyje Pan/Pani ponownie tłumacza automatycznego do tworzenia tekstów w innych językach?

21 respuestas

Figure 24: Attitude towards MT systems regarding future use

Eran muy útiles cuando necesitaba usar las expresiones fijas como refranes, pero no me ayudaron en mantener el sentido del discurso.
No sé qué decir, quizá auydó con la idea en general, pero luego, como la traducción era mala, me llevó tiempo corregirla. Creo que es buena herramienta cuando tienes que escribir un texto que no tiene que ser de muy alta calidad y no te da la gana pensar en el idioma. Pero si quiero curarme un texto es mucho mejor para mi escribirlo en español directamente y aclarar mis dudas con context reverso
La traduccion automatica me ayudo en general, para corregir mis propios errores. Me falto en eso solo la traduccion de las expressiones tipicos de la lengua original.
No me ha ayudado la traducción automática debido a que ya poseo nível de castellano suficiente para evitar el uso de ellos. También puede ser que nunca los uso por la desconfianza que tengo hacia ellos. También ojo que por ejemplo Google Docs ha sugerido que corrigiera ¨rusa¨ a ¨rusia¨ y ¨hablaremos¨ a ¨hablamos¨ - que estoy convencido que lo que he escrito yo es correcto. Si suena natural para los españoles es otro asunto pero en cúanto a grámatica creo que es correcto.
Generalmente no me gusta usar la traducción automatica, prefiero escribir el texto directamente en la lengua extranjera. Es mas dificil para mi corregir el texto despues de una traduccion automatica que escribirlo desde el principio en la lengua extranjera.
pomogło w szybkim tłumaczeniu, które wymaga tylko kilku poprawek
Es mas rapido para usar el traduccion automatica y despues corectar esto que escribir el texto usando solo el diccionario.
Pomocne w znalezieniu bezpośredniego tłumaczenia słów o jednym znaczeniu
La traduccion tradujo palabras que no recordaba o no conocia
la traduccion automatica hacia la mayoria de la traduccion, solo no traduzco mi traduccion del nombre de mi caballo en polaco y me mato entonces cambio el contexto
Przetłumaczyło cały mój tekst tak, że nic bym w nim nie poprawiła.
tłumaczenia pomagają w przypomnieniu słownictwa
Cuando utilizo traductores automaticos lo hago para rapidamente obtener el núcleo de lo que quiero traducir y para esto lo veo muy util, luego corrijo lo que no me suena bien y ya esta, pero para traducir algunas palabras solas prefiero usar diccionarios porque los traductores automaticos me dan solo una respuesta y no se de que contexto viene.
Me ayudó en todo :) Tradujo muy bien todo el texto, entendió el contexto en frases compuestas. Las carencias que he mencionado no influyeron mucho a la calidad de traducción en general. En general, creo que el resultado de traducción automática no suena extraño y poco natural.

Era perfecta para primera versión de la traducción. No tenía que buscar palabras en mi cabeza , porque las ya tenia traducidas
La traducción automática ayudó con el orden correcto de las oraciones y una buena elección de artículos y terminaciones para los sustantivos y adjetivos.
La traducción me ha mostrado frases nuevas o que yo no utilizo, por ejemplo estructuras gramáticas, pues puede ser útil para aprender los idiomas.
En entender lo que ya las herramientas para la traducción automatica ya son muy avanzadas. Era la sorpresa.
Me ha ayudado para ver que tiempo de gramatica puedo usar
Ayuda a encontrar rápidamente las palabras que necesita.
formy czasu przeszłego albo przyszłego - sama nie zawsze jestem pewna; no i oczywiście było szybciej sprawdzać tłumaczenie niż pisać samemu

Figure 25: Extent to which MT systems did or did not help the users

Lo que dije en la pregunta anterior, la herramienta puede ser util para los textos simples, pero no mucho más. En mi caso la traducción era mala, puede ser por mi estilo, no sé. Yo ya sé que si quiero usar google translate, tengo que tener cuidado
Fue un placer participar en el estudio. antes de eso, tenia un poco de miedo, pero al fin olvide sobre todo mi estreso. Tengo mis dedos cruzados por los resultados y, sobre todo, por la autora de este estudio.
El estudio fue muy interesante, me gustó mucho. Me alegro que pude participar en el.
;)
Bardzo ciekawe badanie i chętnie wezmę udział w podobnych w przyszłości.
Era genial y me gustaban todas las ejercicias, aunque algunas eran dificil :)
Un estudio super bueno! Me encanto muchisimo! Muchas gracias :)
Było muy muy muy simpatico :)
zaskakująco dobre tłumaczenie z tekstu polskiego
Una cosa muy interesante para mi - ser conciente de que despues de escribir el texto polaco lo vamos a traducir y corregir he notado que mi estilo se cambió como si intentara escribir en polaco mas facil para que el traductor lo haya traducido bien.
Escribiendo texto en polaco, a veces estaba pensando si debería limitar mí creatividad sabiendo que el siguiente paso de la tarea sería la traducción que podría resultar muy difícil (ya que el texto polaco era difícil). No lo hice y no me arrepentí, porque la traducción automática tradujo todo muy bien. Además, me gustó mucho el tema del texto (viajes) en sí :)
El estudio fue muy interesante y revelador de que las traducciones automáticas son solo una ayuda adicional para nosotros y no una necesidad cuando escribimos textos originales.
A mi gustaba mucho este experimento.
Super doświadczenie, bardzo mi się podobało; myślę że możliwe byłoby jeszcze różnicowanie motywów dla których używa się tłumacza automatycznego albo czasami jako słownik do przetłumaczenia poszczególnych słów (nie kontekstu) . Ja używam go przede wszystkim kiedy mam przed sobą tekst w języku którego zupełnie nie znam (w Internecie) aby mniej więcej zrozumieć o co chodzi. Oczywiście nie zależy mi wtedy na tak dokładnym tłumaczeniu jak osoby które przy użyciu tłumacza piszą teksty w obcym języku.

Figure 26: General comments and observations made by the participants about the experiment