*Nora Aranberri**

What Do Professional Translators Do when Post-Editing for the First Time? First Insight into the Spanish-Basque Language Pair

Abstract

Machine translation post-editing is becoming commonplace and professional translators are often faced with this unknown task with little training and support. Given the different translation processes involved during post-editing, research suggests that untrained translators do not necessarily make good post-editors. Besides, the post-editing activity will be largely influenced by numerous aspects related to the technology and texts used. Training material, therefore, will need to be tailored to the particular conditions under which post-editing is bound to happen. In this work, we provide a first attempt to uncover what activity professional translators carry out when working from Spanish into Basque. Our initial analysis reveals that when working with moderate machine translation output post-editing shifts from the task of identifying and fixing errors, to that of "patchwork" where post-editors identify the machine translated elements to reuse and connect them using their own contributions. Data also reveal that they primarily focus on correcting machine translation errors but often fail to restrain themselves from editing correct structures. Both findings have clear implications for training and are a step forward in tailoring sessions specifically for language combinations of moderate quality.

Keywords

post-editing, guidelines, edit-types, professional translators, Spanish, Basque

1. Introduction

The combination of machine translation (MT) and post-editing (PE) is becoming common practice in the language industry. A 2013 survey by the Common Sense Advisory revealed that 44% of the almost 1,000 language service providers surveyed offered post-editing services (DePalma et al. 2013). The trend is only becoming stronger, as the 2015 survey showed, with post-editing rising from eighth position in 2014 to seventh position in 2015 according to the top10 ranking for the services that grew most (DePalma et al. 2015).

This means that the industry needs skilled professionals in the area. Companies tend to consider translators as the natural post-editors. However, translators associate machine translation with negative concepts such as "*regimentation, dependence, exploitation or impotence*" (Cronin 2013). A review of the tasks involved in translating and post-editing seems to suggest that they differ considerably. If the tasks differ, so will the skills that are needed to address them and specialists will need to be trained. Translation involves "*the transfer of 'meaning' contained in one set of language signs into another set of language signs through competent use of the dictionary and grammar, the process involves a whole set of extra-linguistic criteria also*" (Bassnett 1991: 21). Following what he calls a *minimalist* approach, Pym (2003: 489) defines the translation-specific competences of a good translator as follows:

> 1 – The ability to generate a series of more than one viable target text (TT_1, TT_2 … TT_n) for a pertinent source text (ST);
>
> 2 – The ability to select only one viable TT from this series, quickly and with justified confidence.

* *Nora Aranberri*
IXA Group
Department of Computer Languages and Systems
Euskal Herriko Unibertsitatea (University of the Basque Country)
nora.aranberri@ehu.eus

Of course, he argues, a good translator needs a wider set of skills such as "*grammar, rhetoric, terminology, computer skills, Internet savvy, world knowledge, teamwork cooperation, strategies for getting paid correctly,* […]".

In turn, post-editing involves reviewing and correcting the output of a machine translation system to meet a pre-established quality level. Because both translation and post-editing deal with the adequate transfer of meaning, many of the skills recommended for translators will also be necessary for post-editors. Such is the case for ascertaining the target audience's needs, highlighted by O'Brien (2003). When considering the effect of technology in the translation process, be it translation memories (TM) or machine translation, Pym (2011) argues that both tasks experience the same side effect. He claims that if we simplify the process to three steps where (1) a problem is recognized; (2) alternative solutions are generated; and (3) one solution is selected, technology might speed up the generation part but slow down the selection part. According to him, the use of technology results in the loss of the flow of the text, its linearity. The source text and its possible translations are presented in segments and even paragraph marks are omitted from the translator's working environment. This causes translators to neglect the unity of the text in both computer-aided translation and post-editing.

Even if both translation and post-editing are drawn nearer with the incursion of technology, we could argue that the selection part differs considerably. Translation memory matches are correct translations of similar source segments. On the other hand, machine translation segments are likely to be incorrect translations, or at least translations that have not been checked by a human translator, of that exact source segment. In order to deal with the former, the translator must identify the difference between the stored source segment and the current source segment first, and then replace the differing part in the proposed translation. In order to address the latter, the post-editor must read the source segment and consider to what extent the proposed translation is correct and adequate for the pre-established quality-level and purpose. The idea that translators approach TM and MT proposals differently seems to be confirmed by the results in Teixeira's experiments (2014), where translators reported to be more comfortable working within an environment that provides provenance information about the segments (TM or MT). The author notes that this information seems to "[…]*increase confidence and reduce cognitive load, by giving translators a hint on how to initially approach a suggestion, as they reportedly use different strategies for different kinds of suggestions*" (Teixeira 2014: 55). Focusing on quality, Guerberof's work (2009, 2014) shows that the final quality of the target text is higher when post-editing MT than when translating from scratch and as good as the quality obtained from using fuzzy matches in the range 85-94%. Although these results are tightly linked to the quality of the MT system among other factors such as post-editing/translation experience, they can be taken as an indication that the strategies to be applied in the three tasks (translation from scratch, aided by TM and MT post-editing) might differ.

According to Offersgaard et al. (2008), post-editing requires skills specific to its set-up. In particular, a good post-editor must be able to decide in a matter of seconds whether a machine translated segment is worth editing or whether it would be more efficient to translate it from scratch. Post-editors must be conscious of speed, which means quickly deciding which modification should and should not be made. de Almeida/O'Brien (2010: 2) listed three essential skills that a good post-editor should master to work in the localization market:

> 1 – The ability to identify issues in the raw MT output that need to be addressed and to fix them appropriately. We call these "Essential Changes";
>
> 2 – The ability to carry out the post-editing task with reasonable speed, so as to meet the expectations of daily productivity for this type of activity (approximately 5,000 words post-edited per day, on average);
>
> 3 – The ability to adhere to post-editing guidelines, so as to minimise the number of preferential changes, which are normally outside the scope of PE. We call these "Preferential Changes".

If translation and post-editing are different, we cannot infer that a good translator will be a good post-editor without further training. In an experiment carried out by Offersgaard et al. (2008), they recruited translators specializing in Microsoft documentation to perform post-editing on the same type of texts. Their results revealed that the resulting post-edits did not pass the quality validation process.

Despite the evidence that post-editing and translation differ, there is no standard as to how post-editing should be taught and what type of information helps the most in becoming a good post-editor. Nowadays, in the best of cases, large companies offer minimal training to their translators, but many start post-editing with only the aid of internal post-editing guidelines. In general, these guidelines tend to be relatively vague and translators are left on their own to handle the task (see Section 2).

In the past few years several university translation programmes have acknowledged the need to go beyond the teaching of translation memories (TM) in technology modules and to introduce machine translation into the curriculum. Kenny/Doherty (2014) described a refined syllabus for providing training in statistical machine translation at post-graduate level after a first experience reported in Doherty et al. (2012) showed increased levels of student confidence and knowledge in the area of MT. Flanagan/Christensen (2014) and Konponen (2015) went a step further and included post-editing within their training programmes. Koponen (2015) mainly focused on the intricacies of classroom logistics and student attitudes. Flanagan/Christensen (2014) investigated how postgraduate students interpreted industry-focused post-editing guidelines, and they found that these guidelines can be confusing and not informative enough for student trainees.

Similar to the previous studies, Depraetere (2010) investigated which post-editing guidelines and strategies should be emphasized when training novice translators in the classroom. Rather than displaying an urge to implement preferential changes, the results showed that students trusted the MT output too much, which meant that the students often produced calques. They also failed to be consistent with their edits throughout the text, confirming the claims of Pym (2011). This is in line with the study by de Almeida/O'Brien (2010), who found that translation experience and preferential changes during post-editing correlated inversely.

Following along similar lines of research, we aimed to investigate the technical work done by professional translators who faced a post-editing task for the first time. We analysed the changes professional translators made intuitively when post-editing to help us understand their tendencies in order to learn what post-editing training and guidelines should focus on for this group of experts. In particular, we examined the post-editing results of a small subset collected during the first post-editing workshop for professional translators run for the Spanish-Basque language pair in autumn 2015. Many studies on post-editing have focused on reportedly high machine translation quality, where the quality was judged by automatic scores or human evaluations. An analysis of the post-editing activity of Spanish-Basque allowed us to start uncovering the editing work that might be carried out for moderate MT quality. Challenging the controlled quality of texts, Kliffer (2005) recommends that for training, the quality should not be too high that the texts contain scarcely any errors, and conversely not too low that correction is pointless. Thus we provided translators with MT output from currently available systems of varying quality (see Section 4.3). We believed that this would bring to light differing tasks within post-editing, which depend on the quality level of each proposed translation segment.

2. Post-editing guidelines

When translators face the task of post-editing, they are sometimes given guidelines to follow. These guidelines become the first and only reference for these professionals to complete the task. Company-specific guidelines are private and not available or publishable even for research purposes (de Almeida 2013). This author reported an exception for Microsoft style guides, which were published in the Microsoft Language Portal in 2008 (de Almeida 2013: 38). She pointed out

that among the style guides for the 30 languages covered by Microsoft only a few dealt with post-editing. Spanish was one such case. The post-editing instructions included definitions of MT and the different levels of quality accepted by the company depending on the type of project, possible solutions for lexical, grammatical and other issues to be tackled during post-editing, and indications about acceptable and unacceptable language for the different quality levels, mainly dealing with stylistic considerations. The style guides available nowadays, however, do not include any reference to machine translation or post-editing, following the secrecy trend of other companies.

Although each company is responsible for establishing what exactly is involved in the different *quality-levels* negotiated with a client, we will briefly refer to the two traditional levels distinguished by the post-editing literature, namely, quality similar or equal to human translation (often referred to as *full post-editing*) and good enough quality (often referred to as *light* or *rapid post-editing*) (Loffler-Laurian 1996). Full post-editing involves editing a target text to a high standard of quality whereas light post-editing only requires changes to respect the TL syntax and lexicon, and to structures that hinder comprehension (see Massardo et al. 2016 for more specific guidelines).

de Almeida (2013: 40) argued that post-editing guidelines provided by different companies lack detail, especially taking into account that post-editing may still be a new activity for many translators. Generally, guidelines emphasize the need to avoid unnecessary changes and stress the importance of speed during post-editing. Efficiency is reached by quickly deciding on the minimal changes to make. Guidelines tend to describe the error categories and severity levels to help post-editors decide on their relevance. de Almeida (2013) warned that few practical examples are generally provided and that it is these examples that help post-editors to better understand the categories and severity levels. She added that guidelines do not usually present a clear distinction between the two main types of post-editing either.

In a study on the post-editing of machine translation output for SAP, Schäfer (2003) proposed a definition of the tasks and cognitive skills involved in post-editing, as well as a discussion on a typology of MT errors. The outlined typology was suggested for use with different language pairs, since the author commented that there was a level of similarity among the types of post-editing corrections required for different languages. The author then provided detailed information about the post-editing guidelines developed for SAP projects. The guidelines divided the post-editing process into the following steps: general output check for identifying the main recurring issues in the MT output, such as words to be included in the dictionary; editing the MT output according to the typology of errors provided; proofreading to detect semantic errors and to ensure adequacy of style. The typology classified the errors as: lexical, syntactic, grammatical and due to defective input text. The author provided examples of these categories in different languages and concluded by mentioning that the guidelines were a work in progress, to be complemented with the introduction of controlled language in SAP projects. While the complete post-editing guidelines were not made available, this is a very useful example of how guidelines can be used to help companies make the MT cycle more efficient, and to assist linguists in the post-editing task by providing the necessary knowledge, definitions and clearly-defined error categories to be corrected.

From the descriptions of de Almeida (2013) and Schäfer (2003), we could argue that Microsoft's and SAP's guidelines seem informative for post-editors-to-be. They cover the steps that a post-editor should follow and seem to describe or provide a good set of examples. It is noteworthy how the view and focus of post-editing for each company emerges in the write-up of the guidelines. Microsoft paid particular attention to the description of quality levels and details what exactly each should entail, focusing on acceptability. SAP, in turn, viewed post-editing as the correction of specific errors, which they listed in detail, without considering different levels of quality or more global features.

Allen (2003: 313) emerged as a dissenting voice and warned that unnecessarily detailed and lengthy guidelines may cause confusion.

> [...] much energy can be wasted on (re)creating principles to tell post-editors to fix up the highly frequent, small MT raw output mistakes that unnecessarily add to the cognitive load on these experienced language experts.

As Allen pointed out, we can probably expect translators to easily spot glaring grammatical or semantic errors. However, when dealing with different post-editing levels and concepts such as acceptability, some detail as to what should be considered a light error and where the threshold lies between required style and preferential changes is probably welcomed by translators.

TAUS, a resource centre for the global language and translation industries, published a compilation of guidelines under the title "MT post-editing guidelines" (Massardo et al. 2016). In this report, they claimed that it is advisable to have some common basic guidelines that each company should tailor to its needs and contexts of use. According to TAUS, post-editing guidelines should be dynamic in that translation service providers should analyse the post-edited text and "*identify common over-edit and under-edit mistakes in order to refine post-editing guidelines and determine the workforce training needs to achieve higher productivity"* (Massardo et al. 2016: 6). To help customise the guidelines, Massardo et al. (2016: 8) listed a number of known problem areas by post-editing level which reveals additional considerations for full post-editing as compared to light post-editing. For example, both sets of guidelines address inconsistencies in terminology, morphological, grammatical and word order issues, but full post-editing includes handling lists, tables and other elements, as well as paying attention to proper names. This list is only a first step in differentiating PE levels, as the severity of morphological and grammatical errors, for example, remains to be defined.

What the guidelines reveal is that post-editing focuses on errors and changes. A post-editor is supposed to quickly identify errors and eliminate them according to a pre-established quality level. Translators are trained and have experience in producing texts of the highest possible quality without disregarding the overall purpose of the text. If we compare the two tasks, we see that the new view on correctness, usage and style are bound to interfere and confuse untrained post-editors. These properties, and quality in general, become dynamic, where emphasis is put on the function of the text and quality requirements vary depending on content type, communicative function, end user requirements, context, perishability, or mode of translation generation (O'Brien et al. 2011). Therefore, training material and guidelines should cover these aspects in order to adapt to the perspective of professional translators.

3. Aspects that influence post-editing

We saw that post-editing involves editing incorrect machine translation output efficiently in terms of time. We have recommended that guidelines should specify what edits need to be made. However, not all post-editing tasks are equal. In recent years, researchers have identified many aspects that can turn a post-editing task into a more or less demanding task, namely, MT quality, post-editor skills, translation brief/purpose, sentence length, language structures and language combination (see discussion below). Depending on the configuration of the texts, therefore, the required and expected edits might differ, increasing the complexity of post-editor training and clearly revealing the need for tailored guidelines for each context.

Koehn/Germann (2014) studied the influence the MT system has on post-editing. After testing four different systems, they reported that their best system obtained a productivity gain 20% higher than their worst system. By looking at the long pauses made by the post-editors, they concluded that a better system allowed spending less time solving harder translation problems.

Not all errors made by the MT system are equally easy or difficult to post-edit. This was confirmed by Tatsumi (2009) and Tatsumi/Roturier (2010), for example, who observed that some sentences took longer than expected to post-edit based on the count of edits alone. They reported that source sentence length and structure could explain this phenomenon. They pointed in particular at very long and very short sentences, sentence structure, incomplete sentences, and complex and

compound sentences as candidates for longer post-editing times. This is in line with the perceived post-editing effort as studied by Koponen (2012). When analysing segments rated as particularly difficult by post-editors, she found that these were long sentences requiring considerable reordering, even if in the end not much editing needed to be applied.

In terms of post-editor skills, studies have shown that post-editors differ greatly in speed (Plitt/Masselot 2010, Sousa et al. 2011). In fact, research suggests that even when post-editors tend to address the same errors, they record different speeds and number of keystrokes (Tatsumi/Roturier 2010, Koponen et al. 2012, Koehn/Germann 2014).

Considering different aspects that influence post-editing is not only important for training purposes and to set realistic expectations, but also for post-editing research. In order to allow for comparisons, it is essential that information about the different aspects is provided. This is the case for MT quality. Although more recent research has started to provide more comprehensive data (Koponen 2012), many authors report the name and version of a proprietary tool (Tatsumi/Roturier 2010, O'Brien 2006), mention the type of system used, rule-based or statistical (Depraetere 2010) or describe how the systems were developed without giving explicit evidence of level of quality post-editors were dealing with (Plitt/Masselot 2010).

4. Experimental set-up

Post-editing is no doubt influenced by a large number of variables and no study can account for them all. However, it is good practice to at least refer to such variables to be able to interpret and compare the results better. In this section, we consider the context and methods used to collect our data, providing information about the MT systems, post-editor profile, test set features and data collection.

4.1. Workshop set-up and participant profile

During the autumn of 2015, we ran the first post-editing workshop aimed at professional translators working from Spanish into Basque. Ten participants joined the workshop, all of whom had wide experience in translation (3 to 30 years) but none had ever done machine translation post-editing before. The workshop was intended to serve as a space to experience and discuss post-editing for Basque.

The workshop ran for seven weeks. We held four face-to-face sessions to present theoretical aspects around machine translation and topics on post-editing, and additionally, participants were asked to complete five short online assignments per week. They completed four post-editing tasks where they revised machine translation output and one productivity test where they combined post-editing with translation from scratch. In this paper we will focus on the post-edits of the productivity tests only (see 4.2 for further details on this decision).

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Week 1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Week 2	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓
Week 3	✓	✓	✓	✗	✓	✓	✓	✓	✓	✗
Week 4	✓	✓	✓	✗	✓	✓	✓	✓	✗	✓
Week 5	✓	✓	✓	✗	✓	✓	✓	✓	✗	✓
Week 6	✓	✓	✓	✗	✓	✓	✓	✗	✗	✗
Week 7	✓	✓	✓	✗	✓	✓	✗	✗	✗	✗

Table 1. Productivity tests completed per translator (T) across weeks

Overall, translators performed seven productivity tests. We can see from Table 1 above that most tests were completed, but not all participants completed all the tests. Data collected in week 1

were discarded from the analysis because it was the first time the translators post-edited and used the platform, and reportedly, they were more worried about learning how the platform worked and getting over the fact that they were being timed than completing their post-editing activity. In terms of task characteristics (see Table 2 below), weeks 2-5 use the *itzultzailea*[1] MT system with text extracts from a report. In week 6 they continued translating the same report but used instead Google Translate[2] to obtain the MT suggestions. In week 7, the MT system used was once again *itzultzailea* but this time the text was changed and translators worked with user guides (see Sections 4.2 and 4.3 for further details about the texts and the MT systems).

	MT system	Text type
Week 1	trial (itzultzailea)	trial (report)
Week 2	itzultzailea	Report
Week 3	itzultzailea	Report
Week 4	itzultzailea	Report
Week 5	itzultzailea	Report
Week 6	Google Translate	Report
Week 7	itzultzailea	user guides

Table 2. MT system and text type used for the weekly tasks

The tasks were completed using the TAUS DQF tool.[3] This is a web-based evaluation environment that presents post-editors with the source text segment, the previous and the next source text segments to improve context and alternately MT output for post-editing or a blank space for translating from scratch. Note that in the chosen set-up we did not avail of translation memories. At the back-end, it measures the time spent on each segment. The post-editing analytics provided by the environment include the level of reuse by calculating the proportion of MT words kept in the post-edits.

We provided only basic instructions for post-editors to perform full-post-editing following the definition of human translation quality by TAUS (Massardo et al. 2016: 18) as follows:

> Comprehensible, accurate, stylistically fine, though the style may not be as good as that achieved by a native-speaking human translator. Syntax is normal, grammar and punctuation are correct.
>
> – Aim for grammatically, syntactically and semantically correct translation
> – Ensure that key terminology is correctly translated.
> – Ensure that no information has been accidentally added or omitted.
> – Edit any offensive, inappropriate or culturally unacceptable content
> – Use as much of the raw MT output as possible.
> – Basic rules regarding spelling, punctuation and hyphenation apply
> – Ensure that formatting is correct.

4.2. The working text and the subset for analysis

During the first six weeks of the workshop, participants worked with the report "Sexism in the 2013 games and toys advertising campaign" published in Basque and Spanish by Emakunde, the Basque Institute for Women. The report includes the justification of the report and the theoretical background, the methodology used to collect the data and the analysis of the results. The report was worked on progressively from beginning to end, presenting translators with tasks of around 500 words each. In week 7, the domain was changed to user guides for electrical appliances, where they worked with instructions for a BQ mobile phone and a Bosch washing machine. The change of domain in the last week of the workshop was implemented to see, albeit briefly, whether translators found the need to re-think post-editing strategies used in previous weeks

1 *itzultzailea* is available online at http://www.itzultzailea.euskadi.eus
2 Google Translate is available online at http://translate.google.com
3 Information about the TAUS DQF Tools is accessible online at: https://evaluate.taus.net/evaluate/dqf-tools

while translating the report. A change in domain and text-type may bring a change in MT quality, as well as a change in the level of involvement translators may have towards the text. These user guides were only published in Spanish and no official Basque translations exist. Once again, the texts were split into extracts of 500 words.

Apart from the post-editing data, we collected manual assessments of the work done by the participants to incorporate final quality assessment into the post-editing analysis. Because providing assessments can be time consuming and the post-editing workload was already considerable, we proceeded as follows. Of the segments participants worked on during each of the productivity tests, we randomly selected ten segments in a way that we collected five translations and five post-edits for each participant. We asked participants to assess their colleagues' work on a scale from 1 to 10, where 1 was the lowest possible score and 10 the highest. Rather than asking them to focus on specific aspects such as accuracy or fluency, we instructed them to rate the overall quality of the segments. Because assessment is very subjective, we decided to gather three judgments per segment, that is, each participant would have five translations and five post-edits assessed by three colleagues each. In order to meet these criteria, after collecting all the relevant segments, we compiled a customized assessment set for each participant. Each assessment set included thirty segments, fifteen translations and fifteen post-edits by different colleagues, ensuring that they were not evaluating their own work. It was not possible to avoid the same translator assessing the same source segment multiple times, even though we knew that this might result in certain differences in the translations being penalised harsher than they otherwise would be. We expected that using multiple assessments and not indicating the provenance of the segments would attenuate if not avoid this potential effect. Although translators knew the brief and the context of the translation, they assessed translations at the segment level.

We decided to restrict the analysis to this subset of segments for which we have the post-edits and translations from scratch as well as the automatic and manual quality assessments. The subset contains five post-edits per productivity test completed over weeks 2-7, with a maximum of 30 post-edited segments per translator.[4] Table 3 below shows the variation in the average segment length per productivity test subset. Subset 2 has a higher average, whereas subsets 3, 4, 5 and 6 have a similar average of about 17 words. Again, subset 7 has a considerably lower average of 7 words. The difference in average lengths, particularly those in subsets 2 and 7, affect the number of required changes and therefore, this was taken into account when drawing conclusions on the changes applied during post-editing.

Productivity test subset	2	3	4	5	6	7
Average sentence length	30.70	17.40	16.70	17.75	18.20	7.20

Table 3. Average sentence length per productivity test subset

4.3. The machine translation systems

While the use of machine translation for languages such as English and Spanish, for example, is quite advanced with considerably good results from generic systems such as Google Translate or from customized company-specific systems, the availability and quality of systems lag behind for minority languages such as Basque (Aranberri 2015, Dušek et al. 2016). Still, large multinationals are starting promising attempts to include Basque among the languages covered by their systems and free online systems such as Google Translate are emerging as a potential tool for translators, despite concerns about information security and confidentiality. Such is also the case with *itzultzailea*, a Spanish-Basque system funded by the Basque Government and powered by Lucy Systems. It is a rule-based system that was first made available to the public in 2010 and is main-

4 Note that translators 1-5 and 6-10 performed the opposing post-edits and translations in order to calculate a fairer productivity score.

tained by the company in question. Although exhaustive tests have not been carried out, the experience of the IXA research group[5] shows that it equals or surpasses the quality of Google Translate in a number of domains.

The MT systems used during the workshop were (mainly) *itzultzailea* and Google Translate, both generic systems. Clearly, the quality of a generic system is not comparable to a carefully customized system, but accessing such a system was out of the reach of the organizers. Besides, by working with a generic tool, the participants would be able to experience the current quality of machine translation for Basque and judge its usability for their daily work. Table 4 below shows the average MT quality per productivity subset according to BLEU (Papineni et al. 2002) and TER (Snover et al. 2006).[6] The scores for these metrics were calculated using the freely available online version of Asiya, an evaluation toolkit.[7] We chose to provide these two metrics because they represent two distinct models used in MT evaluation. Although both are string-based metrics, that is, they compute the difference between a machine translated segment and a reference translation at word-form level, BLEU calculates the precision, to put it simply, the number of words in the MT output that match the words in the reference translation, and TER focuses on computing the minimum insertions, deletions, substitutions and shifts required to transform the MT output into the reference. Despite the controversy it generates, BLEU is the most widespread metric within MT research, and TER can be said to mirror more closely the actual task of post-editing. Note that BLEU calculates the similarity between the MT output and the reference translation, and therefore, the higher the score, the more similar they are. However, TER calculates the number of changes required; therefore, the lower the score, the better the MT output is assumed to be. In order to compare the scores more easily, Asiya displays -TER scores, that is, it converts the regular TER scores into negative numbers so that both metrics can be interpreted in the same direction, i.e., the higher the better.

Productivity test subset	Machine translation quality – BLEU	Machine translation quality – -TER
2	0.18973352	-0.61110188
3	0.156408933	-0.617633685
4	0.14556475	-0.71267840
5	0.38790532	-0.59278409
6	0.08816973	-0.79133609
7	NA	NA

Table 4. BLEU and TER automatic machine translation quality scores per productivity test subset

If we look at BLEU scores, we see they are overall quite low, ranging from 0.08 to 0.38. It is not easy to map specific BLEU scores to quality levels, as scores vary greatly depending on the test set and the target language. In fact, strictly speaking, both TER and BLEU can only tell us how similar the MT output is to one specific reference and low scores do not necessarily equal poor quality MT output. We do not know how much improvement in quality an increase of one BLEU point brings. Therefore, rather than assigning specific quality values to scores, the metrics are mainly used to compare sets or systems.

TER scores are more intuitive in the sense that they indicate actual changes the MT output requires to be turned into the reference translation. However, these scores also need to be interpreted with caution, as the optimum combination of changes for the algorithm, which uses no notion of language, is not always the most intuitive combination for a translator, and all changes count equally, that is, no severity hierarchy is applied. Additionally, in our case, we are comparing an

5 The website of the IXA research group is accessible online at: https://ixa.si.ehu.es/Ixa

6 There are no Basque reference translations for the set used in productivity test 7 and therefore automatic metrics could not be calculated.

7 The Asiya Open Toolkit for Automatic Machine Translation (Meta-)Evaluation is available online at http://asiya.cs.upc.edu/

MT output to a reference that was created without the MT output in mind, and therefore, the need for a high number of edits will not necessarily mean that the MT output was incorrect. Rather, TER will represent the edits required to transform the MT output into a valid translation version, without having to be the version that is the closest to the actual MT output. For our subsets, the scores range from -0.59 to -0.79, meaning that for every 100 words, it is necessary to make 59-79 edits.

Despite the inability of the two metrics to establish a specific level of quality, we have compared the scores with those reported in other experiments for Basque (Díaz de Ilarraza et al. 2008, España-Bonet et al. 2011, Labaka et al. 2014, Aranberri et al. 2015, 2016) and confirmed that they are in the same range. Given the impact MT quality has on post-editing opportunities, we hoped to provide workshop participants with MT quality similar to that attainable by state-of-the-art generic systems for the Spanish-Basque language direction such as that reported in Labaka et al. (2014). In terms of system differences, we observed that in week 6, when Google Translate was used, scores were substantially lower, confirming the experience of the IXA research group and the perception of workshop participants, who complained about the decrease in quality after completing the tasks for week 6.

4.4. Methodology for analysing post-editing activity

Many studies on post-editing have so far focused on either analysing the post-editing effort and quality or comparing post-editing with translation from scratch. When investigating effort, and mostly based on the work by Krings (2001), experiments have focused on three aspects: temporal effort, cognitive effort and technical effort.

The temporal aspect is the one that companies are concerned with the most. The quicker a translation is produced, the more profitable it is to post-edit. Temporal effort has been studied by many researchers, including O'Brien (2005), Specia et al. (2009), Tatsumi (2009), Specia (2011) and Carl et al. (2011). Measuring the cognitive effort of post-editing, however, is a more complex task. So far researchers have tackled it with keystroke logging (Krings, 2001, O'Brien 2005, Carl et al. 2011) and gaze data (Carl et al. 2011), such as pauses and fixations (O'Brien 2005), and other techniques such as choice network analysis and think-aloud protocols (O'Brien 2006). All these methods are relatively tangential or subjective and require the analysis of vast amounts of data. Measuring the technical aspect of post-editing effort is more straightforward. Researchers have counted keystrokes and cut-and-paste operations (Krings, 2001, O'Brien 2005, Carl et al. 2011) or measured the edit distance between the raw MT and post-edited version automatically (Tatsumi 2009, Temnikova 2010, Specia/Farzindar 2010, Specia 2011, Blain et al. 2011).

In this work, we do not aim to perform a quality assessment of the post-edited versions or focus on the temporal, cognitive or technical performance, but rather, we aim to investigate the types of changes professional translators introduce when post-editing for the first time in order to describe the intuitive behaviour of these professionals and identify their tendencies for over-editing the machine translation output. The DQF environment does not log keystrokes or visualize the exact changes made to the MT output. Therefore, we decided on a methodology to manually analyse the post-edits and annotate the changes.

There is still no generalized methodology to annotate post-editing activity, but following the example of de Almeida/O'Brien (2010) we customized a typology of changes that best suited our experimental set-up, that is, a typology that combined MT errors and post-editing activity. Similar to de Almeida/O'Brien (2010), we used the three high-level categories Essential changes, Preferential Changes and Essential Changes not Implemented, but restricted the subcategories that are replicated under each of them to grammar and lexical choice. de Almeida/O'Brien (2010) used the main categories from LISA QA, the error typology of the former Localization Industry Standards Association (2009), namely, Mistranslation, Accuracy, Terminology, Language, Style, Country and Consistency; types from GALE's Post Editing Guidelines for GALE Machine Translation

Evaluation (2007), namely, Extra information in MT output, Information missing from MT output, Adjectives, Adverbs, Capitalisation, Determiners, Phrasal ordering, Prepositions, Pronouns, Proper names, Punctuation, Spelling, Verb tense, Decimal points and Quotation marks; and a few additional categories defined by themselves, namely, Format, and the subcategories Gender and Number (under the main category Language).

We considered the LISA QA categories too broad and specific to localisation. We also found that the GALE categories were not comprehensive enough, as not all grammatical categories were included (nouns or conjunctions were not considered) and even for the categories that were included, such as verbs, focus was only paid to specific aspects, such as tense, disregarding other properties such as aspect or person agreement. Furthermore, we aimed to not only record changes but to also uncover why they were made. The categories proposed by LISA QA and GALE by themselves do not indicate the type of modification applied during post-editing. As a result, we proceeded as follows: when annotating the changes, we noted the grammatical category of the word or the name of the structure in question, classified it as either a grammatical or a lexical change, and registered it as a replacement, reordering, addition or deletion, following the type of edit actually introduced by the translators. Given the moderate quality of the MT output, we expected that some segments would require heavy editing and added an auxiliary category for complete edits. We considered that a segment was completely edited when the wording of the post-edited version differed considerably from the suggestions produced by the system, or when it did not follow the main structure proposed in the MT output, e.g., a conditional sentence was transformed into a declarative sentence. Note that the latter case does not necessarily involve a high number of edits.

Identifying a grammatical error or an incorrect word choice is straightforward. When aiming for a human quality translation, therefore, classifying changes to improve these cases as *necessary* is easy. However, deciding whether changing a rare or unnatural structure, or a near-synonym in a particular context is *necessary* or *preferential* can be difficult – for both the post-editor and the evaluator. When these cases emerged during the analysis, we reviewed the reference translation and considered the structures and word choice in these segments as the preferred options. If the reference translation used the structures and lexis that appeared in the MT output, we considered that they did not need to be changed. We classified any change to an element in the MT output that appeared as such in the reference as *preferential*.

5. Analysis of post-edited data

5.1. Analysis of post-editing changes

In this study, the grammatical and lexical edits made to the MT output by the workshop participants reveal valuable weaknesses of the system for developers. Nonetheless, what is interesting in relation to developing training programmes and guidelines for post-editors is to identify which preferential changes translators perform intuitively. A well-trained, efficient post-editor should avoid or at least minimize these changes.

If we consider the types of changes made by the translators, replacements were the most frequent (64%) followed by reorderings (18%) with rare cases of additions (7%) and deletions (11%). The significantly higher share of replacements was expected, especially when working with a rule-based system like *itzultzailea*, which tends to output a potential equivalent for every source word or structure –syntax-based systems often include equivalence rules that go beyond the word level and work at phrase and sentence structures.

Below is a qualitative breakdown of the changes performed by the translators and analysed by the author. We first focus on grammar issues and then on lexical issues. We classified as essential grammar changes those modifications that addressed ungrammaticality or such a rare rendition of a structure that, despite not being able to describe it as ungrammatical, it is odd to a native speaker (see Example 1).

Source: Las ocupaciones-profesiones de ama de casa, peluquería/estética y modelo fueron las más representadas en los anuncios protagonizados por niñas, en correspondencia a los arquetipos que se señalarán más adelante.

Gloss: The professions that were represented the most in advertisements with a leading female role were that of stay-at-home mum, hairdresser/beautician and model, in line with archetypes that will be discussed later.

MT: Etxekoandre-okupazio-bizibideak, ile-apaindegi/estetika eta modeloa neskek protagonista izandako iragarkietan irudikatuenak izan ziren, aurrerago seinalatuko diren arketipoetarako korrespondentzian.

Post-edit: Etxekoandreak, ile-apaintzaile zein estetizistak eta modeloak ziren neskak protagonista izandako iragarkietan irudikatuenak, aurretik seinalatu diren arketipoekin bat etorriz.

Example 1: Example of an essential grammar change where the incorrect ergative mark was changed to an absolutive mark. (The post-edit includes additional changes)

We classified as preferential grammar changes modifications to grammatical renditions that did not alter the meaning of the text and were clearly conveying the message as such (see Example 2).

Source: Ya hemos visto lo que nos dicen los cuentos:
Gloss: We have already seen what stories tell us:
MT: Jada ipuinek esaten digutena ikusi dugu:
Post-edit: Jada ipuinek zer dioten ikusi dugu.

Example 2: Example of a preferential grammar change where a correct relative clause was turned into an indirect question

Essential grammar issues addressed

Grammatical errors in the Basque translations were abundant given the distance between the source and target languages and the level of development of the system (see Table 5 below). Post-editing changes in this category mainly dealt with the different ways to translate compound nouns and represent modifying elements, the complexity of Basque auxiliaries and the large number of postpositions, clause markers and case-markers that the MT systems need to deal with.

Incorrect grammar that required replacements	
Compound nouns	– noun-noun structures, genitive-noun structures
Adjectives	– noun complements, relative clauses, adjectives
Determiners	– demonstratives, numerals, articles
Number	– singular, plural
Main verbs and auxiliaries	– tense, mood, valency
Case markers and postpositions	– ergative, absolutive, dative, genitive, genitive-locative, postpositions
Adverbial clause markers	– purpose, time
Adverbs and postpositions	
Comparative and superlative structures	
Coordinated noun phrases	

Table 5. Types of essential grammar replacements found in post-edits

Reordering errors are bound to appear in MT output as a result of the differing word order between Spanish and Basque and are exacerbated by the relatively free phrase-level order of the latter (see Table 6 below). In particular, translators dealt with internal reorderings of the elements in

noun phrases and the tendency of the *itzultzailea* MT system to position the main verbs at the end of the sentence. This is considered the canonical position for verbs in Basque but it is not favoured in long sentences as it increases the cognitive load of the reader.

Incorrect grammar that required reordering
Internal reordering of noun phrases
Subjects
Objects
Internal reordering of verb phrases
Verbs

Table 6. Types of essential grammar reorderings found in post-edits

Additions were not very frequent among the changes introduced by translators and mainly occurred in segments addressed in week 6, when using Google Translate's output. Being a statistical system, Google Translate relies completely on the word forms present in its training corpus. For agglutinative languages such as Basque, sparsity issues are common, that is, often corpora fail to include possible combinations of lemmas and affixes, or complex verbal forms for the different persons, tenses and paradigms. Therefore, it is not surprising that some affixes or elements of the verbal periphrasis were omitted. Similar to additions, deletions were also rare (see Table 7 below).

Incorrect grammar that required additions	Incorrect grammar that required deletions
addition of case markers	deletion of auxiliaries
addition of genitives and genitive-locatives	deletion of case markers
addition of commas	deletion of determiners
addition of adverbs	deletion of possessive pronouns
addition of auxiliary verbs	deletion of subordinate markers
addition of completive markers	
addition of particles	
addition of verb phrases	

Table 7. Types of essential grammar additions and deletions found in post-edits

Preferential grammar issues addressed

One of the most recurring changes was that of the representation of modifiers, which can be rendered as relative clauses, genitival structures, compound nouns or adjectives (see Table 8 below). Translators did not seem to agree with the MT system's choice and often modified one structure over the other (see Example 3 below). Another set of preferential changes emerged from verbs, where translators applied changes to mood and tense, or even alternated between periphrastic and synthetic forms of verbs. Other changes included replacing postpositions, determiners, adverbs and the number, and varying capitalisation. Both the structures output by the MT system and the ones modified by the translators were grammatically correct and there was no apparent reason to apply edits aside from individual preferences.

> **Source**: No hacen clara apología de la violencia hacia las mujeres, pero responden asegurando que muchas de las denuncias por malos tratos son falsas y que los hombres también son víctimas de violencia en la pareja, pero que esa realidad se oculta en una 'conspiración de género".

Gloss: They don't defend violence towards women, but respond convinced that many of the allegations of assault are false and that men are also victims of violence within the couple, and that this reality is hidden by a "gender conspiracy".

MT: Ez dute emakumeengananzko bortizkeriaren apologia argirik egiten, baina tratu txarrengatiko salaketa asko faltsuak direla eta gizonak bortizkeria-biktimak direla ere bikotearengan, baina errealitate hori genero-konspirazio batean ezkutatzen dela ziurtatuz erantzuten dute. .

Post-edit: Ez dute emakumeen aurkako bortizkeriaren apologia argirik egiten, baina tratu txarrengatiko salaketa asko faltsuak direla eta gizonak bortizkeriaren biktima ere badirela erantzuten dute, baina errealitate hori 'genero-konspirazio' batean ezkutatzen dela ziurtatutzen dute.

Example 3. A preferential grammar change where a compound noun was turned into a genitival structure. (The post-edit includes additional changes)

Preferential replacements of correct grammar	
Modifiers	– relative clauses, compound nouns, genitival structures
Verbs	– mood, tense, periphrastic-synthetic
Postpositions	– demonstratives, numerals, articles
Number	– singular, plural
Adverbs	
Determiners	
Capitalisation	

Table 8. Types of preferential grammar replacements found in post-edits

As mentioned above, Basque has a free phrase-level ordering – except for the sentence focus, which must be located immediately before the verb. This freedom often results in each translator developing preferences for the optimum position of specific phrases so as to ensure that the message of the text is properly transmitted to the reader (see Table 9 below). Most of the preferential reorderings were due to this freedom. The phrases edited by the translators were correct but there was no apparent reason to relocate them. We can only guess that translators probably considered that comprehensibility was increased by applying the changes, although this was not obvious to the evaluator.

Preferential reorderings of correct grammar
reordering of adjective list
reordering of adverb
reordering of attribute
reordering of genitive-locative
reordering of noun phrase
reordering of object
reordering of subject
reordering of verb

Table 9. Types preferential grammar reorderings found in post-edits

Preferential additions and deletions were rare, with changes probably applied to improve the fluency of the segments (see Table 10 below).

Incorrect grammar that required additions	Incorrect grammar that required deletions
addition of demonstratives	deletion of auxiliaries
	deletion of coordination and head

Table 10. Types of preferential grammar additions and deletions found in post-edits

Essential lexical issues addressed

Let us now consider lexical changes. The essential lexical issues addressed during post-editing were replacements of nouns, verbs and adjectives. Often, the MT systems output the incorrect translation equivalent for polysemic words for the particular context and domain of the source text. This was the case for *mayor*, which can mean bigger (*handiago*) or older (*nagusiago*), and *desear*, which can mean desire (*desio*) or want (*nahi*), for example. Also, the systems output word-by-word translations for figurative uses of language that did not work in the target language. For example, in the source sentence the verb *atravesar* (cross, go through) is used to refer to a pattern that occurs at different departments within a company. The direct translation equivalent *zeharkatu* does not work in the target language, and therefore translators needed to find an alternative equivalent that carries the meaning across. An interesting case, specific to the particular context of gender studies, was the translation of the noun *niños* (boys). Spanish uses the masculine gender in the plural to refer to either a group of males or a group of males and females. Since Basque does not display grammatical gender, it tends to have different words to refer to males, females and mixed groups. The MT systems output the gender-neutral word *haurrak* (children) for every instance of the word *niños*, which translators had to painstakingly change throughout the post-editing of the Emakunde report.

Preferential lexical issues addressed

Because we included all changes in structure, syntax and morphology within the grammar category even when some might also carry a change in meaning, here we mainly discuss preferential changes to the lexis, that is, the lemmas chosen by the MT systems for the translations. Interestingly, we found changes in all open-class categories, nouns, adjectives, verbs and adverbs. Translators replaced words that carried the meaning of the source perfectly and were appropriate for the context – the words appeared in the reference translations – with synonyms they seemed to prefer over the MT output. For example, we found replacements such as *joera_izan→ohi_izan* (to tend to be→to usually be), *gai_izan→kapaz_izan* (to be able→to be capable), *bizibide→lanbide* (job→profession), *biziki→sakonki* (very much→deeply), among others. These preferential changes result in correct translations but add an unnecessary load to post-editing in terms of time and technical effort. It remains to be studied how allowing or prohibiting preferential changes affects the overall performance and efficiency of translators and their degree of satisfaction with the final text.

Complete edits and unedited segments

In many cases, the number of edits made to the segments was such that it was impossible – and probably worthless – to track the modifications and the reasoning behind them. These segments were examples of particularly poor MT quality and all translators edited these segments heavily. Interestingly, we observed two trends when dealing with these segments (see Example 4 below). When the MT system failed to start the segment correctly, the whole segment tended to be disfigured and translators tended to disregard it completely. When the MT output started correctly and then deteriorated, translators tended to reuse the beginning of the segment and complete the remaining with their own preferences. Often, a considerable number of single elements were reused in their translations. In line with the conclusions from Koponen (2013), this could suggest that the

output of the MT system is used to set the main structure of the segments, resulting in translations that are more alike as compared to those produced in the former cases.

Source: En los anuncios con protagonista principal masculino puede apreciarse que son la no-interacción, la amistosa, más comunes, seguidos de la no-amistosa.

Gloss: In the advertisements with a leading male role, we can observe that the no-interaction and the friendly interaction are more common, followed by the unfriendly interaction.

MT: Ez-interakzioa, lagunartekoa, direla iragarkietan protagonista nagusi maskulinoarekin komunago estima dezake, ez-lagunarteko jarraituak.

Post-editing: Protagonista nagusia maskulinoa den iragarkietan, ez-interakzioa, adiskidantzazkoa dira ohikoenak, eta, ondoren, ez-adiskidantzazkoa.

Post-editing: Protagonista maskulinoko iragarkietan ikus daiteke ez-interakzioa, lagunartekoa, ohikoagoa dela, eta gero ez-lagunartekoa,.

Post-editing: Protagonista nagusia maskulinoa duten iragarkietan ikus daiteke interakziorik eza, lagunartekoa, direla ohikoenak, atzetik lagunartekoa ez dena dagoelarik.

Post-editing: Protagonista nagusi maskulinoa duten iragarkietan, harreman ohikoena elkarreragin eza da, eta jarraian ez-lagunartekoa.

Source: Sobre un total de 54 anuncios mixtos, un 18% de los eslóganes utilizaron únicamente el masculino.

Gloss: Out of a total of 54 mixed advertisements, 18% of the slogans used the masculine gender only.

MT: 54 iragarki mistoak dira guztira, leloak% 18k bakarrik gizonezkoa erabili.

Post-editing: 54 iragarki mistotako leloen % 18n maskulinoa baino ez zen erabili.

Post-editing: 54 iragarki misto hartuta, esloganen %18tan bakarrik maskulinoa erabili zen.

Post-editing: Guztira 54 iragarki mistoetatik, leloen % 18k maskulinoa bakarrik erabili zuten.

Post-editing: 54 iragarki mistotatik, % 18ren leloek erabiltzen dute maskulino hutsa.

Example 4. Heavily re-edited segments

It is worth noting that a few segments – not more than 15% – required no changes. There is a minimum number of unedited segments shared by all translators, while some translators left one or two additional segments unedited. These segments were either simple sentences with at most one simple subordinate clause or stand-alone short noun phrases (see Example 5 below).

Source: La belleza está en el interior, siempre que seas el hombre.
MT and post-editing: Edertasuna barnean dago, baldin eta gizona bazara.
Gloss: Beauty in is the inside, provided that you are a man

Source: Formatos de fotografías
MT and post-editing: Argazki-formatuak
Gloss: Picture formats

Example 5. Unedited MT output segments

The essential grammar and lexical changes listed reveal the work done by translators that is dependent on both the quality of the MT system used and the language-pair configuration. Aiming for publishable quality and given the text type and domain, the grammatical errors were glaring to translators and they did not hesitate to correct them. Still, preferential changes crop up even in our small subset. Table 11 below shows that essential changes were substantial as compared to preferential changes. If we discard the unchanged segments and complete edits from the count, on average, translators performed 2.5 essential changes per segment and about 1 preferential change

in every other segment. To this, we would add that, on average, about 17% of the segments were unchanged and about 21% were completely edited.

	T1	T2	T3	T5	T6	T7	T8	T9	T10
Essential changes	59	28	55	71	66	44	46	33	30
Preferential changes	16	15	8	18	5	1	2	2	9
Complete edits	6	7	3	4	10	11	5	1	3
Unchanged segments	5	6	7	5	4	3	4	1	3
Total segments post-edited	30	30	30	30	30	25	20	10	15

Table 11. Edit counts per type and translator

5.2. Level of reuse and MT quality

The main aim of a post-editing task is to maximise the reuse of the MT output – to the extent to which the MT output quality allows – to optimize translation productivity. Our test results display a wide range of reuse. We ran Spearman's correlation to determine the relationship between reuse and MT quality as measured by BLEU and TER (for our test-set BLEU and TER scores showed a strong correlation of .69). Not surprisingly, we found moderate correlation (.41) between reuse and BLEU scores and strong correlation (.61) between reuse and TER, meaning that the better the MT output quality, the more translators reused. This seems to indicate that professional translators are able to distinguish between different MT output qualities and reuse accordingly. This is good news for novice post-editors, as it seems that the will to reuse is there. However, we noted that the translators had a tendency to introduce unnecessary preferential changes.

MT output often displayed disfluent and inaccurate language. One of the difficulties reported by the workshop participants was that of properly and efficiently connecting the correct pieces of MT output with the new contributions added by participants themselves. Discussion among workshop participants and data analysis suggest that when the MT quality was only moderate, the work of post-editing changed from correcting mistakes to reusing good sequences and discarding the rest. The first thing participants did was to read the MT output and quickly decide whether to use it or not. When they decided to use it, they needed to identify the exact pieces and sequences they would reuse and think about the connecting pieces they needed to fill in – and delete. This included replacing lexical items, but what seemed to require most effort was sewing everything together so that all the syntactic and morphological requirements were met while the meaning was transferred completely. It would be natural to think that translators who are not used to the task of post-editing would find this additional exercise overwhelming. The qualitative analysis of the post-edits shows that the post-edited segments contain errors, which appear in 3-40% of the segments depending on the translator (see Table 12 below).

	T1	T2	T3	T5	T6	T7	T8	T9	T10
post-edited sentences	30	30	30	30	30	25	20	10	15
sentences with errors	3	7	12	4	1	4	3	4	3
% sentences with errors	10	23	40	13	3	16	15	40	20

Table 12. Post-edited segments with errors per translator

A look at the errors revealed that they were mostly agreement mistakes or incorrectly transferred meaning. They were possibly due to the willingness to reuse the MT system output and/or to work locally, that is, by addressing fixes at sub-sentential units while neglecting the sentence – and broader text – as a complete unit of translation. This probably ties in with the local and global translation strategies defined by Bell (1998), that is, those applied at text-level and those applied at specific segments. Strategies are not applied evenly as post-editors focus on local strategies and overlook global strategies. Similarly, we seem to observe what Pym (2011) describes as one of re-

sults of introducing technology into the translation process: a shift in translation focus which neglects linearity. Our test set shows that professional translators post-editing for the first time with moderate quality MT systems focused excessively on phrase-level strategies and failed to meet the required translation level and language requirements. We classified the errors found in the test set into two categories:

Target-language errors: These included incorrect sentence-level agreements such as incorrect valencies in auxiliary verbs (Example 6) and incorrect case markers (Example 7).

Translation errors: The sentences that contained translation errors were correctly formed in the target language but the meaning was transferred inaccurately. Examples of these were the incorrect arrangement of coordination heads (Example 8) and generic uses of terminology (Example 9).

> **Source**: En algunos colegios se desarrollan estrategias como limitar la posibilidad de jugar al fútbol (a un día a la semana, por ejemplo), sugiriendo alternativas que permitan a niñas y niños jugar juntos o, al menos, disponer del mismo espacio.
>
> **MT**: Ikastetxe batzuetan estrategiak garatzen dira futbolean jokatzeko aukera mugatzea bezala (asteko egun batera, adibidez), neskei eta umeei jokatzea batera baimentzen duten alternatibak iradokiz edo, gutxienez, espazio bera ukatea.
>
> **Post-editing**: Ikastetxe batzuetan estrategia jakin batzuk garatzen dira, hala nola futbolean jokatzeko aukera mugatzea (astean egun batean soilik jolas daiteke, adibidez), neska-mutikoei elkarrekin jolastea baimentzen duten alternatibak iradokiz, edo, gutxienez, espazio bera eduki dezaten.
>
> Comment: Because the verb *permitir-baimendu* (allow) has a subject (*alternativas-alternatibak*-alternatives), a direct object (*jugar-jokatu*-play) and an indirect object (*a niñas y niños-neska-mutikoei*-to boys and girls), the translator should have used a ditransitive auxiliary verb, *dieten*, rather than the transitive *duten*

Example 6. Incorrect valency in auxiliary verb

> **Source**: El presente gráfico muestra que en los anuncios protagonizados por niños el protagonista tiende a ser de mayor edad que en los mixtos o en los protagonizados por niñas.
>
> **MT**: Orain grafikoan erakusten du umeek protagonista izandako iragarkietan protagonistak joera duela mistoetan edo neskek protagonista izandakoetan baino adin handiagoa izateko.
>
> **Post-editing**: Grafiko honek erakusten du mutilek protagonista izandako iragarkietan protagonistak joera duela adinez nagusiagoa izaten mistoetan edo neskak protagonista izandakoetan baino.
>
> Comment: The verb to star or to have a leading role in an advertisement is an intransitive verb in Basque (*protagonista izan*-to be the lead) and therefore the subject should be marked with the absolutive plural marker *–ak* and not the ergative plural marker *–ek*.

Example 7. Incorrect case marker

> **Source**: En el estudio 'Violencia de género en las relaciones de pareja de adolescentes y jóvenes de Bilbao', tanto chicos como chicas se muestran reticentes a reconocer que la violencia de género es un problema social grave.
>
> **MT**: 'Genero-Indarkeria nerabe-bikote erlazioetan eta Bilboko gazteengan' ikerketan, bai mutilak bai neskak genero-indarkeria arazo sozial larria dela aitortzearekin uzkur azaltzen dira.
>
> **Post-editing**: 'Genero-Indarkeria nerabeen bikote erlazioetan eta Bilboko gazteengan' ikerketan, bai mutilak bai neskak ez daude prest genero-indarkeria arazo sozial larritzat jotzeko.
>
> Comment: Coordinating structures are often ambiguous and translators have to resort to common sense to interpret them correctly. Grammatically, the coordination of the MT system is correct ([in young couples] and [in young people from Bilbao]), but in terms of meaning (and confirmed by the reference translation) the translator should have rearranged it to reflect that both couples and young people are from Bilbao.

Example 8. Incorrect coordination

Source: Entre las exhortaciones más comunes en los anuncios protagonizados conjuntamente por niños y niñas, destacó la utilización de 'descubre' y 'diviértete', figurando en un 30% de los cortes.

MT: Haurrak protagonista konjuntzioa iragarkietan admonitions ohikoena artean, 'ezagutu' eta 'fun', mozketak% 30a azaltzeagatik erabilera nabarmendu zuen.

Post-editing: Haurrak protagonista diren iragarkietan ohikoenak dira: 'ezagutu' eta 'jolastu', % 30eko erabilera nabarmendu zen.

Comment: When translating a gender-related article, much care needs to be taken with gender-related words. In this case, the source sentence was referring to cases where both boys and girls appeared in advertisements together. However, the translation using the generic word *haurrak* (children) does not emphasize this aspect.

Example 9. Incorrect valency in auxiliary verb

One could hypothesise that reusing more MT output could have resulted in post-edited sentences of lower quality. To check this, we measured the relationship between the level of reuse and peer assessment scores and saw a very weak correlation (.19). This indicates that the post-edited sentences produced by each translator were of similar quality regardless of the amount of editing done.

6. Conclusions

Current post-editing guidelines often fail to address the views and doubts of translators and the training curricula are still under development, trying to stay up to date with the new discoveries in post-editing processes. In this work, we contributed to the investigation of post-editing activities by studying the work of professional translators who tackled post-editing for the first time.

In particular, we collected post-editing data for the Spanish-Basque language pair. MT systems for this language combination are gradually reaching a level of maturity sufficient to start attracting users' attention. Translation service providers as well as freelance translators are starting to experiment with such systems (Aranberri 2016). In this context, this work presented the first attempt at analysing the post-editing job involved in dealing with freely available generic machine translation systems, whose key aspect is their moderate quality.

According to the analysis of the post-editing activity, translators mainly perform essential changes which include correcting both grammatical and lexical issues. However, the data suggest that they also tend to over-edit the MT output, particularly in cases where the target language offers alternative structures or synonyms to render a particular concept.

The findings seem to suggest that translators identify errors and apply essential changes diligently. However, participants claimed that it would be useful if examples of potential errors were also presented during training and in post-editing guidelines in general. This would allow post-editors to familiarize themselves with the errors and then be able to identify (and solve) such issues more quickly during future post-editing tasks.

The analysis shows that translators also made preferential changes, which can be time-consuming and should therefore be avoided. Yet, a number of questions remain to be answered before suggesting training options. It is not clear whether the number of preferential changes identified in the test set lowers productivity time significantly. It would be useful to know whether the changes were the result of a long thought process or the result of years of translation experience, with preferences emerging fast, almost unconsciously and with minimal cognitive load.

A revealing conclusion from the workshop was that when working with moderate MT quality, the post-editing activity is transformed into a "patchwork" one. Contrary to studies carried out on mainstream language pairs with relatively high MT quality and where the main task is to identify and fix errors, post-editors working on moderate MT quality need to identify the machine translated elements to reuse and connect them using their own contributions. Despite the reported difficulty in sewing all the pieces together while ensuring a correct grammar and transfer of meaning,

our limited analysis seems to indicate that the amount of reuse correlates with translation quality rather than encouraging clumsy translations.

The transformation of post-editing into a "patchwork" activity also has clear implications for training. The findings suggest that post-editors dealing with moderate MT quality need to be made aware that at times looking for errors might not be the most efficient approach to post-editing. Rather, they should aim to read the source segment and its MT proposal, quickly consider the final translation and begin typing, reusing, reordering, adding and deleting word sequences without hesitation. Post-editing guidelines and training should emphasize this shift in perspective. In this scenario machine translation output functions as an aid to translation rather than a proposal that has to be reused at all costs.

Acknowledgements

The author would like to thank the two anonymous reviewers and the editors for their invaluable feedback on the initial manuscript, as well as all workshop participants, and Ane López and Carlos del Olmo, who helped deliver the face-to-face sessions.

7. References

Allen, Jeff 2003: Post-editing. In Somers, H. L. (ed.), *Computers and translation: a translator's guide*. Amsterdam and Philadelphia: John Benjamins Publishing Company, 297-316.

Aranberri, Nora 2016: Is there room for post-editing into Basque? In *Senez* 47, 195-203.

Aranberri, Nora 201: SMT error analysis and mapping to syntactic, semantic and structural fixes. In *Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-9)*, 30-38. Denver: Colorado. June 4, co-located with NAACL 2015.

Aranberri, Nora/Labaka, Gorka/Díaz de Ilarraza, Arantza/Sarasola, Kepa 2015: Exploiting portability to build an RBMT prototype for a new source language. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, 3-10. Antalya: Turkey, May 11-13.

Aranberri, Nora/Labaka, Gorka/Díaz de Ilarraza, Arantza/Sarasola, Kepa. 2016: Ebaluatoia: crowd evaluation for English–Basque machine translation. *Language Resources and Evaluation*, 1-32. doi:10.1007/s10579-016-9335-x.

Bassnett, Susan 1991: *Translation Studies*. Routledge, London and New York.

Bell, Roger I. 1998: Psychological/cognitive approaches. In *Routledge Encyclopedia of Translation Studies*, edited by Mona Baker. London & New York: Routledge.

Blain, Frederic/Senellart, Jean/Schwenk, Holger/Plitt, Mirko/Roturier, Johann 2011: Qualitative analysis of post-editing for high quality machine translation. In *MT Summit XIII: the Thirteenth Machine Translation Summit* [organized by the] Asia-Pacific Association for Machine Translation (AAMT), 164-171. Xiamen: China, September 19-23.

Carl, Michael/Dragsted, Barbara/Elming, Jakob/Hardt Daniel/Jakobsen, Arnt Lykke 2011: The process of post-editing: a pilot study. In *Proceedings of the 8th international NLPSC workshop. Special theme: Human-machine interaction in translation*, 131-142, Copenhagen Business School, Copenhagen Studies in Language 41. Frederiksberg: Samfundslitteratur, August 20-21.

Cronin, Michael 2013: *Translation in the Digital Age*. Oxfordshire, UK: Routledge.

de Almeida, Giselle 2013: *Translating the post-editor: an investigation of post-editing changes and correlations with professional experience across two Romance languages*. Doctoral dissertation, Dublin City University.

de Almeida, Giselle/O'Brien, Sharon 2010. Analysing post-editing performance: correlations with years of translation experience. In *Proceedings of the 14th annual conference of the European association for machine translation*. St. Raphaël: France, May 27-28.

de Sousa, Sheila C. M./Aziz, Wilker/Specia, Lucia 2011: Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. In *Recent Advances in Natural Language Processing (RANLP-2011)*. Hissar: Bulgaria.

DePalma, Donal A./Pielmeier, Hélène/Henderson, Stephen/Stewart, Robert G. 2015: The Language Services Market: 2013. *Common Sense Advisory*. Boston: USA.

DePalma, Donald A./Hegde, Vijayalaxmi/Pielmeier, Hélène/Stewart, Robert G. 2013: The Language Services Market: 2013. *Common Sense Advisory*, Boston: USA.

Depraetere, Ilse 2010: What counts as useful advice in a university post-editing training context? Report on a case study. In *Proceedings of the 14th annual conference of the European association for machine translation*. St. Raphaël: France, May 27-28.

Díaz de Ilarraza, Arantza/Labaka, Gorka/Sarasola, Kepa 2008: Statistical postediting: A valuable method in domain adaptation of RBMT systems for less-resourced languages. In *Proceedings of the Mixing Approaches to Machine Translation Workshop*, 35-40. Donostia-San Sebastián: Spain, February 14.

Doherty, Stephen/Kenny, Dorothy/Way, Andy 2012: Taking statistical machine translation to the student translator. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas*. San Diego: California, October 28-November 1.

Dušek, Ondřej/Popel, Martin/Macketanz, Vivien/Burchardt, Aljoscha/Avramidis, Eleftherios/van Noord, Gertjan/Rodrigues, João/Branco, António/Labaka, António/Simov, Kiril/Popov, Aleksandar 2016: *Report on the third MT pilot and its evaluation. QTLeap Project D2.11* [online]. http://qtleap.eu/wp-content/uploads/2016/11/QTLEAP-2016-D2.11.pdf (accessed 12 June 2017).

España-Bonet, Cristina/Màrquez, Lluis/Labaka, Gorka/Díaz de Ilarraza, Arantza/Sarasola, Kepa 2011: Hybrid machine translation guided by a rule-based system. In *Machine translation summit XIII: proceedings of the 13th machine translation summit*, 554-561. Xiamen: China, September 19-23.

Flanagan, Marian/Christensen, Tina Paulsen 2014: Testing post-editing guidelines: how translation trainees interpret them and how to tailor them for translator training purposes. In *The Interpreter and Translator Trainer* 8 (2), 257-275.

Guerberof Arenas, Ana 2008: Productivity and quality in the post-editing of outputs from translation memories and machine translation. In *Localisation Focus* 7(1), 11-21.

Guerberof Arenas, Ana 2014: Correlations between productivity and quality when post-editing in a professional context. In *Machine Translation* 28, 165-186.

Kenny, Dorothy/Doherty, Stephen 2014: Statistical machine translation in the translation curriculum: overcoming obstacles and empowering translators. In *The Interpreter and Translator Trainer* 8(2), 276-294.

Kliffer, Michael 2005: An experiment in MT post-editing by a class of intermediate/advanced French majors. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation*, 160-165. Budapest: Hungary, May 30-31.

Koehn, Philipp/Germann, Ulrich 2014: The Impact of Machine Translation Quality on Human Post-editing. In *Workshop on Humans and Computer-assisted Translation* 38-46. Gothenburg: Sweden, April 26.

Koponen, Maarit 2012: Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, 227.236. Montreal: Canada, June 7-8.

Koponen, Maarit 2013: This translation is not too bad: an analysis of post-editor choices in a machine-translation post-editing task. In *Proceedings of the Second Workshop on Postediting Technology and Practice*. Nice: France, September 2.

Koponen, Maarit 2015: How to teach machine translation post-editing? Experiences from a post-editing course. In *4th Workshop on Post-Editing Technology and Practice (WPTP4)*, 2-15. Miami: Florida, November 3.

Koponen, Maarit/Aziz, Wiker/Ramos, Luciana/Specia, Lucia 2012: Post-editing time as a measure of cognitive effort. In *Proceedings of Workshop on Post-editing Technology and Practice*, 1-10. San Diego: CA, October 28.

Krings, Hans P. 2001: *Repairing texts: Empirical investigations of machine translation post-editing process*. The Kent State University Press, Kent, OH.

Labaka, Gorka/España-Bonet, Cristina/Màrquez, Lluís/Sarasola, Kepa 2014: A hybrid machine translation architecture guided by syntax. In *Machine translation* 28(2), 91-125.

Loffler-Laurian, A.M. 1996: *La Traduction Automatique*. Lille: Presses Universitaires du Septentrion.

Massardo, Isabella/van der Meer, Jaap/O'Brien, Sharon/Hollowood, Fred/Aranberri, Nora/Drescher, Katrin 2016: *MT post-editing guidelines*. The Netherlands: TAUS Signature Editions.

O'Brien, Sharon 2003: Controlling controlled English. An analysis of several controlled language rule sets. In *Proceedings of EAMT-CLAW, 3*, 105-114. Dublin City University, Ireland, May 14-15.

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

O'Brien, Sharon/Choudhury, Rahzeb/van der Meer, Jaap/Aranberri, Nora 2011: *Dynamic Quality Evaluation Framework*. TAUS.

O'Brien, Sharon 2006: *Machine-translatability and post-editing effort: An empirical study using Translog and Choice Network Analysis*. Doctoral dissertation, Dublin City University.

Offersgaard, Lene/Povlsen, Claus/Almsten, Lisbeth/Maegaard, Bente 2008: Domain specific MT in use. In *Proceedings of the 12th European Association for Machine Translation conference*, 153-154. Hamburg: Germany, September 22-23.

Papineni, Kishor/Roukos, Salim/Ward, Todd/Zhu, Wei- Jing 2002: BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311-318. Philadephia: Pennsylvania, July 6-12.

Plitt, Mirko/Masselot, François 2010: A productivity test of statistical machine translation post-editing in a typical localisation context. In *The Prague Bulletin of Mathematical Linguistics*. 7-16. Prague: Czech Republic: Universita Karlova.

Pym, Anthony 2003: Redefining Translation Competence in an Electronic Age. In Defence of a Minimalist Approach. In *Meta: journal des traducteurs/Meta: Translators' Journal* 48(4), 481-497.

Pym, Anthony 2011: What technology does to translating. In *Translation & Interpreting* 3(1), 1-9.

Schäfer, Falko 2003: MT post-editing: How to shed light on the "unknown task". Experiences made at SAP. In *8th International workshop of the European Association for Machine Translation (EAMT 03)*. Dublin City University, Dublin, Ireland, May 15-17.

Snover, Matthew/Dorr, Bonnie/Schwartz, Richard/Micciulla, Linnea/Makhoul, John 2006: A study of translation edit rate with targeted human annotation. In Proceedings of association for machine translation in the Americas (Vol. 200, No. 6).

Specia, Lucia 2011: Exploiting Objective Annotations for Measuring Translation Post-Editing Effort. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, 73-80. Leuven, Belgium, May 30-31.

Specia, Lucia/Farzindar, Atefeh 2010: Estimating Machine Translation Post-Editing Effort with HTER. In *Proceedings of the Second Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC 10)*, 33-41. Denver: CO, November 4.

Specia, Lucia/Turchi, Marco/Cancedda, Nicola/Dymetman, Marc/Cristianini, Nello 2009: Estimating the Sentence-Level Quality of Machine Translation Systems. In *Proceedings of the 13th Annual Conference of the EAMT*, 28-35. Barcelona: Spain, May 14-15.

Tatsumi, Midori 2009: Correlation between Automatic Evaluation Metric Scores, Post-Editing Speed, and Some Other Factors. In *MT Summit XII: proceedings of the twelfth Machine Translation Summit*, 332-339. Ottawa, Ontario, Canada, August 26-30

Tatsumi, Midori/Roturier, Johann 2010: Source Text Characteristics and Technical and Temporal Post-Editing Effort: What is Their Relationship?. In *Proceedings of the Second Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC 10)*, 43-51. Denver: CO, November 4.

Teixeira, Carlos 2014: Perceived vs. measured performance in the post-editing of suggestions from machine translation and translation memories. In *Third Workshop on Post-Editing Technology and Practice*, 45-59. Vancouver: Canada, October 22-26.

Temnikova, Irina 2010: Cognitive Evaluation Approach for a Controlled Language Post-Editing Experiment. In *LREC 2010: proceedings of the seventh international conference on Language Resources and Evaluation*, 3485-3490. Valletta: Malta, May 17-23.