Comparison of post-editing productivity between professional translators and lay users

Nora Aranberri nora.aranberri@ehu.es
Gorka Labaka gorka.labaka@ehu.es
Arantza Diaz de Ilarraza a.diazdeilarraza@ehu.es
Kepa Sarasola kepa.sarasola@ehu.es
IXA Group, Department of Computer Languages and Systems,
University of the Basque Country, Donostia, 20018, Spain

Abstract

This work compares the post-editing productivity of professional translators and lay users. We integrate an English to Basque MT system within Bologna Translation Service, an end-to-end translation management platform, and perform a producitivity experiment in a real working environment. Six translators and six lay users translate or post-edit two texts from English into Basque. Results suggest that overall, post-editing increases translation throughput for both translators and users, although the latter seem to benefit more from the MT output. We observe that translators and users perceive MT differently. Additionally, a preliminary analysis seems to suggest that familiarity with the domain, source text complexity and MT quality might affect potential productivity gain.

1. Introduction

Thanks to the significant improvement of machine translation (MT) over the past two decades, the translation industry has already started to exploit it, mainly by combining it with post-editing. A good number of recent works report a productivity increase thanks to post-editing of MT output as compared to the traditional human translation (e.g., Guerberof, 2009; Plitt and Masselot, 2010; Garcia, 2011; Pouliquen et al., 2011; Skadiņš et al., 2011; den Bogaert and Sutter, 2013; Green et al., 2013; Läubli et al., 2013).

Most post-editing research is designed with professional translators in mind (even if often post-editors involved in experiments are non-professionals and students). However, it is not only language professionals who can benefit from MT in their daily tasks but also regular users who might need to perform a translation sporadically. In this work, we aim to compare the post-editing productivity between professional translators and lay users. We consider the regular example of professional translators working for a language service provider (LSP) and the particular context of administrative and staff members at the University of the Basque Country. This institution is set in a bilingual cultural context. All legal documentation and administrative communication must be provided in Spanish and Basque and study programmes are offered in both languages in parallel. In this scenario, university employees often find themselves having to produce the same material in two languages, that is, having to translate. This work aims to examine the potential benefit of MT during translation for these users as well as for professionals.

Post-editing research has so far focused on mainstream languages. An added challenge of this work is the use of an English to Basque MT system for post-editing. Research on Basque MT has been ongoing for a few years now (Diaz de Ilarraza et al, 2000a; 2000b; Labaka et al., 2007; España-Bonet et al., 2011; Mayor et al., 2011). However, Basque being a low-resourced language, researchers and developers have found themselves with limited resources to build competitive MT systems and automated translation has not been included within the translation processes of local LSPs yet. To our knowledge, this is the first (open) productivity experiment done for the English to Basque translation direction.

Laurenzi et al. (2013) pointed out the existence of many communities that could benefit greatly from machine translation but, as in the case presented in this work, have not yet started to use it, as authors suggest, either due to lack of awareness or barriers to adoption. The work in Laurenzi et al. (2013) presents a feasibility study to introduce MT coupled with post-editing in local and regional health departments in the United States. It highlights a number of requirements the translation platform should address, such as being intuitive and easy to install, allowing users to share ongoing and completed jobs. Our work builds on this first feasibility study and goes a step forward by assessing the actual translation performance. We identify a suitable tool for our users that is intuitive, easy to access and allows sharing translation resources such as translation memories (TM) or specialized MT engines and measure productivity gain, while comparing it with the performance of professional translators.

2. Experimental Design

In the following sections we describe the platform and the texts used during the experiment, we present the profile of the participants and detail how the productivity test was set up.

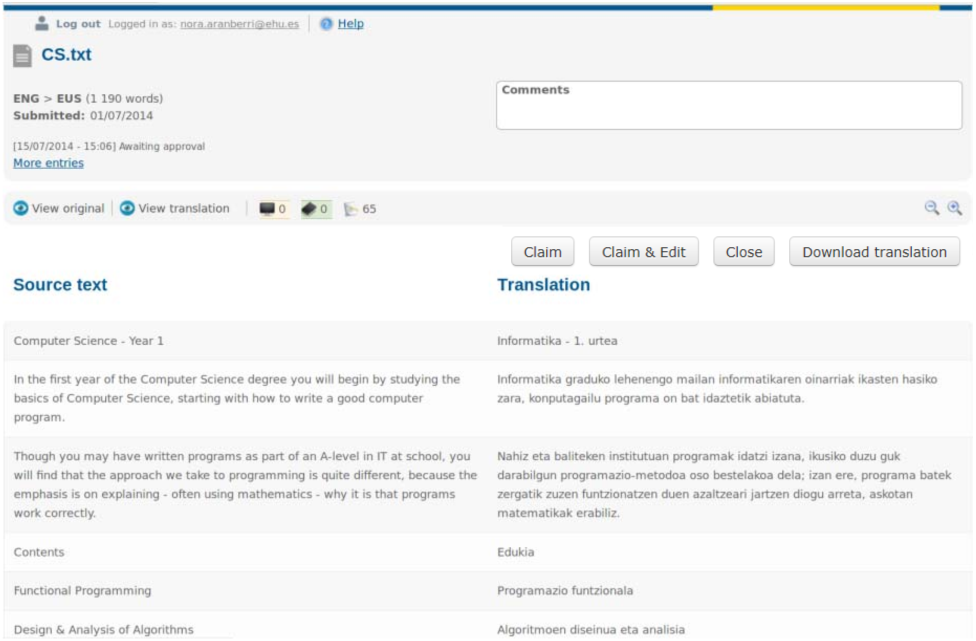
2.1. The Platform

The post-editing environment used in the experiment was the Bologna Translation Service (BTS), the product of an EU-funded ICT PSP 4th Call, Theme 6: Multilingual Web project (ID 270915).[1] It is an end-to-end web-based translation management tool in which users with different roles (manager, requester, reviewer, etc.) participate on-line at different stages of the translation workflow. It couples translation memory (TM) and machine translation (MT) capabilities within a simple work environment. BTS was designed with lay users in mind. The work environment offers a simple layout with a top bar with the main action buttons and job information (see Figure 1). Below, the source text is split into segments and the target side is filled with either TM (fuzzy-)matches or MT candidates for the reviewer to work on. It is a plain tool as opposed to more sophisticated software developed in the CASCAMAT[2] and MateCat[3] projects (Alabau et al. 2013; Federico et al., 2012), which include interactive translation prediction and track post-editing operations.

[1] http://www.bologna-translation.eu/

[2] http://www.casmacat.eu/

[3] http://www.matecat.com/



Logged in as: nora.aranberri@ehu.es

ENG > **EUS** (1 190 words)
Submitted: 01/07/2014

[15/07/2014 - 15:06] Awaiting approval

Source text	Translation
Computer Science - Year 1	Informatika - 1. urtea
In the first year of the Computer Science degree you will begin by studying the basics of Computer Science, starting with how to write a good computer program.	Informatika graduko lehenengo mailan informatikaren oinarriak ikasten hasiko zara, konputagailu programa on bat idaztetik abiatuta.
Though you may have written programs as part of an A-level in IT at school, you will find that the approach we take to programming is quite different, because the emphasis is on explaining - often using mathematics - why it is that programs work correctly.	Nahiz eta baliteken institutuan programak idatzi izana, ikusiko duzu guk darabilgun programazio-metodoa oso bestelakoa dela; izan ere, programa batek zergatik zuzen funtzionatzen duen azaltzeari jartzen diogu arreta, askotan matematikak erabiliz.
Contents	Edukia
Functional Programming	Programazio funtzionala
Design & Analysis of Algorithms	Algoritmoen diseinua eta analisia

Figure 1. Screenshot of the translation environment at BTS.

For the current experiment, the BTS platform was enhanced with an English to Basque MT system. A standard phrase-based statistical machine translation system was built based on Moses using a parallel corpus of 14.58 million English tokens and 12.50 million Basque tokens (1.3 million parallel sentences) which includes localization texts (graphic user interface strings and user documention), academic books and web entertainment data. To address the token mismatch between English (analytic language) and Basque (agglutinative language) tokens, the aligner was fed with segmented words for the agglutinative language. Several segmentation options exist: we can isolate each morpheme, or break each word into lemma and a bag of suffixes; we can establish hand-written rules for segmentation, or let an automatic tool define and process the words unsupervised. Based on the results from Labaka (2010), we opted for the second option and joined together all the suffixes attached to a particular lemma in one separate token. Thus, on splitting a word, we generated, at most, three tokens (prefixes, lemma and suffixes). Moses was trained and optimized on segmented text. Note that when using segmented text for training, the output of the system is also segmented text. Real words are not available to the statistical decoder. This means that a generation postprocess (unsegmentation step) is needed to obtain real word forms. We incorporated a second language model (LM) based on real word forms to be used after the morphological postprocess. We implemented the word form-based LM by using an n-best list, as was done in Oflazer and El-Kahlout (2007). We first asked Moses to generate a translation candidate ranking based on the segmented training explained above. Next, these candidates were postprocessed. We then recalculated the total cost of each candidate by including the cost assigned by the new word form-based LM in the models used during decoding. Finally, the candidate list was re-ranked according to this new total cost. This somehow revises the candidate list to promote the ones that are more likely to be real word-form sequences. The weight for the word form-based LM was optimized at Minimum Error Rate Training (Och, 2003) together with the weights for the rest of the models.

2.2. The Texts

Two texts of around 1200 words each were selected for the experiment. Because the SMT system was trained on science and localization texts among others, it was deemed convenient to use texts from related domains. Text A consists of short 1st year computer science course descriptions from a UK university, and Text B is a collection of six short science articles from www.sciencenews.org, the flagship website by the Society for Science & the Public (SSP), dedicated to public engagement in scientific research and education.

The selection of the texts was somehow also motivated by the prospective profile of the lay user group. Whereas translators are professionals who are trained to handle a large variety of topics, we aimed to engage staff members of the Faculty of Computer Science as lay users. Therefore, we considered that they would feel more comfortable dealing with texts from computing and scientific domains.

Both texts are very similar from a size point of view. Text A consists of 1190 words and 65 sentences, whereas Text B consists of 1196 words and 67 sentences. Moreover, a wide variety of sentence-lengths are present in the texts, ranging from 1 to 51 words in length. Similarly, both texts display a moderate degree of difficulty, as they address specialized topics. In particular, terminology is very significant in both texts. Text-types, in contrast, are different. Text A is mainly descriptive and Text B, although descriptive, tends to be more literary. The literariness might leave some room for creativity in the transations, but at the same time, this might pose extra difficulty, particularly for lay users.

2.3. The Participants

We aim to compare the productivity gain for professional translators and for lay users. To this end, the same experiment was conducted with a group of translators and a group of users. The former consisted of six professional translators who regularly work for the Elhuyar Foundation (see Table 1).[4] They reported a translation experience ranging between 4 and 18 years. Four out of six had never performed post-editing before, whereas two reported having participated in previous MT experiments. They completed the translations required by the experiment as if they were regular jobs with the difference that they used the BTS platform instead of the usual TM tool (SDL Trados).

	T1	T 2	T 3	T 4	T 5	T 6
Translation experience	8 years	12 years	4 years	18 years	20–23 years	8 years
PE experience	experiments	no	no	no	experiments	no

Table 1. Translation and post-editing experience of professional translators.

The lay user group was represented by five lecturers and one post-doc from the Faculty of Computer Science at the University of the Basque Country (see Table 2). They report a level of English ranging from B2 to C1, and their level of Basque ranges from C1 to C2. As mentioned below, the University of the Basque Country is set in a bilingual cultural context and study programs are offered in both Spanish and Basque in parallel (with English becoming more and more a third language of instruction). In this scenario, some lecturers have “bilingual” job positions, which means that they might find themselves teaching the same modules in two languages. As a result, they need to prepare the same study material in both languages. The participants in the lay user group report having little translation experience, and when any, this seems to be mainly from Spanish to Basque or viceversa. They report never

[4] http://www.elhuyar.org/EN

using MT engines for this purpose. In fact, they admit to most often re-writing the material rather than translating it sentence by sentence.

	U1	U2	U3	U4	U5	U6
Level of English	B2	B2	B2	C1	C1	C1
Level of Basque	C1	C1	C1	C2	C1	C2
Translation experience	little Spanish-Basque, no MT	little no MT used	little no MT used	little no MT used	no	re-write of material

Table 2. Translation experience and language proficiency of lay users.

2.4. The Productivity Test

We aimed to measure changes in productivity (if any) by monitoring the difference in the time spent translating the texts with and without the help of our MT system's output. In order to do so, BTS was programmed to count the time participants needed to complete the job. As described in section 2.1, source texts are provided to participants segmented on the left column, and a translation box per segment is opened for him/her to work on on the right column. More precisely, therefore, BTS would record the time the participants spent in each segment by saving the time each translation box was active. This method also opened the possibility for participants to work on a single segment multiple times if necessary. We asked participants to avoid distractions when completing the translation/post-editing job so that extra time would not count towards the time spent on the job. Although we recommended keeping them at a minimum, the possibilities to pause the job, and log in and out of the platform were provided to ensure that saved times were as accurate as possible.

Each participant worked under both setups, with and without the help of the MT output. In each group, texts were assigned to each participant in a way that each text was translated and post-edited three times, and the same text was not assigned to the same participant for both setups.

Simple guidelines were provided to participants with information about how to use the platform, as well as the job they should complete. They were given total freedom as to the resources they could use to perform the job (dictionaries, web searches), the only restriction being that they should not use an external MT engine. Therefore, once participants registered in the BTS platform and were assigned the two tasks, they could decide freely when and where to complete the jobs. They were given 1–2 weeks to complete the tasks. All the previous conditions should help simulate a real translation scenario as much as possible for both professional translators and lay users.

The BTS platform includes a translation memory (TM) feature. During this experiment, however, translators and users would work with an empty TM so that we only focus on the difference in translation throughput (average number of words translated per hour) and no other parameters such as fuzzy-match repetition rates introduce noise in the data.[5] We assume that the use of TMs would affect both setups (translation from scratch and post-editing) equally, and therefore, no effort was made in compiling TMs for this occasion. Nonetheless, the TM feature was activated so that the translations were stored segment-aligned and can serve as parallel data for future NLP-related tasks.

[5] The current version of the BTS does not include fuzzy-match propagation of the translations validated within the project.

3. Results

3.1. Professional Translators

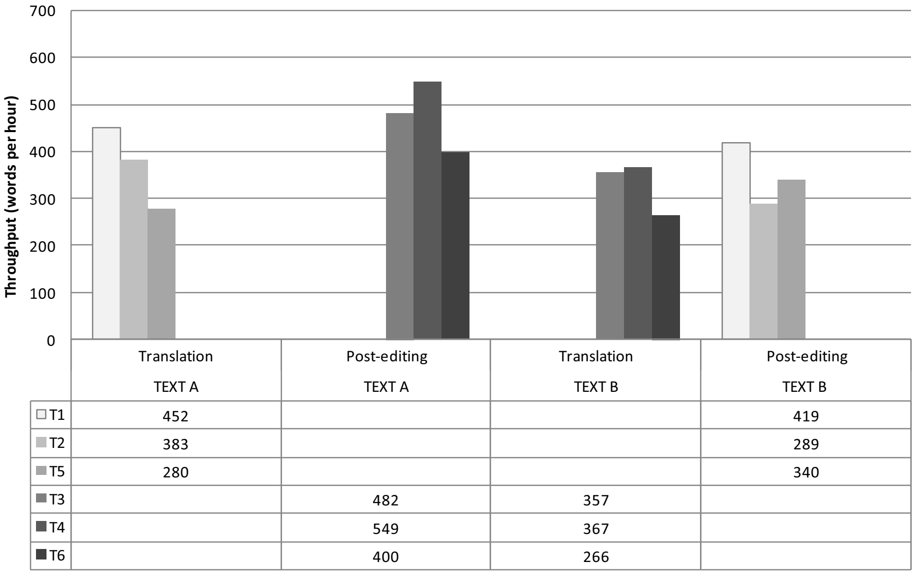
By looking at the average throughput per setup, we see that overall, post-editing our MT system obtains a slightly higher throughput than translating from scratch (17.66%). The experiment shows that, on average, the throughput increased for both texts with the aid of MT, although at different levels. The throughput for Text A increased from 372 words per hour to 477 words per hour (28.22%) and for Text B from 330 words per hour to 350 words per hour (6.06%).

If we look at the performance of individual translators, we observe that T1 and T2 have their throughput lowered with the introduction of MT 7.30% and 24.54%, respectively. However, T3, T4, T5 and T6 increased their throughput 35.01%, 49.59%, 21.43%, and 50.37%, respectively. Table 3 presents the post-editing productivity ratio (PPR), the ratio of the post-editing speed to translation speed (both expressed in words per hour). Figures less than 1 indicate cases where post-editing decreased translation throughput. Figures above 1, in contrast, indicate the ratio in which post-editing increased translation throughput.

	T1	T2	T3	T4	T5	T6
Post-editing productivity ratio	0.93	0.75	1.35	1.49	1.21	1.50

Table 3. Post-editing productivity ratio for professional translators.

If we consider closely the combination of setup, translator and text, we see that it is the three translators who post-edited Text A that benefited from the introduction of MT, as well as the slowest translator for Text A, who post-edited Text B (see Figure 2).



	Translation TEXT A	Post-editing TEXT A	Translation TEXT B	Post-editing TEXT B
□ T1	452			419
■ T2	383			289
■ T5	280			340
■ T3		482	357	
■ T4		549	367	
■ T6		400	266	

Figure 2. Throughput per text, setup and translator.

3.2. Lay User Group

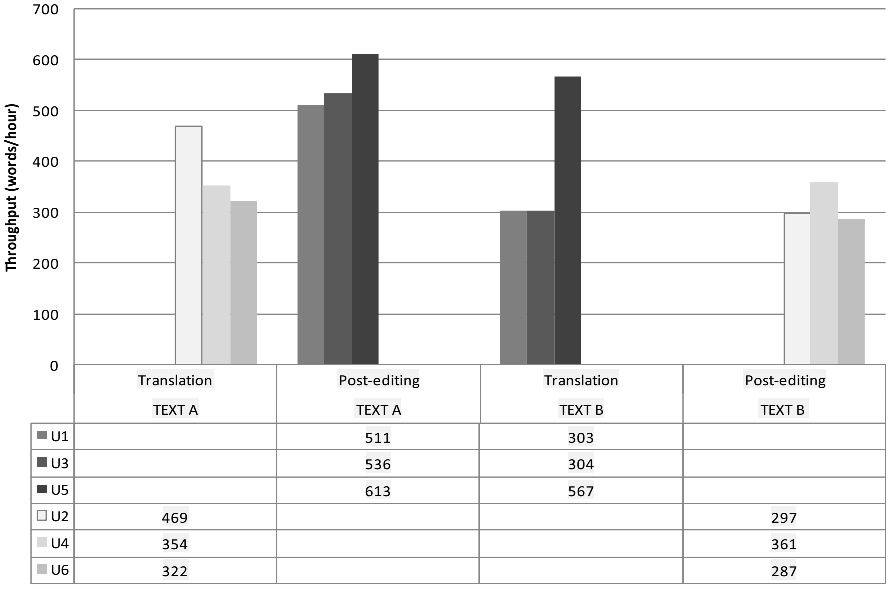
The average post-editing throughput for the lay user group (434 words per hour) also surpassed the average translation throughput (386 words per hour) in 12.43%. Reinforcing the trend observed with translators, post-editing for Text A increases productivity 45.13% whereas post-editing Text B does not (-19.44%).

The performance of individual users shows that U1, U3, U4 and U5 increase translation productivity when using MT output. U2 and U6, in contrast, do not (see Table 4 for PPRs).

	U1	U2	U3	U4	U5	U6
Post-editing productivity ratio	1.69	0.63	1.76	1.02	1.08	0.89

Table 4. Post-editing productivity ratio of lay users.

Once again, if we look closely to the combination of setup, user and text, we observe that the users who benefited most from the use of MT were those who post-edited Text A. U4, who barely increased productivity, post-edited Text B, similarly to U2 and U6, who saw their throughput lowered when post-editing.



	Translation TEXT A	Post-editing TEXT A	Translation TEXT B	Post-editing TEXT B
■ U1		511	303	
■ U3		536	304	
■ U5		613	567	
□ U2	469			297
■ U4	354			361
■ U6	322			287

Figure 3. Throughput per text, setup and user.

4. Discussion

Results show that overall, both professional translators and lay users benefited from the use of our MT engine during translation. Translators have obtained an overall productivity gain of 17.66% and users 12.43%. However, the gain seems to be dependent on the text participants worked on. Translators have benefited more when post-editing Text A (28.22% increase) as opposed to Text B (6.06% increase). Users show an increased productivity of 45.13% when post-editing Text A, but post-editing Text B slows down the job in about 19.44%. The latter is mainly the result of U5's high translation rate for Text B, almost double that of U1 and U3, and the low post-editing performance for Text B.

If we compare the performance of individual participants, we see that four translators and four users improve their throughput when post-editing, whereas two translators and two user do not (these last four post-edited Text B). Most participants benefit from the use of MT but not all.

Given these results, we briefly consider three factors that emerged during the experiment and might shed some light on the outcome of the experiment itself and hint to features that a company might want to consider when exploiting the MT system: attitude towards post-editing and training, source text difficulty and MT quality. Although they are all intertwined, we will consider them separately here.

4.1. Post-editing Skills and Attitude

Firstly, the skill and willingness of translators and users to make use of the MT output might affect the translation process. Post-editing has been claimed to be a different task from that of translating and one that requires different cognitive abilities and practical skills (Krings and Koby, 2001; O'Brien, 2002). Translators, therefore, need to be trained to maximize the potential benefit of post-editing. Several industrial players nowadays offer post-editing courses.[6] These usually provide an overview of MT development for the users to get acquainted with the intricacies of the systems so that they learn to interpret the output, and tips about features and patterns to watch out for in order to maximize the reuse of the output. Similarly, translator training centres have started to introduce machine translation and post-editing content within their curricula.[7] Participants in our experiment had no experience or training in post-editing. This does not allow measuring the maximum post-editing benefit.

The attitude towards post-editing or, more generally, MT might also set the tone for the job and highlight the importance of more objective measurements rather than basing the integration of MT on translator perception only. The mismatch between the translators' perception of productivity and their actual productivity has been previously reported by Autodesk, specifically on the company's follow up work on Plitt and Masselot (2010).[8] To check for this effect, we asked participants to fill in a short questionnaire after completing the tasks. One of the questions asked was whether they thought having the MT output helped them complete the translation. They were asked to mark this on a scale from -5 to 5 where -5 meant that the MT output had greatly hindered their work and 5 meant that the MT output had greatly helped their work (see Table 5). Overall, users were more positive about having the MT output displayed when translating, with only one out of six claiming that it hindered the process. In the case of translators, however, three out of six heavily penalized its use, one reported that it was better than not having it, and two reported some benefit. T1 commented that translation and post-editing required different skills and that should the same time be spent in post-editing and translating, the translation would most probably be of better quality. T6 was the most positive of all with regards to MT and admitted that the output helped in acquiring the terminology but was hopeless with syntax, which needed a complete rework. T2, T3 and

[6] For examples of training courses see TAUS's online course in collaboration with Welocalize at https://evaluation.taus.net/post-editing-course-pricing or SDL's course at http://www.translationzone.com/learning/training/post-editing-machine-translation/index-tab2.html#tabs.

[7] See an example at: MSc in Translation Technology. Dublin City University. http://www.dcu.ie/prospective/deginfo.php?classname=MTT#

[8] See http://langtech.autodesk.com/productivity.html for results of a 2-day translation and post-editing productivity test with 37 participants that Autodesk held in August of 2011.

T4 indicated that the MT output had clearly interfered in their job. T2 reported that MT output slowed down the process considerably because reading, understanding and considering what to reuse from it was very time-consuming. T3 commented that translating from scratch was easier and faster, and that even checking the MT output for terminology would most often not help. T4 claimed that the MT system did not translate the order of the phrases properly, which rendered the translation incomprehensible. Interestingly, T3 and T4 did benefit from post-editing.

U1, U2 and U3 reported that the terminology and certain chunks suggested by the MT system were useful, even when they claim to have reworked the sentences completely. U4 argued that given her lack of familiarity with the domain (Text B), she found it difficult to decide whether the terminology proposed by the MT system was correct, and therefore, she would still look up the terminology in an external source. She commented that MT output could have been a potential benefit should she be familiar with the domain of the text to be translated.

	T1	T2	T3	T4	T5	T6
PE help	0	-5	-4	-5	2	1
	U1	U2	U3	U4	U5	U6
PE help	3	2	2	0	-2	3

Table 5. Perception of MT output help during translation ([-5,5] range, where -5 greatly hinders translation and 5 greatly helps translation).

4.2. Difficulty of Source Texts

Secondly, the difficulty of the texts also needs to be considered. We will highlight two aspects here. Firstly, the familiarity with the domain of the texts; secondly, the linguistic complexity. Professional translators stated they were not familiar with the domains covered by the texts. This probably means that they were not used to the terminology and phraseology of the given domains. In contrast, lay users were a post-doc and lecturers of computer science, the domain covered by Text A. This aspect seems to be reflected in the participants' perception towards text difficulty.

As part of the questionnaire, participants were asked to specify the difficulty of the texts in a scale of 1–5, where 1 was very easy and 5 was very difficult (see Table 6). Translators reported the texts slightly differing in difficulty, but they did not agree which, Text A or Text B, was more difficult. T1 argued that Text A was *more specialized* whereas T5 considered Text B *more difficult to understand and very technical*. T2 found that Text A was easier to translate because *the whole text followed the same thread*. In contrast, T3 commented that Text A was *very abstract and disjointed*, whereas Text B was *believable and interesting*. Users, on the other hand, show a clearer tendency with three out of six identifying Text A as easy to moderate and Text B as difficult to very difficult. All users explicitly commented on their familiarity with Text A.

Questions	T1	T2	T3	T4	T5	T6
Text A difficulty	4	3	3	4	4	4
Text B difficulty	3	5	4	3	3	5
	U1	U2	U3	U4	U5	U6
Text A difficulty	2	4	4	3	2	5
Text B difficulty	4	4	4	5	5	3

Table 6. Perception of text difficulty (1–5 range, where 1 is very easy and 5 is very difficult).

Even when we are aware that many other factors are involved in the process, we turned to a readability – reading difficulty – measurement as a proxy for translation difficulty. We calculated the number of hard words[9], lexical density[10] and Gunning Fog Index (Gunning, 1952) (see Table 7).[11] When comparing both texts, we see that Text A has a slightly lower number of hard words, 11.94% as opposed to 13.64% for Text B. Lexical density is considerably higher for Text B, which means that repetitions are lower. Given that both Text A and Text B have very similar number of words, we conclude, therefore, that Text B has a higher number of different words, making it more complex. Finally, the Fog Index confirms that more years of education are necessary to read Text B. Overall, readability features suggest that Text B is more difficult to read.

	Text A	Text B
Hard Words	142 (11.94%)	162 (13.64%)
Lexical Density	34.57%	53.87%
Fog Index	11.88	12.34

Table 7. Readability-related measurements for Text A and Text B.

Understanding the source text is a vital step in translation, but other factors such as the mapping of the concepts and grammatical features pay an important role. In an attempt to measure both sides of the translation process in terms of complexity, we have also analysed the linguistic complexity of the translations produced by the participants. Based on the linguistic analysis presented in Gonzalez-Dios, et al. (2014), we have calculated the average occurrences of a number of linguistic features (including lexical, morphological, syntactic and pragmatic features) present in the translations and post-edited versions of Texts A and B (see Table 8). We observe that out of the 96 features studied, Text B has a higher number of occurrences for 63 for both translators and users, and Text A for 25 and 17, for translators and users, respectively, having no occurrences for 8 and 6 features. Additioanlly, we have considered the 10 most predictive features for complexity according to the same authors, which include a number of the most predictive features according to Feng et al. (2010), namely, part-of-speech ratios for nouns. We see that Text B appears to be more complex, scoring higher in 7 out of the 10 features.

	Text A Translators	Text A Users	Text B Translators	Text B Users
Number of features analysed	96	96	96	96
Number of features with more hits[12]	25	17	63	63
Number of features with no hits	8	6	8	13
Ratios				
Proper nouns / common nouns	0.01077	0.01830	**0.15715**	0.15672
Appositions / noun phrases	0.04433	0.03040	0.13129	**0.14345**
Appositions / all phrases	0	0.00065	**0.00293**	0.00331
Named entities / common nouns	**0.14422**	0.18222	0.11357	0.11184
Unique lemmas / all lemmas	0.03586	0.01747	0.05376	**0.06150**
Acronyms / all words	**0.24811**	0	0.21422	0.03030
Causative verbs / all verbs	**0.00137**	0.00035	0.00061	0.00030

[9] For readability testing hard words are those with three or more syllables.

[10] Lexical density is provided as the type/token ratio x 100.

[11] The Gunning Fox Index returns the number of years of education that a reader hypothetically needs to understand a particular text. It is calculated by multiplying by 0.4 the sum of the average number of words in the sentences and the percentage of hard words. For instance, the New York Times has an average Fog Index of 11-12.

[12] Counts normalized for 1000 words.

Modal-temporal clauses / subordinate clauses	0	0	**0.00047**	0.00000
Destinative case endings / all case endings	0	0	**0.00085**	0
Connectors of clarification / all connectors	0.16157	0.14564	0.25437	**0.25869**

Table 8. Comparison of linguistic complexity features for Text A and Text B translations and post-edited versions.

A final aspect that is worth noting is text expansion rates. English is an analytic language and Basque is an agglutinative language, which usually means that word-counts contract when translating into Basque. For translators, on average, Text A has contracted to 90.25% and Text B has expanded to 103.85% with respect to the English source. For users, both texts contract but whereas Text A goes down to 85.21%, Text B still remains at a high 98.21%. The fact that an expansion has occurred in Text B might be due to participants tending to over-explain or paraphrase. This might be a result of the complexity of the content.

4.3. MT Quality

MT output quality is also essential in post-editing measurements, a factor that is often neglected when reporting productivity gain. An exception is a seminal work by Koehn and Germann (2014). They studied the relation between MT quality and post-editing, and concluded that differences in post-editing skills might be more decisive than MT quality to foresee productivity gain when comparing systems within the same quality range. Our findings show that the text for which a higher increase in productivity was obtained seems to be slightly easier and better suited for our MT system.

In order to test for MT quality, we calculated BLEU and TER scores on Text A and Text B using the translations and post-editings obtained during the experiment as references (see Table 9). If we consider the scores obtained with the translations as references, we see that Text B obtains a slightly higher BLEU score than Text A, but output for Text A is better according to TER. If we take post-editings or all six versions as references, then Text B seems to get a higher quality output. As expected, we observe that the post-edited versions obtain significantly better BLEU and TER scores as the post-edited versions resemble the MT output more than translations made from scratch. Overall, automatic score results would lead us to conclude that the MT output for Text B might be slightly better, and therefore more reusable.

Translators	Translations		Post-editings		All 6 references	
	BLEU	TER	BLEU	TER	BLEU	TER
Text A	12.33	68.07	23.31	55.70	25.26	55.44
Text B	12.71	71.26	27.26	55.32	28.61	54.07

Users	Translations		Post-editings		All 6 references	
	BLEU	TER	BLEU	TER	BLEU	TER
Text A	8.44	77.78	39.65	52.32	41.00	51.27
Text B	10.45	79.35	29.17	56.41	30.49	56.44

Table 9. BLEU and TER scores for Texts A and B using participants's texts as references.

It is worth noting the difference in BLEU scores between the post-edited versions of translators and users. A higher BLEU score means that more of the MT output was kept in the final versions. Users, therefore, accepted or reused a considerably larger amount of MT output than translators. A possible interpretation is that the MT output was of a relatively good quality and users, domain experts of Text A, were easily able to identify reusable chuncks and exploit terminology much more so than translators.

Finally, to test whether the MT engine was better prepared to address Text A or Text B, we calculated perplexity and out-of-vocabulary (OOV) words. Perplexity is used as a measurement of how well the language model predicts the reference translations. The smaller the perplexity, the more and longer overlap exists between the reference and the language model in the MT system. This measurement shows that the MT engine is better suited to output a correct version for Text A than for Text B (see Table 10). Note that the high perplexity values, calculated per word, are in line with those reported for morphologically rich languages (see Popel and Mareček, 2010).

	Translators			Users	
	Text A	Text B		Text A	Text B
Translation 1	988.220	1274.940	Translation 1	2086.65	1359.14
Translation 2	898.022	1081.580	Translation 2	1423.39	1794.06
Translation 3	776.408	966.309	Translation 3	1686.82	1944.15
Post-editing 1	850.781	909.031	Post-editing 1	842.705	1341.37
Post-editing 2	660.740	909.031	Post-editing 2	1002.280	1300.49
Post-editing 3	688.585	984.311	Post-editing 3	902.018	1173.91

Table 10. Perplexity calculated on 5-grams.

In Table 11 we see the number of OOVs in the training corpus with respect to the source texts. Once again, Text A seems better suited for our MT system, as only 0.7% of the tokens were missing from the training data, as opposed to the 4.9% for Text B. This is yet another feature that hints that MT output for Text A might be of better quality than for Text B.

	Training sentences	Training tokens	Training types	Tokens	Types	OOV tokens	OOV types
Text A	1,290,501	15,798,942	221,172	1292	420	9 (0.7%)	8 (1.9%)
Text B	1,290,501	15,798,942	221,172	1381	645	68 (4.9%)	32 (5.0%)

Table 11. OOV counts for the Text A and B together with information on training data.

5. Conclusions

We have integrated an English to Basque MT system within BTS, an end-to-end translation management platform, and performed a post-editing producitivity experiment in a real working environment to compare the performance of professional translators and (prospective) lay users of BTS. Results suggest that overall post-editing increases translation productivity for both translators and users, although the latter seem to benefit more from the MT output specially when working on their domain of expertise.

We have observed that translators and users perceive MT output differently. Overall, translators seem to find that it interferes and slows down their work. However, users do not show a negative attitude towards it and profit more from it, specially when working on a familiar domain. Although we addressed them separately, we saw that textual complexity and MT quality are connected and seem to affect potential productivity gain. We observed that, although both texts under study were considerably specialised, the text that had higher readability and less linguistic complexity, and that was better fitted for our MT engine obtained a larger increase in productivity gain.

Acknowledgements

We would like to thank the staff members from the Faculty of Computer Science who willingly agreed to participate in the experiment.

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union’s Seventh Framework Programme (FP7/2007/2013) under REA grant agreement nº 302038.

References

Alabau, V., Bonk, R., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., González, J., Koehn, P., Leiva, L., Mesa-Lao, B., Ortiz, D., Saint-Amand, H., Sanchís, G. and Tsoukala, C. (2013). CASMACAT: An open source workbench for advanced computer aided translation. *The Prague Bulletin of Mathematical Linguistics*, 100:101–112.

Diaz de Ilarraza A., Mayor A., Sarasola K. (2000a). Building a Lexicon for an English-Basque Machine translation System from Heteogeneous Wide-Coverage dictionaries. In *Proceedings of MT 2000: machine translation and multilingual applications in the new millennium*, University of Exeter, United Kingdom: 19-22 November 2000, pages 2.1–2.9.

Diaz de Ilarraza A., Mayor A., Sarasola K. (2000b). Reusability of Wide-Coverage Linguistic Resources in the Construction of a Multilingual Machine Translation System. In *Proceedings of MT 2000: machine translation and multilingual applications in the new millennium*, University of Exeter, United Kingdom: 19-22 November 2000, pages 16.1–16.8.

España-Bonet, C., Labaka, G., Diaz de Ilarraza, A., Màrquez, L. and Sarasola, K. (2011). Hybrid Machine Translation Guided by a Rule–Based System. In *Proceedings of the Thirteenth Machine Translation Summit (MT Summit XIII)*, Xiamen, China, pages 554–561.

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

Feng, L., Huenerfauth, M., Jansche, M. and Elhadad, N. (2010). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Coputational Linguistics (COLING): Posters*, pages 276–284, Beijing, China.

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

Gonzalez-Dios, I., Aranzabe, M., Diaz de Ilarraza, A. and Salaberri, H. (2014). Simple or Complex? Assessing the readability of Basque texts. In *Proceedings of the 5th International Conference on Computational Linguistics (COLING)*, Dublin, Ireland, pages 334–344.

Green, S., Heer, J. and Manning, C. (2013). The efficacy of human post-editing for language translation. In *Proceedings of ACM Human Factors in Computing Systems (CHI)*, pages 439–448.

Guerberof, A. (2009). Productivity and quality in MT post-editing. In Proceedings of MT Summit Workshop on New Tools for Translators.

Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill: New York.

Koehn, P. and Germann, U. (2014). The impact of machine translation quality on human post-editing. In *Workshop on Humans and Computer-assissted Translation*, pages 38–46, Gothenburg, Sweeden. Association for Computational Linguistics.

Krings, H. and Koby, G. (eds) (2001). *Repairing Texts: Empirical Investigations of Machine-Translation Post-Editing Processes*. Kent State University Press: Kent, Ohio.

Labaka, G. (2010). EUSMT: Incorporating Linguistic Information into SMT for a Morphologically Rich Language. Its use in SMT-RBMT-EBMT hybridation. PhD Thesis. University of the Basque Country.

Labaka, G., Stroppa, N., Way, A. and Sarasola, K. (2007). Comparing Rule-Based and Data-Driven Approaches to Spanish-to-Basque Machine Translation. In Proceedings of MT-Summit XI, Copenhagen, pages 297–304.

Mayor, A., Alegria, I., Diaz de Ilarraza, A., Labaka, G., Lersundi, M. and Sarasola, K. (2011). Matxin, an open-source rule-based machine translation system for Basque. Machine Translation Journal, 25(1), pages 53–82.

O'Brien, S. (2002). Teaching post-editing: a proposal for course content. In Proceedings of the Sixth EAMT Workshop Teaching Machine Translation, pages 99–106, Manchester, UK.

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 (ACL)*, pages 160–167, Sapporo, Japan.

Oflazer, K. and El-Kahlout, I. D. (2007). Exploring Different Representation Units in English-to-Turkish Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 25–32, Prague, Czech Republic.

Plitt, M. and Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context, *Prague Bulletin of Mathematical Linguistics*, 93:7–16.

Popel, M. and Mareček, D. (2010). Perplexity of n-Gram and Dependency Language Models. In P.Sojka et al. (Eds.): TSD 2010, LNAI 6231, pages 173–180.

Pouliquen, B., Mazenc, C. and Iorio, A. (2011). Tapta: A user-driven translation system for patent documents based on domain-aware statistical machine translation. In *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT*), pages 5–12.

Skadins, R., Purins, M., Skadina, I. and Vasiljevs, A. (2011). Evaluation of SMT in localization to under-resourced inflected language. In *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*, pages 35–40.

Van den Bogaert, J. and De Sutter, N. (2013). Productivity or quality? Let's do both. In Proceedings of the Machine Translation Summit XIV, pages 381–390.