Post-editing Effort and Linguistically Motivated Evaluation of Machine Translation

Sergi Alvarez Vidal

TESI DOCTORAL UPF / ANY 2021

DIRECTOR DE LA TESI
Antoni Oliver i Toni Badia
Deparatment Traducció i Ciències del Llenguatge

A la Damna, l’Etna i l’Alguer, perquè ho sou tot per a mi.

Acknowledgements

Primer de tot, voldria donar les gràcies a la meva dona, la Damna. La seva comprensió i perseverança han fet poosible que tingués el temps necessari que cal invertir per fer una tesi. També vull donar les gràcies a les meves filles, que tot i no entenien gaire bé què volia dir fer una tesi doctoral, m'animaven i respectaven que em tanqués a fer feina quan calia.

I would like to thank my supervisors, Toni Badia and Antoni Oliver, who haven given me all their support throughout this journey. Since the beginning, they offered their help and opened the doors to all what research means. Without them, none of this could have been ever possible. TB, your calm and reflection have helped me understand what I wanted to do. TO, working side-by-side with you has been real fun and a path full of knowledge.

I would also like to thank the two supervisors I had at Dublin City University, Sharon O'Brien and Felix do Carmo. Sharon, thank you for your comments and for adapting to the new situations. Felix, I really appreciate all the warmth and help you gave me. I would also like to thank all the friends and colleagues from DCU who helped me feel at home in Dublin and are an essential part of this journey in one way or another: Eva, Dimi, Carlos, Wine, Alessandra, Joss, Ana and Sheila. Thanks. I would also like to thank Maja Popovic and Antonio Toral, who have always been kind and willing to share their knowledge and expertise.

Finally, I would like to thank Núria Bel for listening to me and for her valuable suggestions, and Elisenda Bernal, who has given all her support as a tutor all these years.

Abstract

The recent improvements in neural MT (NMT) have driven a shift from statistical MT (SMT) to NMT, which has propelled the use of post-editing (PE) in translation workflows. However, many professional translators state that if the quality of the MT output is not good enough, they delete the remaining segments and translate everything from scratch. The problem is that usual automatic measurements do not always indicate the quality of the MT output, especially with high quality outputs, and there is still no clear correlation between PE effort and productivity scores.

We combine quantitative and qualitative methods to study some of the usual automatic metrics used to evaluate the quality of MT output, and compare them to measures of post-editing effort. Then, we study in detail different direct and indirect measures of effort in order to establish a correlation among them. We complement this study with the analysis of translators' perceptions of the task.

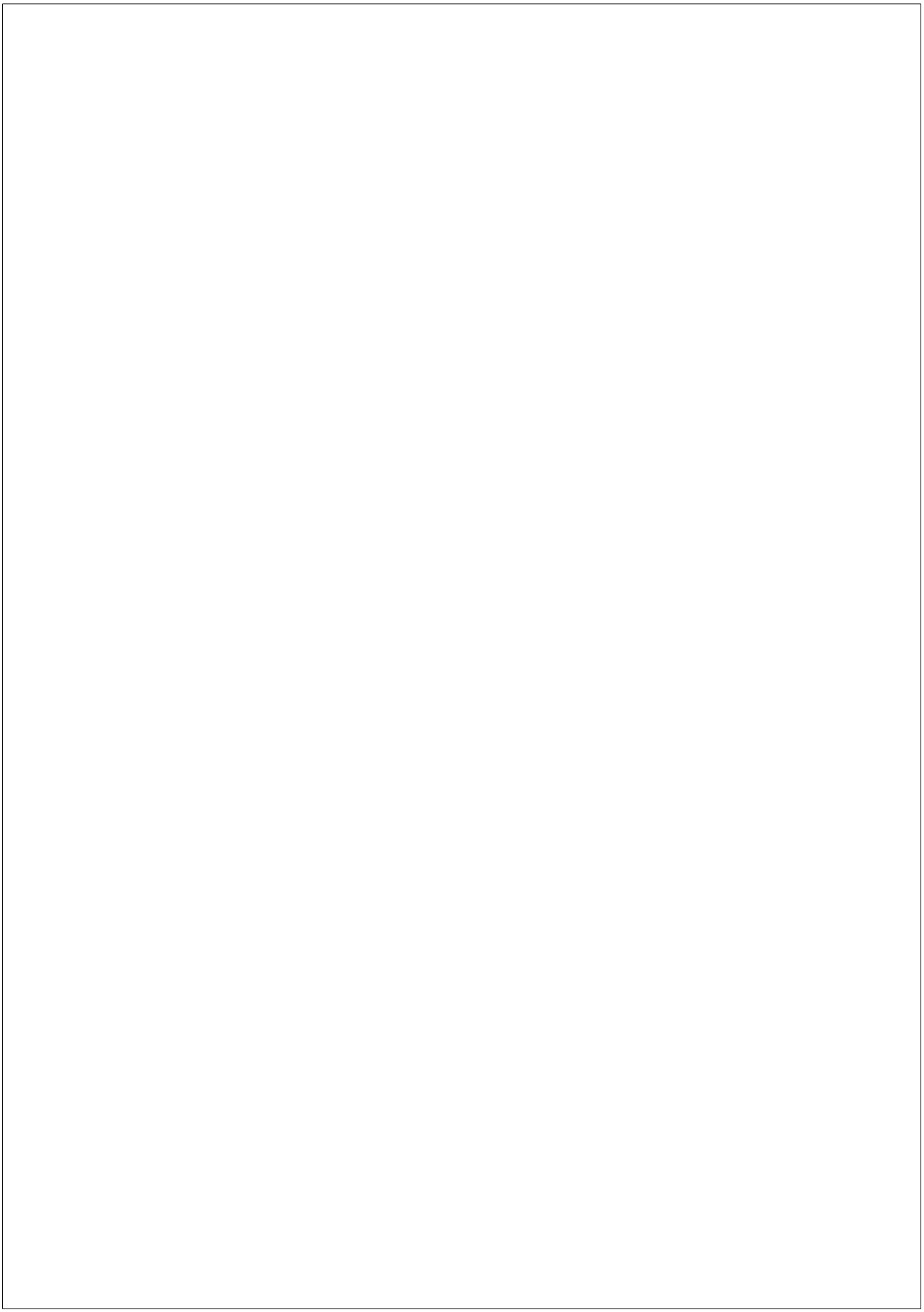
Finally, we conduct a fine-grained analysis of MT errors based on post-editing corrections and suggest an error-based approach to evaluate raw MT output which includes the use of challenge sets.

Resum

Les recents millores en traducció automàtica neuronal (TAN) han provocat un canvi de la traducció automàtica estadística (TAS) a la TAN, que ha incrementat l'ús de la postedició en els fluxos de treball industrials. Tanmateix, molts traductors professionals afirmen que si la qualitat de la TA no és prou bona, eliminen tot el segment i el tradueixen de cap i de nou. El problema és que les mesures automàtiques habituals no sempre indiquen la qualitat de la TA, especialment quan aquesta és bona, i no hi ha una correlació directa entre l'esforç de postedició i les mesures de productivitat.

Combinem mètodes quantitatius i qualitatius per estudiar algunes de les mesures automàtiques més habituals utilitzades per evaluar la qualitat de la TA, i les comparem amb les mesures de l'esforç de postedició. A continuació, estudiem en detall diferent mesures directes i indirectes d'esforç per establir-hi una correlació. Complementem aquest estudi amb l'anàlisi de les percepcions dels traductors que duen a terme aquesta tasca.

Finalment, fem una anàlisis detallada dels errors de TA a partir de les correccions fetes en la postedició i proposem un enfocament basat en errors per evaluar la TA que inclou l'ús d'un conjunt de frases de prova.



Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

Even though all the recent developments in machine translation (MT) have placed post-editing (PE) as a very common practice in the translation industry, in the late 1950s and early 1960s it was a surprisingly hot topic (García, 2012). A clear sight of what could be achieved was presented, but technology lacked the requirements to fulfill the expectations.

However, the ALPAC report (ALPAC, 1966) stated that it would be no real help funding MT research to improve or accelerate the translation process. Obviously, this put an end to large-scale MT funding as it established that MT satisfied a non-existent need, which was already fulfilled with translators.

But despite this setback, some projects continued to exist. For example, Systran was founded in 1968 and soon grouped clients such as General Motors and Caterpillar (García, 2012). METEO, the Canadian MT system for translating weather report from English into French, is another example of MT deployment. In the 80s, the Pan American Health Organization implemented PE to translate their documentation from English into Spanish (Vasconcellos and León, 1985).

The advances in computing capabilities and the irruption of the Internet

in the 1990s increased the use of PE in big institutions like the European Commission (EC). Since then, many companies and international organizations have added PE as part of their usual workflow to manage the translation of many thousands of words with a short turnaround.

MT research also increased, and free online MT services opened the way to new uses of MT output, including different levels of post-editing. This includes light post-editing, in which only most obvious typos, word, and grammatical errors are corrected, and also full post-editing, in which publishable-quality translations are produced. However, this is a blurry distinction which poses in practice many problems to translators.

The change of century consolidated the use of SMT in translation workflows usually for in-domain languages and for technical documents or documents using controlled languages. In mid 2010s, the attention focused on neural machine translation (NMT) due to the promising results obtained in terms of quality. These results have increased the interest in this new paradigm for the translation industry, which has begun to substitute its corpus-based predecessor, statistical machine translation (SMT), for new NMT models.

It has also boosted the incorporation of PE in many translation workflows. A 2016 survey showed that more than half of the Language Service Providers (LSPs) offered PEMT as a service (Lommel and Depalma, 2016) and in the 2018 Language Industry Survey[1] 37% of the respondents reported an increase of MT post-editing and an additional 17% indicated that they had started implementing this practice.

Post-editors "edit, modify and/or correct pre-translated text that has been processed by an MT system from a source language into (a) target language(s)" (Allen, 2003, p. 297). Yet, many professional translators state that after a few segments post-editing MT, they delete the remaining segments and translate everything from scratch if they consider it will take them less time (Parra Escartín and Arcedillo, 2015). Research has com-

[1] http://fit-europe-rc.org/wp-content/uploads/2019/05/2018-Language-Industry-Survey-Report.pdf?x77803

pared the levels of human effort, speed and translation quality obtained with traditional translation and PE. In most cases PE seems to increase productivity (Plitt and Masselot, 2010) but it does not always improve translators' satisfaction or PE effort.

1.1 Motivation

As a way of addressing some of these issues, the purpose of the present thesis is to investigate different potential factors that might be related to the effort implied in PE. This includes the three dimensions stated by Krings (2001): temporal, technical and cognitive.

Our empirical investigation analyzes the correlation between automatic indicators of MT quality and measures of PE effort, which do not always correlate (Koponen, 2016). We will study the most salient elements which affect the PE process and establish the most convenient direct or indirect scores to measure it.

Findings can serve a number of purposes, including predicting the effort posed by the activity, and help determine if a raw MT output presents the quality necessary for PE.

1.2 Goals

The main goal of this thesis is to analyze what factors influence post-editing effort and which automatic, direct or indirect measures of post-editing effort can be used to establish a linguistically-motivated evaluation of the raw MT output. As such, we introduce a number of main goals and specific objectives which we will revisit in Chapter 6.

The **main goals** are:

- Assess the importance of all elements included in the MT process into the global MT quality for post-editing.

- Create an evaluation model enriched with linguistically motivated features to assess the quality of MT sentences in a translation workflow with post-editing.

The **specific goals** are:

- Assess current automatic and manual MT evaluation methods.
- Assess direct and direct PE effort indicators.
- Study the linguistic errors with the major incidence when post-editing.
- Generate an evaluation model for raw MT output which takes into account PE effort.

1.3 Structure of the Thesis

Chapter 2 presents a general literature review and details of the methodologies used throughout the thesis.

Chapter 3 presents the main MT models, their characteristics and evaluation methods. It compares the two most used architectures, SMT and NMT, and tests its usefulness for post-editing.

Chapter 4 explains what post-editing is and the different dimensions of post-editing effort. It details the different direct and indirect ways to measure PE effort. It also presents a tool designed to measure the main effort indicators and includes an experiment conducted to measure the correlation between these measures.

Chapter 5 studies how PE is perceived by translators. It compares the perception of the task with productivity results.

Chapter 6 studies the evaluation based on errors and how it can be used to assess the raw MT output to predict PE effort.

Chapter 7 presents the main conclusions and explains the possible future work to be carried out.

1.4 Publications and Presentations

Part of the work included in this dissertation has been published in peer-review conference proceedings and journal papers or presented at workshops. A list of these publications or presentations follows:

1. Alvarez, S., Oliver, A., and Badia, T. (2019). Does NMT Make a Difference when Post-editing Closely Related Languages? The Case of Spanish-Catalan. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 49–56, Dublin, Ireland. European Association for Machine Translation.

2. Alvarez, S., Oliver, A., and Badia, T. (2019). Implementing MTPE into a real industrial scenario: what do translators need for a fair MT workflow?. Presented at Fair MT: Building ethical and sustainable MT workflows.

3. Oliver, A., Alvarez, S., and Badia, T. (2020). PosEdiOn: Post-Editing Assessment in PythOn. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 403-410).

4. Alvarez, S., Oliver, A., and Badia, T. (2020). Quantitative Analysis of Post-Editing Effort Indicators for NMT. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 411-420).

5. Alvarez, S., Oliver, A., and Badia, T. (2020). Post-editing for Professional Translators: Cheer or Fear?. In *Revista tradumàtica: traducció i tecnologies de la informació i la comunicació*, (18), 49-69.

6. Alvarez, S., Oliver, A., and Badia, T. (2020). Comparing NMT and

PBSMT for Post-editing In-domain Formal Texts: a Case Study. In Tra&Co Group (ed.), *Translation, interpreting, cognition: The way out of the box*, 33–47. Berlin: Language Science Press. (pending publication)

7. Alvarez, S., Oliver, A., and Badia, T. (2020). What Do Post-editors Correct? A Fine-grained Analysis of SMT and NMT Errors. In *Revista tradumàtica: traducció i tecnologies de la informació i la comunicació*, (18). (accepted for publication).

Chapter 2

RESEARCH ON POST-EDITING

In this chapter we include a general approach to the current state-of-the-art and methodologies. We present an overview of the current research regarding MT and post-editing, even though in each of the following chapters of this thesis we develop a description of the relevant research. In Chapter 3, we detail the MT evaluation methods. In Chapter 4 we analyze PE effort and all the evaluation scores used to measure its three dimensions. And in Chapter 4 we focus on the error-based evaluation of MT.

Furthermore, we describe the methods used throughout the thesis and further detailed in the methodology corresponding to each of the experiments.

2.1 Literature Review

Statistical MT (SMT) has been well established as the dominant approach in MT for many years. However, recent developments in neural MT (NMT) have generated much excitement because these systems have out-

performed previous systems in terms of quality in recent evaluation campaigns (Barrault et al., 2019, 2020). These results have driven a technological shift from SMT to NMT in many translation industry scenarios.

MT research has always been linked to the research on its evaluation. Assessing the different MT systems and its output becomes essential both to check the results and to improve future MT models. It includes both manual evaluation and automatic metrics. Manual evaluation, such as sentence ranking, fluency and adequacy and direct assessment (DA) (Graham et al., 2016), produces quite reliable metrics but implies more time and money, and suffers from low inter- and intra-annotator agreements (Turian et al., 2003; Snover et al., 2006).

Automatic metrics produce faster results and include a plethora of measures such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006), (NIST) (Doddington, 2002) and chrF (Popović, 2015). One of the main problems of automatic scores is that they usually compare MT outputs to one or more reference translations and measure the differences. However, there are many possible correct translation for one single document, which could produce many different scores.

Even though automatic scores usually show correlation with human judgments of translation (Coughlin, 2001), it has been repeatedly questioned as a means to assess MT output (Callison-Burch et al., 2006; Wieting et al., 2019; Mathur et al., 2020), especially when comparing high-quality systems (Ma et al., 2019). Furthermore, scores for the same measures are often calculated in different ways, which entails a great variation and produces scores which are not really comparable (Post, 2018). Besides, recent research has shown measures like BLEU tend to underestimate NMT output (Shterionov et al., 2018).

Recently, the problems of current evaluation systems have been on the spotlight due to the claims of parity between NMT and human translations (Hassan et al., 2018; Popel, 2018). A deeper analysis exposed the flaws of the evaluation methods used for these claims, which are the usual ones for all MT systems. For example, current evaluation methods do not

use professional translators as evaluators for manual evaluations and do not take into account the context but only assess the output at the sentence level (Toral et al., 2018; Läubli et al., 2018). Current NMT systems produce high quality outputs which demand fine-grained methods for its evaluation (Läubli et al., 2020).

The consistent increase in quality produced by MT systems, especially NMT models, has been paired with an increase of its use in the translation workflows (Lommel and Depalma, 2016). For example, 2019 Language Industry Survey Report [1] conducted by EUATC [2] pointed out more than half of the companies and translators wanted to increase the use of MT. This includes the use of raw MT output for post-editing by professional translators.

Hence, we should devise evaluation methods which can assess the quality of an MT output depending on the task or the function for which the output is intended (Hovy et al., 2002). In the case of post-editing, we can assess, for example, if the quality of the MT output is good enough to post-edit. In case it contains too many errors, translators often report they find it easier to translate from scratch than to post-edit it (Sanchez-Torron and Koehn, 2016).

Research on this field has usually been focused on the effort translators need to post-edit the MT output, depending on the quality and type of text (Guerberof, 2009a,b; Specia, 2011, 2010), but also on the tools used to post-edit (Castilho et al., 2014; Moorkens and O'Brien, 2013).

It has also studied the impact of post-editing on productivity, basically based on time (O'Brien, 2011b; Parra Escartín and Arcedillo, 2015; Plitt and Masselot, 2010; Sanchez-Torron and Koehn, 2016) and the perceived post-editing effort (Moorkens et al., 2015).

All previous research has included the three dimensions of effort sug-

[1] https://euatc.org/wp-content/uploads/2019/11/2019-Language-Industry-Survey-Report.pdf

[2] https://euatc.org/

gested by Krings (2001): temporal (time spent post-editing), technical (amount of editing work implied in post-editing) and cognitive (mental processes involved in the PE task). All three dimensions are related, but there is not a single measure that includes them all (Moorkens et al., 2015). Temporal effort is the dimension which is most straightforward to measure. Studies consistently show that post-editing is faster than translating from scratch (O'Brien, 2005; Carl et al., 2011; Aranberri et al., 2014; Jia et al., 2019; Läubli et al., 2019).

Technical effort refers to the editing conducted by the translator while post-editing the raw MT output. It can be captured by registering the number of insertions, deletions and re-orderings, and is usually measured with keystroke analysis or key-logging data. It is calculated with indirect measures based on the post-edited product. The most usual scores are HTER (Snover et al., 2006), HBLEU and edit distance.

Finally, cognitive effort is related to the cognitive demand. In educational psychology, effort is one of the elements included in the cognitive load, which involves three variables: mental load, mental effort and performance (Paas et al., 1994; Kirschner, 2002; Paas et al., 2003). According to cognitive load theory, the load implied by a task can be intrinsic and extraneous (Nunes Vieira, 2016). MT output could be linked to the intrinsic cognitive load and the extraneous cognitive load would be related to all external factors.

As cognitive effort cannot be measured directly different proxy measures have been used, which include think aloud protocol (TAP) (Krings, 2001; Nunes Vieira, 2016; Alves, 2003), pause measuring (Lacruz and Shreve, 2014; Lacruz, 2012), subjective ratings (Koponen, 2012; Vieira, 2016; Gaspari and Toral, 2014) eye-tracking (Carl et al., 2011; O'Brien, 2011b; Doherty, 2013) and choice network analysis (Campbell, 1999, 2000).

However, these intrinsic measures of effort are affected by extraneous elements, such as the tools used or the perception of the task (Moorkens et al., 2018; Nunes Vieira and Alonso, 2018). Research shows translators perceive the edition of MT outputs as less productive even when a quanti-

tative analysis shows otherwise (Gaspari and Toral, 2014). They consider MT output to be tedious to post-edit (Moorkens and Brien, 2017) and they prefer to translate from scratch even if this has a negative impact on productivity (Teixeira, 2014).

Another way to assess MT outputs is to analyze the errors it contains by error classification and error analysis. It is useful to assess the quality of the MT output and it can serve to detect patterns of correction in the PE process (Popović, 2018). In fact, post-editing can be understood as an implicit error annotation (Popović and Arčan, 2016).

Recent research (Bentivogli et al., 2018; Castilho et al., 2017c; Klubicka et al., 2017) shows NMT reduces the number of errors compared to SMT, but different errors imply different cognitive effort (Daems et al., 2017b). Hence, studying MT errors can help to shed light on the quality of MT outputs (Koponen et al., 2019).

An alternative evaluation system based on errors which has been long used in MT evaluation is challenge sets (King and Falkedal, 1990; Isahara, 2006; Koh et al., 2001; Arnold et al., 1993). They have a number of advantages: systematicity, control over data, inclusion of negative data and exhaustivity. They have recently regained popularity because they enable to study specific linguistic phenomena in a controlled way, which is increasingly difficult with NMT models (Isabelle et al., 2017; Burchardt et al., 2017; Avramidis et al., 2019).

2.2 Methodology

To evaluate the diverse data collected in the experiments detailed in this thesis, a mixed-methods approach has been applied. That is, both qualitative and quantitative methods have been used. We believe the combination of these two approaches can provide a better understanding of the research problem. Furthermore, whenever possible, we have triangulated the data to cross-check the results obtained.

Regarding the source of data, we also study both product and process. We analyze the textual product that is the outcome of the PE task, but we also seek to understand the underlying cognitive processes to study PE effort. In this way, we can better understand how product and process relate.

To this end, we have used the following methods:

Quantitative methods have been used to calculate to usual MT metrics with one reference, such as BLEU, NIST, WER (see Section 3.6 and Section 3.7). Furthermore, we have also used these methods to study direct (temporal and technical effort) and indirect (HBLEU, HTER, edit distance) measures of effort and the statistical significance there is among some of these measures (see Section 4.5).

Qualitative methods have been used to study in detail the product both from raw MT and from post-edited versions. Mainly, we have used fine-grained error annotation methods (see Section 6.5) to produce a manually-crafted challenge set (see Section 6.6) which can help evaluate linguistic features in raw MT output.

Participant-oriented methods include the use of surveys and follow-up questions in Section 5.3 to gather detailed information on the perceptions of translators regarding post-editing. The goal was to study if there was a negative bias in relation to PE and if that could affect the perception of quality of the MT output related to productivity.

Chapter 3

MACHINE TRANSLATION AND POST-EDITING

3.1 Introduction

In this chapter, we present the different MT technologies currently used and the main automatic and manual methods used to evaluate them. On the one hand, manual methods usually produce reliable results if they are conducted by professional translators, but entail higher costs and time. On the other hand, automatic measures provide an easy and quick way to assess MT systems and outputs, but encounter many problems when a more granular evaluation is necessary.

We also detail two experiments we conducted comparing SMT and NMT outputs for two different language combinations. The goal is not only to evaluate the raw output of these two architectures with both manual and automatic scores, but also to compare the results with measures of PE effort (temporal and technical).

3.2 Machine Translation

Although it has been in the research agenda for many years, currently MT is a major topic with the universal application of the Internet and the acceleration of the integration process of the world economy. MT research has been studied applying two methods: rule-based translation systems and corpus-based methods.

On the one hand, rule-based translation systems include literal translation methods, interlingua and transfer-based methods. On the other hand, corpus-based translation methods include statistical machine translation (SMT) and neural machine translation (NMT). The latter is the most used today because it benefits from the massive computing capacities and the great amount of data currently available. Hereby we explain the main characteristics of the MT systems more often used in industry and research.

3.2.1 Rule-based Machine Translation

One of the first approaches which is still in place, especially for closely related languages or languages that have less available training data (Bayatli et al., 2018), is transfer rule-based machine translation (RBMT). It manipulates linguistic knowledge using handcrafted grammatical and lexical rules that translate from the source to the target language. This approach allows to fine-tune the translation process and control the output produced, although there is a high human cost in formalising all the linguistic rules. In corpus-based systems, this cost is dramatically reduced because the different corpora are combined with a machine learning approach to infer the language and translation model. Some of the best-known examples of this model are Lucy LT[1] (Alonso and Thurmair, 2003) and the open-source Apertium[2] (Forcada et al., 2011).

RBMT technology is currently used in certain commercial applications,

[1] http://www.lucysoftware.com
[2] http://www.apertium.org

for example, in daily applications in the written media, as is the case of Catalan and Spanish versions of certain newspapers in Catalonia[3]. The RBMT technology has also been included as a complement to statistical machine translation (SMT) systems in hybrid systems (Alegria et al., 2008). However, current data-driven methods can be obtained in much less time for a larger number of languages as the linguistic rules necessary for ruled-based models can be costly and time-consuming, and are hardly generalizable to other languages. Quality metrics are usually much better for corpus-based methods for most domains and language combinations.

3.2.2 Statistical Machine Translation

Statistical machine translation (SMT) is data-driven, that is to say, it is based upon corpora, and dominated the research agenda from the mid -2000s (Kenny, 2018). In this method, knowledge is built from different corpora: monolingual target-language data for inducing the language model, and sentence-aligned source–target data for inducing the translation model.

The language model provides a model of the monolingual training corpus and also a method for computing the probability of a previously unseen string using that model. It includes relative frequencies for the substrings occurring in that corpus, which is applied to the assigned probability to a new sentence.

The translation model expresses the probability that a source sentence and the candidate translate are equivalent. It provides a model of the sentence-aligned source–target training corpus, and a method for computing the probability that a source sentence and a target sentence are equivalent using that model. It includes relative frequencies of parallel corpora occurring in the corpus.

When translating new input, these models are used to actually decode the translation. It is treated as a search problem (Lopez, 2008; Hearne and

[3]http://www.elperiodico.com and http://www.lavanguardia.com

Way, 2011), where many translation hypothesis are generated for a single string and the most probabilistically plausible is chosen. In fact, both models are created separately, which differentiates notions of adequacy and fluency that are usually conceived together in human evaluations.

There are two algorithms used to calculate the score. The noisy-channel model is the traditionally used (Brown et al., 1993, 1990) while the log-linear is more flexible and general and includes a scoring with unlimited number of features. In the beginning, words were aligned but later on phrase-alignment heuristics (Koehn, 2010; Och, 2003) were introduced to extract many-to-many alignments, which was called phrase-based statistical machine translation (PBSMT). Throughout the years, many other improvements have been applied, mainly introducing hybrid models (Costa-Jussà et al., 2016).

Even though SMT is currently being substituted by NMT models in most industrial scenarios due to the improved quality results (Castilho et al., 2017a), many companies still use their fine-tuned SMT systems to produce their translations, especially in small companies. However, a detailed comparison of both models usually shows better results for NMT systems (see Sections 3.6 and 3.7 for detailed information on the experiments conducted).

3.2.3 Neural Machine Translation

NMT is not a new architecture, but it could only be applied once the computational limitations had been solved (Cho et al., 2014; Bahdanau et al., 2014). In recent evaluation campaigns, such as the Third Conference on Machine Translation (WMT2018) (Bojar et al., 2018), the Fourth Conference on Machine Translation, (WMT 2019) (Barrault et al., 2019), and the Fifth Conference on Machine Translation (Barrault et al., 2020) (WMT2020), both automatic scores and human evaluations have shown excellent quality results.

NMT is also a corpus-based machine translation, which is trained on huge amounts of corpora usually formed by pairs of source language segments

and their translations, although novel methods based solely on monolingual corpora have been suggested (Artetxe et al., 2018). In this sense, it is similar to the SMT technology, but uses a completely different computational approach: neural networks (Forcada, 2017), which are artificial units similar to neurons, because their output or activation depends on the level of connection and degree of stimuli.

The first NMT models (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014) consist of an encoder and a decoder which are trained jointly. The encoder encodes the source language sentences into a sequence of vectors, which are the hidden representations of the source tokens, and also a meaning vector.

Then the decoder generates the sequence of tokens for the target language taking into account the meaning vector from the encoder. The encoder and the decoder are usually two different neural networks which can have different architectures, and can include a stack of neural networks where multiple layers are used as input for the following layers.

This NMT encoder-decoder architectures obtain good results when translating short sentences, but show more problems when translating longer sentences (Cho et al., 2014). Instead of encoding the entire source sentence into a single fixed context vector, Bahdanau et al. (2014) proposes the attention mechanism, which filters the relevant source words specifically for each output decoder layer and generates a context vector, which is a weighted vector extracted from encoders (Vaswani et al., 2017). This improves the results for longer sentences and has become an essential part of NMT models.

Different neural networks have been applied to encoders and decoders. Recurrent neural networks (RNN) are sequential networks that change as new inputs are introduced, where each state has a direct connection only to the previous state. RNNs present short-term memory and, when they encounter a long-enough sentence, they have problems to carry the information from one step to the following. During back propagation, RNNs suffer from the vanishing gradient problem. The values used to update the

neural networks weights shrink when they backpropagate through time and do not contribute to learning.

As a solution to the vanishing gradient problem, the long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) and the gated recurrent unit (GRU) (Cho et al., 2014) were introduced. Both algorithms use a gating mechanism to control the memorization process.

Another alternative model are convolutional neural networks (Gehring et al., 2017). Instead of using the encoder-decoder model described before, the decoder produces representations of each of the words and takes into account a few words before and after. Contrary to RNNs, CNNs computation can be parallelized, optimization is easier since the number of non-linearities is fixed and independent of the input length and they outperform the LSTM accuracy (Wu et al., 2016).

The transformer architecture (Vaswani et al., 2017) follows mainly the encoder-decoder model with attention passed from encoder to decoder. It employs a self-attention mechanism that allows the encoder and decoder to account for every word included in the entire input sequence. Transformer proposes to encode each position, apply self-attention in both decoder and encoder, and enhance the idea of self-attention by calculating multi-head attention. This improves performance expanding the model's ability to focus on different positions and gives the attention layer multiple sets of weight matrices. There are no recurrent networks, only a fully connected feed-forward network.

Words are usually considered as the basic unit of meaning, so its modelling is straightforward and makes sense to humans. However, there is always a limited vocabulary size which generates the out-of-vocabulary (OOV) problem. This can affect severely the quality of the translation. To solve it, character-level models have been suggested, as languages have a limited number of characters. However, they increase sentence length and make it more complicated for NMT models to capture dependencies (Tang, 2020).

An alternative which includes word-level and character-level models is

subwords. For this solution, the vocabulary includes both words and subwords. When a rare word is encountered, it is segmented into separate subwords. Sennrich et al. (2016) apply the byte pair encoding (BPE) algorithm to generate subwords, which produces better results than character n-gram models.

There have also been different approaches to leverage the grammatical and lexical information included in RBMT into NMT models, focused especially in low-resourced languages. Several studies have included morphological features and parts-of-speech (POS) information as input features to improve translation quality with good results for English to Romanian (Sennrich and Haddow, 2016) and Arabic dialects (Baniata et al., 2018).

Etchegoyhen et al. (2018) analyze different machine translation models including linguistically-motivated word segmentation methods for subword units. Other authors have included syntactic structures such as dependency trees using graph convolutional encoders to improve NMT quality (Bastings et al., 2017). Additionally, MT models combining the hybrid use of RBMT and NMT have been proposed for different language combinations: Sanskrit-Hindi (Singh et al., 2019), English-Basque and English-Irish (Torregrosa et al., 2019) and Breton-French (Sánchez-Cartagena et al., 2020), respectively.

Even though the output of NMT systems is similar to the output produced by SMT and presents some of the problems common to corpus-based MT, there are some new issues that should be taken into account. NMT tends to produce more fluent outputs but, due to the semantic nature of learned representations, produces semantically-motivated errors which are difficult to spot (Forcada, 2017).

Furthermore, the use of sequence subword units (usually not linguistically motivated) improves performance but produces hallucinations, which are translations that are fluent but unrelated to the source (Müller et al., 2019; Wang and Sennrich, 2020) (see Section 6.5 for detailed information on NMT errors).

3.3 Evaluation of Machine Translation

Evaluating machine translation has always been a very active area of research in MT (Hutchins, 2001) because it is an important element for all parties involved: researchers need to know if their new developments really work, industry members want to show their clients how efficient their systems are and final users want to know which system to use.

MT evaluation is normally used for the following purposes (Gimenez, 2008):

- **System Optimization**: The quality of a MT system needs to be assessed in order to improve the system as a whole.
- **Comparison of MT Systems**: Different versions of the same system or different systems need to be compared to assess the fitness for a certain domain or task, training data, etc.
- **Error Analysis**: A detailed error analysis is needed to identify the strengths and weaknesses of the systems.

One of the main issues with MT evaluation is that the goal of the assessment should be clearly stated to produce reliable results (Wilks, 1994).

3.3.1 Manual Evaluation

One of the first manual evaluations was conducted in 1966 as part of the tasks performed by the automatic language processing advisory committee (ALPAC). It reflected intelligibility (measured if the translation could be read easily and was understandable) and fidelity (measured if the translation included the whole meaning intended in the original without distorsions) (Carroll, 1966).

In the 1990s, the Advanced Research Projects Agency (ARPA) created a methodology to evaluate machine translation systems using adequacy and fluency (Church and Hovy, 1993). The evaluator had to assess the adequacy and fluency on a scale 1-to-5. Basically, they were asked to state

the level of correlation in meaning between the source and target segments in the adequacy and the linguistic correction of the target language in the fluency. The final scores were calculated by averaging the judgments over all of the decisions in the translation set (White et al., 1993).

Since then, human judges have usually taken part in the evaluation of MT stating the semantic accuracy and fluency of the MT output against one or more reference translations conducted by professional translators.

Bojar et al. (2016) explain the main human evaluation methods used in the WMT campaigns, which currently serve to test the different MT engines:

- **Fluency and Adequacy**. Annotators are presented with different sentences and they have to rank them on a five-point scale. Fluency measures if a translation is fluent, regardless of the correct meaning, and Adequacy measures if the translation conveys the meaning of the source text. We present the results of a comparison between NMT and PBSMT using this evaluation method in Section 3.6.
- **Sentence Ranking**. Annotators are presented multiple outputs of different systems, along with the source, and are asked to rank them. We present the results of a comparison between NMT and PBSMT using this evaluation method in Section 3.6.
- **Constituent Ranking**. Instead of ranking different translations according to their quality, annotators are asked to rank only identified constituents.
- **Constituent Judgement (Y/N)**. Annotators also assess a constituent, but only provide a binary judgement on the translation.
- **Sentence Comprehension**. Annotators are asked to edit MT output for fluency (without providing the reference), and then (separately) to determine via binary judgement whether those edits result in good translations.
- **Direct Assessment (DA)**. Annotators are asked to provide a direct assessment of the quality of a single MT output compared to

a single reference, using an analog scale. Graham and Liu (2016) and Graham and Baldwin (2014) have shown it is more reliable for evaluation of metrics and it was adopted as the official human evaluation in WMT17. We use it as a comparison metric in Section 4.5.

Translations are intended for human users and, as such, human judgements are the right measure of the quality and problems presented by a translation. Moreover, human perception and knowledge of the real world allows evaluators to assess MT errors and correlate them to their practical importance in a translation (Sanders et al., 2011).

However, manual evaluations usually imply a lot of time and effort, and too often the people who conduct these evaluations have limited knowledge or experience. As a consequence, evaluations can suffer from low inter- and intra-annotator agreements (Turian et al., 2003; Snover et al., 2006). Evaluating MT output is a challenging and complex task, which can lead to tiredness among evaluators. Many elements need to be taken into consideration when conducting the evaluation and too often guidelines are not well defined (Callison-Burch et al., 2007).

This can also be applied to the evaluation of human translations. For Translation Studies the concept of quality is often linked to the different perspectives on translation (Castilho et al., 2018) and for the industry quality is directly linked to customer satisfaction (Drugan, 2013). Recent approaches have abandoned prescriptive evaluations, but there is still considerable divergence (Munday, 2016).

In response to recent claims of parity between NMT systems and human translation (Hassan et al., 2018; Popel, 2018), refuted later by Toral et al. (2018), Läubli et al. (2018) and Läubli et al. (2020) expose some of the problems related with current evaluation systems. First, raters show a clear preference for human translations when document level context is taken into account, which stresses the failures in some of the current evaluation practices in MT. Second, they suggest a new protocol to carry out the evaluations, which implies the use of professional translators, com-

pared to crowdsourcing and direct assessment, where evaluators have little or no expertise in translation. As MT quality improves, translations will become harder to evaluate in terms of quality and this will require experts to conduct the evaluations.

3.3.2 Automatic Evaluation

Automatic MT metrics were developed as a solution to the slowness, subjectivity and high costs of human evaluations.

These automatic measures use three different approaches to evaluate the quality of the MT output: (i) reference proximity, (ii) confidence or quality estimation (QE) metrics, and (iii) performance-based methods (Babych, 2014; Chatzikoumi, 2020).

Reference Proximity Methods

These methods compare the MT output (also called hypothesis) with one or more human translations of the same source text (called references and also known as the gold-standard human translation). The closer the MT output is to the reference, the better the MT output is considered.

The main problem of this approach is that there are many good translations for a given source sentence, which can vary in word choice or in word order even when they use the same words. Thus, every reference used produces different evaluation metrics, which could lead to very different scores for the same output.

One possible solution is to provide multiple references, which could account to a certain level for the variations in linguistic aspects of the translations such as grammar, word order, style and word choice. Dreyer and Marcu (2012), for example, showed how having many references could improve considerably MT metrics for n-gram based evaluations. Even though the use of references as a bag of n-grams does not improve significantly the evaluation results (Doddington, 2002), the combination of recurrence distributions, divergence information and the length of n-grams

from multiple references has shown promising results (Qin and Specia, 2015).

Gimenez (2008) classifies these methods into four separate groups:

a) Edit Distance Measures

WER (Word Error Rate) (Nießen et al., 2000) calculates the Levenshtein distance (Levenshtein, 1966), which is the minimum number of substitutions, deletions and insertions necessary to convert to hypothesis into the reference translation. **PER** (Position-Independent Word Error Rate) (Tillmann et al., 1997) also calculates the edit distance, but regardless of word order. **TER** (Translation Edit Rate) (Snover et al., 2006) calculates the amount of post-editing necessary to match the reference translation, including insertions, deletions, substitutions and shift of phrases. All edits have equal cost.

b) Precision-oriented Measures

The most common methods to measure the distance between the hypothesis and the references are n-gram metrics, which are based on the lexical similarity between a machine translation and one or more human references. Therefore, the calculation of these measures is based on precision. **BLEU** (Bilingual Evaluation Understudy) (Papineni et al., 2002) is currently used as a standard for MT evaluation. It compares 1 to 4 words from the MT output with multiple references and n-gram precision is modified to eliminate repetitions that occur across sentences. It also includes a brevity penalty that down-scales the score for the MT outputs that are shorter in length than the reference. Even though it has shown correlation with human judgments of translation quality in many cases (Coughlin, 2001), some studies have questioned the role of BLEU in MT assessment (Callison-Burch et al., 2006; Wieting et al., 2019; Mathur et al., 2020), especially when comparing high-quality systems (Ma et al., 2019).

Furthermore, there is a lack of consistency in the reporting of BLEU scores. That is, the parameters included in this metric can vary wildly

and many BLEU scores are not really comparable, due mainly to the different tokenization and normalization schemes applied to the reference. (Post, 2018). Besides, recent research has shown it tends to underestimate NMT output (Shterionov et al., 2018) (see Sections 3.6 and 3.7 for further details) and can be severely affected by the outliers and sample size (Mathur et al., 2020)

WMN (Babych and Hartley, 2004) is a variation of BLEU which weights n-grams taking into account their statistical salience estimated from a large monolingual corpus. The National Institute for Standards and Technology (NIST) (Doddington, 2002) created **NIST**. The main difference with BLEU is that NIST performs an arithmetic mean instead of a geometric one. It also takes into account n-grams of length 5 and weights more heavily n-grams which occur less frequently.

RIBES (Isozaki et al., 2010) is a metric based on rank correlation coefficient of word order in the translation and reference. It analyses if the MT system produced the correct word order and has been used to evaluate languages with very different structures.

c) Recall-oriented Measures

Recall-oriented measures calculate the lexical recall, that is, the proportion of lexical units (usually n-grams) in the reference covered by the MT output. **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) (Lin and Och, 2004) calculates the n-grams up to length 4 to compute the lexical recall. **CDER** (Cover/Disjoint Error Rate) (Leusch et al., 2006) is a recall-oriented measure modeling block reordering.

d) Measures Balancing Precision and Recall

These measures include a combination of lexical precision and recall and are mainly based on the F-measure. **chrF** (character n-gram F-score) (Popović, 2015) calculates n-gram precision and recall arithmetically averaged over all n-grams. **GTM** (Melamed et al., 2003) is another F-measure, which is the result of the harmonic mean of precision and recall. It can also be used with multiple references by concatenating the refer-

ences together, and not allowing a match to cross the boundary between references.

METEOR (Banerjee and Lavie, 2005) aligns the MT output to the reference translation using stems, synonyms, and paraphrases, besides exact word matching, and then computes candidate-reference similarity based on the proportion of aligned words in the candidate and in the reference. It assigns different weights to the word matches, taking into account the type of lexical similarity, function and content words. It also integrates a fragmentation penalty for differences in word order. The final METEOR score is a parametrized combination of F-measure and fragmentation penalty. **MPEDA** (Zhang and Gildea, 2007) is another metric based on METEOR. It uses a domain-specific paraphrase database instead of a general one to reduce noisy matches.

Fomicheva and Specia (2019) add an additional group of evaluation methods which includes linguistic representations for comparing the MT output against the reference translation. The goal of these methods is measuring the quality of the MT output by assessing its grammaticality. **UPF-Cobalt** (Fomicheva and Bel, 2016) incorporates a syntactically informed context penalty to penalize lexically similar words that play different roles in the candidate and reference sentences. **DPMFComb** (Yu et al., 2015) is a syntax-based metric that parses the reference translation with a standard parser and trains a new parser on the tree of the reference translation.

Gimenez (2010) suggests a number of measures which assess different linguistic elements. **SP metrics** (Gimenez, 2010) measure the similarities at the level of parts of speech, word lemmas, and base phrase chunks. **DP metrics** study the similarities between dependency trees associated with the MT output and the reference translation. And **CP metrics** study the similarities between constituent parse trees associated with MT outputs and reference translations.

Recently, other evaluation measures have focused on deep representations. **YISI-1** (Lo, 2019) computes the semantic similarity of the different phrases in the MT output with the reference, with the use of con-

textual word embeddings (Devlin et al., 2019). Another proposal is **ESIM** (Mathur et al., 2019), which is a trained neural model that computes sentence representations from BERT embeddings, and then calculates the similarity between the two strings.

Confidence or Quality Estimation Methods

The automatic evaluation of the source and target MT documents without a human reference was first initiated in the 2000s and was called Confidence Estimation (CE). It included source, target and glass-box features and it had a number of possible working applications, such as assessing the revision needed for a segment or text, establishing the readability of a text, the effort required for post-editing, selecting among different translations for the same original segment, etc.

Early work in quality estimation (QE) built on the concept of confidence estimation that had been developed for speech recognition (Blatz et al., 2004). These systems usually relied on system-dependent features, and focused on measuring how confident the system was with respect to the proposed translation rather than measuring the quality level of the final translation. Later on, research in QE focused only on black-box system independent features that were based on the source and target sentences. The goal was to predict the quality of a certain MT output only with the input and the output text but without any other information about the expected output.

Since the Workshop on Statistical Machine Translation (WMT) celebrated in 2012 (Callison-Burch et al., 2012), QE has been included as a shared-task to evaluate the different QE systems proposed against a certain dataset delivered. Generally, QE is conceived as a machine learning task that uses different algorithms to induce various models according to relevant parameters. Once the model is built, it is used to estimate the quality of machine-translated texts. In these shared-tasks, participants can use a 17 features baseline developed by Specia et al. (2013). Later versions, include document-level quality estimation (Specia et al., 2015), neural-

based architectures (Ive et al., 2018) and a Pytorch-based open source framework (Kepler et al., 2019).

Performance-based Methods

The idea behind these methods is that the quality of an MT output would affect how well someone can carry out a task based on the text translated. In automated MT evaluation, these methods are calculated as evaluation scores for clearly-defined tasks for a specific MT system. Systems are presumed to perform worse with worse-quality MT outputs.

Examples include metrics based on performance of syntactic parsers (Rajman and Hartley, 2001) or raw counts and the recall score for the named entity recognition task from MT output (Babych and Hartley, 2004). These methods, however, have more limited applications mainly because they are linked to a specific text type. It can also include the evaluation chekpoints: linguistically motivated features, such as ambiguous words and noun or prepositional phrases, which are automatically extracted from parallel sentences (Zhou et al., 2008).

As we have seen, there are many different automatic MT evaluation scores. These automated evaluation metrics can be useful for certain purposes, but are problematic as a measure of translation quality for any MT system (Babych and Hartley, 2004) for a number of reasons:

- As previously mentioned, it tend to produce different estimations for different MT architectures (Callison-Burch et al., 2006).
- These metrics are applied regardless of the purpose of the MT output, as if there was a universal score of translation quality.
- Score values are only meaningful compared with previous versions of the same system or similar systems, but cannot be directly interpreted as an exact measure of translation quality.
- When assessing MT models with high quality, certain automatic scores tend to loose sensitivity (Babych and Hartley, 2008).

It is obvious that automatic metrics are much more cost-effective than human evaluations but they also lose granularity. Furthermore, they are mainly based on the comparison to a reference translation. The quality of the MT output is established on the basis of how closely it matches the human-translated reference, while it is clear that many possible good translation solutions are possible for the same source text. Therefore, automatic scores seem objective but are based on questionable assumptions (Castilho et al., 2018).

3.4 MT Evaluation for Post-editing

With the improvement in MT quality, this technology has increased its presence both in translation industry workflows and in society in general. The array of traditional and emerging uses of MT include, among others, the direct use of raw MT for gisting purposes, light and full post-editing.

Therefore, we should devise evaluation methods which can assess the quality of an MT output depending on the task or the function for which the output is intended (Hovy et al., 2002). In the case of post-editing, we can assess, for example, if the quality of the MT output is good enough to post-edit. In case it contains too many errors, translators often report they find it easier to translate from scratch than to post-edit it (Sanchez-Torron and Koehn, 2016).

However, academia and industry have different goals in this evaluation and usually analyze different elements (Drugan, 2013). On the one hand, academia has usually focused on the concept of quality, which is directly linked to theoretical assumptions on the different views of translation and imply different ways of assessing it (House, 2014).

It has also focused its research on the effort translators require to post-edit the MT output, depending on the quality and type of text (Guerberof, 2009a,b; Specia, 2011, 2010), but also taking into account the tools used to edit the output and its usability for translators (Castilho et al., 2014; Moorkens and O'Brien, 2013). In his seminal work, Krings (2001) di-

vided PE effort into three separate but interrelated categories: temporal, technical, and cognitive (see Chapter 4 for a complete description of their characteristics and of the metrics used), which have been used for all studies thereafter.

On the other hand, the evaluation models used in the industry are mainly focused on productivity (O'Brien, 2011b; Parra Escartín and Arcedillo, 2015; Plitt and Masselot, 2010; Sanchez-Torron and Koehn, 2016) and are predominantly error-based, where errors are counted and weighted according to severity.

3.5 SMT versus NMT for Post-editing

Even though SMT has been well established as the dominant approach in MT for many years, recent developments in NMT have driven a technological shift from SMT to NMT in many translation industry scenarios. This makes it necessary to compare how these two different technologies affect evaluation metrics, both automatic and manual.

One of the first complete papers studying the impact of SMT and NMT in post-editing was Bentivogli et al. (2016). They carried out a small scale study on post-editing NMT and SMT outputs of English to German translated TED talks. They concluded that NMT in general terms decreased the post-editing effort, but degraded faster than SMT with sentence length. One of the main strengths of NMT was reodering of the target sentence.

Some years later, Bentivogli et al. (2018) extended the scope of their previous paper increasing the number of systems analyzed, adding an extra language pair and conducting a three-category error analysis on the results. They confirmed the increase in quality for NMT systems and concluded most errors produced in NMT outputs were lexical, especially proper nouns.

Wu et al. (2016) evaluated the quality of NMT and SMT, in this case using

BLEU and human scores for machine-translated Wikipedia entries. Results showed that NMT systems outperformed and improved the quality of MT results. Junczys-Dowmunt et al. (2016) confirmed this diagnostic when they studied 30 different translation directions from the United Nations Parallel Corpus, as have the recent Findings on the Conference on Machine Translation (Bojar et al., 2018; Barrault et al., 2019, 2020).

Toral and Sánchez-Cartagena (2017) broadened the scope of Bentivogli et al. (2016) adding different language combinations and metrics, and they concluded that although NMT yielded better quality results in general, it was negatively affected by sentence length, and the improvement of the results was not always perceivable in all language pairs.

Castilho et al. (2017b) discussed three studies using automatic and human evaluation methods. One of them included in-domain formal texts for chemical patent titles and abstracts. In addition to the automatic metrics, two reviewers assessed 100 random segments to rank the better translation and to identify the translation errors. Automatic evaluation didn't give clear results, but SMT system was ranked better than NMT in human evaluation.

Castilho et al. (2017c) reported on a comparative study of PBSMT and NMT, with four language pairs and different automatic metrics and human evaluation methods. It highlighted some strengths and weaknesses of NMT, which in general yielded better results. It focused especially on post-editing and used the PET interface (Aziz et al., 2012) to compare educational domain output from both systems using different metrics.

The authors concluded that NMT reduced word order errors and improved fluency for certain language pairs, so fewer segments required post-editing, especially because there was a reduction in the number of morphological errors. However, they didn't detect a decrease in PE effort nor a clear improvement in omission and mistranslation errors.

Mutal et al. (2019) compared the output quality of a SMT and NMT engine which had been customised for the Swiss Post's Language Service using the same training data. They investigated how professional transla-

tors perceived the differences between the MT output and a human reference in terms of deletions, substitutions, insertions and word order.

Translators considered deletions the most important errors and presented less agreement for NMT errors, which may show errors in this architecture are more difficult to identify. They also detected NMT capacity to produce correct paraphrases, which could account for the subestimation in quality often showed with BLEU for NMT systems.

Esperança-Rodier et al. (2017) focused on SMT and NMT systems and their impact on the translator's activity. The authors carried a comparative quantitative analysis, based on BLEU, TER and METEOR of two in-house systems from French to English. Then, they qualitatively analysed translation errors from linguistic criteria. NMT performed just slightly better than SMT but the authors warned about the limited training of their NMT model.

Lohar et al. (2019) evaluated user-generated context, specifically tweets translated from German into English with NMT and PSMT systems. The systems were evaluated using BLEU, TER and METEOR. Results showed a worse performance of NMT system when training data was small. However, when both models were trained with additional data, NMT model showed a rapid improvement, while PSMT results remained more o less the same.

Isabelle et al. (2017) proposed a detailed translation evaluation and error analysis of the two MT methods with a challenge set approach (see Section 6.4 for a detailed explanation). They manually crafted a small set of sentences to evaluate a system's capacity to solve a particular linguistically-motivated problems in the source language. In their experiments, except for idioms, NMT performed better than SMT.

There has also been some research studying low-resource language combinations. Dowling et al. (2018) compared SMT and NMT for English to Irish in the public administration domain. The results showed that for languages with less resources like Irish, out-of-the-box NMT systems didn't work as well as tailor-made domain-specific SMT systems.

Skadina and Pinnis (2017) compared English to Latvian SMT and NMT for narrow domains. They showed that with a small amount of data for training, SMT systems performed better than NMT systems. The SMT systems learned better terminology and phrases specific for the domain. However, in the news domain, results were better for NMT approaches.

Mahata et al. (2018) studied the differences in SMT and NMT for English-Hindi and English-Bengali language pairs. They tested both systems for simple sentences as well as sentences of other complexities. The quality of the translation was assessed both with automatic metrics such as BLEU and TER and manual evaluation. They observed that NMT outperformed SMT in the case of simple sentences whereas SMT outperformed NMT where all types of sentences were studied.

In the case of related languages, Costa-Jussà (2017) analyzed automatic metrics and human scores for NMT and SMT from Spanish into Catalan. She concluded that NMT quality results were better both for automatic metrics and for human evaluation for all in-domain sets, but PBSMT results were better for general domain ones.

In the following pages we describe the experiments we conducted to compare the post-editing of SMT and NMT outputs for English to Spanish and Spanish to Catalan using both human and automatic metrics, and temporal and technical effort.

3.6 Experiment 1: Comparing SMT and NMT Output for English into Spanish

We conducted two separate experiments with English-Spanish medical texts to study the differences of post-editing NMT and phrase-based SMT (PBSMT) outputs for formal in-domain texts. We compared the usual automatic scores for MT with direct and indirect PE effort metrics. Mainly, we studied translators' perception regarding quality, and fluency and accuracy, and analyzed temporal and technical post-editing effort.

Our objectives with these experiments were threefold:

- Determine which MT method (PBSMT or NMT) yielded better results for post-editing in-domain formal texts.
- Analyze the relation between human and automatic metrics for post-editing.
- Study translators perception as a prospective measure of PE effort.

3.6.1 Methodology

In the first experiment, we conducted different surveys for the human evaluation of two MT systems. In the second experiment, we recorded temporal and technical effort of professional translators while they post-edited.

Translation Ranking

For the first experiment, we used two surveys. In the first survey, participants had to answer some questions about their previous experience in the translation industry. The survey was open both to students and professional translators as we were mainly interested in the perception of quality. Then, participants had to rank the translation of 40 segments (human translation, NMT and PBSMT), which had no context and were randomized to avoid bias. They were selected so there were no repeated translations and all had a minimum length of 100 characters. Then we applied a script to ensure there was a minimum editing distance of 15% between the human-PBSMT, human-NMT and PBSMT-NMT solutions. This reduced the number of segments from 230 to 145. We hand-picked 40 segments without typos nor any other problem.

Fluency and Adequacy

In the second survey, we presented the same English segments as in the previous task. First of all, participants had to answer some questions

about their previous experience in the translation industry. The participants were both students and professional translators who were not necessarily the same as in the translation ranking assessment. Afterwards, they had to evaluate the fluency and adequacy of the proposed translation on a four-point Likert scale. The translation was either PBSMT or NMT chosen randomly without any knowledge of the participants. The goal was to assess fluency and adequacy for in-domain formal texts.

PE Time and Technical Effort

In the second experiment, participants had to post-edit 41 segments from a 2018 medical paper. They had to carry out the task in PET (Aziz et al., 2012)[4], a computer-assisted translation tool that supports post-editing (see Section 4.3 for a full list of tools used to collect measures of PE effort). It was used with its default settings. It logs both post-editing time and edits (keystrokes, insertions and deletions, that is, technical effort). Four professional translators with more than two years of experience post-editing carried out the task: two of them post-edited the PBSMT output and the other two post-edited the NMT output.

3.6.2 MT Systems and Training Corpus

MT Systems

We used ModernMT (Germann et al., 2016) version 2.4 to train the PBSMT and NMT systems. This version allows to train both statistical and neural MT systems. We used the default options for this version. One of the salient characteristics of ModernMT is the fact that it can take into account the context of the sentence to be translated. In the evaluation results we show figures for both cases: with or without taking the context into account. In the experiments we took context to be the previous and the next segment (except for the first segment and the last segment, where we took into account the next segment and the previous segment only,

[4]http://wilkeraziz.github.io/dcs-site/pet/index.html

respectively). A short context is usually enough to calculate the context vector used by ModernMT.

In order to help contextualise the results of the experiment, we decided to use two MT systems as reference to compare the results with the ones of the systems we trained. As reference MT systems we chose Apertium (Forcada et al., 2011), a shallow transfer MT system, and Google Translate, a neural MT system for the English-Spanish language pair.

Data: Medical Corpus

To train the system we compiled all, to our knowledge, publicly available corpora in the English-Spanish pair. We also created several corpora from websites with medical content:

- The EMEA[5] (*European Medicines Agency*) corpus.
- The IBECS[6] (*Spanish Bibliographical Index in Health Sciences*) corpus.
- Medline Plus[7]: we have compiled our own corpus from the web and we have combined this with the corpus compiled in MeSpEn.
- MSDManuals[8] English-Spanish corpus, compiled for this project under permission of the copyright holders.
- Portal Clínic[9] English-Spanish corpus, compiled by us for this project.
- The PubMed[10] corpus.
- The UFAL Medical Corpus[11] v1.0.

[5]http://opus.nlpl.eu/EMEA.php
[6]http://ibecs.isciii.es
[7]https://medlineplus.gov/
[8]https://www.msdmanuals.com/
[9]https://portal.hospitalclinic.org
[10]https://www.ncbi.nlm.nih.gov/pubmed/
[11]https://ufal.mff.cuni.cz/ufal_medical_corpus

We also treated as a corpus glossaries and glossary-like databases containing a lot of useful terms and expressions in the medical domain. Namely:

- English-Spanish glossary from MeSpEn.
- ICD10-en-es: ICD10 is the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list by the World Health Organization (WHO).
- Snomed-CT: SNOMED Clinical Terms is a systematically organized computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting.

With all the corpora and glossaries we created an in-domain training corpus of 2,836,580 segments and entries. We splitted the corpus in two parts: 99% of the segments for training, and the remaining 1% for testing.

We also used other general corpora for training the two MT systems, namely:

- The Scielo corpus (*Scientific Electronic Library Online*), that is formed by complete full text articles from scientific journals of Latin America, South Africa and Spain. As these articles are not necessarily in the area of medicine, we consider this corpus as general.
- Europarl corpus[12] (Koehn, 2005) obtained from Opus Corpus (Tiedemann, 2012a).
- Global Voices corpus, parallel corpus of news stories from the web site Global Voices[13] compiled and provided by CASMACAT[14] and obtained from Opus Corpus.
- News Commentary: a parallel corpus of news commentaries pro-

[12]http://www.statmt.org/europarl/
[13]https://globalvoices.org/
[14]http://casmacat.eu/corpus/global-voices.html

vided by WMT for training SMT. The source is also taken from CASMACAT and downloaded from Opus Corpus.

In Table 3.1 the sizes of all corpora and glossaries used for training the MT systems are shown. The figures are calculated eliminating all the repeated source segment - target segment pairs in the corpora.

Corpus	**Segments/Entries**	**Tokens eng**	**Tokens spa**
EMEA	366,769	5,327,963	6,008,543
IBECS	628,798	13,432,096	14,879,220
MedLine Plus	15,689	209,074	234,660
MSD Manuals	241,336	3,719,933	4,467,906
Portal Clinic	8,797	159,717	169,294
PubMed	320,475	2,752,139	3,035,737
UFAL	258,701	3,202,162	3,437,936
Glossary MeSpEn	125,645	286,257	348,415
ICD10-en-es	5,202	25,460	30,580
SnowMedCT Denom.	887,492	3,509,062	4,457,681
SnowMedCT Def.	4,268	177,861	184,574
In-Domain	**2,836,580**	**32,479,955**	**36,893,257**
Scielo	741,407	17,464,256	19,305,165
Europarl	1,961,672	50,008,219	52,489,142
Global Voices	559,418	10,717,938	11,496,683
News Commentary	259,412	5,898,912	6,903,975
Out-of-Domain	**3,521,363**	**84,087,899**	**90,193,659**

Table 3.1: Size of the corpora and glossaries used to create the corpus to train the MT systems.

3.6.3 Automatic Evaluation of the MT Systems

In Table 3.2 we can observe the evaluation values of the trained systems using MTEval[15] along with Apertium and Google Translate. This soft-

[15]https://github.com/odashi/mteval

ware allows to calculate BLEU, NIST, RIBES and WER using only one reference. We have used all the test set of the corpus. As shown in the table, the systems trained in the experiment obtain better results in all metrics than the reference systems used, except for the Google Translate system, which obtains a slightly better NIST result than the MMT Phrase-Based system without context and a better WER result than the two MMT Phrase-Based systems. The MMT Neural system performs consistently better than the MMT Phrase-Based system. In the MMT Neural system we do not see any significant difference between the results obtained when trained with or without context.

MT system	**BLEU**	**NIST**	**RIBES**	**WER**
Apertium	0.193	6.442	0.713	0.703
Google T.	0.402	9.632	0.809	0.530
MMT P.B. no context	0.424	9.536	0.814	0.638
MMT P.B. context	0.445	9.801	0.819	0.621
MMT Neural no cont.	0.504	11.106	**0.837**	0.485
MMT Neural context	**0.506**	**11.141**	0.836	**0.481**

Table 3.2: Results of the automatic evaluation using mteval.

3.6.4 Results

Translation ranking

29 people answered the survey. From those, 86.21% had previous experience as translators and 58.62% had worked on post-editing tasks. Confirming the initial hypothesis, most respondents preferred the human translation. However, this percentage was only of 60.52%. The second most preferred translation was NMT, with 25.17%, and PBSMT was only considered the best translation for 14.31% of the segments. We calculated inter annotator agreement using Fleiss Kappa (Fleiss et al., 1971), which showed a fair agreement among the annotators (κ=0.36). These results

were statistically significant in a one-way ANOVA comparison (p<.05).

Evaluation	Human	NMT	PBSMT
EN-ES (40)	60.52%	25.17%	14.31%

Table 3.3: Results of the human-NMT-PBSMT ranking survey.

Although the survey was conducted on a fairly small number of sentences, it seems to point in two directions: NMT is far from achieving the quality of human translation for medical texts, and NMT yields better translations than PBSMT. We conducted a manual analysis of the sentences in which NMT or PBSMT were selected as the best translation. It was observed that the main reason for the selection was terminology precision and fluency of the MT output.

Fluency and Adequacy

In the second experiment, eleven people answered the survey. Seven of them were translators with more than two years of experience and only four of them were students. Both fluency and adequacy obtained a higher rate for NMT after calculating the mean for both MT systems. We calculated inter annotator agreement using Fleiss Kappa (Fleiss et al., 1971). For fluency, it showed a poor agreement among the annotators (κ=0.01). Results were statistically significant in a one-way ANOVA comparison, with an f-ratio value of 2.75586 and a p-value of 0.04856 (significance at p<.05). For adequacy, there was also a poor agreement among annotators. These results weren't statistically significant, with an f-ratio value of 0.96767 and a p-value of 0.412816 (p<.05).

If we take a closer look at the sentences that had to be assessed, PBSMT segments often contain morphological problems (e.g. concordance) that we cannot spot on NMT segments, as in the following example. This way the general higher ratings for fluency and adequacy of the NMT system are confirmed.

Source: Craniopharyngioma had more hormone deficiencies (...).

PBSMT: Craneofaringioma tenían más déficits hormonales (...).

System	Fluency	Adequacy
PBSMT	2.28	2.24
NMT	2.46	2.50

Table 3.4: Results of the ranking survey.

PE time and technical effort

Results for the post-editing task by professional translators have been grouped in temporal effort and technical effort (see Tables 3.5 and 3.6). In both cases, the mean for PBSMT is higher, though only technical effort shows a statistically significant difference (in a T-test with a p-value of 0.002054). It should also be highlighted the considerable difference in time and keylogging between the translators, especially for the two professionals who post-edited PBSMT (as indicated by the standard deviation in Tables 3.5 and 3.6).

System	Mean	Std. Deviation
PBSMT	88.75	44.59
NMT	79.25	33.43

Table 3.5: Temporal post-editing effort (secs/segment).

System	Mean	Std. Deviation
PBSMT	130.68	39.63
NMT	54.99	16.90

Table 3.6: Technical effort (keystrokes/segment).

3.6.5 Discussion

Although the number of segments analyzed is quite small, for this language combination and text type, there seems to be a clear preference for

human translations, which are considered better in more than half of the cases. Regarding MT engines, NMT presents a better score in fluency and adequacy. This corresponds with the higher results in all automatic metrics.

Regarding temporal and technical effort, there is also a reduction on both means for NMT outputs. However, the reduction in temporal effort is much lower than the one we can see in technical effort. Hence, even though NMT produces more fluent results and implies less typing when post-editing, this improvement does not always entail a proportional reduction of the temporal effort for professional translators. This is probably due to the difficulty of spotting errors in more fluent outputs.

3.7 Experiment 2: Comparing SMT and NMT Output for Catalan into Spanish

For many years now, (closely) related languages have been post-edited using rule-based and phrase-based machine translation (MT) systems because they present less challenges due to their morphological and syntactic similarities. Due to the good quality results obtained with NMT, we analyze the performance of this approach compared to phrase-based statistical MT (PBSMT) on in-domain and general domain documents. We use standard automatic measures and temporal and technical effort to assess if NMT yields a real improvement when it comes to post-editing the Spanish-Catalan language pair.

We tried to answer to the following research questions:

- Which MT method (PBSMT or NMT) yields better results for post-editing Spanish into Catalan?
- How do post-editing measures correlate with automatic metrics?
- How does the domain and the formality of the texts affect the post-editing performance between Spanish and Catalan?

3.7.1 Methodology

We carried two experiments to assess the correlation of MT metrics with the post-editing time and technical effort. The participants were students in their last year of the Degree in Translation and Language Sciences at Universitat Pompeu Fabra (UPF). They post-edited during a task organized as part of a course syllabus on Localization taught by one of the authors.

They all acknowledged a C2 level in both languages. Although students may not be experienced professionals, the participants had translated into this specific language combination during their translation degree program, and had received specific post-editing (PE) training during the aforementioned course before carrying out the PE task. For these reasons, we can consider them semiprofessionals (Englund Dimitrova, 2005).

In the first experiment, 12 participants post-edited a short text (441 words, 14 segments) from Spanish into Catalan translated with three different MT systems: an in-domain PBSMT Moses and an in-domain NMT Marian we had previously trained, and NMT Google Translate. The text was a passage from a UE document, which presented a lot of fixed syntactic structures, and technical content.

They had to carry the task using PET (Aziz et al., 2012) (see Section 3.6 and 4.3 for further information). It logs both post-editing time and edits (keystrokes, insertions and deletions, that is, technical effort). As it was a short text, they were asked to post-edit it without any pauses. The main characteristics of the post-editing tool were also explained before beginning the task.

In the second experiment, the same 12 participants post-edited a general domain short text (379 words, 17 segments) from Spanish into Catalan translated with our general purpose PBSMT Moses, our NMT Marian and NMT Google Translate systems. The text was a fragment from a piece of news appeared in the newspaper *El País* on April 4th, 2019. They post-

edited the text with the same tool and the same conditions as in the first experiment.

In order to avoid bias, participants never post-edited the same text twice. We divided the 12 post-editors into groups of 4 people. All the members of each group post-edited the in-domain text translated with an MT system. They also post-edited the general text output for the same MT system.

3.7.2 MT Systems and Training Corpus

For our experiments, we have trained two statistical and two neural machine translation systems: one for a general domain and the other for the Administrative/Legislative domain.

Corpora

For the general domain we have combined three corpora:

- A self-compiled corpus from Spanish-Catalan bilingual newspapers
- The GlobalVoices corpus (Tiedemann, 2012b)
- The Open Subtitles 2018 corpus (Lison and Tiedemann, 2016)

The systems for the Administrative/Legislative domain have been trained with the corpus from the Official Diary of the Catalan Government (Oliver, 2017). The Catalan part of the corpora has been normalized according to the new orthographic rules of Catalan. This step has been performed in an automatic way.

In Table 3.7 the sizes of the training corpora are shown. A small part of the corpus (1000 segments) has been reserved for optimization (statistical) and validation (neural). Another set (1000 segments) has been reserved for evaluation. So there are no common segments in the train, validation and evaluation subcorpora.

The corpora have been pre-processed (tokenized, truecased and cleaned) with the standard tools distributed in Moses[16]. The same pre-processed corpora have been used for training the statistical and the neural systems.

Corpus	Segments	Tokens es	Tokens ca
DOGC	6,943,595	155,233,465	157,000,914
General	4,163,009	93,489,848	93,538,673

Table 3.7: Size of the training corpora

PBSMT system

For the statistical system we have used Moses (Koehn et al., 2007) and trained a system for each of the corpora. We have used a language model of order 5. For the alignment we have used mgiza with grow-diag-final-and.

NMT system

For the neural machine translation system we have used Marian[17] (Junczys-Dowmunt et al., 2018). We have trained the systems using an RNN-based encoder-decoder model with attention mechanism (s2s), layer normalization, tied embeddings, deep encoders of depth 4, residual connectors and LSTM cells (following the example of the Marian tutorial[18]).

3.7.3 Automatic Evaluation of the MT Systems

The systems have been automatically evaluated using mteval[19] to obtain the values for BLEU, NIST and WER. Table 3.8 shows the evaluation

[16]http://www.statmt.org/moses/
[17]https://marian-nmt.github.io
[18]https://marian-nmt.github.io/examples/mtm2017/complex/
[19]https://github.com/odashi/mteval

System	BLEU	NIST	WER
NMT Marian Admin.	0.845	13.055	0.1424
PBSMT Moses Admin.	**0.896**	**13.458**	**0.0881**
Google Translate Admin.	0.869	13.279	0.0918
NMT Marian General	0.767	12.426	0.185
PBSMT Moses General	0.812	12.799	0.171
Google Translate General	**0.826**	**12.980**	**0,121**

Table 3.8: Automatic evaluation figures

figures for all the MT systems used. As a reference, we also include the metrics for Google Translate[20] for the same evaluation sets.

3.7.4 Results

Automatic measures

To assess the quality of the MT systems, we included some of the most commonly used automatic evaluation metrics (see Section 3.3.2 for a complete list of all metrics). The BLEU metric (Papineni et al., 2002) and the closely related NIST (Doddington, 2002) are based on n-gram. The word error rate (WER), which is based on the Levenshtein distance (Levenshtein, 1966), calculates the minimum number of substitutions, deletions and insertions that have to be performed to convert the generated text into the reference text.

For all the measurements, our NMT Marian system had the worst rates (see Table 3.8). However, our PBSMT Moses model had 0.027 BLEU points more than Google Translate for in-domain texts. In the general domain, Google Translate was better rated. That is why we decided to include Google Translate as part of the post-editing tasks.

[20]Translations were performed on April 9th, 2019

Domain	System	Mean	Std. Deviation
In-domain (UE)	Marian	50.89	11.78
	Moses	73.70	29.60
	Google	**34.68**	10.88
General domain	Marian	33.71	2.75
	Moses	42.94	13.96
	Google	**32.93**	12.65

Table 3.9: Temporal post-editing effort (secs/segment)

Domain	System	Mean	Std. Deviation
In-domain (UE)	Marian	64.55	65.75
	Moses	12.09	10.50
	Google	**2.23**	1.38
General domain	Marian	37.99	31.91
	Moses	**16.43**	1.62
	Google	27.34	37.88

Table 3.10: Technical post-editing effort (keystrokes/segment)

Domain	System	Mean	Std. Deviation
In-domain (UE)	Marian	42.85	0.71
	Moses	53.57	1.50
	Google	**85.71**	1.32
General domain	Marian	20.59	1.12
	Moses	20.58	1.12
	Google	**39.70**	0.83

Table 3.11: Percentage of unmodified segments

Post-editing time and effort

For the in-domain (Administrative/Legislative) PE task, our NMT Marian model was the one that supposed higher PE technical effort. In fact, as we can see in the manual evaluation (see example 2, Table 3.13), errors include adding hallucinations (Wang and Sennrich, 2020) to the target text which do not convey the meaning of the source text.

Google Translate has a very low rate, which is statistically significant, and correlates to the number of unmodified segments (see Table 3.11). This correlates to the results obtained by Shterionov et al. (2018), where the automatic quality evaluation scores indicated that the PBSMT engines performed better, but the human reviewers showed the opposite result.

In the case of temporal effort, the Moses system was the one that took longer, even though it shows the best results for all the automatic metrics. Our Moses system had 0.027 BLEU points more than Google Translate in the automatic evaluation. However, post-editors spent less time post-editing the Google Translate output (see Table 3.9).

For the general post-editing task, automatic metrics correlate to temporal but not to technical effort. The Google Translate output, which showed a 0.014 increase in BLEU, was translated using far more keystrokes per segment. However, it should be noted the high standard deviation in this case, as in the case of the Marian output.

Another interesting figure is the number of unmodified segments (see Table 3.11). In this case Google Translate results are far better than Moses, both for in-domain and general domain, which seems to indicate that NMT produces more fluent sentences.

Manual analysis

The goal of the manual analysis is to complement the information provided by the measures in previous sections. Following Farrús et al. (2010), we have used a taxonomy in which errors are reported according to the different linguistic levels involved: orthographic, morphological, lexical,

semantic and syntactic, and according to the specific cases that can be found in the post-editing tasks from Spanish into Catalan.

Domain	System	Ort.	Morph.	Lex	Sem.	Synt.	Total
In-domain (UE)	Marian	0	0	2	18	0	**20**
	Moses	2	0	2	0	2	**6**
	Google	0	0	0	0	1	**1**
General domain	Marian	0	0	8	5	3	**16**
	Moses	9	12	2	0	5	**28**
	Google	0	11	1	0	3	**15**

Table 3.12: Number of errors according to the linguistic level

Table 3.12 shows the error rates corresponding to all outputs and Table 3.13 includes several translation examples from the three systems for the general domain test set. In general, the examples included show the advantages of the two NMT models (especially Google Translate) compared to the PBSMT system, in the following terms:

1. There is a **better use of pronouns** in the NMT versions. In this case, the Marian output generates the best version (for example, translates correctly the pronoun *el* into Catalan as it can be seen it the first example).

2. There is a **better use of prepositions** in the Marian output. For example, *el* is used before the year instead of *en* as in Spanish.

3. There is **a better integrity of meaning** in the Google Translate version. One of the recurrent problems of our Marian version was the addition of extra information or the mistranslations, like in the third example, where it ads *d'arreu del món*. The Moses version also ads *basca* (it's the only time Moses adds extra information).

4. The Google Translate version **is more fluent**. Even though the Moses output generally includes all the source information, it sometimes truncates the sentences.

1	ES	[...] lo pidió prestado al dueño en 1890 [...]
	Marian	[...] **el** va demanar prestat al propietari **el** 1890 [...]
	Moses	[...] **ho** va demanar prestat **el** propietari en 1890 [...]
	Google	[...] va demanar prestat **a l'amo en** 1890 [...]
2	ES	Es un Lefaucheux [...]
	Marian	És **un lladre** [...]
	Moses	És **un Lefaucheux** [...]
	Google	És **un Lefaucheux** [...]
3	ES	[...] un prado de la localidad de Auvers-sur-Oise [...]
	Marian	[...] un **enclavament** de la localitat **d'arreu del món** [...]
	Moses	[...] un prat de la localitat **basca** d'Auvers-sud-Oise [...]
	Google	[...] un prat de la localitat d'Auvers-sud-Oise [...]
4	ES	[...] intentaron trabajar juntos en Arlés, al sur de Francia.
	Marian	[...] van intentar treballar junts a **Espanya**, al sud de França.
	Moses	[...] van intentar treballar junts, a **Arle. Al sud** de França
	Google	[...] van intentar treballar junts a Arles, al sud de França.
5	ES	De la pistola no volvió a saberse nada hasta 1965 [...]
	Marian	De la pistola no es va tornar a saber res **fins al 1965** i [...]
	Moses	De la pistola no va tornar a saber res **fins 1965.**
	Google	De la pistola no va tornar a saber res **fins a 1965** i [...]

Table 3.13: Translation examples

5. NMT achieves **a better syntactic organization** that produces a more understandable sentence with less mistakes.

3.7.5 Discussion

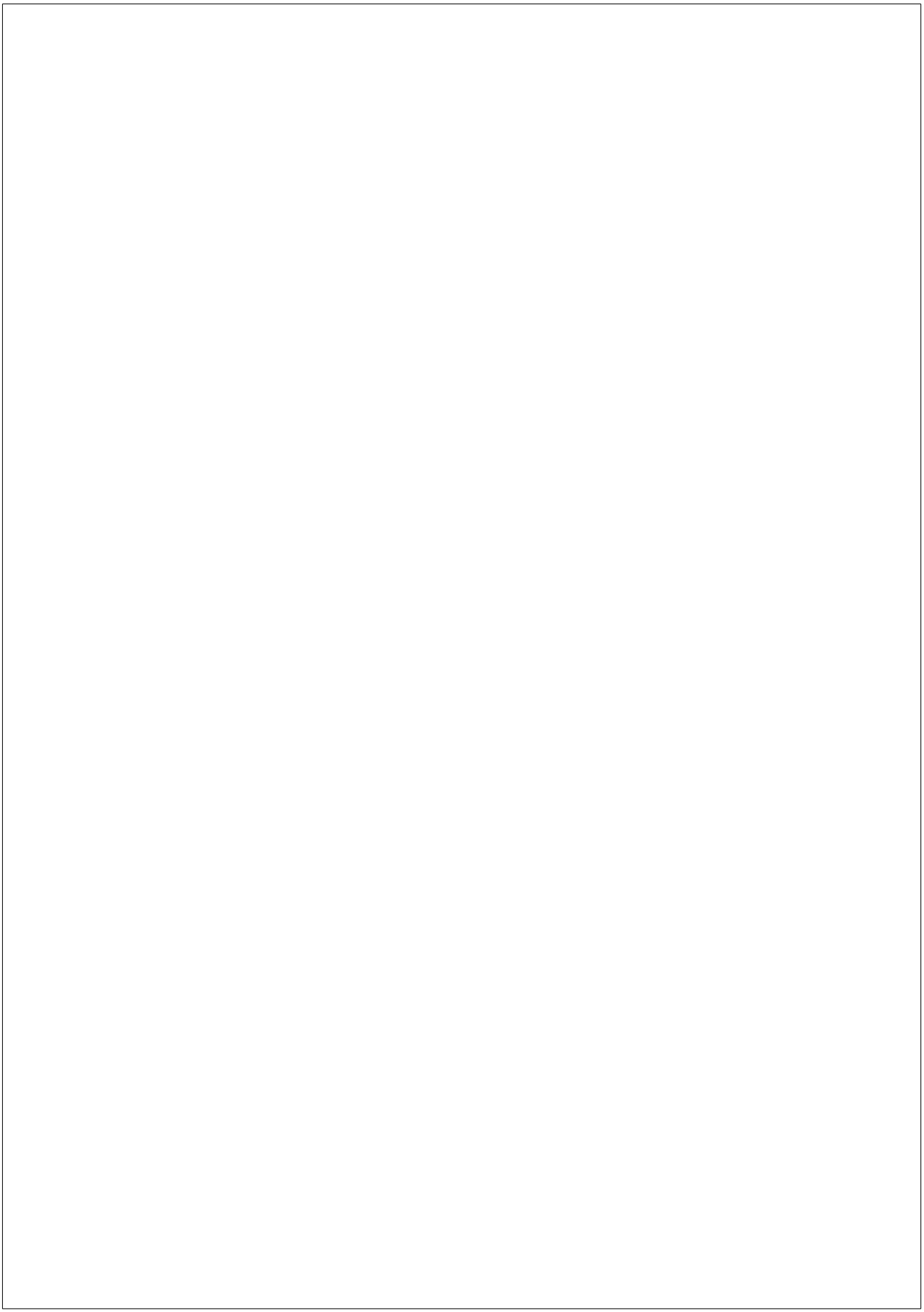
These experiments shows a comparison between PBSMT and NMT for general and in-domain documents from Spanish into Catalan. Automatic metrics show better results for PBSMT with in-domain texts. However, Google Translate NMT system has a better rate when translating general domain sentences.

Regarding post-editing, for this study, text types, and language pair results show an improvement of unmodified segments and temporal effort for NMT systems. For the in-domain text, with a lower BLUE rate, both technical and temporal effort, as well as the number of unmodified segments and translation errors, show a clear improvement of Google Translate. The manual analysis also confirms that NMT systems tend to solve some of the usual problems of PBSMT systems when translating closely related languages. However, as it is shown in the translation from our NMT Marian system, a lower quality in NMT systems tends to produce unreliable translation outputs, which complicate the post-editing process.

3.8 Conclusion

In this chapter we have compared raw NMT and PBSMT outputs used for post-editing. We have studied the correlation of MT automatic metrics with measures of PE effort for the English-Spanish and Spanish-Catalan language combinations.

In general, NMT shows an improvement both in the automatic scores and in temporal and technical effort. However, these measures do not always have a clear correlation. Furthermore, sometimes automatic scores tend to underestimate the NMT output.



Chapter 4

MEASURES OF POST-EDITING EFFORT

4.1 Introduction

All the metrics presented in the previous chapter, which are often used to measure post-editability, focus on the MT raw output. Translation studies have also used methodologies from fields like psychology to study the cognitive processes implied in the PE task. We believe understanding and using these methodologies to measure the actual effort can help us better understand PE and can be useful to assess the quality of the MT output for PE.

In this chapter, we describe all the dimensions included in PE effort and the different scores and tools devised to measure it. We also present two experiments designed to research how all these measures of effort interrelate and can be used to describe the translators' cognitive processes. In Section 4.4 we introduce a new tool designed to easily measure direct and indirect PE effort indicators. In Section 4.5 we study how different measures of effort correlate to each other in order to use them as indicators of the cognitive effort.

4.2 Post-editing Effort

Krings (2001) offered in his seminal work a definition of PE effort which is still applied in all related research. It used three separated but interrelated types of effort: temporal, technical and cognitive. Research has shown cognitive effort correlates with technical and temporal PE effort (Moorkens et al., 2015), even though it is not correlated directly with the number of edits (Koponen, 2012) or the time spent.

4.2.1 Temporal Effort

Temporal effort measures the time spent post-editing the MT output. As time can be directly linked to productivity, it is a key aspect in current industrial scenarios, where there is a need to reduce costs (Guerberof, 2009a; Sosoni and Rogers, 2013) and shorten time cycles, and it is often used to calculate the rate applied to PE tasks.

Research consistently shows that post-editing is faster than translating from scratch (O'Brien, 2005; Carl et al., 2011; Aranberri et al., 2014; Jia et al., 2019; Läubli et al., 2019), although for general language texts some studies see no significant improvement in speed (Screen, 2017). In general terms, PE does not have a negative impact on quality (Plitt and Masselot, 2010).

However, recent research has studied if translationese also affects PE: a phenomenon which has been called post-editese. Results show PE documents are simpler and more normalised, with a higher degree of interference from the source language than human translations (Daems et al., 2017a; Castilho et al., 2019; Toral, 2019).

Initially, in order to study temporal effort, translators were asked to record themselves the time spent, but this method was highly unreliable. Current research can measure productivity with usual CAT tool or other tailored tools developed to measure different dimensions of PE effort (see Section 4.3 for a comprehensive list of products currently used).

4.2.2 Technical Effort

Technical effort refers to the amount of editing work that is involved in the PE process, and can be captured by registering the number of insertions, deletions and re-orderings carried out by the translator while post-editing. It can be measured with keystroke analysis or key-logging data, which usually requires the translators to post-edit using a specific software (see Section 4.3).

This dimension of effort is the one usually used to calculate indirect measures based on the post-edited product. Even though these measures do not account for the different revisions and modifications conducted on segments, they compare the final post-edited version (instead of one or more references translated from scratch) with the raw MT output to produce a numeric score.

Some of these metrics are listed below:

Human-mediated Translation Edit Rate (HTER) (Snover et al., 2006) calculates the Translation Edit Rate (TER) using the post-edited version as the reference translation to score the MT output. TER is an edit-distance measure that allows block movement of words, called shifts. These movements have the same cost as insertions, deletions or substitutions. It uses "a greedy search to select the words to be shifted, as well as further constraints on the words to be shifted. These constraints are intended to simulate the way in which a human editor might choose the words to shift" (Snover et al., 2006). It equals the number of edits divided by the average number of reference words.

Some years later, Snover et al. (2009) also proposed TER-Plus, which includes morphology, synonymy and paraphrases and incorporates tunable parameters.

HBLEU is the BLEU metric (see 3.3.2 for a complete definition) often used to assess MT quality but using the PE version as single reference.

The **edit distance** (Levenshtein algorithm) is a measure of similarity which

calculates the number of deletions, insertions, substitution and (sometimes) changes in position required to transform one segment into another, and can be calculated at the character level or the word level.

In this case, we compare the raw MT output with the final post-edited version to produce the absolute edit distance. A matrix is initialized measuring in the (m,n)-cell the Levenshtein distance between the m-character prefix of one with the n-prefix of the other word. The cost is usually set to 1 for each of the operations (Levenshtein, 1966; Rani and Singh, 2018).

4.2.3 Cognitive Effort

The two aforementioned dimensions can be measured directly, but only account for a part of the PE effort. It is important "not only the ratio of quantity and quality to time but also the cognitive effort expended; and the higher the effort, the lower the productivity." (O'Brien, 2011b, p. 198)

Cognitive effort is related to the mental processes involved in PE task and can be traced to cognitive psychology. It includes reading the texts, thinking about the translation from the source and the suggested MT solutions, correcting the mistranslations or omissions, and revising the final version of the translation produced. The cognitive processes and decisions implied in post-editing a text may not involve any edits and, on the contrary, many errors of the MT raw output can include repetitive and mechanical tasks which imply a lot of time but very little cognitive effort.

Effort has been also extensively studied in educational psychology. The research in this field studies how cognitive demand influences learning tasks and how the design of learning materials and environments can reduce this load. Effort is one of the elements included in the cognitive load, which involves three variables: mental load, mental effort and performance (Paas et al., 1994; Kirschner, 2002; Paas et al., 2003). Mental load is related to the intrinsic difficulty posed by the task. Mental effort is linked to the total amount of cognitive resources which are expended by individual users when performing the task. And performance is the result of the interaction of the two previous variables.

According to cognitive load theory, the load implied by a task can be intrinsic and extraneous. This theory deals with learning materials, but a parallel can be drawn with the load implied in post-editing (Nunes Vieira, 2016). MT output could be linked to the intrinsic cognitive load, which is the main goal of this thesis (see Section 4.4 and 4.5 for measures of this load). And extraneous cognitive load would be related to all external factors, such as the tool used, the working conditions of translators or any negative bias, for example (see Section 5.3 to have detailed information on perception and bias).

Since cognitive effort cannot be measured directly, different proxy measures have been used. Krings (2001) proposed **Think-Aloud Protocol (TAP)** to study the extent of the cognitive processes implied in PE. He realized TAP could be useful to understand conscious processes but was unable to account for all the automatic processes carried out. However, Nunes Vieira (2016) found there was a strong correlation between TAP ratings and other measures of effort. Although it has been acknowledged as a useful methodology, it has received certain criticism, mainly that there are two processes that interfere with each other (Toury, 2012) and that the artificiality of the process implies that subjects do not explain all thoughts to the researchers (House, 1988).

After being introduced as part of Translation Process Research (TPR), **eye-tracking** tools have also been used to measure eye positions and eye movements when studying PE effort (Carl et al., 2011; O'Brien, 2011b; Doherty, 2013). They count the number and duration of fixations, when the eyes are relatively still (Moorkens, 2018). It has been used in combination with key-logging data and retrospective think-aloud protocols (Alves, 2003) and triangulated with technical and temporal effort with pause analysis to obtain more reliable results (O'Brien, 2007). Other studies have used a wider range of physiological measures such as pupil dilatation, galvanic skin response, blood pressure, heart rate (variability) and respiration to account for cognitive load (Herbig et al., 2019).

Even though sometime pauses and interruptions while translating can be attributed to external factors, it would be natural to assume that these

pauses are indications of cognitive effort (Lacruz, 2017). O'Brien (2006) analyzed some of the factors which influence pauses and could relate them to common locations such as before or after phrases or sentences. She also introduced the concept of pause ratio, which is the total time in pause divided by the total time of post-editing. She studied the impact of Negative Translatability Indicators (NTI) on PE by studying pause ratio, but results were not conclusive or even contradicted the hypothesis that NTI would increase pause ratio.

Lacruz (2012) suggested a new metric, average pause ratio (APR), which is computed as the average time per pause divided by the average post-editing time per word. The author realized that challenging edits were often accompanied by clusters of short pauses, which did not affect the pause ratio. However, when these pauses were present, APR was slightly lower, which would be an indication of a higher PE effort. Further research confirms the relation between APR and cognitive effort with a pause threshold of 5000 ms (Carl et al., 2008; Liu and Du, 2014).

Lacruz and Shreve (2014) also introduced another metric, pause to word ratio (PWR), which is the number of post-editing pauses divided by the number of words in the MT segment. It showed high correlation with other effort metrics regardless of the pause threshold selected.

Subjective ratings have also been used to estimate the cognitive effort, as a higher cognitive load is supposed to generate a higher sensation of effort (Vieira, 2016). Koponen (2012) compared perceived technical PE effort with actual technical effort. Participants rated worse longer segments, even if few modifications were needed, which suggests length affects cognitive effort. She also associated certain parts of speech and word order to higher perceived PE effort.

Gaspari and Toral (2014) compared the perception of translators after carrying out a PE task with actual productivity. Results showed participants always preferred translation from scratch and had a negative bias regarding their performance. Teixeira (2014) found there was no direct correlation in all cases between temporal and technical effort with par-

ticipants' post-task ratings of their own performance (see Section 5 for results of experiments comparing translators' perception and productivity when post-editing).

Choice Network Analysis (CNA) (Campbell, 1999, 2000; Hale and Campbell, 2002) was suggested to solve the problems posed by TAP. Thus, CNA is a method that enables to estimate the difficulty of certain passages of text based on the complexity of choices available to translators. When all translators translate a sentence the same way, it is assumed the passage has little difficulty. And the contrary happens for passages where translators suggest many different translation solutions.

O'Brien (2005) points out a potential weakness of the analysis: the differences could be attributed to translator individuality and creativity. However, she suggests this method could work better for PE because translators are already offered a version, which they can keep or decide to modify if it is not correct. Furthermore, it can be triangulated with other cognitive, temporal or technical effort indicators. Her experiments show a correlation between CNA and other indicators of effort, which clearly suggests this methodology could be useful to spot segments with high cognitive demand (see Section 6.5).

4.3 Tools to Measure PE Effort

In order to analyze the different components of post-editing effort, it becomes paramount to use tools that are able to log time, keyboarding, and other potential indicators of cognitive effort (e.g. gaze data). Currently there is a proliferation of these tools (Vieira, 2013), mainly because each research project has specific requirements and focuses on the aforementioned dimensions of effort.

Some studies have focused on productivity in real industrial scenarios and they have worked with commercial tools already used to post-edit. Federico et al. (2012) report on a field test conducted in SDL Trados Studio with 12 professional translators translating from English into German

and Italian. Even though they used a plugin to measure time, not all obtained time measurements were reliable and had to remove roughly 30% of translated words before carrying out the analyses.

Läubli et al. (2013) carried out a small-scale experiment involving six translation students and four short texts. They used the translation workbench Across with screen recordings to obtain precise time measurements. Parra Escartín and Arcedillo (2015) also used a commercial tool, MemoQ, to keep track of the time spent post-editing each segment. They showed a productivity gain for post-edited segments of 24%.

Other examples of tools which are part of real translation workflows are the Qualitivity[1] plugin, which can be added to SDL Trados to measure post-editing effort, and Dynamic Quality Framework,[2] a tool developed by TAUS which can be used as a standalone benchmark or as an SDL Trados plugin.

Alternatively, we find iOmegat[3], which is an instrumented version of the open source translation tool OmegaT, created in collaboration between Welocalize, John Moran, and the Centre for Next Generation Localization (CNGL).

There has also been EU-funded research to develop open-source workbenches to help improve quantitative measures of effort (CASMACAT[4] and Matecat[5]).

Other research tools collect gaze data, which can be used as a proxy to study cognitive effort. Tobii Pro Lab is the commercial Windows-oriented eye-tracking software that accompanies Tobii eye trackers [6]. It can calculate a variety of eye-tracking metrics and create visual representations of the data.

[1] https://appstore.sdl.com/language/app/qualitivity/612/
[2] https://www.taus.net/dqf
[3] https://www.adaptcentre.ie/downloads/license/iOmegaT_Available_to_License.pdf
[4] https://www.casmacat.eu
[5] https://www.matecat.com
[6] https://www.tobii.com

Another similar product is Translog-II (Carl, 2012), which is a Windows-oriented program that records user activity data (UAD), that is, all the keystrokes and gaze movements. It is meant specifically for translation process research (TPR) and it offers the possibility of further processing all the data collected with the scripts included in the TPR database of the Centre for Research and Innovation in Translation and Technology (CRITT TPR-DB). Even though these tools collect extensive information, they have specific and demanding settings which are not suitable for all experimental scenarios.

Some products devised for a specific experiment are not made available to the public afterwards (Plitt and Masselot, 2010; Green et al., 2013). Other tools focus on obtaining as much information as possible with an easy-to-use product. For example, TransCenter (Denkowski and Lavie, 2012) is an open-source, web-based tool that allows users to carry out PE tasks and logs time and keyboard/mouse activity at a sentence level.

Another tool useful for quantitative investigations specifically designed for post-editing is PET (Aziz et al., 2012). It can also be accessed from any platform, although it is based in Java, which can sometimes be challenging for end-users who need to open the tool from their desktop computers to work on the files. In fact, we used it for the experiments described in Section 3.6 and 3.7 and many of the translators had problems executing the tool, mainly due to compatibility issues with the Java versions.

In addition to recording time and effort indicators at a segment level, PET also allows users to perform evaluation tasks on different customizable scales and criteria (error annotation and translation ranking). The data file with all the information is saved in an xml file. However, it does not offer graphics or any other visual information with the results nor does it include an analyzer which can produce multiple automatic metrics.

In order to collect data from the PE effort in an easy way which can eventually help determine the quality of the MT raw output for a posterior post-editing, we have developed PosEdiOn, which is explained in Section

4.4. In Section 4.5 we describe the first of the experiments conducted with this tool to study measures of PE effort.

4.4 PosEdiOn: Post-Editing Assessment in Python

Professional translators usually use commercial products to translate and post-edit. In the 2018 Language Industry Survey[7] conducted by EUATC, Elia, FIT Europe, GALA and LINDWeb, SDL Trados[8] was the most used product with more than half of the market quota, followed by MemoQ,[9] Memsource,[10] Wordfast,[11] and Across.[12] However, these existing post-editing environments have a restricted availability and flexibility, which can be challenging for conducting the experiments.

Instead of trying to reproduce the working conditions of translators, which vary greatly among individuals, other tools establish controlled conditions in order to obtain non-biased data. In this context, translators use a post-editing tool that records all the post-editing information. The tool should be easily accessed from any platform and must have an easy-to-use interface.

We present PosEdiOn[13], a simple standalone tool that allows post-editing of MT output and records information of the post-editing effort (time and keystrokes) at sentence-level. It also includes different evaluation scores that the user can interpret easily to assess the post-editing process (such as edit distance, HBLEU and also HTER). As it does not depend on any specific CAT tool, it allows the collection of post-editing data in a controlled way. It can be used by professionals to assess the convenience of

[7] http://fit-europe-rc.org/wp-content/uploads/2019/05/2018-Language-Industry-Survey-Report.pdf

[8] https://www.sdl.com/

[9] https://www.memoq.com

[10] https://www.memsource.com

[11] https://www.wordfast.net

[12] https://www.across.net

[13] https://github.com/aoliverg/PosEdiOn

post-editing a certain MT output and by researchers to study post-editing effort.

4.4.1 PosEdiOn

Characteristics

PosEdiOn is a post-editing tool developed mainly to collect information on different implicit and explicit effort indicators. It records time and keystrokes, and it also calculates some of the main indirect effort estimation measures (HTER, HBLEU and edit distance). It produces a file with the raw measurements but it also includes a results file with visually structured information that can be easily understood by any user.

It was developed completely in Python3 and it works in any platform which has Python installed. Translators tend to work from home with a great variety of platforms and devices, and do not always have the computer skills to solve any compatibility errors they may encounter with the tools they are about to use. Therefore, a Windows executable file is also available, which allows to run PosEdiOn without the need of installing the Python interpreter.

Files and tasks

PosEdiOn is designed to facilitate the distribution of post-editing tasks in an easy and error-free way. The user receives a zip compressed folder with all the needed elements:

- The PosEdiOn program itself, usually as a Python file. Optionally, a Windows executable can be also used. In this case, sending the zipped file by e-mail can cause problems as some mail providers block attachments with executable files. Alternatively, a link to the zipped file can be used to distribute the post-editing tasks.
- The configuration file (*config.yaml*) that provides all the information necessary for the post-editing task. See section 4.4.3.

- The post-editing task itself as a tab delimited plain text file. The text file is structured in six fields: segment count, segment id, source text, machine translated text, post-edited text and segment status.

  For translation tasks, only the source text is compulsory and the target text can be left empty. In this case, the translator will be presented only with the source text. For post-editing tasks, both source text and machine translated text are compulsory and the post-editor will be presented with the source text and the output text from MT. Each time a segment is validated, this file and the status of the segment are updated.

Once the compressed file is received, it must be unzipped. After executing the program, the task is directly presented. When the translator begins to work on the new task, a new file (*actions.txt* or any other file name stated in the configuration file) is created. All actions including keystrokes, mouse actions and button clicks are stored in this file along with the time it is performed. An example can be seen in the following figure:

```
START 1 1-0 2020-02-22 22:28:04.979308
F 1 1-0 020-02-22 22:28:04.996692 Focus_in
M 1 1-0 2020-02-22 22:28:08.840216 Mouse.button1
F 1 1-0 2020-02-22 22:28:08.840857 Focus_in
K 1 1-0 2020-02-22 22:28:09.742533 Key.letter.u 1.6
M 1 1-0 2020-02-22 22:28:13.129137 Mouse.button1
OUT 1 1-0 2020-02-22 22:28:23.827548
IN 2 1-0 2020-02-22 22:28:23.829034
K 2 1-0 2020-02-22 22:28:25.018297 Command.CtrlReturn 1.8
OUT 2 1-0 2020-02-22 22:28:25.020480
IN 3 1-0 2020-02-22 22:28:25.046122
K 3 1-0 2020-02-22 22:28:29.602347 Key.navigation 2.5
....
```

Figure 4.1: File with the actions recorded

All analysis and measurements can be obtained from this actions file.

Each line contains several information fields separated by tabs:

- The first field provides information about the kind of action. The actions are: START (task is started); PAUSE (task is paused); EXIT (user exits the application); RESTART (user restarts the task); IN (user enters into a segment); OUT (user exits a segment); K (keyboard action); M (mouse action); C (command action); B (user clicks a button on the application); F (application loses or gains focus); CLEAR (user clears all the content of the translation); RESTORE (user restores the content of the translation).
- The second field indicates the segment number.
- The third field indicates the segment id.
- The fourth field gives the time and date of the event.
- Some actions have a sixth field which provides more detailed information about the event. For example, the key pressed, the text copied or pasted, and so on.
- Key actions have another field indicating the position in the target text where the key is pressed.

The user can pause and even stop the task and close the PosEdiOn program. Once the task is restarted, the new data will be appended to the existing actions file.

When the task is finished, the folder containing the program should be compressed again and sent back to the person who has to carry out the analysis.

4.4.2 User Interface

The interface displays the source and target language segments one on top of the other. Figure 4.2 shows the PosEdiOn interface, where the upper window contains the source segment and the lower window enables the translator to edit the text. Translators can see a wider context using the

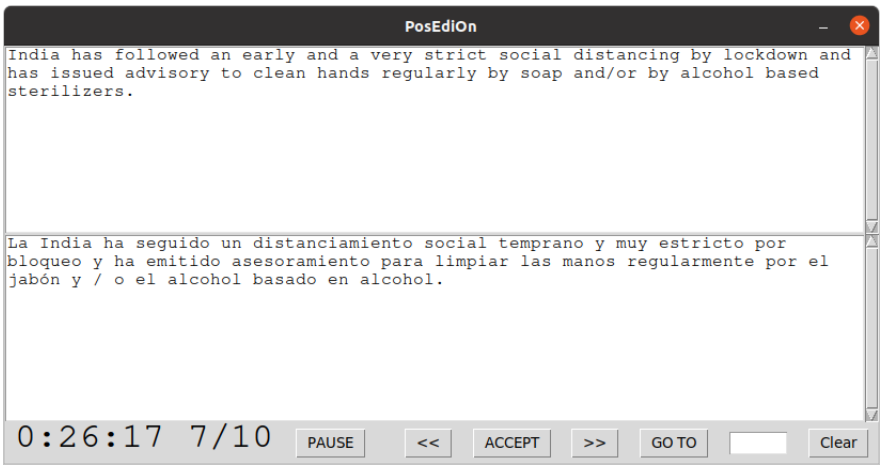


Figure 4.2: PosEdiOn interface

toolbar buttons located on the lower part, which can be used to move along the whole document.

Each unit is translated/edited one at a time and navigation through the different segments of the document can be achieved in four ways:

- Once the translator has finished post-editing a segment, he needs to validate it using the Ctrl+Enter keys. When this is done, the tool moves automatically to the next segment.
- To validate a segment, the user can also use the ACCEPT button. Once pressed, it also moves to the next segment.
- Using the << or >> buttons in the toolbar located at the lower part of the screen.
- Using the GO TO box, where you can write the number of the segment you want to move to.

Once a segment is accepted, its background turns green. The user can mark a segment as validated (green) using Ctrl+g; or he can change the state to undone (white background) using Ctrl+w. Segments can also be

marked as red (Ctrl+r) to indicate a problematic status. Red segments can be reached directly using Ctrl+s.

4.4.3 Customization

In order to facilitate customization, certain elements can be modified in the *config.yaml* file without having to access the Python script.

As shown in Figure 4.3, users can customize the following elements:

- The name of the text file containing the translation or the post-editing task.
- The name of the actions file, where all the information containing the user’s actions is stored.
- The source and target language codes.
- The size of the tool’s window. Both the height and the width can be changed.
- Whether the source segment text can be edited or not. The edits introduced in the source segment are not registered by the tool. If the source segments can be edited, users can select and copy fragments of the source text.
- The size and the type of font used for the source and target segments.
- Whether or not to show the chronometer.
- The set of characters to be considered as symbols or punctuation. It also includes up to three user-defined groups of characters. In the example, a user-defined group called *mathematical* (containing symbols of mathematical operations) is defined.

```
Text:
    file: test-Google-1.txt

Actions:
    file: actions.txt

Languages:
    source: eng
    target: spa

Size:
    height: 10
    width: 80

Behaviour:
    allowEditSL: True

Font:
    font: courier 12

Chronometer:
    status: show
    #possible values: show / hide

Definition:
    symbols: "! @ # $ % ^ & ( ) _ { } [ ] ’ ? ¿ ! ¡ < >"
    punctuation: ", : ; ."
    nameuserdef1: mathematical
    userdef1: "+ - * / ="
    nameuserdef2: None
    userdef2: None
    nameuserdef3: None
    userdef3: None
```

Figure 4.3: View of the customizable elements

4.4.4 PosEdiOn Analyzer

PosEdiOn has a companion program, PosEdiOn analyzer, that performs different analyses on the PosEdiOn project files and offers a wide range of measurements. More specifically, it can calculate:

- Time spent editing each segment.
- HTER (Snover et al., 2006), the TER value comparing the raw MT output with the post-edited segment. A value of HTER is provided for each segment. The value of TER is calculated using tercom.[14]
- HBLEU, a BLEU (Papineni et al., 2002) value obtained comparing the raw MT output with the post-edited segment.
- HEd, an edit distance (Leveshtein distance) value calculated comparing the raw MT output with the post-edited segment.
- Keystrokes for each segment.

To calculate the normalization of time, HEd (and eventually Ed) and keystrokes values, users can chose three different criteria: segment, token or character. All these values are provided both for each segment and for the whole document. On top of that, the mean and standard deviation are also calculated.

Users can choose to prune results. The pruning is based on a maximum value of normalized time and on a maximum value of normalized keystrokes. These maximum values are calculated with the mean value and two times the standard deviation.

All segments with a normalized time greater than the maximum, or with a normalized number of keystrokes greater than the maximum, are not taken into account to calculate the pruned values of all scores. The results are provided as numeric values and with a visual presentation of the results following the ideas of the Vis-Eval Metric Viewer (Steele and Specia, 2018).

[14] https://github.com/jhclark/tercom

```
Filepath:
  path_in: /home/user/directory
  path_out: /home/user/directory
Files:
   results: results.txt

Measures:
   bysegment: True
   normalization: tokens
   #one of segment, token, char
   HTER: True
   HBLEU: True
   HEd: True
   round_time: 2
   round_keys: 2
   round_hTER: 4
   round_hBLEU: 4
   round_hEd: 2
   round_other: 1

Graph:
    create: True
    #one of True, False
    type: bar
    #one of bar, pie
    measure: HTER
    #one of HBLEU, HEd, HTER
    pruned: True
    #one of True, False
```

Figure 4.4: Yaml configuration file

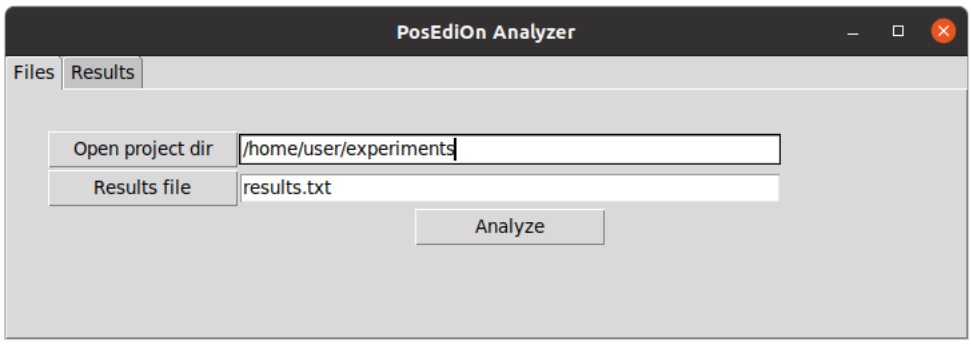


Figure 4.5: PosEdiOn analyzer interface

Configuration

The configuration of the tool is performed using a *yaml* configuration file (*config-analyzer.yaml*) as shown in Figure 4.4.

The file paths including the location of the project and all the results can be specified. The name of the results file can also be customized by adding a prefix, a suffix and an extension to the current name of the project. If no prefix, suffix or extension is required, any of these fields can be left blank.

The measurements can also be customized, and users can decide whether or not to show measurements by segment, the normalization criteria, which measurements will be calculated and shown, as well as the number of decimal points. Remember that the values of TER, BLEU and Ed will be calculated and shown only if a reference file is provided, regardless of the values in the configuration file.

Use of PosEdiOn analyzer

PosEdiOn analyzer has a simple GUI interface (see Figure 4.5) where we can set the input project and the results files if we want to override the parameters given in the configuration file.

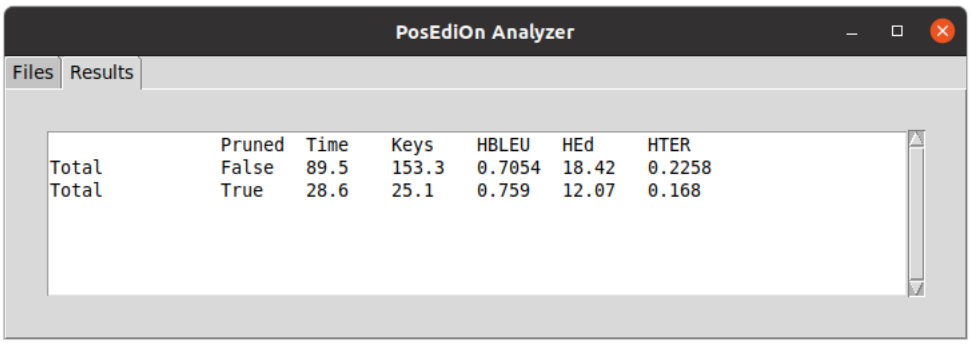


Figure 4.6: Results tab in PosEdiOn analyzer

Count	ID	Pruned	Time	Keys	HBLEU	HEd	HTER
1	1-0	False	132.1	78.6	0.66	20.61	0.2857
2	1-0	False	18.7	18.1	0.8336	10.84	0.1143
3	1-0	False	39.4	46.4	0.5066	31.37	0.3902
4	1-0	False	17.4	0.4	1.0	0.0	0.0
5	1-0	False	18.4	4.0	0.9235	1.7	0.0286

Figure 4.7: Detailed information for each segment

4.4.5 Results

In the Results tab of the GUI interface we can observe a summary of the results (see Figure 4.6).

In the results file we can find detailed information for each segment (see Figure 4.7). The information includes: segment number, segment ID, whether this value is pruned or not, time normalized, number of keys pressed normalized, HBLEU, HEd and HTER.

PosEdiOn is able to generate graphics using the data (as the one shown in Figure 4.8) created from the pruned HTER values. The user can choose which data should be used to generate graphics and the type of graphic in the configuration file.

The results are stored in a tabulated text file, so they can be easily imported into any spreadsheet to perform further calculations.

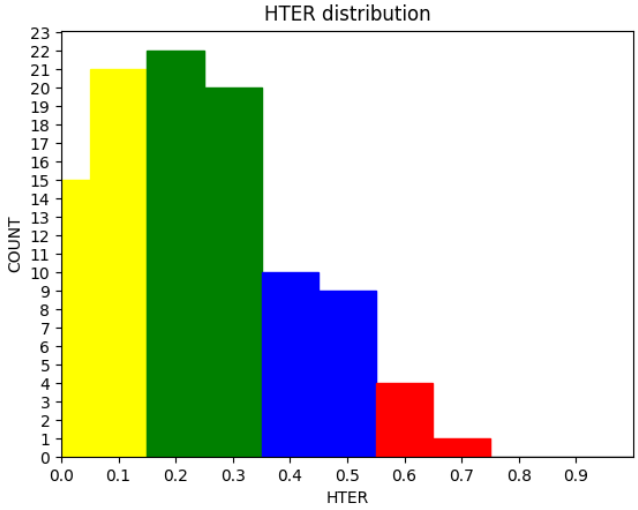


Figure 4.8: Graphic of the pruned HTER distribution

4.4.6 Discussion

PosEdiOn is a tool designed to perform evaluations of post-editing tasks. It also includes its companion program PosEdiOn analyzer, which allows users to easily analyze the data obtained with PosEdiOn. Both programs are released under a free license (GNU GPL v3) and can be freely downloaded from the Github page created for the project.[15]

We have used this tool to study different measures of post-editing effort (see Section 4.5). Both programs are developed in Python3 and they can be easily adapted and improved. As the data are stored as tabbed text files, they can be easily processed or imported into any spreadsheet program to perform further analysis or data visualization.

[15] https://github.com/aoliverg/PosEdiOn

4.5 Experiment 3: Quantitative Analysis of PE effort indicators

We present a quantitative analysis of different PE effort indicators for two NMT systems (transformer and seq2seq) for English-Spanish in-domain medical documents. We compare both systems and study the correlation between PE time and other scores which can be considered proxies for PE effort (as detailed in 4.2.2). Results show less PE effort for the transformer NMT model and a high correlation between PE time and keystrokes.

4.5.1 Methodology

We conducted a preliminary comparative quantitative analysis of different post-editing effort indicators (technical and temporal). We used two NMT systems (transformer and s2s) trained with the corpora described in the following section to translate from English into Spanish three texts (1468, 631 and 2247 words respectively) from the medical domain.

Four professional translators with at least one year of post-editing experience carried out the task: two of them post-edited the s2s output (T1 and T2) and the other two, the transformer output (T3 and T4). They were asked to produce publishable-quality translations. As we wanted to reduce the external variables as much as possible, they all used PosEdiOn (see Section 4.4 for a detailed description of the tool). It logs both post-editing time and edits (keystrokes, insertions and deletions, that is, technical effort). The main characteristics of the post-editing tool were also explained to them before starting the task.

In order to avoid any bias, translators never post-edited the same text twice. However, they were told a NMT system was used to produce the output. They were payed their usual rate and had a two-week deadline. Two of them expressed concerns about the tool, as they preferred to work with their usual tools. However, they didn't think it would affect the final quality of their job or their usual working speed. While post-editing, they

could search for all the needed information in order to produce the final translation. They could also pause the post-editing task whenever they wanted.

4.5.2 MT Systems and Training Corpus

MT Systems

In addition to the two NMT models we also trained a Moses model with the same amount of corpus, and even used Google translate. We assessed the resulting engines with standard automatic metrics (see Table 4.1). The best scores for BLEU (Papineni et al., 2002) were obtained by the Moses engine, even though WER was better for the two NMT systems. This is in line with the results of recent research, which has shown certain automatic metrics tend to underestimate NMT systems (Shterionov et al., 2018; Alvarez et al., 2019).

Additionally, we conducted a manual evaluation of a 30 segment sample for the three MT outputs employing monolingual direct assessment (DA) of translation adequacy (Graham and Liu, 2016; Graham and Baldwin, 2014). We used this DA setup because it simplifies the task of translation assessment (usually done as a bilingual task) into a simpler monolingual assessment task. We obtained the results averaging the assessment of two annotators and the NMT systems received higher marks.

System	BLEU	NIST	WER	DA
Marian S2S	0.3601	7.6142	0.6893	64
Marian Transformer	0.3616	7.3863	0.6334	68
Moses	0.3942	7.8146	0.7386	46
Google Translate	0.3304	7.1197	0.7788	56

Table 4.1: Automatic and DA evaluation figures

As it can be seen in Table 4.1, DA assessment classified Moses as the worst rated. Therefore, we decided to include only the two NMT systems for the post-editing tasks.

For the NMT systems we used Marian[16] (Junczys-Dowmunt et al., 2018). We trained two systems (see Section 3.2.3 for a detailed description of different NMT architectures). For the first one (1) we used an RNN-based encoder-decoder model with attention mechanism (s2s), layer normalization, tied embeddings, deep encoders of depth 4, residual connectors and LSTM cells. For the second one (2), the transformer, we used the configuration in the example of the Marian documentation[17], that is, 6 layer encoder and 6 layer decoder, tied embeddings for source, target and output layer, label smoothing, learn rate warm-up and cool down.

Training Corpus

To train the system, we have partly used the corpora prepared for 3.6 but we have increased the number of tokens adding extra publicly available corpora in the English-Spanish pair:

- Biomedical translation repository (BMTR)[18]
- Medline abstracts training data provided by Biomedical Translation Task 2019[19]
- The UFAL Medical Corpus[20] v1.0.
- The Khresmoi development data[21]
- The IBECS[22] (*Spanish Bibliographical Index in Health Sciences*) corpus.
- The SciELO corpus[23]

[16]https://marian-nmt.github.io
[17]https://github.com/marian-nmt/marian-examples/tree/master/transformer
[18]https://github.com/biomedical-translation-corpora/corpora
[19]http://www.statmt.org/wmt19/biomedical-translation-task.html
[20]https://ufal.mff.cuni.cz/ufal_medical_corpus
[21]https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2122
[22]http://ibecs.isciii.es
[23]https://sites.google.com/view/felipe-soares/datasets

Corpus	Segments/Entries	Tokens eng	Tokens spa
BMTR	816,544	14,726,693	16,836,428
Medline Abstracts	100,797	1,772,461	1,964,860
UFAL	258,701	3,202,162	3,437,936
Kreshmoi	1,500	28,454	32,158
IBECS	72,168	13,575,418	15,014,299
SciELO	741,407	17,464,256	19,305,165
MedLine	140,479	1,649,869	1,846,374
MSD Manuals	241,336	3,719,933	4,467,906
EMEA	366,769	5,327,963	6,008,543
Portal Clinic	8,797	159,717	169,294
Glossary MeSpEn	125,645	-	-
ICD10-en-es	5,202	-	-
SnowMedCT Denom.	887,492	-	-1
SnowMedCT Def.	4,268	177,861	184,574
Total	**4,430,765**	**66,147,518**	**74,663,550**

Table 4.2: Size of the corpora and glossaries used to create the corpus to train the MT systems

- The EMEA[24] (*European Medicines Agency*) corpus.

We have also created several corpora from websites with medical content:

- Medline Plus[25]: we have compiled our own corpus from the web and we have combined this with the corpus compiled in MeSpEn.
- MSD Manuals[26] English-Spanish corpus, compiled for this project under permission of the copyright holders.
- Portal Clínic[27] English-Spanish corpus, compiled by us.

We have also used several glossaries and glossary-like databases treating

[24]http://opus.nlpl.eu/EMEA.php
[25]https://medlineplus.gov/
[26]https://www.msdmanuals.com/
[27]https://portal.hospitalclinic.org

them as corpora. These resources contain a lot of useful terms and expressions in the medical domain. Namely, we have used the English-Spanish glossary from MeSpEn, the 10th revision of the International Statistical Classification of ICD and SnowMedCT.

With all the corpora and glossaries we have created an in-domain training corpus of 4,430,765 segments and entries.

In Table 4.2 the size of all corpora and glossaries used for training the MT systems is shown. Figures are calculated eliminating all the repeated source segment-target segment pairs in the corpora.

4.5.3 Results

PE effort indicators

Once translators finished post-editing, we calculated the following task-specific (PE based) metrics (showed in Table 4.3):

- **PETpT**, PE time in seconds normalised by the length of the target segment in tokens.
- **HTER** (Snover et al., 2006), the TER value comparing the raw MT output with the post-edited segment.
- **HBLEU**, the BLEU (Papineni et al., 2002) score obtained by comparing the raw MT output with the post-edited segment.
- **HED**, an edit distance value (Levenshtein distance) calculated comparing the raw MT output with the post-edited segment.
- **Keystrokes** normalized by the number of tokens.

	T1 (S2S)		T2 (S2S)		T3 (T)		T4 (T)	
	mean	st. dev.	mean	st. dev.	mean	st. dev.	mean	st. dev.
HTER	0.16	0.12	0.11	0.09	0.17	0.17	0.12	0.17
HBLEU	0.53	0.27	0.65	0.27	0.56	0.29	0.67	0.33
HED	1.28	1.19	0.84	0.94	1.56	2.04	1.09	2.07
Keys/tok	6.36	28.25	3.38	5.25	7.53	27.62	5.91	25.59
PETpT	9.19	33.97	4.61	8.56	4.57	12.22	3.03	8.69

Table 4.3: PE-based metrics (mean and standard deviation) for the task

"output" — 2021/5/1 — 14:54 — page 79 — #93

	S2S NMT		Transf. NMT	
	mean	st. dev.	mean	st. dev.
HTER	0.13	0.10	**0.11**	0.09
HBLEU	0.59	0.27	**0.65**	0.27
HED	1.06	1.06	**0.84**	0.94
Keys/tok	4.87	16.75	**3.38**	5.25
PETpT	6.90	21.26	**4.61**	8.56

Table 4.4: Total PE-based metrics for each NMT model

In order to avoid outliers, we didn't include those segments in which (normalized) time or (normalized) keystrokes doubled the mean plus the standard deviation of the total time or number of keystrokes. As it usually happens in this type of tasks, post-editing shows considerable variation among translators. For the seg2seg model, translators showed a difference of 4.58 PETpT between them. This difference was reduced to 1.54 in the case of the transformer model. However, if we check the total figures for each system (see Table 4.4), PE time is clearly reduced for the transformer model, as are all the other scores.

We also used the distribution-agnostic Kolmogorov–Smirnov test to compare the distribution of PETpT for the two translators of each NMT model. We found there was no clear distribution (considering $p < 0.05$). This would seem to indicate the need to increase the number of translators for any given post-editing test to obtain a more representative mean.

Post-editor	**Unmodified seg.**
T1 (S2S)	22
T2 (S2S)	31
T3 (T)	19
T4 (T)	58

Table 4.5: Unmodified segments after post-editing

Another interesting figure to understand PE effort is the number of unmodified segments. Even though that does not mean those segments im-

	T1 (S2S)	T2 (S2S)	T3 (T)	T4 (T)	ALL
HTER	0.309*	0.545*	0.418*	0.00705*	0.49*
HBLEU	-0.072	-0.209	-0.148	-0.370*	-0.21*
HED	0.043*	0.706	0.0770*	0.809*	0.66
Keys	**0.823***	**0.868***	**0.824***	**0.822***	**0.82***

Table 4.6: Spearman's correlation with time as a gold standard for different effort indicators (*p<0.001)

ply no PE effort, it could give an indication of MT output. Table 4.5 shows the number of unmodified segments per translators from a total of 224 segments. There is not a clear tendency for any MT system, but rather a preference corresponding to the individual translator, especially T4, who didn't modify a high number of segments, which correlates to the low PE time recorded.

We also checked PETpT related to segment length, as research has shown longer segments tend to imply higher PE effort (Bentivogli et al., 2016). We studied segments with more than 35 token to see if PETpT or any other PE effort indicator increased. We could find no statistically significant evidence linking segment length to translators' effort in our experiments. This could indicate newer NMT models do not always reduce MT quality in longer segments.

Correlation between scores

Once established the overall results per model, we tried to identify which metric produced scores that were closest to the total time spent per segment. We calculated Spearman's correlation coefficient between the total amount of time and all other metrics.

As it can be seen in Table 4.6, the best overall correlation is found with the number of keys (see Figure 4.9) for all translators and for the total results, followed by the calculated edit distance. Most of the results show a statistically significant correlation, especially those relating to the number

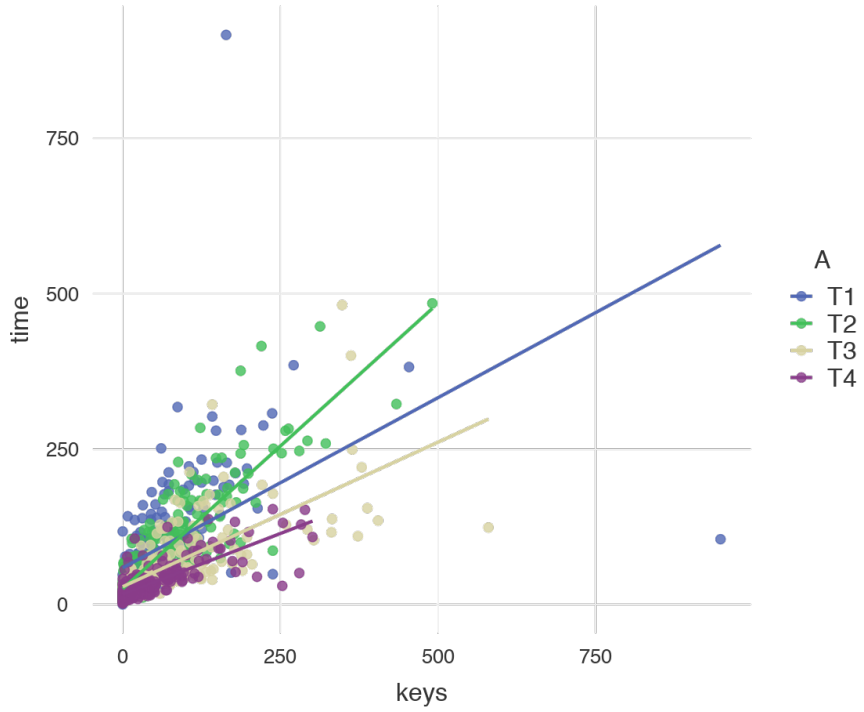


Figure 4.9: Scatter plot of keystrokes and time for all of the translators

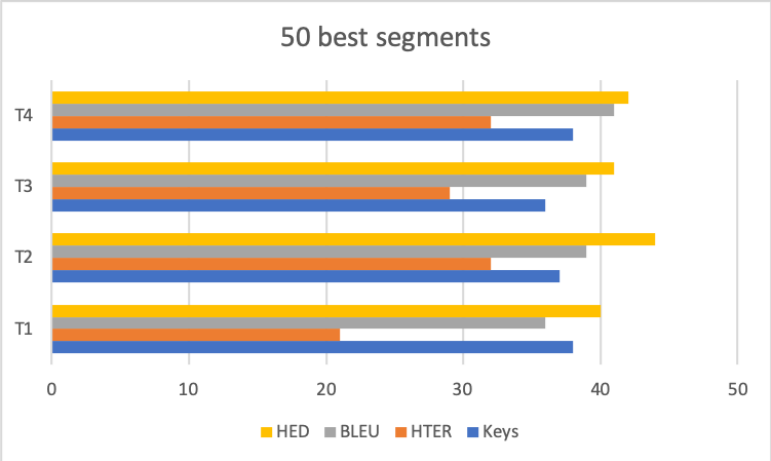


Figure 4.10: Correlation of best segments

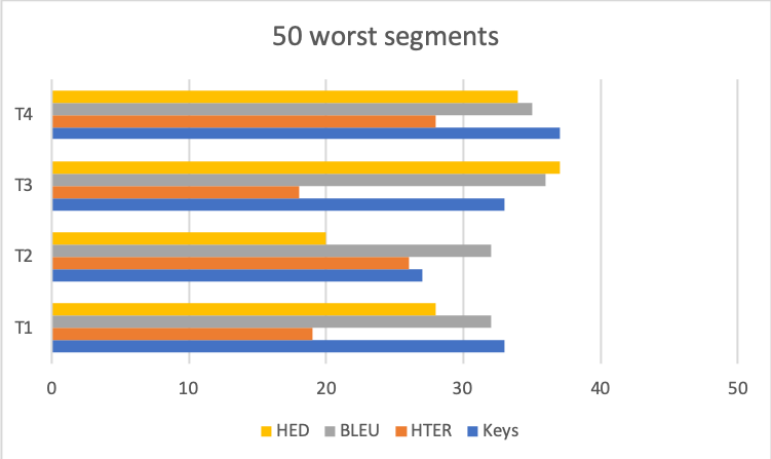


Figure 4.11: Correlation of worst segments

of keystrokes (*p<0.001). This is in line with the results reported by previous work (Scarton et al., 2019; Graham et al., 2016) that found no clear correlation between temporal effort and the most frequent metrics, even though the number of keystrokes was the metric more closely related.

Tails distribution

There was a lack of correlation between the distribution of PE time among translators, and between this indicator and the others. We wanted to take a closer look at the best and worse segments to analyse if the correlation improved. We counted the number of common segments between the 50 best and worst time segments and all other metrics calculated.

As it can be seen in Figures 4.10 and 4.11, there is a better correlation for the segments in which less time was spent post-editing. Furthermore, the edit distance shows the best correlation in these cases. For the segments with the higher time recorded, correlation is notably reduced in all cases and the edit distance and the number of keystrokes show a higher correlation.

4.5.4 Discussion

There is a need of reliable metrics to evaluate MT quality in order to produce outputs which translators can post-edit without too much effort. Our experiments have shown that no single PE indicator can provide all the information necessary to assess the quality of the MT output. PE time provides a measurable useful information, even though it does not always correspond with other PE metrics and includes a great variation among translators.

However, temporal effort presents a strong correlation with keystrokes (technical effort). From the indirect measures of effort, edit distance is the one that shows a better correlation.

In industrial scenarios, the quality of a certain MT output is usually linked to PE time. The results of our experiment suggest the analysis of temporal

effort can indicate the quality of the MT output, but there is a need to include other scores to assess the convenience of post-editing a certain raw MT output.

4.6 Conclusion

In this chapter we have presented the main dimensions of PE effort and the most frequent scores used to measure them. We have also presented a simple tool to measure most frequent direct and indirect indicators of PE effort.

We have used this tool to measure PE effort for medical texts for the English-Spanish combination for NMT models. Results show PE effort is lower for transformer models. Furthermore, when studying the correlation of temporal effort with indirect measures of effort, edit distance is the indicator which shows a better correlation. However, results suggest the need for a more fine-grained analysis of PE modifications to understand how raw MT output affects effort.

Chapter 5

TRANSLATORS' PERCEPTION OF POST-EDITING

5.1 Introduction

In the last decade, the translation market place has suffered important changes that have affected the translation profession. Both the rapid globalization propelled by the neoliberal policies and the recent global economic crisis have increased the effort from LSPs and customers to reduce costs in the translation workflow, which has had an important impact both on translators' revenues and working practices (Moorkens et al., 2018).

Furthermore, the increased technologisation means that many translators need to include different working routines into their professions, which in some cases may limit the scope of their work (Nunes Vieira and Alonso, 2018). PE has been included as part of the translation workflow because it increases productivity, and it is seen as a way to minimize human intervention and thus minimize costs (Guerberof, 2009a; Sosoni and Rogers, 2013).

From a cognitive perspective, according to the Cognitive Load Theory (see Section 4.2.3 for a further information), extraneous factors influence the cognitive processes (Nunes Vieira, 2016). Thus, understanding if there is a negative perception or bias regarding PE can help us better understand subjective factors that may influence PE effort.

This chapter presents a study of translators' perceptions on post-editing. First, we outline the current research on the subject. Then, in Section 5.3 we present a pilot study which tries to analyze the possible negative bias regarding PE and its relation with productivity, both for novel and experienced translators.

5.2 Research on Translators' Perceptions

As we have seen in previous chapters, research on PE has mainly focused on measuring the post-editing effort related to MT output quality (Guerberof, 2009b; Specia, 2010, 2011) and productivity. Less studies focus on the perceptions of professional translators who post-edit.

Currently, very few translators ignore tools such as term banks, translation memory systems and quality checkers in the daily translation tasks. However, studies show that translators still regard the use of technology mainly as a threat (Katan, 2009). LeBlanc (2013) reported on translators' perception of TMs and described the main advantages and disadvantages. Although translators admitted it helped increase productivity and reduced repetitive work, their main concern was that it was a barrier for creativity and made translators increasingly passive and lazy.

Regarding the use of MT, research shows that translators perceive they are less productive post-editing, even when a quantitative analysis shows otherwise (Gaspari and Toral, 2014). They consider MT output to be tedious to post-edit (Moorkens and Brien, 2017) and they prefer to translate from scratch even if this has a negative impact on productivity (Teixeira, 2014).

Guerberof (2013) surveyed the perception of MT post-editing among current post-editors. The majority of the 27 respondents were translators already familiar with post-editing who showed mixed answers. There was not a clear rejection to use MT and they were mainly satisfied with their jobs as post-editors.

However, Läubli and Orrego Carmona (2017) analyzed posts on social media as a way to understand how translators felt about MT. They showed a negative general perception and a disconnection between the research and the translation community.

Cadwell et al. (2016) interviewed translators working at the European Commission's Directorate-General for Translation (DGT) to better understand the factors involved in the translators' adoption and non-adoption of MT during their translation tasks. They had a broadly positive attitude to MT because they believed (a) it increased speed and productivity, (b) the MT output had good quality, (c) it served as inspiration and (d) reduced typing or clicking.

However, the main reasons not to use MT were (a) the perceived poor quality of MT output, (b) the negative influence it had on the translator's abilities, (c) the fear it awakened among them and (d) the extra attention needed from the translators when post-editing.

In a follow-up study with translation companies, Cadwell et al. (2018) highlighted mainly the same concerns, but translators also explained they were worried about the fairness of monetary compensation for the post-editing tasks.

Finally, Nunes Vieira and Alonso (2018) carried out comprehensive interviews with different actors in the translation industry from Great Britain and Spain to enquire about several problematic aspects of MT use in translation processes. The interviews suggested some key problems regarding MT that are not exclusively technological, such as different notions of translation quality, lack of transparency, lack of training and pricing pressures.

5.3 Experiment 4: Perceptions of Post-Editing from Professional Translators: A Case Study

We conducted a pilot study which has two objectives: firstly, to assess the attitude of translators who are post-editing MT output for the first time before and after the task, and to relate it to productivity rates in an industrial scenario. Secondly, to compare the results with the perceptions of current professional post-editors.

To achieve these objectives, we set four professional translators to the task of post-editing and translating from scratch from Spanish into English general domain texts. Translators answered different questions regarding their perceptions and attitude before and after post-editing, and we recorded the number of keystrokes and the time spent while performing the tasks. Additionally, 50 participants answered an online survey we conducted addressed to professional post-editors regarding their perception of the job.

5.3.1 Methodology

The aim of our pilot study was to analyze translators' perception when post-editing for the first time and compare the results to their productivity measured as technical and temporal effort, which are correlated with cognitive effort (Moorkens et al., 2015). To this end, we worked with an LSP called Incyta [1]. Four experienced translators without previous post-editing experience who usually collaborated with the LSP participated in the experiment translating from Spanish into English.

First of all, translators answered a short questionnaire before post-editing to assess their attitudes towards the task. Then, we selected three general domain documents with similar terminology and type-token ratio. They translated from scratch a document of 2437 words. Then, they post-edited two documents of 2189 and 1920 words, respectively, that had been trans-

[1] https://www.incyta.com/

lated using DeepL[2]. The texts were provided by the LSP and included enough segments to assess both perception and productivity from translators. In both tasks, they were asked to produce printable-quality translations.

Instead of trying to reproduce the working conditions of translators, which could vary greatly among individuals, we used PET (Aziz et al., 2012) (which has already been described and used in Section 3.6 and Section 3.7). Although PET could slow down at first translators' productivity, it leveled all variables related to their working environment. Moreover, a week before beginning the task, we sent them the tool and delivered detailed information on its use, together with a short text they could use to test it. We used this testing period to answer questions and solve usability issues so that all translators could have a fair knowledge of the tool before beginning the task.

Once the task was finished, translators answered another questionnaire with questions about their perception of the task and, in some cases, they were asked to answer follow-up questions by mail to clarify some of the answers.

Additionally, we prepared a larger survey targeting experienced post-editors to study their perception of the task. We compared their answers with the opinions expressed by translators in the case study described beforehand who post-edited for the first time. For the survey, we used the web-based SurveyMonkey[3] platform because it allows to create online surveys that can be easily distributed and also allows to analyze and summarize all data collected in different formats.

Even though surveys have been used in previous research to assess post-editors' perceptions, our questions on the current survey focused on how experience and training could affect the perception of the task, and what elements could help improve this perception.

[2]https://www.deepl.com/translator
[3]https://www.surveymonkey.com/

We published a 33-question survey targeted exclusively to translators who had already worked as post-editors. We tested the wording of the questions in a pilot study with two professional translators to ensure that there was no ambiguity, so that all answers could provide enough data for a complete analysis. Once the survey was ready, we published it on Linkedin and sent it to three translation associations. 50 participants answered the survey.

5.3.2 Post-editing Task

Language Service Provider

Incyta is a Barcelona-based Language Service Provider (LSP) founded in 1993. It is also the provider of Lucy Software, a commercial rule-based machine translation engine. It has previous experience in post-editing and it is currently working with the Spanish-Catalan, Spanish-Galician and Spanish-Portuguese language combinations, mainly to translate news on a daily basis for different newspapers. In order to post-edit these language combinations, they use their own commercial MT software.

The company was interested in introducing PE for some new language combinations because there had been an increase in the demand. It was planning to implement PE for Spanish into English for some news workflows in 2021.

After a quality assessment it had conducted internally, it had decided to use NMT for this new language combination. The translators with whom the company usually worked were quite skeptical about the quality of the MT output and the economic repercussions it would have in their earnings and were not willing to begin post-editing.

And on the other hand, the LSP wanted to continue the current collaboration with these translators due to the good quality of their work. With this goal in mind, the LSP wanted to give translators the opportunity to post-edit at their full rate and collect their opinions regarding quality and productivity. For post-editing tasks, the company envisaged to pay trans-

lators 70% of their current rate based on their previous experience working with closely-related languages, although they were paid their regular rate for this experiment.

Perceptions before post-editing

We sent the four translators taking part in the experiment a questionnaire before they began the task so that we could know their current use of technology and MT and their attitude towards post-editing.

These were the questions:

Q1. How many years of experience as a professional translator do you have?
Q2. How long have you worked with this LSP?
Q3. Have you worked before correcting translation outputs?
Q4. Do you use computer-assisted translation (CAT) tools?
Q5. Are you a regular consumer of machine translation?
Q6. How do you feel about post-editing machine translation output?
Q7. What are your expectations regarding post-editing?
Q8. What do you think the MT quality is going to be?
Q9. What do you think your global experience post-editing is going to be?
Q10. To which of these tasks do you think it will be similar? (Possible answers: A. Reviewing human translations; B. Translating with fuzzy matches; C. Translating from scratch).

The four professional translators (T1, T2, T3 and T4) who carried out the task had extensive experience translating (12, 16, 15 and 18 years, respectively) (Q1) and had worked more than three years with this LSP translating from Spanish into English (Q2). T1 and T2 had never used CAT tools before, while T3 and T4 used them only sometimes for certain specific projects (Q4).

We also asked them about their previous experience correcting human outputs, as research has shown translators often relate it to post-editing

(Guerberof, 2013). T1 and T4 had done corrections of human translations before, but T2 and T3 did only translations (Q3). Although none of them were regular consumers of MT in their daily lives, they believed the general quality of MT had improved considerably in recent years (Q5).

Regarding their attitude towards post-editing (Q6), none of the translators were looking forward to it. They recognised MT was "getting better, but still cannot compare to a (decent) human translation" (T2). T4 showed concerns MT will take over the industry and T3 thought it would take as much time as translating from scratch and "can influence my own translation."

When asked about their expectations (Q7), T1 thought it would be like reviewing translations by non-natives, where "sentence structure and context acquire special importance". T2 thought it was not going to be an enjoyable job and she also had "ethical conflicts with my profession disappearing and only becoming post-editing, which is more poorly paid and frankly less fun and creative". Only T3 highlighted the quality of the MT output as a key factor. If the quality is good, it will be a "positive experience".

Q8 and Q9 offered the participants a Likert scale where 1 was "Very bad" and 5 was "Excellent". The majority of the translators thought the quality of the MT output was going to be very good (4), but their experience post-editing was only going to be good (3).

In the last question (Q10), from the three options offered as answers, T1 and T3 thought post-editing would be similar to reviewing human translations and T2 and T4 believed it would be like translating with fuzzy matches. None of them believed post-editing would be similar to translating from scratch.

Although their general attitudes regarding post-editing were mainly negative, they did not think the post-editing experience was going to be bad nor did they perceive the use of a translation tool as negative. In fact, some of the fears they expressed were more related to rates and other market practices.

Perceptions after Post-editing

Once they had finished post-editing and translating from scratch, we sent them another questionnaire to collect information on their perceptions, mainly to understand what the main difficulties had been and if their opinions had changed after carrying out the task.

We asked them the following questions:

Q1. Grade the global post-editing experience
Q2. Would you be willing to post-edit on a regular basis?
Q3. What did you like best about post-editing?
Q4. What did you like less about post-editing?
Q5. Do you think following some training would improve your productivity post-editing?
Q6. Do you think having more information about the MT engine would improve your productivity post-editing?
Q7. What is your assessment of the MT quality?
Q8. What were the main errors it produced?
Q9. Do you think post-editing is similar to revising human translations?
Q10. Did you find some errors difficult to spot?
Q11. Do you think you had a higher productivity than translating from scratch?
Q.12 Do you think the final translation had the same quality?
Q13. Are you as satisfied with the result as if it had been translated from scratch?

For the first questions (Q1), translators were offered again a Likert scale where 1 was "Very bad" and 5 was "Excellent". T1 and T3 thought the experience had been good (3), T2 thought it had been bad (2) and T4 qualified it as excellent (5). Except T2, who was quite disappointed with the experience, the other three translators would be willing to post-edit in a regular basis (Q2), but only if the "rate was right". T1 stressed "there is nothing enjoyable (to me) about post-editing, whereas translating is enjoyable," even though they recognised the improved quality of the MT output.

As positive elements (Q3) they thought post-editing saved them time typing and they did not have to correct basic mistakes. On the downsides (Q4), they thought it "constrained creativity when reformulating sentences" and it was "total roteness" because of the lack of creativity. These opinions coincide mostly with the ones expressed in previous studies regarding the adoption of MT by translators (Cadwell et al., 2016).

Three of the translators believed that a proper training would improve their performance post-editing (Q5) and two of them thought it would be positive to have information on the MT engine (Q6). Regarding the MT quality (Q7), the mean rating was 3.75 out of 5, even though they found some important errors (Q8) while post-editing. T1 and T3 highlighted the high number of grammatical errors, while T2 thought the main errors were "too-literal translation of the sentences" and missing nuances.

T4 also detected some inconsistencies (e.g. *pliego* had been translated both as *document* and *specifications*; *sobre* had been translated as *about* and *envelope*) and some words that had been badly translated (e.g. *unión temporal* had been translated as *temporary union* instead of *joint venture*; *garantía* had been translated as *security* instead of *bid bond*).

All four translators agreed that post-editing was not similar to revising human translations (Q9) because the errors were of "a different nature". They also explained (Q10) that some errors were hard to spot because there was a lack of uniformity.

There were not usual errors, such as "typos and spelling mistakes and I had to pay special attention to the actual translation". However, recent research has shown that for specialized texts MT output and texts translated from scratch present similar errors (Fischer and Läubli, 2020).

All four translators agreed that productivity was higher when post-editing (Q11) but this was not the only important factor. They stated a lower degree of satisfaction ("I become a 5th-grade teacher correcting essays, and that is not the profession I signed up for!"). They also agreed their final product was of similar quality as if they had translated it from scratch (Q12).

	T1	T2	T3	T4	Mean
From scratch	935.51	1994.36	486.32	560.59	994.19
PE	1246.92	3209.27	880.27	753.62	1522.54

Table 5.1: Temporal effort in words per hour

	T1	T2	T3	T4	Mean
From scratch	4.47	6.83	24.47	17.06	13.21
PE	3.38	1.24	4.66	5.70	3.74

Table 5.2: Technical effort in characters per word

They were proud of the results (Q13) after all the corrections had been introduced. In fact, they all agreed that the post-edited translations they produced were as good as they would have been if they had translated them from scratch.

In general, their experience was better than they had expected. They found the MT quality to be good enough, although mistakes were sometimes difficult to spot.

Productivity Results

We analyzed the technical and temporal effort collected during the translation with PET to calculate the productivity differences between post-editing and translating from scratch. As it can be seen in Table 5.1, although there is a great variability among translators, the mean shows there is an increase of 53.14% in productivity in words per hour if we compare the translation from scratch and the post-editing task.

This increase ranges from 33.29% in the case of T1 to a 81.01% in the case of T3. If we consider exclusively the productivity figures, the rate reduction of 30% suggested by the LSP could be considered in line with these results.

Regarding the technical effort, calculated in keystrokes per word, Table 5.2 shows there is a reduction of 71.69%. T1 shows the lowest reduction with 24.38% while T2 shows the highest decrease with 81.84%. As post-editors have to correct the MT output instead of typing the whole translation from scratch, there is much less typing involved.

5.3.3 Survey for Post-editors

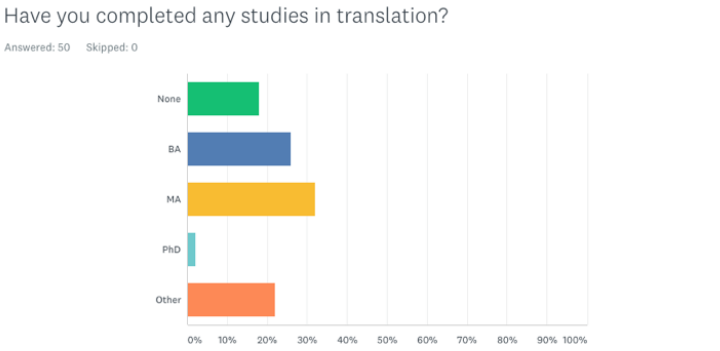
In our case study, translators who post-edited for the first time showed in general a negative perception of the task even though there was an increase in their productivity. We prepared a larger survey to ask translators with experience in post-editing what their opinions were.

We wanted to know if the knowledge and expertise gained through training and experience had affected their current post-editing practices and also what their general working conditions were in relation to rates, professional satisfaction and their working environment. Even though only 50 post-editors participated in the survey, the answers can be used to obtain a fair picture of the current perception of post-editors regarding their job.

In the first question (Q1), we asked them to introduce a user ID in order to identify them. In the second question (Q2), we asked participants about their working language pairs. Most of them worked with European languages such as English, German, Spanish, Italian and Portuguese, which are common language combinations in MT engines.

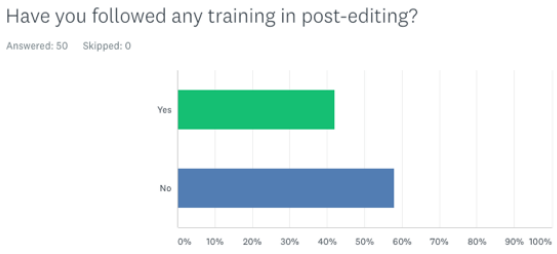
Then, we asked them if they had followed studies in translation (Q3) (see Figure 5.1) and if they had any training in post-editing (Q4) (see Figure 5.2). The majority of the participants had completed translation studies at university (60%). However, only some of them had followed some training or instructions on post-editing (42%).

To all of those who had received some sort of training, we asked them to state which one and give their opinion about the quality of the training (Q5). Most of them explained they were only given instructions about the



ANSWER CHOICES	RESPONSES	
None	18.00%	9
BA	26.00%	13
MA	32.00%	16
PhD	2.00%	1
Other	22.00%	11
TOTAL		50

Figure 5.1: Answers regarding studies in translation (Q3)



ANSWER CHOICES	RESPONSES	
Yes	42.00%	21
No	58.00%	29
TOTAL		50

Figure 5.2: Answers regarding training in post-editing (Q4)

post-editing process and the final quality required (52%).

Some of them received training by the LSP they were working for (17%) when they first began post-editing, and the remaining translators (31%) followed a more formal course, such as the ones developed by SDL or TAUS. All participants who followed these courses found them helpful. All respondents who had not followed any training agreed it would have been a great help when they began post-editing (Q6).

Currently many university translation programs have acknowledged the need to go beyond the teaching of translation memories (TM) in technology modules and to include post-editing courses across the curriculum because this task requires a specific set of skills (O'Brien, 2002; Kenny and Doherty, 2014; Mellinger, 2017), which can be grouped in three main competences: core, linguistic and instrumental (Torrejón and Rico, 2013).

Some authors have also highlighted the importance of using tailored post-editing guidelines that express without ambiguity the goals of the task to be performed (Flanagan and Christensen, 2014; Hu and Cadwell, 2016).

Then, we asked about the amount of experience they had translating (Q7) and post-editing (Q8). As we can see in Figures 5.3 and 5.4, 43 participants (86%) had more than three years of experience translating (86%).

However, only 17 respondents (34%) had a similar amount of experience post-editing, while the majority of them (50%) had been post-editing for only between one and three years. This could be in part due to the recent increase in the demand of post-editing in the market (Lommel and Depalma, 2016), as it reduces costs and increases productivity (Plitt and Masselot, 2010).

Even so, when inquired about the percentage that post-editing represented in their whole workload (Q9), for most of them (70%) it was less than 20%. And only 44% of the respondents stated their post-editing workload had increased (a 20% on average) in the last few years (Q10).

We also asked about the type of texts they post-edited (Q11) (see Figure 5.5), which were usually technical or medical, domains in which post-

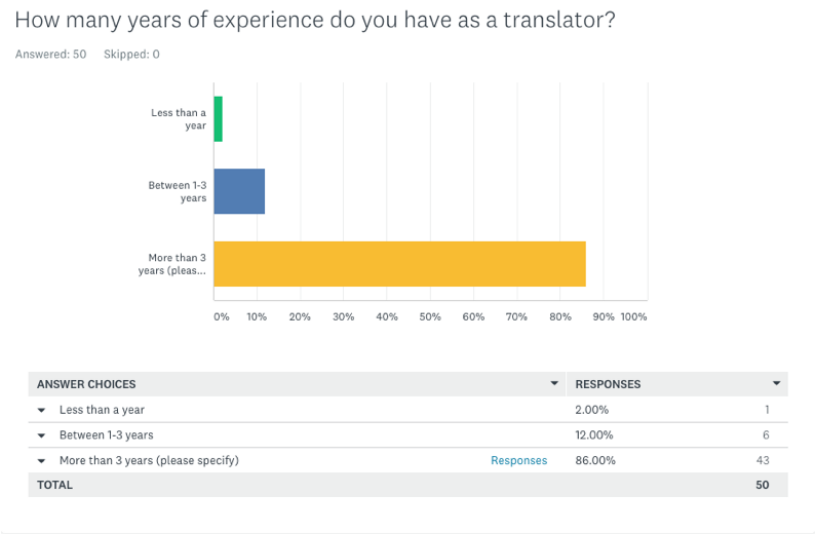


Figure 5.3: Answers regarding their experience as translators (Q7)

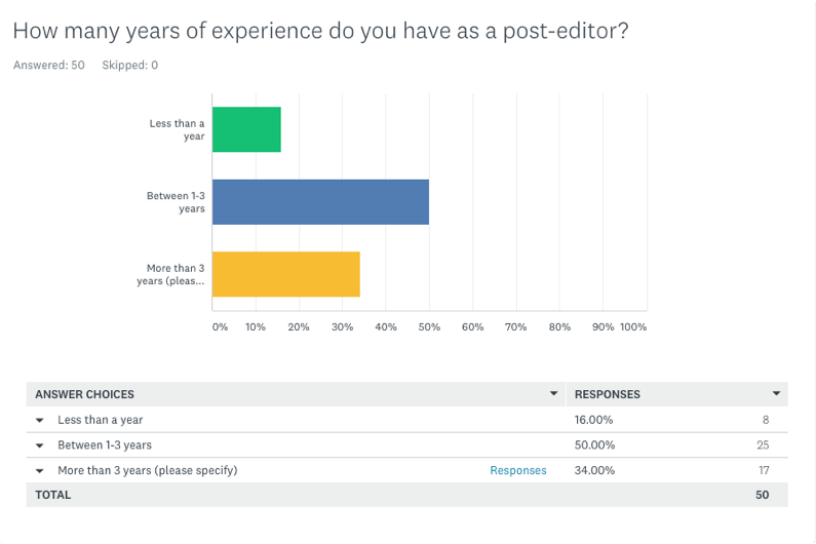


Figure 5.4: Answers regarding their experience as post-editors (Q8)

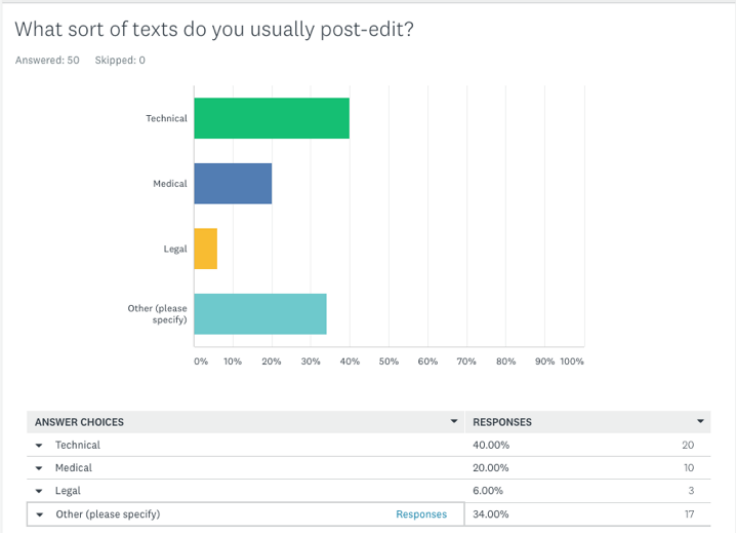


Figure 5.5: Types of documents they post-edited (Q11)

editing has traditionally achieved better results (Aymerich, 2005; Kirchhoff et al., 2011). Very few translators (5%) post-edited general domain documents.

When asked to select among different statements which described with more accuracy their progress they had experienced post-editing along time, 70% of the participants agreed that experience had led them to better detect MT errors (Q12) (see Figure 5.6) and post-edit faster (Q13) (see Figure 5.7), which is in line with the results obtained in previous studies relating experience with higher efficiency (Moorkens and O’Brien, 2015), even though some participants stressed the great variation of errors found in the MT outputs.

However, regarding the effort it entailed in relation to translating (Q14), 50% considered it required more effort than translating using translation memories and 22% believed it was a task that entailed more effort than revising human translations.

ANSWER CHOICES		RESPONSES	
I can spot MT errors better now.		70.00%	35
I feel that there has been no change in my capacity to spot MT errors with experience.		22.00%	11
I can spot less MT errors now because I'm used to them.		4.00%	2
I don't know (describe your impressions)	Responses	4.00%	2
TOTAL			50

Figure 5.6: Multiple-choice question regarding MT error detection (Q12)

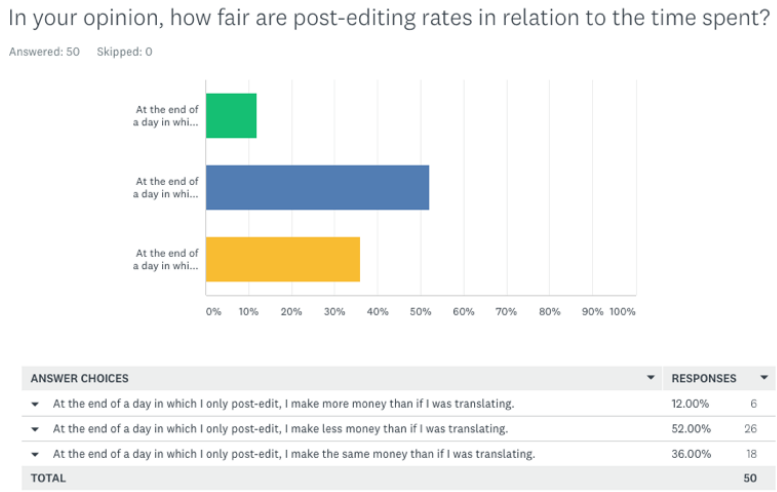
ANSWER CHOICES		RESPONSES	
Experience has helped me post-edit faster.		70.00%	35
I feel that there has been no change in my post-editing speed with experience.		14.00%	7
I post-edit slower now.		2.00%	1
I don't know (describe your impressions)	Responses	14.00%	7
TOTAL			50

Figure 5.7: Multiple-choice question regarding post-editing experience (Q13)

In order to obtain a better picture of their current working situation, we asked about post-editing rates. As we can see in Figure 5.8, participants in the survey mostly believed that at the end of a day in which they only post-edited, they earned less money (52%) or the same amount of money (36%) than if they had been translating (Q15). This fact can be linked to the effort both LSPs and customers have done to reduce costs since the 2008 crisis, which has negatively affected translating rates (Moorkens, 2017).

Regarding the quality of the MT they had to post-edit (Q16), most of them thought it was acceptable though it needed many editions (44%), or they even believed that it was of borderline quality (22%). Another 22% of the respondents even erased the whole MT output and translated from scratch in certain segments throughout the task although they were only paid for post-editing. It is well known that the quality of the MT output is a key element in post-editing as it affects the productivity gain (Garcia, 2011).

The majority of participants clearly stated that post-editing rates were not



ANSWER CHOICES	RESPONSES	
At the end of a day in which I only post-edit, I make more money than if I was translating.	12.00%	6
At the end of a day in which I only post-edit, I make less money than if I was translating.	52.00%	26
At the end of a day in which I only post-edit, I make the same money than if I was translating.	36.00%	18
TOTAL		50

Figure 5.8: Correlation between post-editing rates and time spent (Q15)

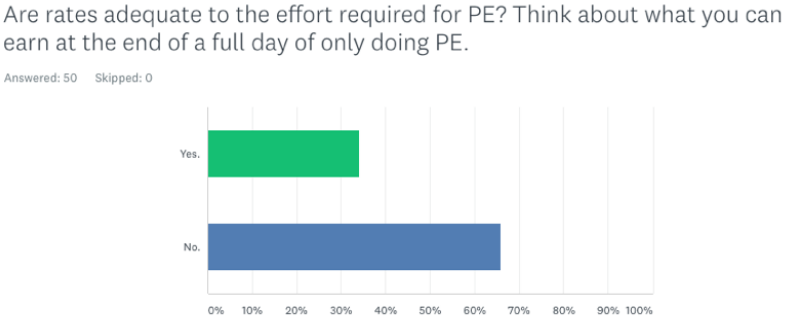


Figure 5.9: Correlation of effort and rates (Q17)

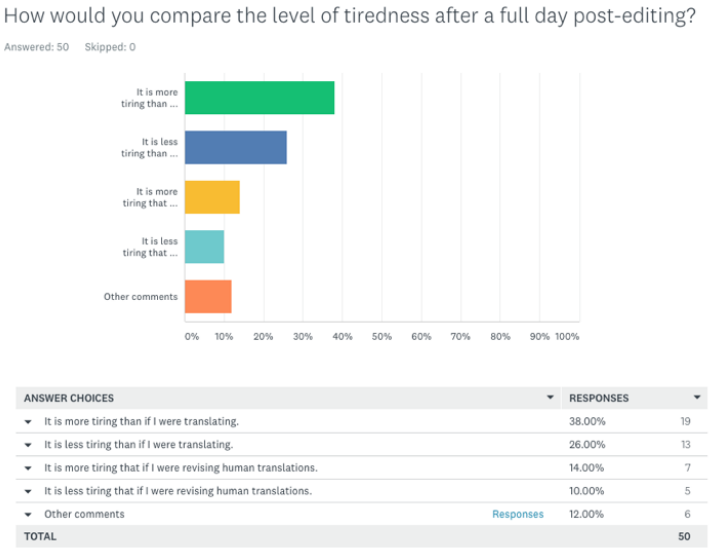


Figure 5.10: Level of tiredness implied in post-editing (Q18)

currently adequate to the effort it entailed (66%) (Q17) (see Figure 5.9) because it was more tiring than translating (38.78%) and than revising human translations (14.29%) (Q18) (see Figure 5.10). Regarding the estimation method to calculate the rate they were paid (Q19), nearly half of them preferred being paid by word (44%) although 26% felt comfortable with both payment methods.

Regarding the tools they used, all translators who answered the survey used the same tool to translate (Q20) and to post-edit (Q21). They mainly used SDL Trados Studio (55%), followed by MemoQ (16%) (Q22). As it can be seen in Figure 5.11, most of them explained the tool they were currently using was the best suited for post-editing (74%) (Q23).

However, Moorkens and Brien (2017) concluded after an extensive survey that currently post-editing was not well-supported by existing tools and there was a need to study new specifications for user interfaces (UIs) that better supported the post-editing task.

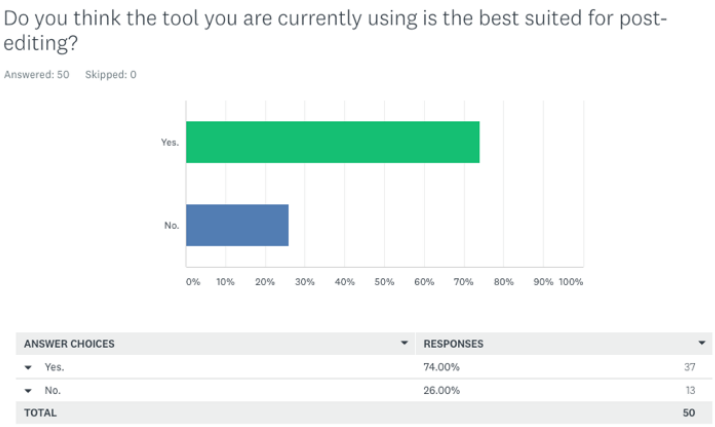


Figure 5.11: Suitability of current tool for post-editing (Q23)

To get the participants insight, we asked them if they would like to add any additional functionality for post-editing to the tool they were currently using (Q24). 20% of the respondents suggested that the propagation of post-editor’s corrections would be very useful and could save the more repetitive and edit-intensive tasks, although three post-editors stressed the fact that errors are not always the same, especially when translating lexical elements. Another post-editor thought it would be useful to include measurements of post-editing effort while translating, instead of having to wait until the post-editing had been finished. This could be useful to give the post-editors some insight regarding their progress while post-editing.

When asked about any additional element they would incorporate to the UI (Q25), participants mainly suggested adding more shortcuts for tag insertions, providing automatic corrections and including tools to help rearrange words in a sentence.

Another important element when post-editing is the MT system used to produce the output. Post-editors who answered the survey did not usually (38%) or never (26%) receive any information regarding the MT system (Q26). However, this could be a useful piece of information as recent re-

search has shown that different MT models produce different types of errors (Klubicka et al., 2017). In fact, most respondents (67%) believed that if translators were trained on understanding how MT works they would feel more confident post-editing (Q27).

In the following questions, we enquired participants to rate their satisfaction level with the translation tasks (Q28) and post-editing tasks (Q29) giving a mark from 0 ("Very bad") to 100 ("Excellent"). For translation, the mean rating was 83, while post-editing obtained a 56.

Additionally, we asked them to explain the main reasons for the previous rating of translating (Q30) and post-editing (Q31). They thought translating boosted creativity and gave translators the chance to work with different text types. However, respondents showed more concerns about post-editing. They found it was more boring and repetitive. They believed that having to correct computer-generated errors tended to be tedious, as they usually "have to correct as little as possible to be profitable, so we do not aim for the best quality". One respondent even suggested that "the hardest is to remember what is genuine in the language."

Finally, we asked them if they thought their voices were heard in industrial workflows (Q32) and what suggestions would they make to improve this workflow (Q33). Except in two cases, they all thought their voices were never listened to, which correlates to previous surveys analyzing the current working conditions of professional translators (Nunes Vieira and Alonso, 2018).

They mainly stressed the importance of increasing post-editing rates, which were considered low, and improving the quality of the MT output. Post-editors also made reference to the tight deadlines in the current translation marketplace and the possibility of correcting formal or repetitive mistakes in the text before post-editing began.

Post-editors found the task was somewhat repetitive but their main claims were related to negative working conditions and not the task itself. However, they highlighted the improvements related to proper training and experience.

5.3.4 Discussion

Translators participating in the case study who post-edited for the first time showed prejudices and a general negative attitude before post-editing. In part it was due to the specific characteristics of the task but also because of other external elements such as rates and the future of the translation profession.

However, once they had finished post-editing, their opinions were not so negative and most of them would be willing to post-edit on a regular basis even though they all enjoyed translating more. The main challenges of post-editing were related to the constraints it imposes, mainly to creativity. Another important problem was the unpredictable errors in the MT output, which were sometimes difficult to spot. Regarding productivity, post-editing reduced in half the time spent by word.

Experienced post-editors also considered this task to be more repetitive, more tiring and less paid than translating from scratch. However, they highlighted post-editing productivity increases with experience and proper training. Moreover, post-editing reduces typing, which usually helps to increase productivity. In general, translators are less satisfied with post-editing than with translation from scratch.

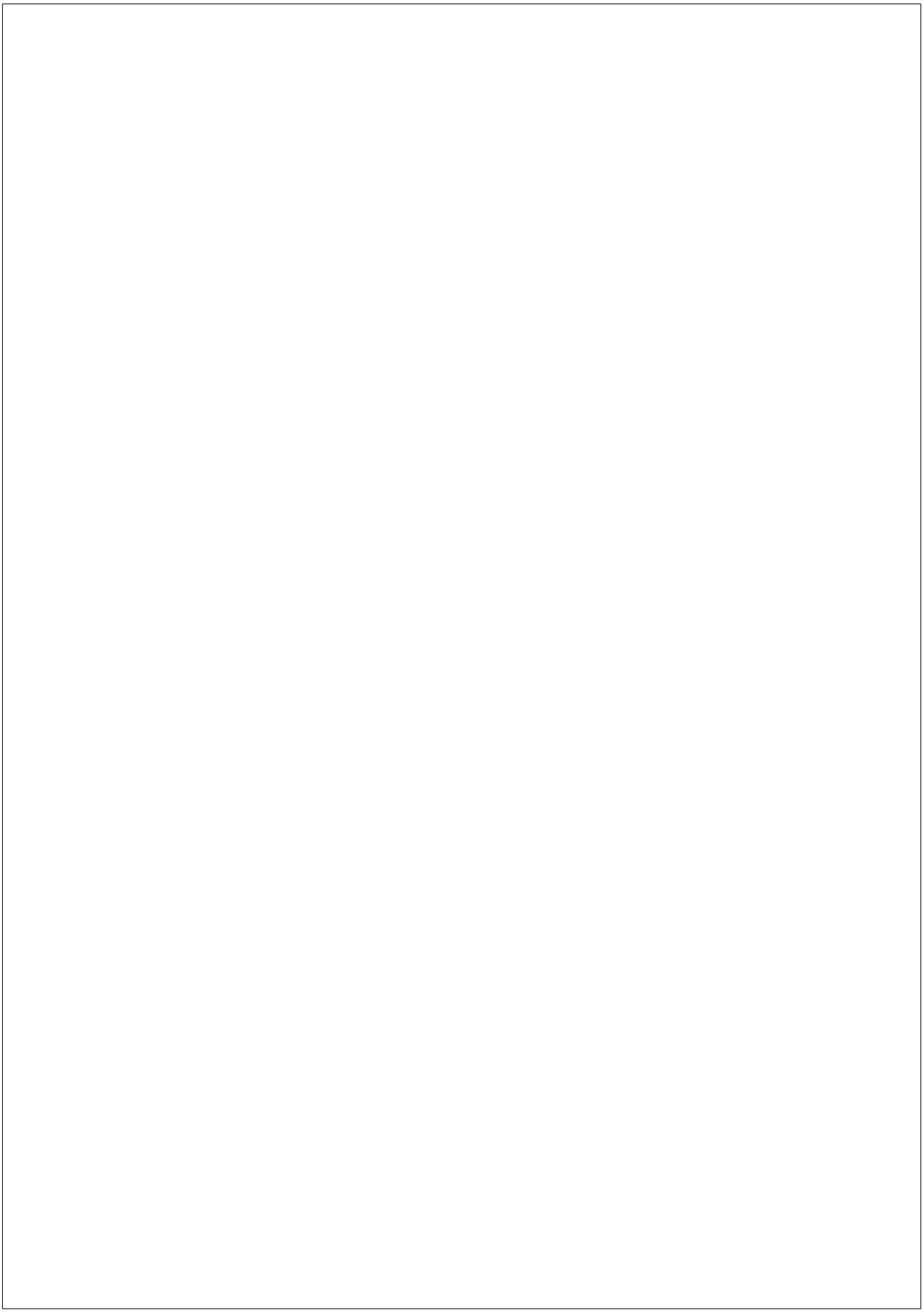
According to the opinions of the participants, training and experience are key elements to post-editing with more confidence. There is also a need for a more fluent communication throughout the translation workflow, mainly to clarify aspects such as the final quality demanded and the origin of the MT output.

As Nunes Vieira and Alonso (2018) point out, one of the main problems for translators is the lack of communication from project managers regarding what they are asked to do and how to do it. The focus of the translation workflow should be the human translator and MT should be used and perceived as a helpful tool that they can use to improve a human-centered process.

5.4 Conclusion

In this chapter we have studied how post-editing is perceived by translators. There is a generalized bias against this task, due in part to extraneous factors, such as rates and a lack of communication in the translation workflow. But also due to intrinsic factors, that is, to the quality of MT output.

Hence, an evaluation of the raw output that takes into account frequent problems and linguistic errors could be useful to improve the post-editing process.



Chapter 6

EVALUATION BASED ON LINGUISTIC ERRORS

6.1 Introduction

In previous chapters, we have described the evaluation of MT output and its correlation with PE effort. When assessing the usefulness of MT models for PE, it is also essential to analyze the most frequent errors and how they affect the task. Although recent research (Bentivogli et al., 2018; Castilho et al., 2017c; Klubicka et al., 2017) and our previous experiments (see Section 3.6 and 3.7) show NMT reduces the number of errors compared to SMT, each error type affects differently the PE effort (Daems et al., 2017b; Koponen, 2012).

In this chapter, we present error classification and the main error taxonomies used to conduct this task. We also introduce challenge sets as an evaluation method based on linguistic errors. In Section 6.5 we explain a fine-grained analysis we conducted to annotate MT errors based on post-edited corrections. In Section 6.6 we suggest a methodology that combines the creation of error-based challenge sets with indirect measures of effort to assess raw MT outputs.

6.2 Error Classification

A widely used method for human evaluation is error classification, ideally accompanied by error analysis. The goal is to identify and classify the errors in a translation to provide a better understanding of the MT output. It has been often used in the past for system development and improvement.

It can also be used to compare two different systems by finding the error distribution in different outputs. In the case of PE, it can be used to assess the quality of the MT output, it can serve as a tool to detect patterns of correction in the PE process, or it can be linked to particular preferences of post-editors/revisers (Popović, 2018).

Some automatic or semiautomatic tools have been developed to conduct this task. Addicter (Zeman et al., 2011) is a tool for the automatic detection and display of common translation errors which uses a first-order Markov model for aligning reference words with hypothesis words. Hjerson (Popovic, 2011) uses WER alignments and compares the sets of words identified as erroneous due to a mismatch with the reference.

Both tools classify errors into different categories similar to the ones suggested by Vilar et al. (2006) (see Section 6.3). There has even been a proposal to merge both tools into a pipeline (Berka et al., 2012). These automatic tools can also be used as a first classification step before a manual annotation (Vardaro et al., 2019).

However, error classification is usually conducted manually because currently available tools are still not able to distinguish detailed error classes, and are prone to confusions between mistranslations, omissions and additions. Manual annotation can provide a fine-grained analysis of the errors produced.

This task is usually developed by annotators who identify the errors in the MT output with or without a reference translation. One of the main problems of manual annotation is the low inter-annotator agreement (IAA), in part due to the different understanding of quality problems among annotators (Lommel et al., 2014).

With the widespread use of PE in the translation workflow, the analysis of PE corrections is receiving more and more attention. It is usually conducted as a separate task from PE, even though these two tasks are highly related.

For any error annotation task, a clearly detailed taxonomy should be defined beforehand. The errors should be grouped into significant categories which are relevant for the task as well as for the language combination involved. Error categories should cover all possible linguistic problems found in the MT output and should be sufficiently detailed. In the next section, we present some of the most usual taxonomies.

6.3 Error Taxonomies

Flanagan (1994) proposed one of the first classification systems for errors (see Table 6.1 for a detailed description of all the categories) in MT output designed for use by potential MT users, rather than MT developers. The error categories were designed according to criteria which are important to the user, such as improvability and intelligibility. The author modified certain categories depending on the language combination.

Elliott et al. (2004) devised an adaptable categorization scheme for the French-English language combination. It focused on fluency errors, but also enabled the detection of adequacy errors, without access to the source text. It stemmed from the need to identify error types in MT output to guide automated evaluation.

Vilar et al. (2006) suggested one of the first error classifications focused

Category	Description
Spelling	Misspelled word
Not Found Word	Word not in dictionary
Accent	Incorrect accent
Capitalization	Incorrect upper or lower case
Elision	Illegal elision or elision not made
Verb inflection	Incorrectly formed verb, or wrong tense
Noun inflection	Incorrectly formed noun
Other inflection	Incorrectly formed adjective or adverb
Rearrangement	Sentence elements ordered incorrectly
Category	Category error (e.g. noun vs. verb)
Pronoun	Wrong, absent or unneeded pronoun
Article	Absent or unneeded article
Preposition	Incorrect, absent or unneeded preposition
Negative	Negative particles not properly placed or absent
Conjunction	Failure to reconstruct parallel constituents after conjunction, or failure to identify boundaries of conjoined units
Agreement	Incorrect agreement between subject-verb, noun-adjective, past participle agreement with preceding direct object, etc.
Clause boundary	Failure to identify clause boundary, or clause boundary unnecessarily added
Word Selection	Word selection error (single word)
Expression	Incorrect translation of multi-word expression

Table 6.1: English to French error categories suggested by Flanagan (1994).

on explicit error categories and analysis. In order to identify the problems produced by SMT systems, different language combinations were analyzed (Chinese-to-English, Spanish-to-English and English-to-Spanish). The classification proposed had a hierarchical structure and it was based on the error typology used for refinement of rule-based systems (Carbonell and Lavie, 2005). It grouped errors into five big categories: missing words, word order, incorrect words, unknown words, and punctuation.

Farrús et al. (2010) designed an error taxonomy with five broad categories for SMT outputs from the Catalan-Spanish language combination. They correlated the different categories with human evaluations and noticed that semantic errors influenced the most in the perception of quality. This same scheme was used to develop an automatic linguistic-based evaluation metric (Comelles et al., 2012).

Federico et al. (2014) used a similar taxonomy focused on detecting MT errors for translations from English into Arabic, Chinese and Russian. The final goal was to study the impact of different error types on the overall quality score using mixed-effect models.

Kirchhoff et al. (2012) presented another detailed taxonomy used to annotate English to Spanish translations. After applying a conjoint analysis to study relations, results show the most annoying errors are word order and word sense errors.

Stymne and Ahrenberg (2012) used another hierarchical error scheme and was the first work researching inter-annotator agreement. They studied how guidelines could affect agreement among different annotators. Results showed a 25% agreement without guidelines, which increased to 40% when guidelines were delivered. The difference was also relevant for simpler taxonomies. In this case, there was an agreement of 65% without guidelines versus an agreement of 80% with guidelines.

Costa et al. (2015) reported an error taxonomy tailored for Romance languages. In their study, highly ranked sentences clearly showed low number of grammatical errors, and a high inter-annotator agreement between two annotators was reported.

Translation industry has also developed error taxonomies which have been included in quality metrics. Many companies use for their evaluations error-based models that seek to "identify errors, classify them, allocate them to a severity level and apply penalty points with a view to deciding whether or not the translation meets a specific pass mark." (O'Brien, 2011a, p. 58)

The LISA QA metric [1] was initially designed to promote the best translation and localization methods for the software and hardware industries. Although it is no longer in use, its methods are still used in translation quality evaluation. This metric includes three severity levels, but there is no weighting. It consists of a set of 20, 25 or 123 error categories, depending on how they are counted.

The SAE J2450 metric originated in the automotive industry. It includes seven primary error categories which cover such areas as terminology, meaning, structure, spelling, punctuation, completeness, etc. and two severity levels. In contrast to LISA, it focuses on linguistic quality and includes no formatting or style issues. It also includes two meta-rules to help evaluators make a decision in case of ambiguity.

The TAUS Dynamic Quality Framework (DQF) [2] uses different tools, which include an error taxonomy, for the evaluation of translation quality. It was recently harmonized with the Multidimensional Quality Metrics (MQM) (Lommel et al., 2014) to offer translation professionals and researchers a unified model.

The Multidimensional Quality Metrics (MQM) was developed as part of the QTLaunchPad [3] project (funded by the European Union) to address the shortcomings of previous quality evaluation systems. This framework offers a flexible system for annotating errors and provides a list of error types that can be correlated to specific errors in the MT output (see Figure

[1] http://producthelp.sdl.com/SDL_TMS_2011/en/Creating_and_Maintaining_Organizations/Managing_QA_Models/LISA_QA_Model.htm

[2] https://www.taus.net/data-for-ai/dqf

[3] http://www.qt21.eu/launchpad/

6.1 for MQM core categories).

It contains 114 issue types derived from a careful study of existing translation quality evaluation metrics and represents a generalized superset of the issues that can be found in current metrics and tools. These types are organized in a hierarchy, and range from broad to detailed, allowing you to obtain as much or as little detail as you need. They cover translation issues (such as mistranslations), design, and other problems with the target language to suitability of the translated text for the target market.

MQM is based on translation specifications (based on the ASTM F2575 for describing translation project parameters) to define translation requirements. It has become a very popular framework both for the translation industry and research and, in fact, it was conceived as an update of the aforementioned LISA QA metric, which was widely used in the localization industry.

It is multi-dimensional and allows users to select issues that measure translation quality in multiple dimensions: Accuracy, Fluency, Design (layout and formatting), Verity (a new way of dealing with the suitability of text for the target locale and audience), and Internationalization (issues related to whether the source content was properly prepared for translation); and assess the quality for each dimension (see Section 6.5 for a detailed explanation of the MQM metrics used in our experiment).

It has been used in recent research comparing different MT engines. Klubicka et al. (2017); Klubička et al. (2018) compared the errors produced by an English-Croatian pure phrase-based, factored phrase-based and NMT system performing a manual evaluation via error annotation of the systems' outputs. Two annotators used a metric compliant with MQM (multidimensional quality metrics) and results showed that NMT reduced considerably the number of errors.

Ye and Toral (2020) also conducted a fine-grained human evaluation to compare the transformer model and recurrent approaches to neural MT for the English-Chinese combination. They followed a tailored MQM taxonomy and observed the transformer produced an overall better trans-

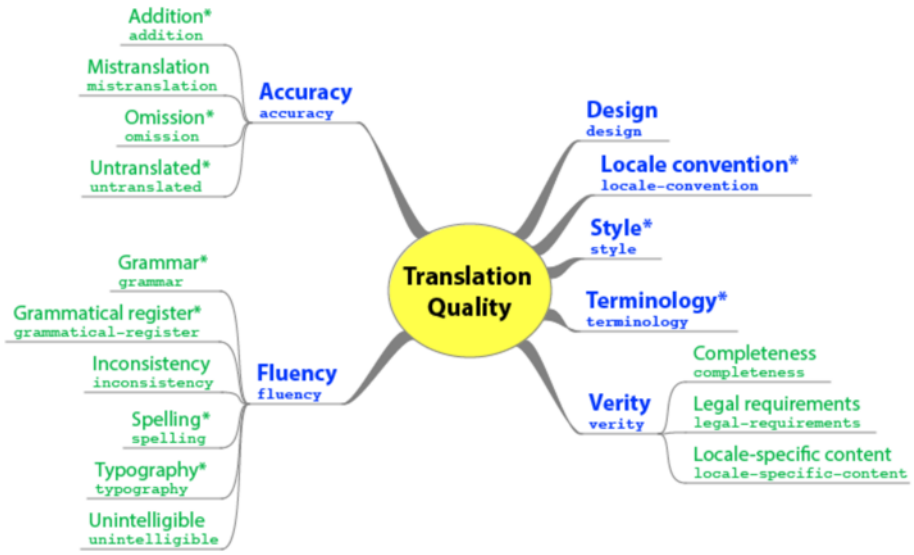


Figure 6.1: Graphical representation of the MQM core error categories (Lommel et al., 2015)

lation reducing the number of errors related to accuracy, fluency and comprehensibility.

6.4 Challenge Sets

Benchmark datasets are usually drawn from text corpora and used in NLP to evaluate system performance. However, although it may reflect the natural distribution of phenomena in language, some of these datasets may not capture a wide range of phenomena.

Challenge sets or test suites are an alternative evaluation system which targets specific problems and has been long used in MT evaluation (King and Falkedal, 1990; Isahara, 2006; Koh et al., 2001; Arnold et al., 1993). It could be defined as a ”representative set of isolated or in-context sentences, each hand or (semi)automatically designed in order to evaluate a

system's capacity to translate a specific linguistic phenomenon" (Popović and Castilho, 2019). Lehmann et al. (1996) noted several key properties of challenge sets: systematicity, control over data, inclusion of negative data and exhaustivity.

After some decades where there was a rise of large-scale quantitative evaluation methods, challenge sets have regained certain popularity. A good example of this is the "Additional Test Suites" in the framework of the WMT translation shared task, which was first included in 2018 [4].

Many of the challenge sets developed include the properties explained by Lehmann et al. (1996), although ill-formed or negative examples are currently not used so much. According to Belinkov and Glass (2019), challenge sets can be categorized using the following criteria:

- the task they seek to evaluate
- the linguistic phenomena they aim to study
- the language(s) they target
- the size
- the method of construction
- the evaluation method

Challenge sets are usually designed to evaluate one or more linguistic phenomena. First sets were focused on exhaustivity (Lehmann et al., 1996), but currently they study specific properties of interest. The size of the challenge sets also vary a lot depending on the construction process. Manually created sets are smaller and automatically built datasets range from several thousands to close to a hundred thousand (Linzen et al., 2016).

Popović and Castilho (2019) conducted a Workshop focused on challenge sets within the Machine Translation Summit 2019. They suggested three types of sets.

[4] https://www.statmt.org/wmt18/translation-task.html

The first group concentrates on specific phenomena. Some of them have focused on grammatical issues. Sennrich (2017) suggested the evaluation of NMT by using contrastive translation pairs. This method introduces a specific type of error automatically in reference sentences. Then it checks if the conditional probability model of the NMT system prefers the original reference or the modified version. The author determined a character-based model was able to improve generalization on unseen words, but at the same time introduced new grammatical errors. Cinková and Bojar (2018) used a test suite to study the grammatical contrasts generated between Czech and English.

Others have concentrated on morphological divergences. For example, Burlot and Yvon (2017) studied morphology by including 14 morphological properties. Or have studied discourse phenomena (Bawden et al., 2018; Voita et al., 2019).

The second group studies the ambiguity or variations presented by specific elements. It can include the study of pronouns (Guillou and Hardmeier, 2016; Guillou et al., 2018; Müller et al., 2018), lexical ambiguity of nouns (Rios Gonzales et al., 2017; Rios et al., 2018; Raganato et al., 2019), ambiguous conjunctions (Popovic and Castilho, 2019; Popović, 2019) or gender bias (Stanovsky et al., 2019).

The third group includes a large taxonomy with different categories. Isabelle et al. (2017) assessed the performance of NMT systems compared to PBSMT. It showed NMT worked better in all cases even when there was a small difference in BLEU scores. They manually created an English to French challenge set of difficult examples using expert linguistic knowledge. It contained 108 sentences and was crafted using areas where the source and target differ, focusing on morpho-syntactic, lexico-syntactic, and syntactic divergences.

The authors pointed out there could be errors elsewhere in the test sentence, but each test sentence focused on one specific phenomenon. Moreover, they kept the test short to avoid problems which could arise from the interaction of different linguistic phenomena. They used three bilingual

native speakers as annotators to make binary judgements of the issues tested.

Burchardt et al. (2017) developed a manually built test suite that included a large range of linguistic phenomena for the English-German language pair. The outputs from different MT systems were compared to a post-edited version of the output and were manually evaluated a professional linguist.

Avramidis et al. (2019) developed a challenge set based on the previous paper focusing on 107 phenomena organized in 14 linguistic categories:

- Ambiguity
- composition
- Coordination and Ellipsis
- False Friends
- Function Word
- LDD and Interrogatives
- Multiword Expressions
- Named Entity and Terminology
- Negation
- Non-verbal agreement
- Punctuation
- Subordination
- Verb/Tense/Aspect/Mood
- Verb valency

They compiled or produced manually a total of 5,560 sentences, with 20 to 180 instances for every phenomenon. For the evaluation process, they combined regular expressions that could be automatically validated. If

an output did not match a positive or a negative regular expression, the system issued a warning which needed to be manually validated. This was an iterative process devised for a constant update of the set.

Results showed that BLEU scores obtained for the different systems seemed correlated with the macro-average accuracy obtained for all systems except one.

6.5 Experiment 5: A Fine-grained Analysis of SMT and NMT Errors

Error annotation has been used to study the quality of the MT products (Vilar et al., 2006; Costa et al., 2015) and to investigate whether an MT output is fit for post-editing (Denkowski and Lavie, 2012). Recetly, it has also been used to better understand the difference between SMT and NMT errors (Klubicka et al., 2017).

We present a pilot study of a fine-grained analysis of MT errors based on post-editors corrections for an English to Spanish medical text translated with SMT and NMT. We will compare the errors found on these two different MT outputs and compare the versions produced by the different translators, mainly to study the segments where there is a greater divergence among translators.

Our goal is twofold: study the type of errors post-edited for SMT and NMT for this type of text and language combination and analyse the differences among translators post-editing the same MT output using Choice Network Analysis (CNA).

6.5.1 Methodology

To conduct the error analysis, we used the data collected for the experiment explained in Section 3.6, which we briefly summarize in the following lines. Four professional translators specialized in the medical domain

and with more than 3 years of professional experience post-edited 41 segments from a 2018 medical paper to produce a publishable-quality version using a computer-assisted translation tool. Two of them post-edited the SMT output and the other two the NMT output.

Following Popović and Arčan (2016), for the error annotation we considered post-editing as an implicit error annotation (Popović and Arčan, 2016). However, we did not assess the correction of post-edited modifications (Koponen et al., 2019), but considered all corrections as errors.

One annotator with previous experience in marking MT errors manually annotated the four post-edited versions using the MQM (Lommel et al., 2014) taxonomy (see Section 6.5.2 for further details on the MQM version used) because it is a popular framework both in research and the translation industry.

We also included a different weight for every error according to its severity in line with MQM instructions. And following Klubička et al. (2018), we counted the number of words corresponding to each error.

On the other hand, we studied the 10 segments of each version in which there were more differences among translators to see if there were certain error types which were more cognitively demanding following the Choice Network Analysis (CNA) (Campbell, 1999) (see Section 4.2.3 for a detailed explanation of CNA and how it can be linked to cognitive effort).

6.5.2 MQM Adaptation

MQM framework offers the possibility of describing and defining custom translation quality metrics. Its goal is to provide a flexible vocabulary of quality issue types and a way to use these elements in order to generate quality scores. Instead of imposing a unique metric for all situations, it provides a detailed catalog of different quality issue types, including standardized names and definitions, that can be used to describe particular metrics for specific tasks.

The hierarchical structure groups errors into different major issues (such

as Fluency and Accuracy) which can be further specified into detailed error types. This enables different levels of granularity, from a coarse analysis to a fine-grained metric, and also facilitates the customization of the framework for different language combinations.

For example, if the analysis focuses on grammar errors, this category can be further specified to include a detailed error description for all the MT output issues encountered. It also includes a guide for the annotators using the MQ framework, and a decision tree designed to standardize the categorization process.

For our analysis, we used four main categories: Accuracy, Fluency, Style, and Terminology:

Terminology includes the specific terms related to the domain of the specialized text analyzed, in this case medicine. Even though in some cases it can coincide with a mistranslation or an omission (which would be part of the Accuracy category), in this category we only included errors which were clearly related to terminological problems from the medical domain.

Style groups all modifications introduced by the post-editor which can be considered unnecessary or stylistic. It includes all preferential choices of the different translators when post-editing but also modifications which, even though helped to better understand the text, cannot be considered an error.

Accuracy groups errors which entail adding or removing some part of the source text information. These errors are usually the ones with the biggest impact on the MT output as they usually create critical problems in meaning (see Figure 6.2).

Fluency includes errors which have an impact on the quality of the target text, for example, all grammar mistakes produced by the MT system (see Figure 6.3). We have further detailed this category to specify the corresponding type of errors. Apart from punctuation, capital letters and spelling, we have grouped errors mainly taking into account the grammatical category of the error detected. Furthermore, we have included word

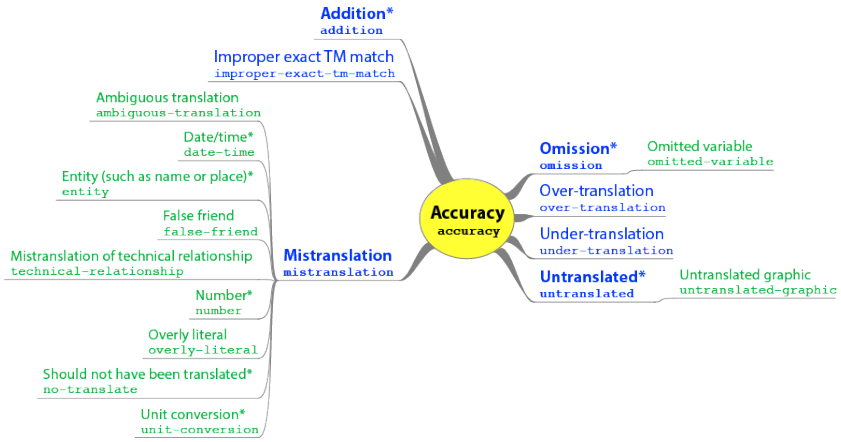


Figure 6.2: Detail of the Accuracy error category (Lommel et al., 2015)

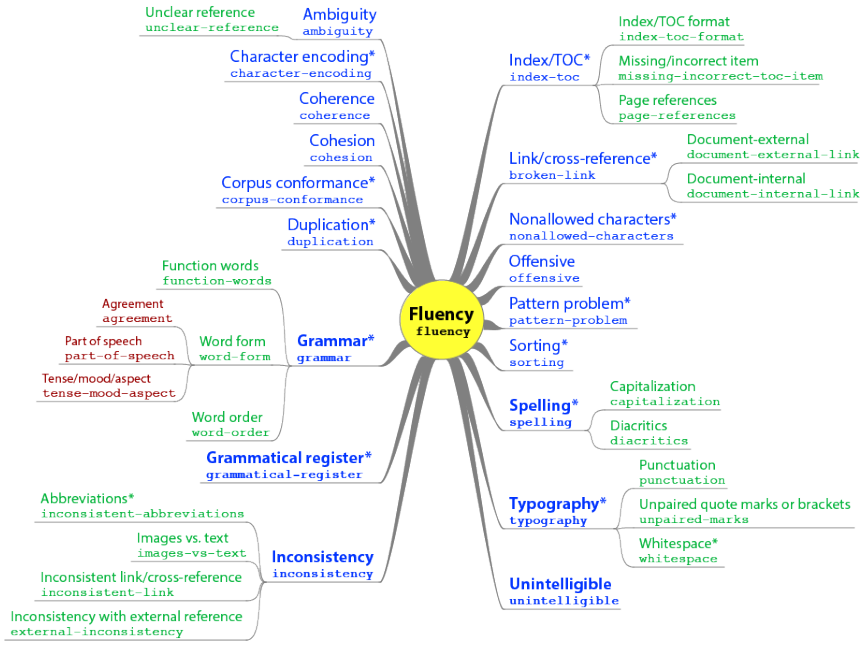


Figure 6.3: Detail of the Fluency error category (Lommel et al., 2015)

order (which also includes the modification of the syntactical order of the sentence) and what we have called co-reference.

This category usually includes references within the same sentence or the previous sentence which the MT system has repeated, but that should have been omitted or mentioned with another sort of reference. That is, taking into account the context, there is a redundant translation. Although it could be argued this is more of a stylistic problem, we believe it generates a fluency problem in the MT output.

For instance, in the following segment "pacientes" was removed the second time it appears, as in Spanish lexical repetitions should be avoided within the same sentence if possible.

Source: Sixty-nine patients had local recurrence and 17 patients showed [...].
MT output: Sesenta y nueve pacientes presentaron recaída local y 17 pacientes presentaron [...]
PE version: Sesenta y nueve pacientes (80%) presentaron recaída local y 17 presentaron [...]

6.5.3 Results

Once the annotation process was completed, we calculated the number of corrections per each category and the mean for each MT system. As it can be seen in Table 6.2, there is a great divergence between the translators who post-edited the SMT output. In fact, PE1 introduced very few modifications. The results of the translators post-editing the NMT version are more alike, although PE4 detected many more terminology errors.

The mean of all results shows that less errors were corrected in the NMT output, although the difference is not statistically significant. The most relevant divergence in errors corresponds to accuracy, where NMT presented no untranslated elements from the source and reduced in more than half the mistranslations. In the following sentences we can see examples of the untranslated elements in the SMT compared with the NMT output:

Error Type	SMT PE1	SMT PE2	SMT MEAN	NMT PE 3	NMT PE4	NMT MEAN
Accuracy	**46**	**72**	**59**	**30**	**41**	**35.5**
Mistranslation	24	34	29	18	19	18.5
Omission	6	7	6.5	13	8	10.5
Addition	6	11	8.5	2	14	8
Untranslated	10	20	15			
Fluency	**34**	**74**	**54**	**52**	**55**	**53.5**
Punctuation	5	5	5	6	7	6.5
Verb	4	8	6	4	8	6
Word order	5	6	5.5	12	7	9.5
Prepositions	5	19	12	6	14	10
Capital letters	1		0.5	3		1.5
Concordance	6	10	8	5	2	3.5
Possessive		1	0.5			
Articles	5	15	10	7	15	10.5
Spelling				1	1	1
Connectors	2	2	2	1		0.5
Pronouns		3	1.5	2	1	1.5
Co-references	1	5	3	5		2.5
Style	**3**	**39**	**21**	**24**	**19**	**21.5**
Terminology	**11**	**19**	**15**	**41**	**20**	**30.5**
TOTALS	**94**	**204**	**149**	**147**	**135**	**141**

Table 6.2: Number of errors post-edited by each translator.

Source: [...] and the overall long-term survival rate is [...].
SMT: [...] y la supervivencia global es [...].
NMT: [...] y la supervivencia global a largo plazo es [...].

Source: Study Population.
SMT: Población.
NMT: Población del estudio.

As it was a medical text, a considerable number of errors were produced by the use of the wrong terminology as it has been described in previous research (Hayakawa and Arase, 2020). However, even though the two MT models were trained with the same data, translators corrected more terminology issues in the NMT version.

If we remove from the total results the errors attributed to style, which in most cases correspond to an elective correction introduced into the MT output, results also show NMT output produced less errors (128 errors for SMT versus 119.5 for NMT), with a correlation similar as the one obtained in the total results.

We also included a weight for each of the errors annotated according to the severity of the error. We used the four categories included in MQM and the definitions suggested by O'Brien (2011a):

- **Neutral**: Corresponds to stylistic corrections which do not really imply an error and it also includes fluency corrections that do not have a negative impact on the MT output.
- **Minor**: Noticeable errors that do not have a negative impact on meaning and are not confusing or misleading.
- **Major**: Errors that are considered to have a negative impact on meaning.
- **Critical**: Errors which have major effects on the overall meaning, and can compromise product usability, and consumer safety and health.

As we can see in Table 6.3, critical errors were clearly reduced in the

MT system and post-editor	Neutral	Minor	Major	Critical
SMT PE1	10	42	31	11
SMT PE2	34	105	51	13
NMT PE3	22	87	33	5
NMT PE4	19	70	40	6

Table 6.3: Severity of the annotated errors post-edited by each translator.

MT system and post-editor	Error ratio
SMT PE1	16.9%
SMT PE2	34.3%
NMT PE3	25.8%
NMT PE4	20.4%

Table 6.4: Error ratio for each post-edited version.

two NMT post-edited versions, which seems to indicate that NMT was able to convey better the meaning of the source text. These results can be directly linked to the accuracy errors detected in both systems, in which NMT showed a better performance in reproducing the whole meaning of the source segment into the target.

Finally, we counted the number of words corresponding to each error corrected to calculate the error ratio (Klubička et al., 2018). For each version we divided the number of words that contain an error by the total number of words included in the final post-edited version:

Error ratio = Words with errors / Total number of words

As we can see in Table 6.4, the percentage of errors is consistent with the global number of errors annotated in each post-edited version. Even though there is a big variability among the SMT versions, the mean of the corrections introduced by the two post-editors (51.2%) is slightly higher

Most frequent errors	N. of repetitions
word order	6
terminology	5
co-reference	5
prepositions	5

Table 6.5: Most frequent errors in segments with higher variability

than the mean corresponding to the translators who post-edited the NMT output (46.2%).

After annotating all four post-edited versions and counting the number of errors per each category, we analyzed the variability among the post-edited versions. To do so, we selected the ten segments from each pair of versions where there was a greater divergence in the number of errors.

Then, we studied the errors annotated in those segments to see if there were errors which could be associated with the greater variability. According to the Choice Network Analysis (CNA) (Campbell 2000), errors for which translators' versions diverge the most are an indication of higher cognitive effort. Table 6.5 includes the most frequent errors and the number of times they appeared.

The variability in the use of terminology can be related to the complexity of the domain (medical texts). However, the other three error categories would suggest errors in which there is less agreement among translators and thus imply higher cognitive effort. As it can be seen in the following example, one of the translators modified the sentence order suggested in the SMT output and the other translator kept it unmodified:

Source: Distant metastasis was defined as the detection of malignant tissue in metastatic lesions [...].
SMT: Se definió metástasis a distancia como la detección de tejido maligno en lesiones metastásicas [...]
PE1: La metástasis a distancia se definió como la detección de tejido maligno en lesiones metastásicas [...]

PE2: Se definió a la metástasis a distancia como la detección de tejido maligno en lesiones metastásicas [...]

6.5.4 Discussion

PE is a practice that will increase in the near future and it is necessary to understand translators' corrections in order to ensure an adequate post-editing process. Error analysis will be a useful tool to achieve it. In our analysis for an English to Spanish medical text, the NMT reduced slightly the number of errors, especially the critical ones related to omissions or mistranslations from the source text. However, there was not a significant reduction in the other types of errors found.

Regarding variability, word order, co-reference and prepositions errors appeared with more frequency in the segments in which there was more variability among translators. This would suggest these errors entail a higher cognitive effort.

6.6 Experiment 6:Methodology for an Error-based Evaluation of MT output

The rapid increase in quality led by NMT systems has put additional pressure on traditional automatic metrics like BLEU. As NMT continues to improve, these metrics will lose part of their effectiveness. Moreover, one of the disadvantages of NMT is its opacity, as it is often difficult to understand the motivation of certain errors produced by theses systems.

As we have seen in the previous experiment, a fine-grained analysis of MT errors can be useful to understand what is corrected during the post-editing process. Hence, we suggest an error-based approach to create a methodology that is useful to evaluate the linguistic errors present in a raw MT output.

It includes the use of a challenge set to assess how an MT engine solves

specific linguistic issues known to be problematic. Challenge sets or test suites are an evaluation method which targets specific problems and has been long used in MT evaluation. It is a set of sentences designed to evaluate the capability of an MT system to translate specific linguistic phenomena (see Section 6.4 for further information).

It can be used to assess the convenience of using a specific MT engine for post-editing or to compare two different MT outputs to check which is more suitable for further post-editing. Even though it can be applied to any number of linguistic combinations, we describe in detail the methodology to create a pilot challenge set for English to Spanish. Once created, we test it on three general domain publicly-available engines: Google Translate[5], Bing[6] and DeepL[7].

6.6.1 Methodology

Following Isabelle et al. (2017), we used a manually created challenge set based on the most frequent errors in the English-Spanish language combination, which have been shown in Section 6.5. Even though it mainly includes errors related to specific linguistic problems, we believe it could provide insightful information regarding the linguistic quality of the raw MT output.

For the creation of the challenge set, we followed three steps: compilation of sentences including potential errors, selection of sentences according to the error frequency previously annotated, and fine-tuning or modification of the sentences included in the challenge set.

First of all, we created a corpus with raw MT outputs and post-edited versions of different general domain texts (basically humanistic). The corpus contained 122.323 tokens. Then we selected only those segments with the higher edit distance. This way, we could focus exclusively on those

[5]https://translate.google.es
[6]https://www.bing.com/translator
[7]https://www.deepl.com/translator

segments with the highest concentration of errors to locate the targeted errors.

Then, we selected 40 sentences which included the group of errors annotated in Section 6.5. It is important to know which errors are more frequent for the specific language combination we are working with in order to produce a useful challenge set. Some sentences were repeated because each time a different error was targeted. We kept a similar proportion of errors as the one stated in the aforementioned section for the NMT models but we increased the presence of error categories which also implied a great cognitive effort according to the Choice Network Analysis (CNA) we conducted after the error annotation.

Finally, we revised the original sentences and modified some of them to increase the level of difficulty.

The error categories included in the challenge set are the following (see Chapter 8 for a complete list of all the sentences included):

- **Word order**: Many correct solutions are acceptable in the target language. However, in the examples chosen, we focus on sentences in which the position of the different elements in the sentence is essential for conveying the meaning properly.
- **Prepositions**: As they do not contain semantic information, they do not modify the accuracy of the translation, but are an indication of the linguistic precision of the translations proposed.
- **Verb**: Tenses and passive voice are a problematic issue when translating, especially when talking in the past.
- **Concordance**: It generates more problems when the subject is far away from the verb.
- **Co-reference**: It refers to elements which have been mentioned before in the sentence and would need to be changed into pronouns or referenced without constantly repeating them.
- **Mistranslations**: It includes elements which have been wrongly

translated and generate a major o critical issue for the accuracy of the translation. They are mainly polysemic nouns which have different translations depending on the context of the sentence.

- **Nouns**: We have included compound nouns to see how the system solves complex noun formation.
- **Possessives**: In English they are used more frequently than in Spanish and the MT outputs tend to reproduce the source language use.
- **Punctuation**: Even though it is not one of the most frequents problems, we have included one sentence with punctuation problems in our challenge set.

6.6.2 Results

We produced a challenge set which includes the main linguistic errors for the English-Spanish language combination, which had been previously annotated. The methodology described enables the creation of error-based sets which can be easily used to evaluate the main known issues of specific language combinations.

The challenge set does not generate a global evaluation of the MT output. But focusing on how the MT engine solves complex linguistic problems is a clear indication of the general quality of the raw MT output.

To test the challenge set, we translated the sentences it contains with three online MT engines: Google Translate, DeepL and Bing. For each sentence included in the set, we checked only if the error type was solved correctly and did not take into account how the rest of the sentence had been translated. This made it fast and easy to revise.

In table 6.6, we can see the number of errors annotated for each MT engine when translating the challenge set. DeepL produced a much higher score and translated correctly nearly half of the sentences from the set, while Google Translate and Bing produced worse results.

MT engine	Correct translations	Errors
Google Translate	12	28
DeepL	**19**	21
Bing	10	30

Table 6.6: Number of correct and incorrect sentences for the different MT engines

Hence, we could say that for this language combination DeepL solves much better the linguistic problems posed in the challenge set.

6.6.3 Discussion

Even though the challenge set we used had only 40 sentences, it included some of the most frequent errors produced by MT engines for the English-Spanish combination. The use of a challenge set to evaluate raw MT output is an easy and direct way which can give a fair idea of the linguistic quality of the output, even though it is not an exhaustive evaluation of all elements included in the output.

If we are working with in-domain documents, we could include an additional set to evaluate the terminology adequacy or the phrasing usual for that domain. The challenge set can also be increased with further examples and some of the errors could be automatized using regular expressions.

Furthermore, the results can be also triangulated with some of the indirect measures studied in Section 4.5, which could be easily obtained with PosEdiOn (see Section 4.4 for further details).

6.7 Conclusion

Error classification and analysis has been long used as a tool to better understand MT output. We applied the MQM taxonomy to study post-

editors corrections from English into Spanish. It showed NTM reduced the number of errors, especially the critical ones.

We used the information gathered from error annotation to generate a challenge set focused on the most frequent errors for this language combination. Instead of evaluating the raw MT output with n-gram-based metrics, we suggest a methodology to build challenge sets focusing on complex linguistic issues for a specific language combination. Challenge sets are easy and quick to build and can be tailored to specific domains and needs.

Studying how an MT engine solves complex linguistic issues can give a fair idea of the quality of the MT output, for example, to assess the convenience of post-editing.

Chapter 7

CONCLUSION

7.1 Final Remarks

Currently, raw MT output is increasingly used as part of post-editing workflows. However, if the quality of MT is not good enough, post-editors tend to erase the MT output and translate all the segment from scratch. Hence, it is essential to understand how raw MT output can be assessed for an adequate post-editing.

In this thesis, we studied the different approaches used to evaluate MT and calculate the PE effort translators need to post-edit a certain MT output. Automatic metrics are often used as a quick way to obtain quality scores of raw MT outputs. However, determining the quality necessary to post-edit implies considering all the dimensions included in post-editing, both internal and extraneous.

There is currently a technological shift in all industrial workflows from SMT to NMT. This new technology improves MT quality in most scenarios, but it is necessary to study its flaws and limitations for post-editing. To do so, we evaluated the most frequent manual and automatic metrics both for SMT and NMT.

We realized current automatic scores give a general idea of the quality of the MT output for post-editing, but are unable to deliver a fine-grained assessment, and in some cases contradictions between metrics can be found. In Section 3.7 we showed the SMT model we trained obtained better automatic scores, even though both the indicators of manual evaluation and PE effort obtained better punctuation for NMT.

Even though the quality of NMT has improved in relation to SMT, in Section 3.6 we show a human evaluation conducted by translators clearly preferred human translations (60.52%). However, NMT yielded better translations than PBSMT. We also conducted a manual analysis of the sentences in which NMT or PBSMT were selected as the best translation. It was observed that the main reason for the selection was terminology precision and fluency of the MT output.

To assess MT output for post-editing, PE effort needs to be studied. In Section 4.4 we developed PosEdiOn, an easy-to-use tool which allows to record technical and temporal effort and produces in a quick and visual way the most frequent indirect measures of effort, such as HBLEU, edit distance and HTER.

In Section 4.5, we used this tool to compare two different NMT models (transformer and s2s) for English-Spanish medical texts. PE effort was lower for the transformer model, even though a great diversity among the different translators was recorded. Divergences were greater for segments where there were more errors in the raw MT output but much less when segments did not require modifications. Moreover, temporal and technical effort showed a strong correlation and the edit distance was the indirect measure of effort which showed the greatest correlation with temporal effort.

Then, we studied extraneous factors which can affect the PE effort in Section 5.3. First, we detected a generalized negative bias against post-editing, which can be in part rooted in the economical aspects and the tight experienced by translators. Even so, in our experiment translators' production increased an average of 53.14% when post-editing compared

to translating from scratch. When experienced post-editors were asked about their current professional situation and if it had evolved during time, most of them agreed that experience and training improved the perception of the task and its results.

Then, we considered an error-based approach to evaluate the linguistic quality of the MT output for post-editing. First of all, in Section 6.5 we studied the most frequent MT errors for SMT and NMT models. We conducted and error annotation task following the MQM taxonomy. For English-Spanish medical texts, NMT produced less number of errors and less critical errors. Then, we studied which errors entailed more effort according to Choice Network Analysis (CNA). Our analysis suggested word order, prepositions and verbs are the errors which entail more cognitive effort.

In Section 6.6 we suggested the use of a challenge set for the English-Spanish language combination. It was manually created taking into account the most frequent errors annotated in the aforementioned Section. It can help assess the linguistic quality of an MT output for post-editing. It can also be triangulated with other automatic metrics, such as indirect metrics of effort.

Returning to the specific goals stated at the beginning of the thesis, there are a number of affirmations we can now make:

Assess current automatic and manual MT evaluation methods. Although they are useful to compare MT systems and provide a coarse evaluation, they show problems when the systems are of good quality and they do not always correlate to PE effort. Automatic scores tend to underestimate NMT in relation to SMT, and human perception of quality does not directly correlate to the temporal or technical effort necessary to post-edit the raw MT output.

Assess direct and indirect PE effort indicators. The different dimensions of PE effort are correlated but there is currently no single measure that includes them all. However, we can use proxy measures to correlate some of the measures. We have shown the correlation between temporal

and technical effort, and suggested edit distance and the best correlated proxy indicator.

Study the linguistic errors with the major incidence when post-editing. For English to Spanish medical documents NMT produces less errors than SMT. It produces more fluent outputs with less critical errors. Moreover, SMT and NMT often produce errors of different nature that require different approaches to post-edit. Experiments also suggest that certain errors, such as word order, prepositions and verbs, entail more PE effort.

Generate an evaluation model for raw MT output which takes into account PE effort. We suggest a challenge set based on frequent MT errors that takes into account frequent errors for the English-Spanish language combination. Instead of providing an absolute measure for quality, we propose a measure that targets specific linguistic problems to assess how an MT system tackles with them. Then, we suggest triangulating the results with the edit distance obtained from a test post-editing.

7.2 Limitations and Future Work

All the experiments presented in this thesis include only two language combinations, one for the error analysis. Furthermore, they were pilot studies, with a limited number of post-editors and document length studied. We believe if we increase the number of translators and text length for our future tests, experiments could provide more definite results.

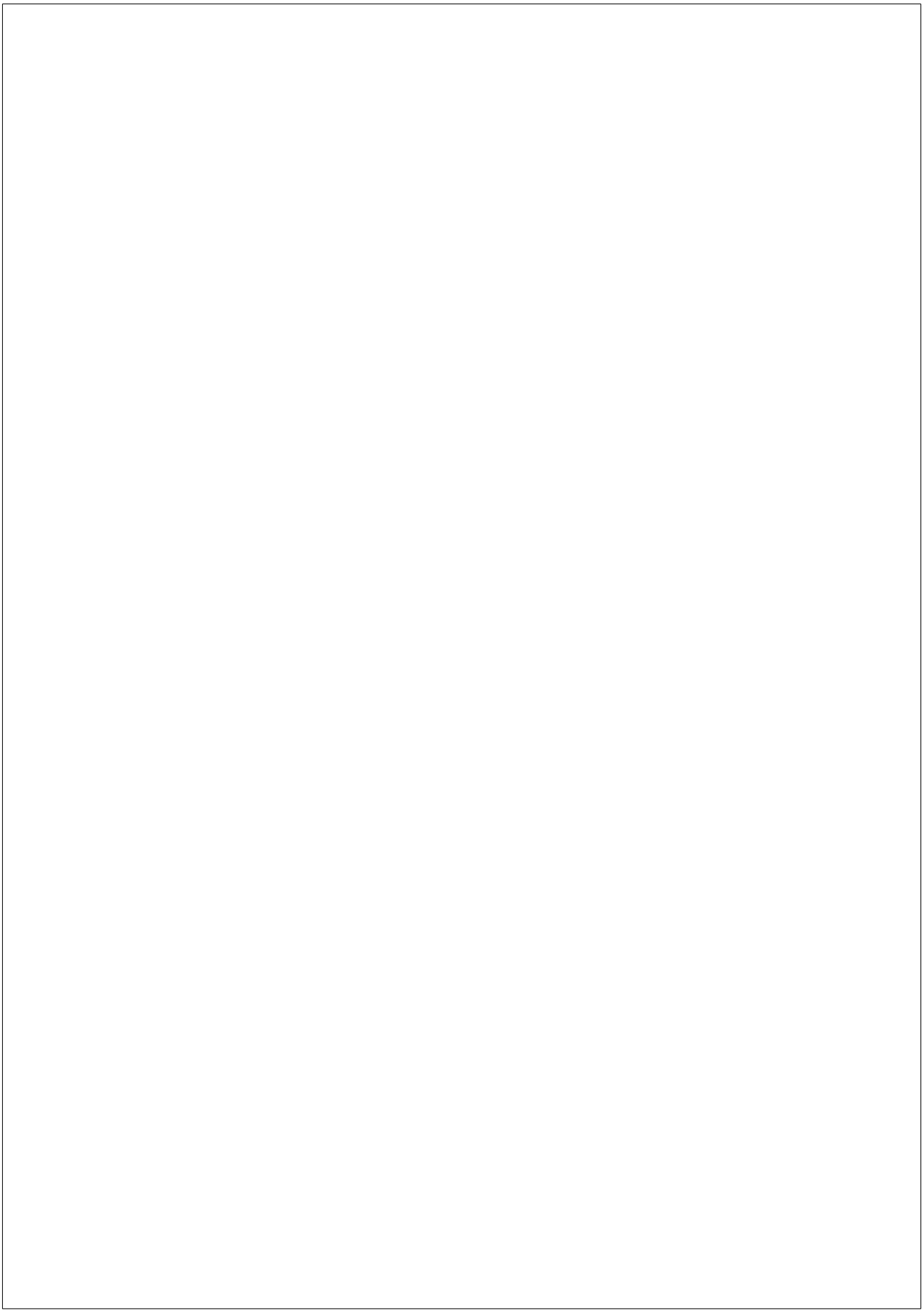
Regarding our future work, we think PosEdiOn is a useful tool with a great potential to study PE effort and indirect metrics of effort. Not only can it help understand the correlation among different measures, but it can also be useful to study the PE process, for example, studying the number of revisions of a same sentence translators conduct. We also plan on adding new features which can help annotate errors easily, for example, and use it as a training tool for post-editing courses.

In terms of automatic evaluation of MT, these scores need to be contextu-

alized in order to increase their precision. They can be triangulated with complementary measures and establish different terminological and linguistic checking points to help complement current available measures. In this line, challenge sets can be a potential field of research which can study specific aspects of NMT outputs, such as the use of pronouns or verb tenses.

We also intend to further research the evaluation of PE effort by studying aspects of the translation process which have not been covered in this thesis. For example, how the domain and length of a text affect PE effort. Even though variability will always be present, narrowing the key elements which most influence PE effort can help determine the best way to assess this effort.

Depicting PE effort can be useful to design automatic measures which are multidimensional and take into account different scores tailored to correlate with PE effort. Furthermore, current MT quality leads to new use scenarios, such as gist translation or MT for second-language learning. We intend to explore these new uses of MT and study how the evaluation of MT output can improve its usability and suitability.



Chapter 8

APPENDIX 1: CHALLENGE SET

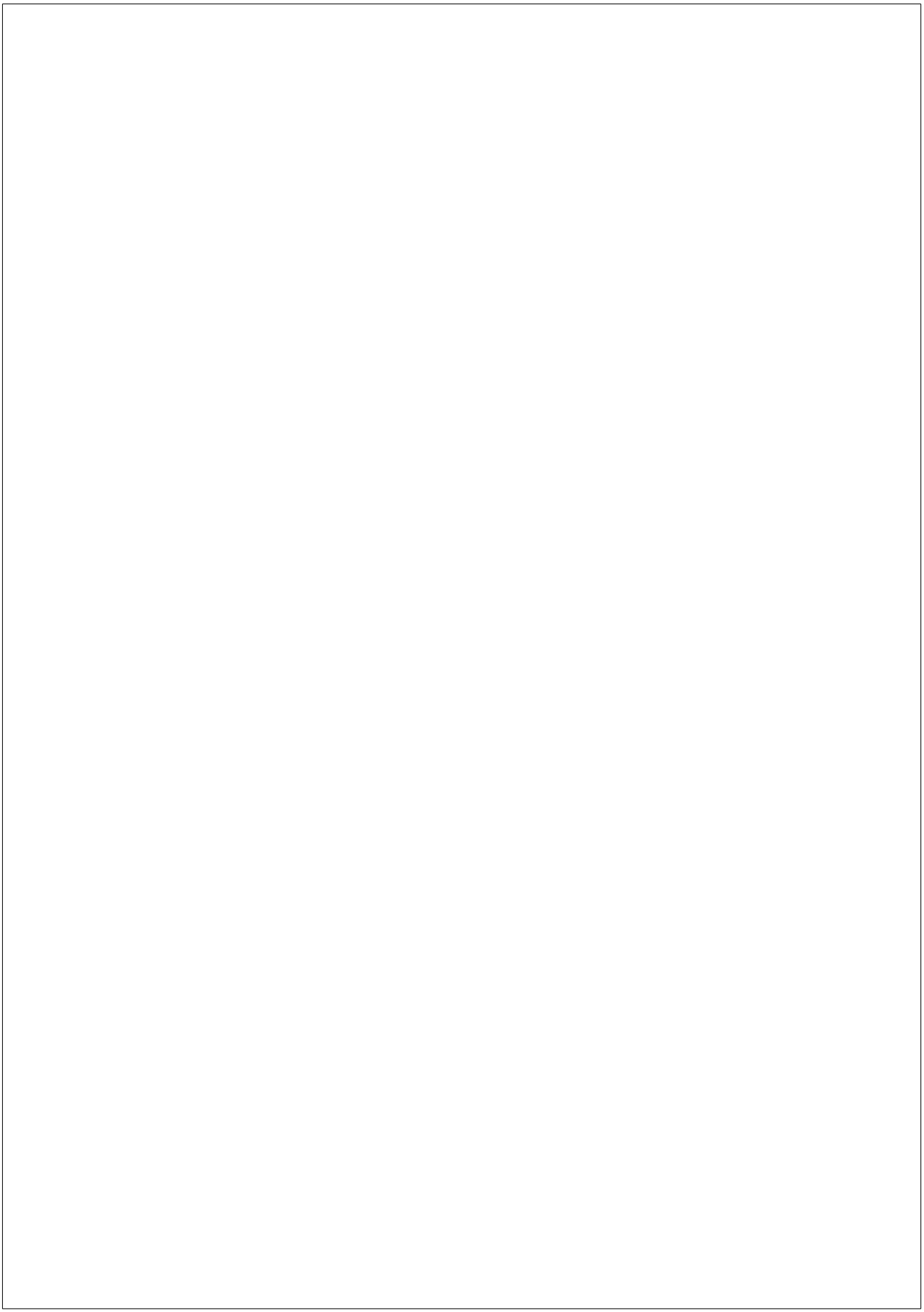
Challenge set sentence	**Error type**
With the help of executive producer, Gilligan came up with an idea inspired by two episodes of The Dick Van Dyke Show, called ”The Night the Roof Fell In” and ”Mary and her friends”, **in which** the main characters, Rob and Laura Petrie, tell different versions of a fight they have had.	concordance
There were two **patients** with severe problems, three **patients** with acute symptoms and only one **patient** without symptoms.	co-reference
Ackbar **was originally planned** to be more conventionally humanoid, but after creator George Lucas decided to make him an alien.	Word order
Mulder returns to the motel room; after Scully has left, he eats **her pizza** and realizes that he has been drugged.	possessive
Navaras **defects** to Hamilcar	verb
He valued the artefacts **at** 17,000 guilders.	preposition
It was clear that the negotiations **were beyond** Humbert at this point.	mistranslation
Of the 86 **people**, 43 **people** stated that they had known other **people** while still married to someone else.	co-reference
Mago occupied the harbor with 150 ships and encamped 60,000 infantry in the part of the city on the **Sicilian mainland**.	mistranslation
He used **his** hands and also **his** legs to return to **his** normal position.	possessive
No plans of a third expedition ever came near to success however, and Humbert decided to return to Italy to live on his military pension.	word order
Fan service must take into account guarantees.	mistranslation
Experts estimate that it took 29,135 **man-hours** to erect Chetro Ketl alone.	compound noun

The canyon bottomlands were further eroded, exposing Menefee Shale bedrock; **this** was subsequently buried under roughly 125 feet (38 m) of sediment.	co-reference
Three **to** four members of Parliament have chosen this option.	preposition
In all situations there were individuals who chose the correct option (3% **to** 4%)	preposition
Mulder returns to the motel room; after Scully has left, he sees **her** bag, he eats **his** pizza, drinks **his** beer, and realizes that he has been drugged.	possessive
Following the confirmation at the Disney Investors Presentation in December that the next film was a go, Lucasfilm and director James Mangold **look to have their sights set** on Harrison Ford's first new co-star in the next installment of the Indiana Jones franchise.	word order
Sources tell Deadline that Phoebe Waller-Bridge is set to co-star opposite Ford in the fifth installment, with **Ford** returning as everyone's favorite fedora-wearing, whip-slinging archaeologist.	co-reference
From infancy, the trilingual Philip **bounced** among his European relatives.	verb
In radio interviews after the 2020 presidential election, West suggested Texas could **vote to again become** a republic, as it was before joining the United States in 1845.	verb
This is something that was written into the Texas Constitution," the former congressman said in one late December radio broadcast.	word order
Or it was promised to Texas when we became part of the United States of America– that if **we** voted and decided, **we** could go back to being our own republic."	pronoun
Four dead gray whales have **washed ashore** San Francisco Bay Area beaches in the last nine days, with experts saying on Friday one had been struck by a ship. They were trying to determine how the other three had died.	verb

The department’s police chief, Amalia Arradondo, stated in her declaration his act was a “murder,” and said at trial that Chauvin’s maneuver “in no way, shape, or form is anything that is **by policy**,”	mistranslation
The department’s police chief, Amalia Arradondo, stated in **her** declaration **his** act was a “murder,” and said at trial that Chauvin’s maneuver “in no way, shape, or form is anything that is by policy,”	possessive
Hotez became a regular on radio talk shows that would **reach people least likely to trust the vaccines**	word order
The group of three adult foreign students from the north of Eslovenia proved quite difficult to understand though, since there were no interpreters available at that moment	word order
An important opportunity for all of us came when a large collection of Etruscan antiquities **was offered for sale**.	word order
All the objects in the list offered by d’Anasty **had been** hidden for at least a quarter of a century	concordance
Humbert was **at this point** promoted to lieutenant-colonel.	connector
Important smaller carnivores include bobcats, badgers, foxes, and two species of skunk.	word order
Local Vancouver **comedian, actor and reporter** Montse Bordes played the coroner.	concordance
Local Vancouver **comedian, actor and reporter** Montse Bordes played the coroner.	punctuation
Only 50 survived the first winter in town out of the about 100 who arrived on the Mayflower.	word order
When you are told the same motto 3 **to** 4 times and 30 **to** 50% of the times you don’t think it is a valuable insight.	preposition
It was clear that the negotiations were **beyond** Humbert at this point.	preposition

Four dead gray whales have washed ashore San Francisco Bay Area beaches in the last nine days, with experts saying on Friday one had been struck by a ship. They were trying to determine how **the other three** had died.	pronoun
The carcass of a 41ft (12.5 meter) adult female gray whale **landed** at San Francisco’s Crissy Field on 31 March	verb
Those events are going to cause stress hormones to pour out into your body, specifically things like adrenaline	co-reference

Table 8.1: Sentences included in the challenge set



Bibliography

Alegria, I., Casillas, A., de Ilarraza, A. D., Igartua, J., Labaka, G., Laskurain, B., Lersundi, M., Mayor, A., Sarasola, K., and Saralegi, X. (2008). Mixing Approaches to MT for Basque: Selecting the Best Output from RBMT, EBMT and SMT. In *MATMT2008 workshop: Mixing Approaches to Machine Translation. pp.27-34. ISBN 978-612-2224-7*.

Allen, J. H. (2003). Post-editing. In Sommer, H., editor, *Computers and Translation: A Translator's Guide*, pages 297–317. John Benjamin, Amsterdam.

Alonso, J. A. and Thurmair, G. (2003). The Compendium Translator Systems. In *Proceedings of the Ninth Machine Translation Summit*. European Association for Machine Translation.

ALPAC (1966). Language and Computers in Translation and Linguistics. *A Report by the Automatic Language Processing Advisory Committee*, National Research Council Publication 1416.

Alvarez, S., Oliver, A., and Badia, T. (2019). Does NMT Make a Difference when Post-editing Closely Related Languages? The Case of Spanish-Catalan. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 49–56, Dublin, Ireland. European Association for Machine Translation.

Alves, F. (2003). Tradução, cognição e contextualização: triangulando a interface processo-produto no desempenho de tradutores novatos.

DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada, 19.

Aranberri, N., Labaka, G., Ilarraza, A., and Sarasola, K. (2014). Comparison of Post-Editing Productivity between Professional Translators and Lay Users. In *Proceedings of the Third Workshop on Post-Editing Technology and Practice (WPTP - 3)*, Vancouver, Canada.

Arnold, D., Moffat, D., Sadler, L., and Way, A. (1993). Automatic Test Suite Generation. *Machine Translation*, 8(1):29–38.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018). Unsupervised Neural Machine Translation. In *ICLR 2018 Conference*.

Avramidis, E., Macketanz, V., Strohriegel, U., and Uszkoreit, H. (2019). Linguistic Evaluation of German-English Machine Translation Using a Test Suite. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 445–454, Florence, Italy. Association for Computational Linguistics.

Aymerich, J. (2005). Using Machine Translation for Fast, Inexpensive, and Accurate Health Information Assimilation and Dissemination: Experiences at the Pan American Health Organization. In *9th World Congress on Health Information and Libraries*.

Aziz, W., Sousa, S. C. M. D., and Specia, L. (2012). PET: A Tool for Post-editing and Assessing Machine Translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12*, pages 3982–3987.

Babych, B. (2014). Automated MT Evaluation Metrics and their Limitations. *Tradumàtica: tecnologies de la traducció*, pages 464–478.

Babych, B. and Hartley, A. (2008). Sensitivity of Automated MT Evaluation Metrics on Higher Quality MT Output: BLEU vs Task-Based Evaluation Methods. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Babych, B. and Hartley, T. (2004). Extending the BLEU MT Evaluation Method with Frequency Weightings. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 621–628, Barcelona, Spain.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR 2018 Conference*.

Banerjee, S. and Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Baniata, L. H., Park, S., and Park, S.-B. (2018). A Multitask-Based Neural Machine Translation Model with Part-of-Speech Tags Integration for Arabic Dialects. *Applied Sciences*, 8(12).

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 Conference on Machine Translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019). Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Bastings, J., Titov, I., Aziz, W., Marcheggiani, D., and Sima'an, K. (2017). Graph Convolutional Encoders for Syntax-aware Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1957–1967, Copenhagen, Denmark. Association for Computational Linguistics.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018). Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Bayatli, S., Kurnaz, S., Salimzianov, I., Washington, J. N., and Tyers, F. M. (2018). Rule-based Machine Translation from Kazakh to Turkish. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 49–58, Alacant, Spain.

Belinkov, Y. and Glass, J. (2019). Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267. Association for Computational Linguistics.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2018). Neural versus Phrase-based MT Quality: An In-depth Analysis on English–German and English–French. *Computer Speech Language*, 49:52–70.

Berka, J., Bojar, O., Fishel, M., Popović, M., and Zeman, D. (2012). Automatic MT Error Analysis: Hjerson Helping Addicter. In *Proceedings of the Eighth International Conference on Language Resources and*

Evaluation (LREC'12), pages 2158–2163, Istanbul, Turkey. European Language Resources Association (ELRA).

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). Confidence estimation for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland. COLING.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., and Monz, C. (2018). Findings of the 2018 Conference on Machine Translation (WMT18). 2:272–303.

Bojar, O., Federmann, C., Haddow, B., Koehn, P., Post, M., and Specia, L. (2016). Ten Years of WMT Evaluation Campaigns: Lessons Learnt. In *Proceedings of the LREC 2016 Workshop "Translation Evaluation – From Fragmented Tools and Data Sets to an Integrated Ecosystem"*, pages 27–34. LREC 2016 Workshop "Translation Evaluation – From Fragmented Tools and Data Sets to an Integrated Ecosystem", LREC 2016 ; Conference date: 24-05-2016 Through 24-05-2016.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Burchardt, A., Macketanz, V., Dehdari, J., Heigold, G., Peter, J., and Williams, P. (2017). A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. *The Prague Bulletin of Mathematical Linguistics*, 108:159 – 170.

Burlot, F. and Yvon, F. (2017). Evaluating the Morphological Competence of Machine Translation Systems. In *Proceedings of the Second*

Conference on Machine Translation, pages 43–55, Copenhagen, Denmark. Association for Computational Linguistics.

Cadwell, P., Castilho, S., O'Brien, S., and Mitchell, L. (2016). Human Factors in Machine Translation and Post-editing among Institutional Translators. *Translation Spaces*, 5:222–243.

Cadwell, P., O'Brien, S., and Teixeira, C. S. C. (2018). Resistance and Accommodation: Factors for the (Non-) Adoption of Machine Translation among Professional Translators. *Perspectives*, 26(3):301–321.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the Role of Bleu in Machine Translation Research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Campbell, S. (1999). A Cognitive Approach to Source Text Difficulty in Translation. *Target. International Journal of Translation Studies*, 11(1):33–63.

Campbell, S. (2000). Choice Network Analysis in Translation research. In Olohan, M., editor, *Intercultural faultlines: Research models in translation studies: Textual and cognitive aspects*, pages 29–42. St. Jerome, Manchester.

Carbonell, J. and Lavie, A. (2005). A Framework for Interactive and Automatic Refinement of Transfer-Based Machine Translation. *European Association for Machine Translation, EAMT 2005 - 10th Annual Conference*.

Carl, M. (2012). Translog-II: a Program for Recording User Activity Data for Empirical Reading and Writing Research. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 4108–4112, Istanbul, Turkey. European Language Resources Association (ELRA).

Carl, M., Dragsted, B., Elming, J., Hardt, D., and Jakobsen, A. (2011). The process of post-editing: A pilot study. *Copenhagen studies in language*, pages 131–142.

Carl, M., Jakobsen, A. L., and Jensen, K. (2008). Studying human translation behavior with user-activity data. In *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science, NLPCS 2008*, pages 114–123.

Carroll, J. B. (1966). An experiment in evaluating the quality of translations. *Mechanical Translation*, 9:55–66.

Castilho, S., Doherty, S., Gaspari, F., and Moorkens, J. (2018). *Approaches to Human and Machine Translation Quality Assessment*, pages 9–38. Springer International Publishing, Cham.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., and Way, A. (2017a). Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108:109–120.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., and Way, A. (2017b). Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics NUMBER*, pages 109–120.

Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Sosoni, V., Georgakopoulou, Y., Lohar, P., Way, A., Barone, M., and Gialama, M. (2017c). A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. In *Proceedings of MT Summit XVI, vol.1: Research Track*, pages 116–131.

Castilho, S., Resende, N., and Mitkov, R. (2019). What influences the features of post-editese? a preliminary study. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 19–27, Varna, Bulgaria. Incoma Ltd., Shoumen, Bulgaria.

Castilho, S., Salis, C., Alves, F., Brien, S. O., Salis, C., and Brien, M. O. (2014). Does post-editing increase usability? A study with Brazilian Portuguese as Target Language. (2010).

Chatzikoumi, E. (2020). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Doha, Qatar. Association for Computational Linguistics.

Church, K. W. and Hovy, E. H. (1993). Good Applications for Crummy Machine Translation. *Machine Translation*, 8(4):239–258.

Cinková, S. and Bojar, O. (2018). Testsuite on Czech–English grammatical contrasts. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 561–569, Belgium, Brussels. Association for Computational Linguistics.

Comelles, E., Atserias, J., Arranz, V., and Castellón, I. (2012). VERTa: Linguistic features in MT evaluation. In *Proceedings of the Eighth*

International Conference on Language Resources and Evaluation (LREC'12), pages 3944–3950, Istanbul, Turkey. European Language Resources Association (ELRA).

Costa, Â., Ling, W., Luís, T., Correia, R., and Coheur, L. (2015). A linguistically motivated taxonomy for Machine Translation error analysis. *Machine Translation*, 29(2):127–161.

Costa-Jussà, M. R., Rapp, R., Lambert, P., Eberle, K., Banchs, R. E., and Babych, B. (2016). *Hybrid Approaches to Machine Translation*. Springer Publishing Company, Incorporated, 1st edition.

Costa-Jussà, M. R. (2017). Why Catalan-Spanish Neural Machine Translation? Analysis, Comparison and Combination with Standard Rule and Phrase-based Technologies. *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 55–62.

Coughlin, D. (2001). Correlating automated and human assessments of machine translation quality. *Proceedings of MT Summit IX*.

Daems, J., De Clercq, O., and Macken, L. (2017a). Translationese and post-editese : how comparable is comparable quality? *LINGUISTICA ANTVERPIENSIA NEW SERIES-THEMES IN TRANSLATION STUDIES*, 16:89–103.

Daems, J., Vandepitte, S., Hartsuiker, R. J., and Macken, L. (2017b). Identifying the machine translation error types with the greatest impact on post-editing effort. *Frontiers in Psychology*, 8:1282.

Denkowski, M. and Lavie, A. (2012). Transcenter: Web-based translation research suite. In *AMTA 2012 Workshop on Post-Editing Technology and Practice Demo Session*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145.

Doherty, S. (2013). *Investigating the Effects of Controlled Language on the Reading and Language on the Reading and Comprehension of Machine Translated Texts: A Mixed-Methods Approach using Eye Tracking*. PhD thesis.

Dowling, M., Lynn, T., Poncelas, A., and Way, A. (2018). SMT versus NMT: Preliminary comparisons for Irish. In *Proceedings of the Workshop on Technologies for MT of Low Resource Languages, LoResMT@AMTA 2018, Boston, MA, USA, March 21, 2018*, pages 12–20.

Dreyer, M. and Marcu, D. (2012). HyTER: Meaning-equivalent semantics for translation evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171, Montréal, Canada. Association for Computational Linguistics.

Drugan, J. (2013). Translation quality : Importance and definitions. In *Quality in Professional Translation : Assessment and Improvement*, pages 35–80. Bloomsbury, London, 1 edition.

Elliott, D., Hartley, A., and Atwell, E. (2004). A Fluency Error Categorization Scheme to Guide Automated Machine Translation Evaluation. pages 64–73.

Englund Dimitrova, B. (2005). *Expertise and explicitation in the translation process*. John Benjamins Publishing Company, Amsterdam.

Esperança-Rodier, E., Rossi, C., Bérard, A., and Besacier, L. (2017). Evaluation of NMT and SMT systems: A study on uses and perceptions. In *39th Conference Translating and the Computer*, Londres, United Kingdom.

Etchegoyhen, T., Martínez Garcia, E., Azpeitia, A., Labaka, G., Alegria, I., Cortes Etxabe, I., Jauregi Carrera, A., Ellakuria Santos, I., Martin, M., and Calonge, E. (2018). Neural machine translation of basque.

Farrús, M., Costa-jussà, M. R., Mariño, J. B., and Fonollosa, J. A. R. (2010). Linguistic-based evaluation criteria to identify statistical machine translation errors. In *Proceedings of the 14th Annual conference of the European Association for Machine Translation*, Saint Raphaël, France. European Association for Machine Translation.

Farrús, M., Costa-jussa, M., Bernardo Mariño Acebal, J., and Fonollosa, J. (2010). Linguistic-based evaluation criteria to identify statistical machine translation errors. *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*.

Federico, M., Cattelan, A., and Trombetti, M. (2012). Measuring user productivity in machine translation enhanced computer assisted translation.

Federico, M., Negri, M., Bentivogli, L., and Turchi, M. (2014). Assessing the impact of translation errors on machine translation quality with mixed-effects models. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1643–1653, Doha, Qatar. Association for Computational Linguistics.

Fischer, L. and Läubli, S. (2020). What's the difference between professional human and machine translation? a blind multi-language study on domain-specific MT. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 215–224, Lisboa, Portugal. European Association for Machine Translation.

Flanagan, M. (1994). Error classification for MT evaluation. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, USA.

Flanagan, M. and Christensen, T. P. (2014). Testing post-editing guidelines: how translation trainees interpret them and how to tailor them for translator training purposes. *The Interpreter and Translator Trainer*, 8(2):257–275.

Fleiss, J. et al. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Fomicheva, M. and Bel, N. (2016). Using Contextual Information for Machine Translation Evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2755–2761, Portorož, Slovenia. European Language Resources Association (ELRA).

Fomicheva, M. and Specia, L. (2019). Taking MT Evaluation Metrics to Extremes: Beyond Correlation with Human Judgments. *Computational Linguistics*, 45(3):515–558.

Forcada, M. L. (2017). Making sense of neural machine translation. *Translation Spaces. A multidisciplinary, multimedia, and multilingual journal of translation*, 6(2):291–309.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Garcia, I. (2011). Translating by post-editing: is it the way forward? *Machine Translation*, 25(3):217–237.

García, I. (2012). A Brief History of Postediting and of Research on Postediting. *Revista Anglo Saxonica*, 3(3):291–310.

Gaspari, F. and Toral, A. (2014). Perception vs reality: Measuring machine translation post-editing productivity. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: Workshop on Post-editing Technology and Practice (WPTP3)*.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia. PMLR.

Germann, U., Barbu, E., Bentivogli, L., Bertoldi, N., Bogoychev, N., Buck, C., Caroselli, D., Carvalho, L., Cattelan, A., Cattoni, R., et al. (2016). Modern mt: A new open-source machine translation platform for the translation industry.

Gimenez, J. (2008). *EMPIRICAL MACHINE TRANSLATION AND ITS EVALUATION*. PhD thesis, Universitat Politècnica de Catalunya.

Gimenez, J. (2010). Asiya: An open toolkit for automatic machine translation (meta-)evaluation. *The Prague Bulletin of Mathematical Linguistics*, 94.

Graham, Y. and Baldwin, T. (2014). Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar. Association for Computational Linguistics.

Graham, Y., Baldwin, T., Dowling, M., Eskevich, M., Lynn, T., and Tounsi, L. (2016). Is all that glitters in machine translation quality estimation really gold? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan. The COLING 2016 Organizing Committee.

Graham, Y. and Liu, Q. (2016). Achieving accurate conclusions in evaluation of automatic machine translation metrics. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA. Association for Computational Linguistics.

Green, S., Heer, J., and Manning, C. D. (2013). The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI 13*. ACM Press.

Guerberof, A. (2009a). Productivity and Quality in MT Post-editing. *Proceedings of MT Summit XII*, page 8pp.

Guerberof, A. (2009b). Productivity and Quality in the Post-editing of Outputs from Translation Memories and Machine Translation. *The International Journal of Localisation*, 7(1):11–21.

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

Guillou, L. and Hardmeier, C. (2016). PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia. European Language Resources Association (ELRA).

Guillou, L., Hardmeier, C., Lapshinova-Koltunski, E., and Loáiciga, S. (2018). A Pronoun Test Suite Evaluation of the English–German MT Systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

Hale, S. and Campbell, S. (2002). The interaction between text difficulty and translation accuracy. *Babel*, 48:14–33.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (2018). Achieving human parity on automatic chinese to english news translation.

Hayakawa, T. and Arase, Y. (2020). Fine-grained error analysis on English-to-Japanese machine translation in the medical domain. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 155–164, Lisboa, Portugal. European Association for Machine Translation.

Hearne, M. and Way, A. (2011). Statistical machine translation: A guide for linguists and translators. *Language and Linguistics Compass*, 5(5):205–226.

Herbig, N., Pal, S., Vela, M., Krüger, A., and van Genabith, J. (2019). Multi-modal indicators for estimating perceived cognitive load in post-editing of machine translation. *Machine Translation (MT)*, 33:1–25.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

House, J. (1988). Talking to oneself or thinking with others? on using different thinking aloud methods in translation. *Fremdsprachen lehren und lernen*, 17:84–98.

House, J. (2014). *Translation Quality Assessment: Past and Present*, pages 241–264. Palgrave Macmillan UK, London.

Hovy, E., King, M., and Popescu-Belis, A. (2002). Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, 17(1):43–75.

Hu, K. and Cadwell, P. (2016). A Comparative Study of Post-editing Guidelines. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 34206–353.

Hutchins, W. J. (2001). Machine translation over fifty years. *HISTOIRE, EPISTEMOLOGIE, LANGAGE, TOME XXII, FASC. 1 (2001)*, 23:7–31.

Isabelle, P., Cherry, C., and Foster, G. (2017). A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Isahara, H. (2006). JEIDA ' s Test-Sets for Quality Evaluation of MT Systems– Technical Evaluation from the Developer ' s Point of View–. In *Proceedings of MT Summit V*.

Isozaki, H., Hirao, T., Duh, K., Sudoh, K., and Tsukada, H. (2010). Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

Ive, J., Blain, F., and Specia, L. (2018). deepQuest: A framework for neural-based quality estimation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3146–3157, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jia, Y., Carl, M., and Wang, X. (2019). How does the post-editing of neural machine translation compare with from-scratch translation? a product and process study. *The Journal of Specialised Translation*, pages 60–86.

Junczys-Dowmunt, M., Dwojak, T., and Hoang, H. (2016). Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. *CoRR*, abs/1610.01108.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A.,

Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast Neural Machine Translation in C++. *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent Continuous Translation Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Katan, D. (2009). Occupation or Profession: A survey of the Translators' World. *Translation and Interpreting Studies*, 4:187–209.

Kenny, D. (2018). Machine translation. In Rawling, P. and Wilson, P., editors, *The Routledge Handbook of Translation and Philosophy*, chapter 26. Routledge.

Kenny, D. and Doherty, S. (2014). Statistical Machine Translation in the Translation Curriculum: Overcoming Obstacles and Empowering Translators, volume = 8, journal = The Interpreter and Translator Trainer, doi = 10.1080/1750399X.2014.936112. pages 276–294.

Kepler, F., Trénous, J., Treviso, M., Vera, M., and Martins, A. F. T. (2019). OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

King, M. and Falkedal, K. (1990). Using Test Suites in Evaluation of Machine Translation Systems. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 2*, COLING '90, page 211–216, USA. Association for Computational Linguistics.

Kirchhoff, K., Capurro, D., and Turner, A. (2012). Evaluating User Preferences in Machine Translation Using Conjoint Analysis. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 119–126, Trento, Italy. European Association for Machine Translation.

Kirchhoff, K., Turner, A., Axelrod, A., and Saavedra, F. (2011). Application of Statistical Machine Translation to Public Health Information: A Feasibility Study. *Journal of the American Medical Informatics Association : JAMIA*, 18:473–8.

Kirschner, P. A. (2002). Cognitive load theory: implications of cognitive load theory on the design of learning. *Learning and Instruction*, 12(1):1–10.

Klubicka, F., Toral, A., and Sánchez-Cartagena, V. M. (2017). Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation. *Prague Bulletin of Mathematic. Linguistics*, 108:121–132.

Klubička, F., Toral, A., and Sánchez-Cartagena, V. M. (2018). Quantitative fine-grained human evaluation of machine translation systems: a case study on English to Croatian. *Machine Translation*, 32(3):195–215.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT summit*, volume 5, pages 79–86.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, USA, 1st edition.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Koh, S., Maeng, J., Lee, J.-Y., Chae, Y.-S., and Choi, K.-S. (2001). A test suite for evaluation of English-to-Korean machine translation systems. In *MT Summit Conference*.

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

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

Koponen, M., Salmi, L., and Nikulin, M. (2019). A Product and Process Analysis of Post-editor Corrections on Neural, Statistical and Rule-based Machine Translation Output. *Machine Translation*.

Krings, H. P. (2001). *Repairing texts: Empirical investigations of machine translation post-editing process*. The Kent State University Press, Kent, OH.

Lacruz, I. (2012). Average Pause Ratio as an Indicator of Cognitive Effort in Post-Editing : A Case Study. In O'Brien, S., Simard, M., and Specia, L., editors, *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice (WTTP)*, San Diego.

Lacruz, I. (2017). *Cognitive Effort in Translation, Editing, and Post-editing*, chapter 21, pages 386–401. John Wiley and Sons, Ltd.

Lacruz, I. and Shreve, M. G. (2014). *Pauses and cognitive effort in post-editing*, pages 246–273. Cambridge Scholars Publishing.

Läubli, S., Castilho, S., Neubig, G., Sennrich, R., Shen, Q., and Toral, A. (2020). A Set of Recommendations for Assessing Human–Machine Parity in Language Translation. *Journal of Artificial Intelligence Research*, 67.

Läubli, S., Sennrich, R., and Volk, M. (2018). Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

LeBlanc, M. (2013). Translators on Translation Memory (TM). Results of an Ethnographic Study in Three Translation Services and Agencies. *The International Journal of Translation and Interpreting Research*, 5.

Lehmann, S., Oepen, S., Regnier-prost, S., Netter, K., Lux, V., Klein, J., Falkedal, K., Fouvry, F., Estival, D., Dauphin, E., Compagnion, H., Bauer, J., Balkan, L., and Arnold, D. (1996). Tsnlp - test suites for natural language processing. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Leusch, G., Ueffing, N., and Ney, H. (2006). CDER: Efficient MT evaluation using block movements. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*.

Lin, C.-Y. and Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Liu, Y. and Du, M. (2014). A Multiple Case Study of Chinese-English Translation Strategies. *Studies in Literature and Language*, 9:58–65.

Lo, C.-k. (2019). YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Lohar, P., Popovic, M., Afli, H., and Way, A. (2019). A systematic comparison between smt and nmt on translating user-generated content. In *Proceedings of CICLing 2019, the 20th International Conference on Computational Linguistics and Intelligent Text Processing*, La Rochelle, France.

Lommel, A., Burchardt, A., Görög, A., Uszkoreit, H., and Melby, A. K. (2015). Multidimensional Quality Metrics (MQM) Issue Types.

Lommel, A., Popovic, M., and Burchardt, A. (2014). Assessing Inter-Annotator Agreement for Translation Error Annotation. In *LREC Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*.

Lommel, A. R. and Depalma, D. A. (2016). Europe's Leading Role in Machine Translation How Europe Is Driving the Shift to MT. Technical report.

Lopez, A. (2008). Statistical machine translation. *ACM Comput. Surv.*, 40(3).

Läubli, S., Amrhein, C., Düggelin, P., Gonzalez, B., Zwahlen, A., and Volk, M. (2019). Post-editing productivity with neural machine translation: An empirical assessment of speed and quality in the banking and finance domain.

Läubli, S., Fishel, M., Massey, G., Ehrensberger-Dow, M., and Volk, M. (2013). Assessing post-editing efficiency in a realistic translation environment.

Läubli, S. and Orrego Carmona, D. (2017). When google translate is better than some human colleagues, those people are no longer colleagues. In *Proceedings of the 39th Conference Translation and the Computer*.

Ma, Q., Wei, J., Bojar, O., and Graham, Y. (2019). Results of the WMT19 metrics shared task: Segment-level and strong MT systems

pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Mahata, K. S., Mandal, S., Das, D., and Bandyopadhyay, S. (2018). Smt vs nmt: A comparison over hindi bengali simple sentences. *arXiv: Computation and Language*.

Mathur, N., Baldwin, T., and Cohn, T. (2019). Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808, Florence, Italy. Association for Computational Linguistics.

Mathur, N., Baldwin, T., and Cohn, T. (2020). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Melamed, I. D., Green, R., and Turian, J. P. (2003). Precision and recall of machine translation. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*, pages 61–63.

Mellinger, C. D. (2017). Translators and Machine Translation: Knowledge and Skills Gaps in Translator Pedagogy. *The Interpreter and Translator Trainer*, 11(4):280–293.

Moorkens, J. (2017). Under pressure: translation in times of austerity. *Perspectives*, 25(3):464–477.

Moorkens, J. (2018). *Chapter 4. Eye tracking as a measure of cognitive effort for post-editing of machine translation*, pages 55–70.

Moorkens, J. and Brien, S. (2017). *Assessing User Interface Needs of Post-Editors of Machine Translation*, chapter 6, pages 109–130. Routledge.

Moorkens, J., O 'brien, S., Da Silva, I. A. L., De, N. B., Fonseca, L., Alves, F., and De Lima Fonseca, N. B. (2015). Correlations of Perceived Post-editing Effort with Measurements of Actual Effort. *Machine Translation*, 29:267–284.

Moorkens, J. and O'Brien, S. (2013). User attitudes to the post-editing interface. In *Proceedings of Machine Translation Summit XIV: Second Workshop on Post-editing Technology and Practice, Nice, France*, pages 19–25.

Moorkens, J. and O'Brien, S. (2015). Post-Editing Evaluations: Trade-offs between Novice and Professional Participants. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 75–81, Antalya, Turkey.

Moorkens, J., Toral, A., Castilho, S., and Way, A. (2018). Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7:240–262.

Müller, M., Rios, A., Voita, E., and Sennrich, R. (2018). A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Munday, J. (2016). *Introducing Translation Studies*. Routledge, Fifth Edition. — Milton Park ; New York : Routledge, 2016. —.

Mutal, J., Volkart, L., Bouillon, P., Girletti, S., and Estrella, P. (2019). Differences between SMT and NMT output - a translators' point of view. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 75–81, Varna, Bulgaria. Incoma Ltd., Shoumen, Bulgaria.

Müller, M., Rios, A., and Sennrich, R. (2019). Domain robustness in neural machine translation.

Nießen, S., Och, F. J., Leusch, G., and Ney, H. (2000). An evaluation tool for machine translation: Fast evaluation for MT research. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Nunes Vieira, L. (2016). Cognitive effort in post-editing of machine translation: Evidence from eye movements, subjective ratings, and think-aloud protocols.

Nunes Vieira, L. and Alonso, E. (2018). *The use of machine translation in human translation workflows: Practices, perceptions and knowledge exchange*. Institute of Translation and Interpreting.

O'Brien, S. (2002). Teaching Post-editing: A Proposal for Course Content. In *Proceedings of the 6th EAMT Workshop: Teaching Machine Translation*, Manchester, England. European Association for Machine Translation.

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

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

O'Brien, S. (2007). An empirical investigation of temporal and technical post-editing effort. *Translation and Interpreting Studies*, 2:83–136.

O'Brien, S. (2011a). Towards a Dynamic Quality Evaluation Model for Translation. *The Journal of Specialised Translation*, 17:55–77.

O'Brien, S. (2011b). Towards Predicting Post-editing Productivity. *Machine Translation*, 25(3):197–215.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

Oliver, A. (2017). El corpus paral·lel del diari oficial de la generalitat de catalunya: compilació, anàlisi i exemples d'ús. *Zeitschrift für Katalanistik*, 30:269–291.

Paas, F., Tuovinen, J., Tabbers, H., and Van Gerven, P. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist - EDUC PSYCHOL*, 38:63–71.

Paas, F., Van Merrienboer, J. J. G., and Adam, J. (1994). Measurement of cognitive load in instructional research. *Perceptual and motor skills*, 79:419–30.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40Th Annual Meeting on Machine Translation*, (July):311–318.

Parra Escartín, C. and Arcedillo, M. (2015). A Fuzzier Approach to Machine Translation Evaluation: A Pilot Study on Post-editing Productivity and Automated Metrics in Commercial Settings. *Proceedings of the ACL 2015 Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, 1(2010):40–45.

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

Popel, M. (2018). CUNI transformer neural MT system for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 482–487, Belgium, Brussels. Association for Computational Linguistics.

Popovic, M. (2011). Hjerson: An open source tool for automatic error classification of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 96:59–68.

Popović, M. (2015). chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Popović, M. (2018). Error Classification and Analysis for Machine Translation Quality Assessment. pages 129–158.

Popović, M. (2019). Evaluating Conjunction Disambiguation on English-to-German and French-to-German WMT 2019 Translation Hypotheses. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 464–469, Florence, Italy. Association for Computational Linguistics.

Popović, M. and Arčan, M. (2016). PE2rr corpus: Manual error annotation of automatically pre-annotated MT post-edits. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 27–32, Portorož, Slovenia. European Language Resources Association (ELRA).

Popovic, M. and Castilho, S. (2019). Are Ambiguous Conjunctions Problematic for Machine Translation? In *2th Conference on Recent Advances in Natural Language Processing*.

Popović, M. and Castilho, S. (2019). Challenge test sets for MT evaluation. In *Proceedings of Machine Translation Summit XVII Volume 3: Tutorial Abstracts*, Dublin, Ireland. European Association for Machine Translation.

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Qin, Y. and Specia, L. (2015). Truly exploring multiple references for machine translation evaluation. In El-Kahlout, I. D., Özkan, M., Sánchez-Martínez, F., Ramírez-Sánchez, G., Hollowood, F., and Way, A., editors, *Proceedings of the 18th Annual Conference of the European Association for Machine Translation, EAMT 2015, Antalya, Turkey, May 11 - 13, 2015*. European Association for Machine Translation.

Raganato, A., Scherrer, Y., and Tiedemann, J. (2019). The MuCoW Test Suite at WMT 2019: Automatically Harvested Multilingual Contrastive Word Sense Disambiguation Test Sets for Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy. Association for Computational Linguistics.

Rajman, M. and Hartley, T. (2001). Automatically predicting mt systems rankings compatible with fluency, adequacy or informativeness scores.

Rani, S. and Singh, J. (2018). Enhancing Levenshtein's Edit Distance Algorithm for Evaluating Document Similarity. In Sharma, R., Mantri, A., and Dua, S., editors, *Computing, Analytics and Networks*, pages 72–80, Singapore. Springer Singapore.

Rios, A., Müller, M., and Sennrich, R. (2018). The Word Sense Disambiguation Test Suite at WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 588–596, Belgium, Brussels. Association for Computational Linguistics.

Rios Gonzales, A., Mascarell, L., and Sennrich, R. (2017). Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Sánchez-Cartagena, V. M., Forcada, M. L., and Sánchez-Martínez, F. (2020). A Multi-source Approach for Breton–French Hybrid Machine

Translation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 61–70, Lisboa, Portugal. European Association for Machine Translation.

Sanchez-Torron, M. and Koehn, P. (2016). Machine Translation Quality and Post-Editor Productivity. In *Proceedings of AMTA 2016*, pages 16–26.

Sanders, G., Przybocki, M., Madnani, N., and Snover, M. (2011). Human subjective judgments. In Oliver, J., Christianson, C., and McCary, J., editors, *Handbook of Natural Language Processing and Machine Translation*, pages 759–759. Springer, Oxford.

Scarton, C., Forcada, M. L., Esplà-Gomis, M., and Specia, L. (2019). Estimating post-editing effort: a study on human judgements, task-based and reference-based metrics of MT quality. volume abs/1910.06204.

Screen, B. (2017). Machine translation and welsh: Analysing free statistical machine translation for the professional translation of an under-researched language pair. *The Journal of Specialised Translation*, 28:218–244.

Sennrich, R. (2017). How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

Sennrich, R. and Haddow, B. (2016). Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shterionov, D., Superbo, R., Nagle, P., Casanellas, L., O'Dowd, T., and Way, A. (2018). Human versus automatic quality evaluation of nmt and pbsmt. *Machine Translation*, 32(3):217–235.

Singh, M., Kumar, R., and Chana, I. (2019). Improving neural machine translation using rule-based machine translation. In *2019 7th International Conference on Smart Computing Communications (ICSCC)*, pages 1–5.

Skadina, I. and Pinnis, M. (2017). NMT or SMT: Case Study of a Narrow-domain English-Latvian Post-editing Project. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 373–383. Asian Federation of Natural Language Processing.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation.

Snover, M., Madnani, N., Dorr, B. J., and Schwartz, R. (2009). Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. In *WMT*, volume 30, page 259–268.

Sosoni, V. and Rogers, M. (2013). Translation in an Age of Austerity: From Riches to Pauper, or Not? *mTm Journal*, pages 109–120.

Specia, L. (2010). Combining Confidence Estimation and Reference-based Metrics for Segment-level MT Evaluation. In *The Ninth Conference of the Association for Machine Translation in the Americas*.

Specia, L. (2011). Exploiting Objective Annotations for Measuring Translation Post-editing Effort. In *Proceedings of the European Association for Machine Translation*, number May, pages 73–80.

Specia, L., Paetzold, G., and Scarton, C. (2015). Multi-level translation quality prediction with QuEst++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Specia, L., Shah, K., de Souza, J. G., and Cohn, T. (2013). QuEst - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria. Association for Computational Linguistics.

Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Steele, D. and Specia, L. (2018). Vis-eval metric viewer: A visualisation tool for inspecting and evaluating metric scores of machine translation output. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 71–75, New Orleans, Louisiana. Association for Computational Linguistics.

Stymne, S. and Ahrenberg, L. (2012). On the Practice of Error Analysis for Machine Translation Evaluation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1785–1790, Istanbul, Turkey. European Language Resources Association (ELRA).

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Tang, G. (2020). *Understanding Neural Machine Translation: An Investigation into Linguistic Phenomena and Attention Mechanisms*. PhD thesis, Acta Universitatis Upsaliensis.

Teixeira, C. (2014). Perceived vs. measured performance in the post-editing of suggestions from machine translation and translation memories. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: Workshop on Post-editing Technology and Practice (WPTP3)*.

Tiedemann, J. (2012a). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Tiedemann, J. (2012b). Parallel Data, Tools and Interfaces in OPUS. In *Lrec*, volume 2012, pages 2214–2218.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H. (1997). Accelerated dp based search for statistical translation. In *In European Conf. on Speech Communication and Technology*, pages 2667–2670.

Toral, A. (2019). Post-editese: an exacerbated translationese. pages 273–281. Machine Translation Summit XVII ; Conference date: 19-08-2019 Through 23-08-2019.

Toral, A., Castilho, S., Hu, K., and Way, A. (2018). Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Toral, A. and Sánchez-Cartagena, V. M. (2017). A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. *CoRR*, abs/1701.02901.

Torregrosa, D., Pasricha, N., Masoud, M., Chakravarthi, B. R., Alonso, J., Casas, N., and Arcan, M. (2019). Leveraging rule-based machine translation knowledge for under-resourced neural machine translation models. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 125–133, Dublin, Ireland. European Association for Machine Translation.

Torrejón, E. and Rico, C. (2013). Skills and Profile of the New Role of the Translator as MT Post-editor. *Tradumàtica: tecnologies de la traducció*, page 166.

Toury, G. (2012). *Descriptive translation studies–and beyond: Revised Edition*, volume 100. John Benjamins Publishing.

Turian, J., Shen, L., and Dan Melamed, I. (2003). Evaluation of Machine Translation and its Evaluation. *Proceedings of MT Summit IX*, pages 386–393.

Vardaro, J., Schaeffer, M., and Hansen-Schirra, S. (2019). Comparing the quality of neural machine translation and professional post-editing. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3.

Vasconcellos, M. and León, M. (1985). Spanam and engspan: Machine translation at the pan american health organization. *Comput. Linguist.*, 11(2–3):122–136.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *NIPS*.

Vieira, L. N. (2013). An evaluation of tools for post-editing research: the current picture and further needs. In *Proceedings of the MT Summit XIV Workshop on Post-editing Technology and Practice (WPTP-2)*, pages 93–101, Nice.

Vieira, L. N. (2016). How Do Measures of Cognitive Effort Relate to Each Other? A Multivariate Analysis of Post-editing Process Data. *Machine Translation*, 30(1):41–62.

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

Voita, E., Sennrich, R., and Titov, I. (2019). When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Wang, C. and Sennrich, R. (2020). On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

White, J. S., O'Connell, T. A., and Carlson, L. M. (1993). Evaluation of Machine Translation. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, page 206–210, USA. Association for Computational Linguistics.

Wieting, J., Berg-Kirkpatrick, T., Gimpel, K., and Neubig, G. (2019). Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Wilks, Y. (1994). Keynote traditions in the evaluation of mt // mt evaluation. basis for future directions. pages 1–3.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A.,

Johnson, M., Liu, X., Kaiser, , Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.

Ye, Y. and Toral, A. (2020). Fine-grained human evaluation of transformer and recurrent approaches to neural machine translation for English-to-Chinese. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 125–134, Lisboa, Portugal. European Association for Machine Translation.

Yu, H., Ma, Q., Wu, X., and Liu, Q. (2015). Casict-dcu participation in wmt2015 metrics task. pages 417–421.

Zeman, D., Fishel, M., Berka, J., and Bojar, O. (2011). Addicter: What is wrong with my translations? *The Prague Bulletin of Mathematical Linguistics*, 96.

Zhang, H. and Gildea, D. (2007). Factorization of Synchronous Context-Free Grammars in Linear Time. In *Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 25–32, Rochester, New York. Association for Computational Linguistics.

Zhou, M., Wang, B., Liu, S., Li, M., Zhang, D., and Zhao, T. (2008). Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1121–1128, Manchester, UK. Coling 2008 Organizing Committee.