Investigating the post-editing effort associated with machine-translated metaphors: a process-driven analysis

Arlene Koglin, Federal University of Pelotas (UFPel) and Laboratory for Experimentation in Translation (LETRA/UFMG)
Rossana Cunha, Federal University of Minas Gerais (UFMG) and Laboratory for Experimentation in Translation (LETRA/UFMG)

ABSTRACT

This paper reports on a study that analyses the impact of two different machine translation (MT) outputs on the cognitive effort required to post-edit machine-translated metaphors by means of eye tracking and think-aloud protocols. We hypothesise that the statistical MT output would have a positive effect on reducing cognitive effort. In order to test this hypothesis, a post-editing experiment was conducted with two different groups of participants. Each experimental group had two post-editing tasks using the language pair English into Brazilian Portuguese. On Task 1 (T1), participants were asked to post-edit a Google machine-translated output whereas on Task 2 (T2) the same participants were assigned to post-edit a Systran machine translated output. Data collection was conducted under the experimental paradigm of data triangulation in translation process research. Data analysis focuses on eye tracking data related to fixation duration and pupil dilation as well as think-aloud protocols. This analysis shows that the cognitive effort required to post-edit the pure statistical MT output might be lower in comparison to the hybrid output when conventional metaphors are machine translated.

KEYWORDS

Statistical machine translation (SMT), rule-based machine translation (RBMT), hybrid machine translation, post-editing MT, cognitive effort, eye tracking, think-aloud protocols, machine-translated metaphors, fixation duration, pupil dilation.

1. Introduction

Despite the continuous development and advances in machine translation (MT) output quality, the idea of fully automatic high-quality translation (FAHQT) has currently been replaced by the more practical use of human-aided machine translation (HAMT) within restricted environments by means of post-editing. According to the Draft of the European Standard for Translation Services (Joscelyne 2006: 5), post-editing refers to the "examination and correction of the text resulting from an automatic or semi-automatic machine system (machine translation, translation memory) to ensure it complies with the natural laws of grammar, punctuation, spelling and meaning."

While human translation errors are unpredictable, MT errors often follow a pattern depending on language pair, type of text and engine. So far, there have been three main approaches used in MT systems: rule-based machine translation (RMBT), statistical machine translation (SMT) and more recently neural machine translation (NMT).

A rule-based system requires syntax analysis, semantic analysis, syntax generation and semantic generation; an RBMT MT engine is built on algorithms that analyse the syntax of the source text and uses rules to transfer the meaning to the target language by building a sentence.

SMT systems, on the other hand, use algorithms to establish probabilities between segments in a source and target language document to propose translation candidates. These systems employ a statistical model, based on the analysis of a corpus.

NMT systems have emerged as a revolutionary paradigm that has outstripped SMT's achievements as a result of three main factors: the evolution of hardware, the higher capacity of processing, and the pioneering work of Bahdanau *et al.* (2014) and Sutskever *et al.* (2014). NMT systems employ an encoder-decoder approach, that is to say, an encoder is responsible for converting input words into a sequence of contextualized representations and a decoder produces an output sequence of words (Koehn 2017: 66).

From a practical point of view, the main obstacle to successful MT is to implement in the system the ability to deal with discourse and textual characteristics which are more complex and context-dependent such as ambiguities, anaphoric reference, and figurative language (Alfaro and Dias 1998). The MT output quality as well as the system capacity to find translation solutions to the segments will also vary according to the text type and MT system architecture. As a result, it is possible to reasonably predict which segments will not be successfully machine translated and therefore will require human intervention by means of post-editing.

Metaphors would be expected to pose difficulties for MT due to the fact that their translation relies on the re-creation of various logical and inferential properties of the source text in the target text (Gutt 1992). Because of that, it has often been assumed that texts rich with metaphorical language are not suited for machine translation; however, there is no firm evidence of the influence of such content on post-editing effort.

This paper seeks to address this gap. To do so, the study compares metaphors machine-translated by a hybrid system and an SMT system in order to analyse the impact of the respective systems on post-editing effort. We hypothesise that the pure statistically based system will translate metaphors more successfully and therefore will require less post-editing effort. The paper compares a hybrid MT and an SMT system due to the fact that the pilot experiment of this study was designed and conducted in 2013, therefore, the MT outputs were generated when both MT providers (Google Translate and Systran) had not yet switched to NMT. The 2013 MT outputs were used to conduct all the experiments so that data would be comparable, and we do not see the absence of NMT in

the analysis as an issue. Metaphors are a complex linguistic phenomenon still under-explored in post-editing tasks that in our view merit investigation in the context of machine translation based on all system architectures.

2. Literature review

2.1. Machine translation

Recent technological advances have allowed the development and improvement of several computerised tools such as spell checkers, grammar correctors, dictionaries, glossaries, terminology databases, and translation memories. This progress permitted the reduction of the time spent on translation tasks as well as considerably reducing manual tasks. At the same time, the increasing use of machine translation has led to a greater number of studies in the area.

Machine translation, or MT, results from the automatic process of translating one natural language to another with the use of computer systems (Baker and Saldanha 2012: 162). Such systems employ distinct architectures or approaches, and usually generate a raw output translation that serves as a starting point for human intervention - or post-editing.

According to Liu and Zhang (2015: 116), the raw MT output brings a significant improvement in translators' work efficiency. Moorkens *et al.* (2015: 267) add the benefits in translation productivity when introducing post-editing and when MT quality is sufficient.

MT systems can be classified according to their approach or architecture, such as example-based (EBMT), free/open-source (FOMT), pragmatics-based (PBMT), rule-based (RBMT), statistical (SMT), hybrid (RBMT and SMT engines), and Neural (NMT) (Chan 2015: xxix, Koehn 2017). In this study, we will consider two types of systems: hybrid and (pure) SMT.

SMT systems base their approach on calculating the probability that a target sentence is the translation of a given source sentence, also called the translation model. According to Chan (2015: 110): "an RBMT system uses rules to direct the process of MT." The author adds that these rules are encoded by linguistics experts (specification of rules for morphology, syntax, lexis etc.).

Based on the aforementioned classification, Systran applies a hybrid technology where RBMT components are developed by adopting linguistic resources for each language/language pair and using simple or multiword 'lexical entries' as customised disambiguation rules (Dugast *et al.* 2007). Systran combines an SMT module with the predictability and language consistency of an RBMT module at each stage of the process (analysis,

transfer, post-editing) in order to improve the translation quality (Systran 2009).

On the other hand, Google Translate, as used in this experiment, is a pure SMT system, based on the statistical analysis of multilingual corpora and using English as a pivot language to support MT between tens of other languages, also known as Interlingua (Chan 2015: 111). Instead of creating rule-based algorithms, the system analyses the probability of correlation between the segments of the different language pairs based on corpora available to the system.

More recently, both systems (Google Translate and Systran) have updated to the NMT approach, applying an encoder-decoder architecture. Although there has been a steep increase in MT systems and their output quality, its practical use by the translation industry still relies on the concept of post-editing.

2.2. Post-editing

According to ISO 18587 (2017), post-editing is the process of the analysis and correction of text resulting from an automatic or semiautomatic translation to ensure its compliance with grammar, punctuation, spelling and meaning. Post-editing quality levels vary greatly and will largely depend on the translation use. Typically, users of MT will ask for one of two different degrees of post-editing: light (or fast) post-editing or full post-editing.

Light post-editing results from the minimum number of changes and typing, so it involves essential corrections only. It is used for texts that are needed urgently and will have an internal, perishable use. Full post-editing, the focus of this study, reaches quality similar to human translation.

In order to meet the primary requirement of post-editing, that is, a balance between productivity and quality, some general guidelines should be followed. O'Brien *et al.* (2009) propose five fundamental guidelines: (a) Retain as much raw translation as possible, (b) Do not hesitate too long over a problem, (c) Do not worry if style is repetitive, (d) Do not embark on time-consuming research, and (e) Make changes only where necessary, i.e., nonsensical, wrong or ambiguous chunks.

According to Mesa-Lao (2013: 15-16), the guidelines for full post-editing need to meet varying customer expectations, therefore some of them might be added or eliminated when working in the field. These guidelines are employed in order to improve quality relative to post-edited output effort (see next section) and they are aimed at resolving MT errors. However, they cannot be considered as standard guidelines due to their vagueness. Differentiating between essential changes and preferential

changes can be quite challenging in some segments, as shown by Koglin (2015). In her study, some participants verbalised their difficulty in deciding when changes were absolutely necessary or only stylistic.

The effort to achieve certain levels of quality will be determined by the output quality that the engine is able to achieve. While productivity is directly related to the level of quality of the raw MT output, post-editing effort is inversely related to productivity, i.e. the higher the effort, the lower the productivity.

2.3. Cognitive effort in post-editing metaphors

Time spent to post-edit a text is the most easily measured and therefore visible aspect of post-editing effort; however, it can be approached in other ways (Krings 2001), such as analysing keystroke logging operations (temporal effort) or detecting MT errors and making a decision to correct them (cognitive effort).

The cognitive effort refers to the "type and extent of those cognitive processes that must be activated in order to remedy a given deficiency in a machine translation" (Krings 2001: 179). According to Krings (2001), the cognitive effort is complex to measure because it requires special tools such as Translog or eye trackers, which do not measure cognitive effort directly, but are assumed to provide measures that represent it.

In translation process-driven studies, the cognitive aspects of post-editing effort have been approached with the help of keystroke logging (Krings 2001; O'Brien 2005; Carl *et al.* 2011) and eye tracking data (Carl *et al.* 2011) in order to measure cognitive effort in terms of fixations and pupil dilation (Koglin 2015). Additionally, O'Brien (2005) and Koglin (2015) have also experimented with the use of think-aloud protocols (TAPs) to investigate cognitive aspects of post-editing.

In a 2004 study, O'Brien relates translatability with post-editing effort using a controlled language (CL) strategy in order to improve the quality of the source text (ST). The author applies translatability indicators by to estimate the suitability of the ST for MT into the target text (TT). Keyboard monitoring (Translog) and Choice Network Analysis (CNA) are used to measure the cognitive effort while the participants complete the translation task. The results show that post-editing effort decreases when CL is used, even though more research is still necessary to investigate problematic sentences such as gerunds or proper nouns.

There has been a great deal of research aimed at determining the feasibility of post-editing and at predicting post-editing effort based on MT output quality. Nevertheless, considerably less is known about the cognitive effort required to post-edit text types rich in metaphorical utterances. Most research on metaphors attempts to understand the

aspects of metaphor interpretation (Gibbs 1994, 2006, 2010; Gibbs and Herbert 2006; Gibbs and Tendahl 2006; Tendahl and Gibbs 2008; Gibbs *et al.* 2011), but fewer studies focus on metaphor post-editing.

From the perspective of Relevance Theory, metaphor is referred to as a "kind of 'loose use' in which, typically, the logical properties of the representation (mental or public) are inapplicable but which gives rise to a range of weak implicatures" (Carston 2008: 378). The implicatures are communicated assumptions which are derived solely via processes of pragmatic inference.

In this approach, the concept of 'loose use' refers to the "use of a representation (whether mental or linguistic) to represent another representation (whether mental or linguistic) with which it is in a relation of non-literal resemblance" (Carston 2008: 378). The interpretation of "loose use", i.e., metaphor interpretation, is accessed by narrowing or broadening its emergent properties in order to derive metaphor concepts.

Based on the results of an empirical investigation, Barsalou (1983, 1987) introduced the notion of *ad hoc* concept to explain metaphor interpretation. According to him, the main difference between *ad hoc* categories and ordinary categories is the fact that the first ones can vary whereas the second ones are more stable. As a result of that distinction resulting from Barsalou's findings, Carston (2010) has also included the *ad hoc* concept in metaphor interpretation.

According to Carston (2010), this additional cognitive component could explain how complex inferences are made during metaphor interpretation. Despite this important move, Relevance Theory still cannot fully explain creative metaphor interpretation compared to conventional metaphor interpretation, i.e., *ad hoc* concepts derived by loosening cannot account for innovative or extended metaphors. As stated by Romero and Soria (2014), sometimes complex concepts must be pragmatically adjusted in order to determine their contribution to the explicature, i.e., an ostensively communicated assumption which is inferentially developed from one of the logical forms.

Besides the pragmatic adjustments proposed by the Relevance Theory approach, in the case of post-editing machine translated metaphors, their interpretation will be conditioned not only to the source metaphor but also to the machine output generated for it. Differently from human translation, the cognitive process of post-editing involves first reading a segment of MT output, then comparing it against a segment in the source text, and next correcting and/or revising the MT output (Carl *et al.* 2011). Therefore, we believe that metaphor interpretation in this context and consequently cognitive effort required to post-edit machine translated metaphors will be affected by both linguistic stimuli.

Regarding the treatment of metaphors by a translating machine, some potential issues may pose difficulties for the machine. Regardless of the MT system, one of the main challenges for the machine is that the meaning of a metaphor cannot always be deducible from the meanings of its individual words.

In the classical approach RBMT, an MT system based on linguistic information about source and target languages basically retrieved from dictionaries and grammars, it seems unlikely that metaphors will be machine translated successfully since their constitution and interpretation is socially situated and they have clear pragmatic purposes (Gibbs *et al.* 2011). Additionally, one of the main shortcomings of this approach is its difficulty to process ambiguities and idiomatic expressions.

With respect to the pure SMT approach, which uses algorithms to establish probabilities between segments in a source and target language in order to propose translation candidates, it seems more likely to find a satisfactory translation for metaphors except when they are creative or not included in the corpora used by the MT system. However, Salton *et al.* (2014) have conducted a study to evaluate the impact of idioms on the SMT process using the language pair English/Brazilian-Portuguese. Their results showed that on sentences containing idioms the SMT system achieved a poorer performance, i.e., they had worse BLEU scores (cf. BLEU as in Papineni *et al.* 2002) when compared with sentences that did not contain idioms because it is difficult for the system to process an expression containing idiomatic or literal usage.

It should be noted that metaphors are widely used in different text types and they usually present challenging properties for an MT system, such as morphosyntactic variations or idiomatic and literal (non-idiomatic) usages (Salton *et al.* 2014: 36).

3. Data collection

To address the gap in the literature regarding the post-editing effort required to post-edit texts rich with metaphors, we conducted this study with a journalistic text to better understand the impact of different MT systems on MT output and post-editing effort. We expected the SMT engine to require less post-editing effort than the rule based one.

3.1. Participants

There were two groups of post-editing participants labelled as PE and PEm. One of them (PE) had 14 undergraduate students recruited at Federal University of Minas Gerais (UFMG) to take part in the experiment whereas the other group (PEm) consisted of 10 undergraduate students who volunteered at Federal University of Ouro Preto (UFOP).

The PE group consisted of male (61.5%) and female (38.5%) undergraduate students in Translation whereas the PEm sample had 80% male and 20% female participants. The PEm group consisted of undergraduate students in Translation (70%) and in Modern Languages: English (30%)

They all were native speakers of Brazilian Portuguese and considered English as their second language. Participants self-reported this information on a survey they were asked to complete before the experiment. The participants had no professional experience with post-editing, but the ones recruited at UFMG attended a 15-week post-editing course that was part of the regular undergraduate course while the participants who volunteered at UFOP were taught a 4-hour post-editing course given by the same instructor from UFMG.

3.2. Experimental design

Building on the experimental paradigm of data triangulation in translation process research, part of the experiment was conducted at Federal University of Ouro Preto (UFOP). The other part was carried out at the Laboratory for Experimentation in Translation (LETRA) using eye tracking, keystroke logging, and retrospective think-aloud protocols (TAPs).

First, all of the participants were asked to complete a short typing task in order to become familiar with all the keys on the keyboard. Next, half of the total participants of each group (UFMG and UFOP) were asked to post-edit a target text that was machine translated using Google Translate in Task 1 (T1) and to post-edit a Systran machine translated output in Task 2 (T2). The other half of the participants were asked to post-edit the same source text in a different order, i.e., Systran machine translated output in T1 and Google Translate in T2.

At the end of each task, all participants were asked to record both free and guided think-aloud protocols. In the free protocol, they were told to think aloud while their full post-editing process was replayed on Translog-II screen. In the guided protocol, they were asked two questions related to metaphor interpretation and its subsequent post-editing decision-making process.

3.3. Material

Both post-editing tasks were performed using the same source text, i.e., a 224-word journalistic text about the Tea Party Movement (see Appendix 1).

3.4. Procedure

In both post-editing experiments, participants identified with odd numbers were systematically assigned to receive Google Translate output for task 1 and Systran output for task 2. Participants identified with even numbers, on the other hand, had the opposite order of stimuli.

3.5. Apparatus

The participants recruited at Federal University of Minas Gerais (UFMG) were seated in front of a Tobii T60 eye tracker at a distance of 55 to 65 cm from the monitor whereas the participants recruited at Federal University of Ouro Preto (UFOP) were seated at a Tobii TX300 eye tracker at a similar distance.

Both Translog-II and Tobii Studio 3.2 were calibrated. Translog-II enabled participants to view the source text in the upper half part of the window and the machine output in the lower half part of the window. This software has been specially designed for process-driven studies because it enables the tracking of keyboard activity and mouse clicks.

4. Data analysis

For the purposes of this paper, the analysis of MT output impact on post-editing effort will focus on both TAPs and eye tracking data related to pupil dilation and fixation duration in five areas of interest (AOIs). All five AOIs contained metaphors (cf. Steen *et al.* 2010), as follows: The Party Pork Binge (metaphor 1 – henceforth M1), pork barrel spending (metaphor 2 – henceforth M2), poster child (metaphor 3 – henceforth M3), spending trough (metaphor 4 – henceforth M4) and bring home the bacon (metaphor 5 – henceforth M5).

The study aimed at analysing the impact of hybrid MT and SMT outputs on the cognitive effort required to post-edit machine-translated metaphors, so cognitive effort was measured by the mean fixation duration and pupil dilation on the AOIs. Additionally, verbal recordings were used to interpret and have a thorough understanding of quantitative data.

Due to poor quality eye tracking data, one participant from the post-editing experiment at UFMG (PE Group) was discarded for the purposes of this analysis. The threshold set for eye tracking data quality was 70% of time spent looking at the eye tracker screen (cf. O'Brien 2009).

All statistical analysis of quantitative data was performed using SPSS statistical software. The cut-off point for significance level was set at 0.05.

5. Results and discussion

The findings reported in this section have been triangulated from eye tracking data (fixation duration and pupil dilation) and verbal recordings made at the laboratory while participants were engaged in post-editing tasks. Longer fixations and increased pupil dilation represent higher cognitive effort.

Figure 1 provides the mean fixation duration distributed on Google Translate and Systran's output in five AOIs, i.e., five metaphors (M1, M2, M3, M4, M5) during task 1 and task 2 (T1 and T2) of PEm Group.

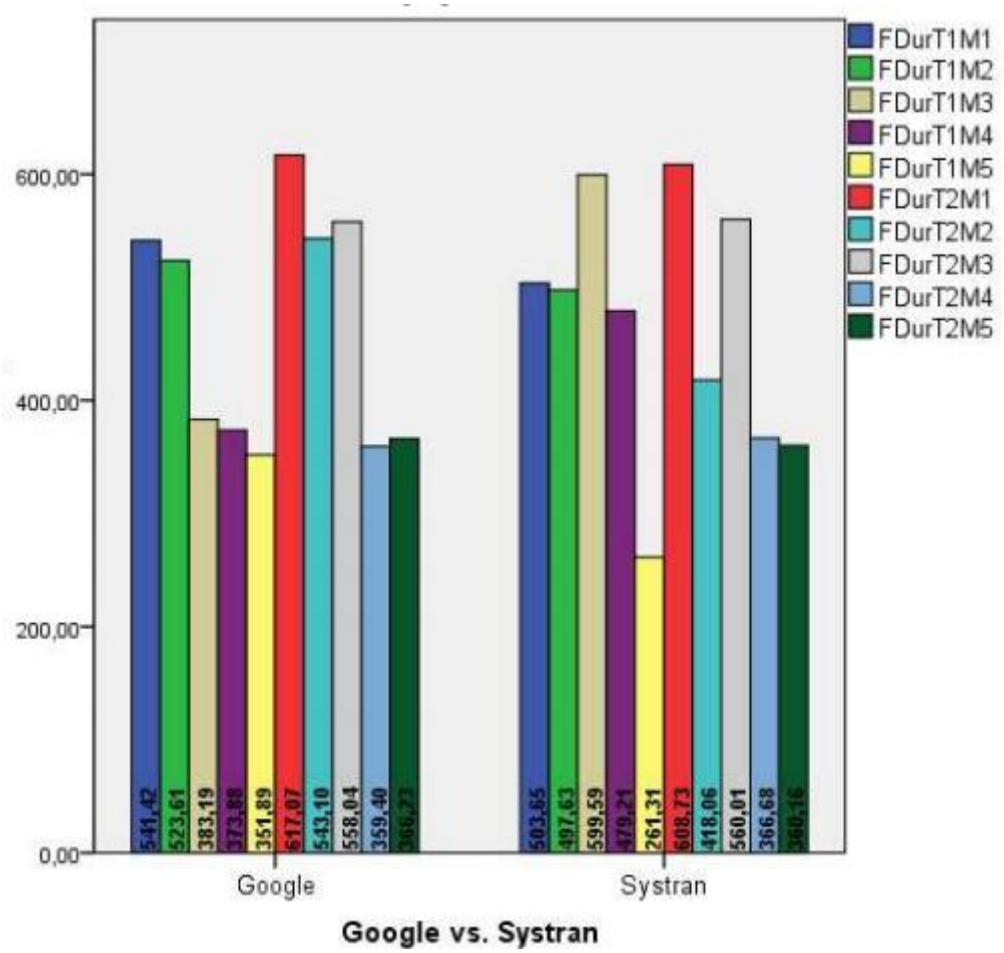

Figure 1. Mean fixation duration (ms) distribution on each MT system (Google Translate vs. Systran) in M1, M2, M3, M4, M5 during T1 and T2 of PEm Group.

As can be seen in Figure 1, the mean fixation duration in task 1 is lower when post-editing the output generated by Google Translate for two metaphors: M3 and M4. Systran's output, on the other hand, had lower fixation duration on the other metaphors: M1, M2 and M5. In task 2, the mean fixation duration followed a similar pattern.

Regarding the impact of the machine translation system on the cognitive effort required to post-edit metaphors in the PEm Group, the Mann-Whitney Test showed that fixation duration *probably* had a marginally

significant difference on M2 (Z = -1.78, p = 0.07) for the outputs generated by both Google Translate and Systran. The same metaphor required less effort to be post-edited in Task 2. These results seem to indicate an impact of the MT system on the cognitive effort, but that effect is not strong. Despite this weakness, the results are still useful because they are aligned with those found in previous studies (Costa-Jussà *et al.* 2012; Sreelekha 2017) as well as giving insight into the cognitive processes the participants go through.

As for the think-aloud protocols, our analysis suggests that the MT system architecture might have some influence as freely verbalized by one participant: "I like some of the solutions generated by this MT output. I think they were good" (P03)[1]. Although the participant is not referring to M2 specifically, his verbalisation suggests the second output (Systran) had a higher quality and therefore could have had a positive effect on the cognitive effort required to post-edit it.

Nevertheless, this interpretation should be considered with caution because task 2 might also have been influenced by the facilitating effect of task 1. This explanation is supported by the free verbalisation of participant P06 as follows: "Well, regarding the second task, I cannot say that it was easier, but it was simpler to identify the errors. I am not sure about the reason, whether it was because of the output or because I have already identified my difficulties in the first task[2]." However, later on, the same participant adds: "In my opinion, this output did not need so many corrections[3]."

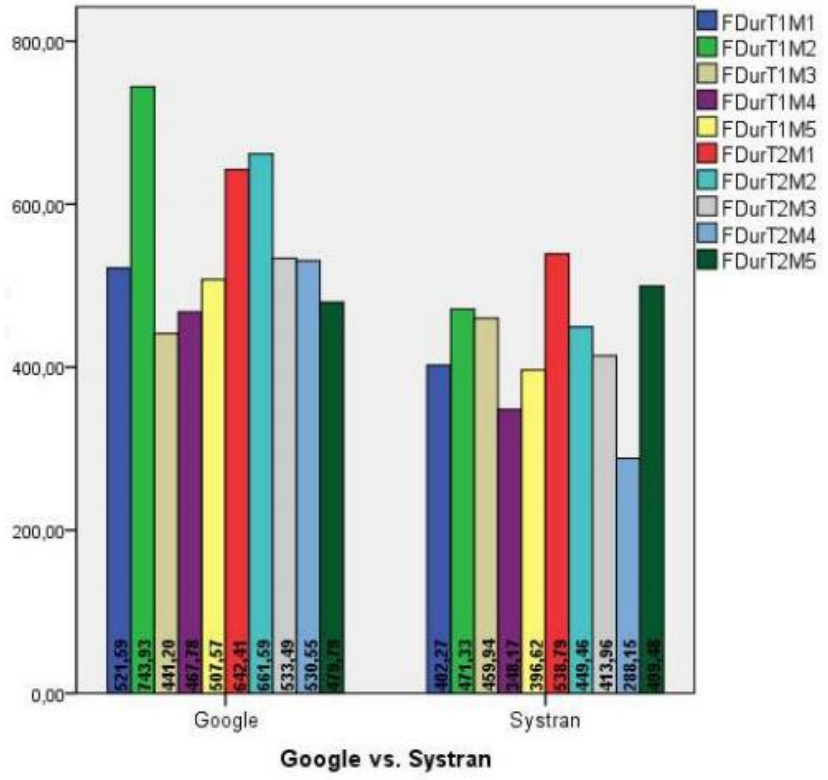

Figure 2. Mean fixation duration distribution on each MT system (Google Translate vs. Systran) in M1, M2, M3, M4, M5 during T1 and T2 of PE Group.

The analysis of this verbalisation suggests that both the MT system and the task had a positive effect on reducing post-editing effort in task 2.

Figure 2 will provide the results generated by PE Group (UFMG) and it shows fixation duration in milliseconds (ms) on tasks 1 and 2.

From Figure 2, we can see that the fixation duration is lower for the output generated by Google Translate only on M3 whereas Systran's output had a lower average fixation duration on M1, M2, M4 and M5. With respect to task 2, only M5 had a lower fixation duration when the output generated by Google Translate was post-edited. On the one hand, the result observed in task 1 might be considered unexpected if the MT systems are considered. Due to its hybrid model[4], it was likely that Systran would provide a more literal translation. As a result, the output could require more post-editing effort, especially in creative metaphors.

On the other hand, the more literal and less fluent translation provided by Systran, as opposed to Google's partial translation of some metaphors, might have been less cognitively demanding for the translators interpreting creative metaphors. Systran's output may have provided at least a hint of interpretation whereas it is very unlikely that Google Translate[5], a pure statistical system, would provide an automatic high-quality translation for creative metaphors, i.e. an MT output requiring few or no changes.

A closer examination of think-aloud protocols revealed that despite the fact that task 2 could be impacted by the facilitating effect of task 1, the MT system also influenced the cognitive effort required to post-edit metaphors as stated by P02 in task 2:

> It was easier to post-edit this time because I have already post-edited the same text[6]. On the other hand, this MT output was much better. Some of its translation options were even better than my own previous solutions to the same segment. And this MT system provided a cleaner text compared to the previous one, which makes post-editing much easier with this raw translation. This output required very little intervention[7].

The participant verbalisation clearly shows that the second MT output generated by Google Translate had a higher level of both readability and quality. Because of that, the participant mentions that fewer changes were required and consequently less effort to post-edit the pure statistical MT output. Interestingly, the participant admits that some of the solutions provided by the MT were even better that his/her decision-making during T1 post-editing. Additionally, he/she suggests there was a facilitating effect from task 1, but places less emphasis on it.

With respect to the impact of the MT system in the PE Group, the Mann-Whitney Test showed that fixation duration was *probably* marginally

significant on M2 (Z = -1.71, p = 0.08) and there were significant differences on M4 (Z = -2,00, p = 0,04) for both Systran and Google Translate. Although significance is not assured for M2, this result regarding the probable effect of MT system provides useful insights about the cognitive aspects involved in post-editing machine-translated metaphors. In addition, M2 and M4 were precisely the metaphors with no decrease in cognitive effort while post-editing them in task 2.

This result suggests that the cognitive effort required to post-edit M1, M3 and M5 could have been affected by the MT system architecture since no reduction of cognitive effort in task 2 has been observed. Therefore, the assumption of a facilitating effect on task 2 was not confirmed for these metaphors.

When investigating pupil dilation as a metric for measuring cognitive effort, Koglin (2015) has found a significant positive correlation between fixation duration and pupil dilation, so let us now turn to pupil dilation data to analyse the cognitive effort required to post-edit the same metaphors.

Figure 3 shows the average pupil dilation (mm) for each MT system, Google Translate and Systran. Unlike fixation duration, Figure 3 presents the average of both experimental groups (PE and PEm) because, during raw data processing, the different eye tracker's frequencies have been normalised in order to adjust data values.

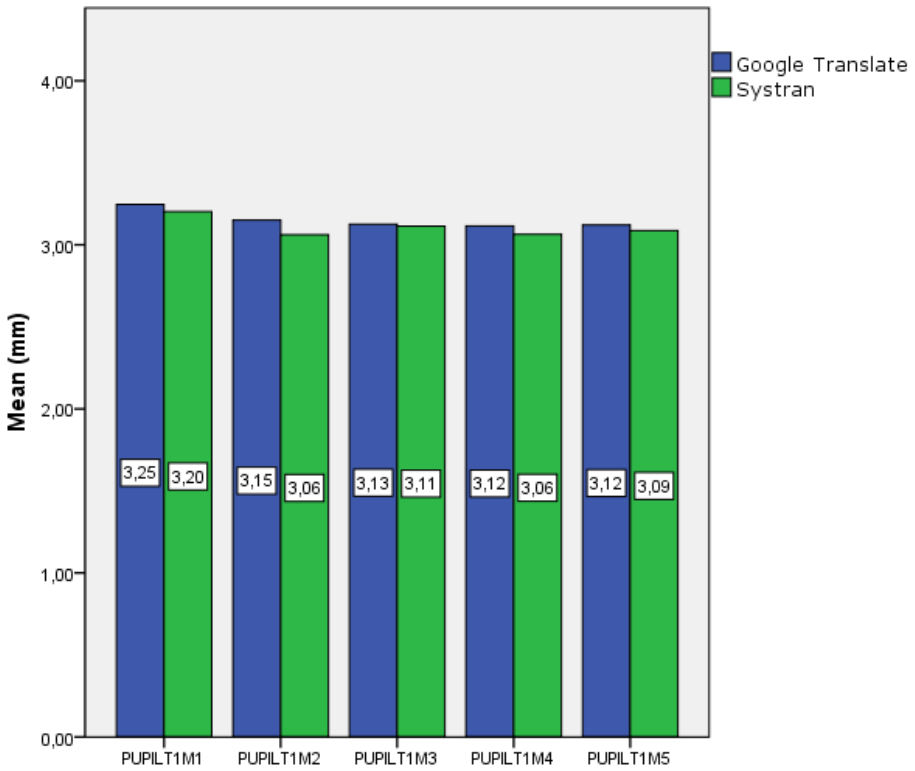

Figure 3. Pupil dilation (mm) by MT system (Google Translate vs. Systran) in M1, M2, M3, M4, M5 during T1 of both groups (PE and PEm).

Figure 3 shows that the mean pupil dilation is higher for all metaphors machine translated by Google Translate, on task 1, when both groups are analysed. This result is compatible with task 1 if we consider that task 2 might have been affected by the facilitating effect. Even though Systran has presented a more literal MT output for metaphors, its output might have contributed to interpreting this trope. The relationship between a conventional metaphor and its propositional form "most of the times is not completely arbitrary and, in many cases, its meaning is recovered through the meaning of the expression's individual constituents: the linguistic form serves as a clue to make inferences[8]" (Bylaardt 2006: 140).

This explanation is supported by pupil dilation analysis of task 2 as can be seen in Figure 4.

Figure 4 shows a slightly different trend regarding post-editing effort, i.e., the pupil has a higher dilation only on M3 when the pure statistical MT output was post-edited. However, this result may due to the facilitating effect of task 2 and not the MT system impact. Further analysis with a higher number of participants and non-metaphorical areas of interest should be conducted in order to have a more thorough understanding of different MT system impact on post-editing effort.

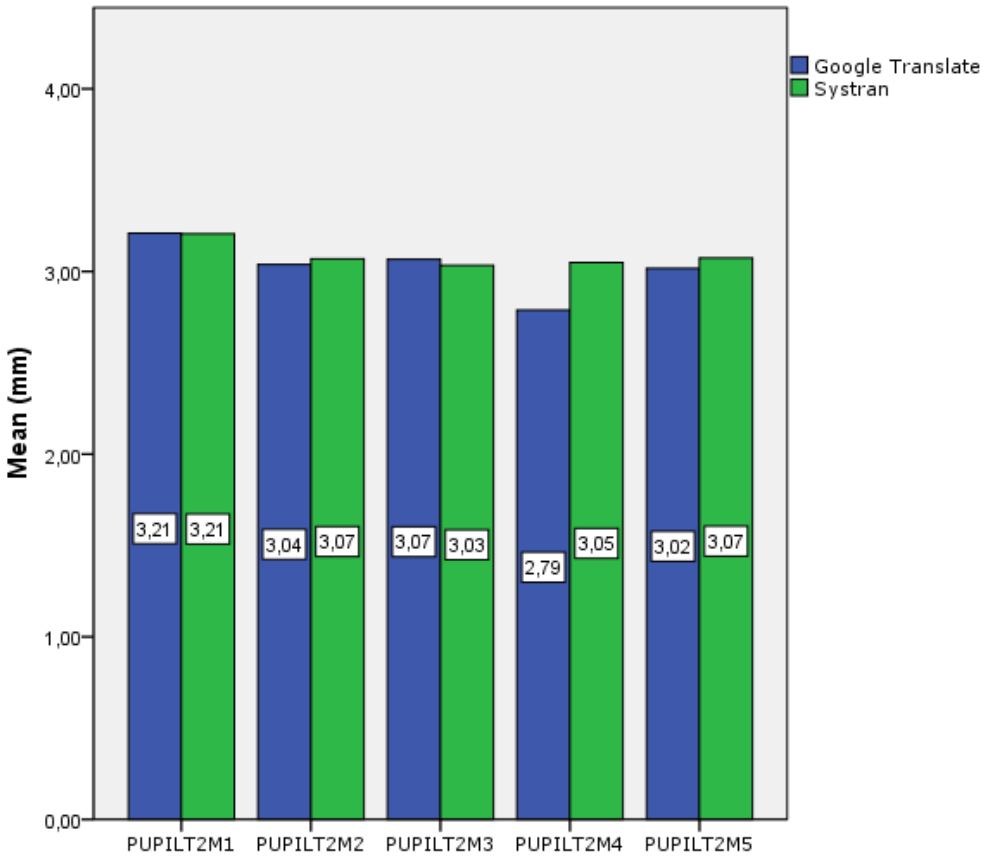

Figure 4. Pupil dilation (mm) by MT system (Google Translate vs. Systran) in M1, M2, M3, M4, M5 during T2 of both groups (PE and PEm).

6. Conclusions and future work

The present investigation has analysed the cognitive effort required to post-edit hybrid MT output and pure statistical MT output of metaphors in newspaper texts. The findings of this study suggest that post-editing pure statistical MT output might be less effortful when conventional metaphors are analysed whereas hybrid MT system might provide some clues for making inferences in creative metaphors and, therefore, could have a positive effect on the post-editing effort.

These findings on the comparative performance of SMT and hybrid MT by means of post-editing effort analysis are consistent with results from previous studies (Costa-Jussà *et al.* 2012; Sreelekha 2017). The findings also contribute to additional evidence on the cognitive effort required to post-edit machine translated metaphors in journalistic texts. However, due to the relatively small sample size and groups comprised of students, caution must be applied, as the findings might not be generalizable to professional post-editors.

In the future, we plan to experiment with participants that are more experienced and with non-metaphors to see if the results differ because of post-editing experience or as a consequence of literal versus figurative language. Additionally, we are planning to replicate the experiment by adding a neural MT output in order to test whether it affects the post-editing effort required to edit metaphorical language.

Acknowledgments

We would like to thank all the participants who contributed to this research project. The authors also would like to thank the National Council for Scientific and Technological Development (CNPq) for their financial support and the Laboratory for Experimentation in Translation (LETRA/UFMG) for providing infrastructure and team support. We should also acknowledge the journal editors and reviewers for their valuable comments.

References

- **Alfaro, Carolina and Maria Carmelita P. Dias** (1998). "Tradução Automática: uma ferramenta de auxílio ao tradutor." Cadernos de Tradução 1(3), 369-390.

- **Bahdanau, Dzmitry, Cho, Kyunghyun and Yoshua Bengio** (2014). "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473.*

- **Baker, Mona and Saldanha, Gabriela** (eds) (2012). *Routledge Encyclopedia of Translation Studies.* New York: Routledge.

- **Barsalou, Lawrence W.** (1983). "Ad hoc categories." *Memory & cognition* 11(3), 211-227.

- **—** (1987). "The instability of graded structure: Implications for the nature of concepts." Ulric Neisser (ed.) (1987). *Concepts and conceptual development: Ecological and intellectual factors in categorization*. Cambridge: Cambridge University Press, 101-140.

- **Bylaardt, Taciana** (2006). "A tradução de expressões idiomáticas à luz da Relevância." Fábio Alves and José Luiz Gonçalves (eds) (2006). *Relevância em tradução: perspectivas teóricas aplicadas.* Belo Horizonte: Editora UFMG.

- **Carl, Michael, Dragsted, Barbara, Elming, Jakob, Hardt, Daniel and Arnt Lykke Jakobsen** (2011). "The Process of Post-Editing: A pilot study." Bernadette Sharp, Michael Zock, Michael Carl, Arnt Lykke Jakobsen (eds) (2011). *Proceedings of the 8th International NLPCS Workshop. Special Theme: Human-Machine Interaction in Translation*. Copenhagen: Samfundslitteratur, 131-142.

- **Carston, Robyn** (2008). *Thoughts and utterances: The pragmatics of explicit communication*. Berlin: John Wiley & Sons.

- **—** (2010). "XIII—Metaphor: Ad hoc concepts, literal meaning and mental images." *Proceedings of the Aristotelian Society* 110(3), 295-321.

- **Chan, Sin-wai** (ed.) (2015). *Routledge Encyclopedia of Translation Technology*. London/New York: Routledge.

- **Costa-Jussà, Marta R., Farrús, Mireia, Mariño, José B. and José A.R. Fonollosa** (2012). "Study and comparison of rule-based and statistical Catalan-Spanish machine translation systems." *Computing and informatics* 31(2), 245-270.

- **Dugast, Loïc, Senellart, Jean and Philipp Koehn** (2007). "Statistical post-editing on SYSTRAN's rule-based translation system." *Proceedings of the Second Workshop on Statistical Machine Translation, Prague, June 2007*. Association for Computational Linguistics, 220-223. http://delivery.acm.org/10.1145/1630000/1626387/p220-dugast.pdf?ip=95.121.87.71&id=1626387&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437& acm =1543707140_b74394836989f05f1e21d51450622e22 (consulted 01.12.2018).

- **Gibbs, Raymond W.** (1994). *The poetics of mind: Figurative thought, language, and understanding.* Cambridge: Cambridge University Press.

- **—** (2006). "Metaphor interpretation as embodied simulation." *Mind & Language* 21(3), 434-458.

- **—** (2010). "The dynamic complexities of metaphor interpretation." *D.E.L.T.A* 26(SPE), 657-677. *Lodz Papers in Pragmatics* 1(7), 3-28.

- **Gibbs, Raymond W. and Herbert L. Colston** (2006). "Figurative language." Matthew Traxler and Morton Ann Gernsbacher (eds) (2006). *Handbook of psycholinguistics*. San Diego: Elsevier, 835-861.

- **Gibbs, Raymond W. and Markus Tendahl** (2006). "Cognitive effort and effects in metaphor comprehension: Relevance theory and psycholinguistics." *Mind & Language* 21(3), 379-403.

- **Gibbs, Raymond, Markus Tendahl and Lacey Okonski** (2011). "Inferring pragmatic messages from metaphor.2 *Lodz Papers in Pragmatics* 1(7), 3-28.

- **Google Translate**. http://translate.google.com (consulted 18.10.2013).

- **Gutt, Ernst-August** (1992). *Relevance theory: A guide to successful communication in translation*. Dallas: Summer Institute of Linguistics.

- **ISO 18587** (2017). *Translation services – Post-editing of machine translation output – Requirements.* Geneva: International Organization for Standardization.

- **Joscelyne, Andrew** (2006). *Best-Practices in Post-Editing.* Translation Automation User Society (TAUS) Special Report. https://www.taus.net/think-tank/reports (consulted 31.12.2014).

- **Koglin, Arlene** (2015). *Efeitos cognitivos e esforço de processamento de metáforas em tarefas de pós-edição e de tradução humana: uma investigação processual à luz da teoria da relevância.* [Cognitive effort and effects in the post-editing of machine translated metaphors compared to the translation of metaphors: a process-driven analysis in the light of Relevance Theory.] PhD Thesis. Federal University of Minas Gerais.

- **Koehn, Philipp** (2017). "Neural Machine Translation." *Statistical Machine Translation*. Chapter 13. Johns Hopkins University. *arXiv preprint arXiv:1709.07809*

- **Koponen, Maarit** (2016). *Machine Translation Post-editing and Effort. Empirical Studies on the Post-editing Process.* PhD Thesis. University of Helsinki.

- **Krings, Hans P.** (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-editing Processes* (Geoffrey Koby, ed.). Kent: Kent State University Press.

- **Liu, Qun and Xiaojun Zhang** (2015). "Machine Translation: General." Sin-wai Chan (ed.) (2015). *Routledge Encyclopedia of Translation Technology*. New York: Routledge, 105-119.

- **Mesa-Lao, Bartolomé** (2013). "Introduction to post-editing - The CasMaCat GUI." SEECAT. https://sites.google.com/site/centretranslationinnovation/seecat (consulted 04.04.2014).

- **Moorkens, Joss, O'Brien, Sharon, da Silva, Igor A.L., de Lima Fonseca, Norma B. and Fabio Alves** (2015). "Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation* 29(3-4), 267-284.

- **O'Brien, Sharon** (2004). "Machine translatability and post-editing effort: How do they relate." *Translating and the Computer* 26.

- **—** (2005). "Methodologies for measuring the correlations between post-editing effort and machine translatability." *Machine translation* 19(1), 37-58.

- **—** (2009). "Eye-tracking in translation process research: methodological challenges and solutions." Inger Mees, Fabio Alves and Susan Göpferich (eds) (2009). Methodology, technology and innovation in translation process research: a tribute to Arnt Lykke Jakobsen. Copenhagen: Samfundslitteratur, 251-266.

- **O'Brien, Sharon, Johann Roturier and G. D. Almeida** (2009). "Post-Editing MT Output. Views for the researcher, trainer, publisher and practitioner." Tutorial presented at *MT Summit XII: The twelfth Machine Translation Summit*, *Ottawa, 26-30 August 2009*. http://www.mt-archive.info/MTS-2009-OBrien-ppt.pdf (consulted 01.12.2018).

- **Papineni, Kishore, Roukos, Salim, Ward, Todd and Wei-Jing Zhu** (2002). "BLEU: A method for automatic evaluation of machine translation." *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (*ACL), Philadelphia, July 2002*, 311–318. https://www.aclweb.org/anthology/P02-1040.pdf (consulted 15.10.2018).

- **Romero, Esther and Belén Soria** (2014). "Relevance Theory and metaphor." *Linguagem em (Dis)curso* 14(3), 489-509.

- **Salton, Giancarlo, Ross, Robert and John Kelleher** (2014). "An Empirical Study of the Impact of Idioms on Phrase Based Statistical Machine Translation of English to Brazilian-Portuguese." *Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra), Gothenburg, Sweden, April 27, 2014*. Association for Computational Linguistics), 36-41. http://www.aclweb.org/anthology/W14-1007 (consulted 15.10.2018).

- **Sreelekha, S.** (2017). "Statistical Vs Rule Based Machine Translation; A Case Study on Indian Language Perspective." *arXiv preprint arXiv:1708.04559*.

- **Sutskever, Ilya, Vinyals, Oriol and Quoc V. Le** (2014). "Sequence to sequence learning with neural networks." *Advances in neural information processing systems*, 3104-3112. *arXiv preprint arXiv:1409.3215.*

- **Systran**. http://www.systransoft.com (consulted 28.11.2013).

- **—** (2009). *Report Document 2009*. http://www.systransoft.be/download/annual-reports/systran-annual-report-2009.pdf (consulted 20.12.2013).

- **Tendahl, Markus and Raymond W. Gibbs** (2008). "Complementary perspectives on metaphor: Cognitive linguistics and relevance theory." *Journal of pragmatics* 40(11), 1823-1864.

Biographies

Arlene Koglin is a lecturer at Federal University of Pelotas (UFPel) and a researcher at Laboratory for Experimentation in Translation (LETRA/UFMG). She holds a PhD in Translation Studies from the Federal University of Minas Gerais (UFMG) and a Master's degree in Translation Studies from the Federal University of Santa Catarina (UFSC). Her current research and publications focus on post-editing of machine-translated texts, translation process, metaphor, cognitive effort and eye tracking. She has experience teaching post-editing and translation to university students.

Email: arlenekoglin@yahoo.com.br

Rossana Cunha is a PhD Student at Federal University of Minas Gerais (UFMG) and a researcher at Laboratory for Experimentation in Translation (LETRA/UFMG). She holds a BSc in Computer Science (Federal University of Para), a BA in English and an MA in Translation Studies (Federal University of Santa Catarina). She is also a software developer and was a Research Assistant at Swansea University. Her areas of interest and research include computational linguistics, natural language generation, corpus-based translation studies, translation technologies and human-computer interaction.

Email: rossanacs@gmail.com

Appendix 1. Source Text[9]

The Tea Party Pork Binge

They brought the nation to the brink of default over spending, but a Newsweek investigation shows Tea Party lawmakers grabbing billions from the government trough. Plus, view the letters submitted by the 'Dirty Dozen.' House Majority Leader Eric Cantor, the Republican leadership's tether to the Tea Party, flutters the hearts of the government-bashing, budget- slicing faithful with his relentless attacks on runaway federal spending. To Cantor, an $8 billion high-speed rail connecting Las Vegas to Disneyland is wasteful *"pork-barrel spending."*

The Virginia Republican set up the "You Cut" Web site to demonstrate how easy it is to slash government programs. And he made the Department of Housing and Urban Development the *poster child* for waste when he disclosed that the agency was paying for housing for Ph.D.s. But away from the cameras, Cantor sometimes pulls right up to the *spending trough*, including the very stimulus law he panned in public. [...]

As the government showdown over debt continues—the so-called congressional supercommittee negotiating cuts has been floundering for weeks—Newsweek found about five dozen of the most fiscally conservative Republicans, from Tea Party freshmen like Allen West to anti-spending presidential candidates like Rick Perry and Ron Paul, trying to gobble up the very largesse they publicly disown, in the time-honored, budget-busting tradition of *bringing home the bacon* for local constituents.

Appendix 2. Google Translate output[10]

O Tea Binge Pork Partido

Eles trouxeram a nação à beira da inadimplência sobre os gastos, mas uma investigação Newsweek mostra legisladores Tea Party agarrando bilhões da calha do governo. Além disso, visualizar as cartas apresentadas pela 'Dirty Dozen'. Casa Líder da Maioria Eric Cantor, amarrar a liderança republicana para o Tea Party, palpita o coração do governo-bashing, orçamento-corte fiel com seus ataques implacáveis sobre os gastos federais em fuga. Para Cantor, 8 bilhões de dólares ferroviária de alta velocidade ligando Las Vegas a Disneyland é um desperdício *"Os gastos de porco barril."* O Republicano da Virgínia criou o "Você Cut" site para demonstrar como é fácil de cortar programas de governo. E ele fez o Departamento de Habitação e Desenvolvimento Urbano *a criança do poster* para os resíduos, quando ele revelou que a agência estava pagando por habitação para doutores. Mas longe das câmeras, Cantor, por vezes, puxa até o *vale de gastos*, incluindo a lei de estímulo muito, ele criticou em público. [...]

Como o confronto do governo sobre a dívida continua a supercommittee-so-called do Congresso negociando cortes foi tropeçando por semana-Newsweek encontrados cerca de cinco dezenas dos republicanos mais conservadores fiscais, a partir de calouros Tea Party como Allen West para anti-gastos candidatos presidenciais como Rick Perry e Ron Paul, tentando engolir a generosidade muito que repudiar publicamente, no time-honored, tradição orçamento-rebentando de *trazer para casa o bacon* para constituintes local.

Appendix 3. Systran output[11]

O frenesi da carne de porco do tea party

Trouxeram a nação ao limiar do defeito sobre a despesa, mas os legisladores de um tea party das mostras da investigação de Newsweek que agarram biliões da calha do governo. O sinal de adição, vê as letras submetidas "pela dúzia suja." Abrigue o cantor de Eric do líder da maioria, o baraço da liderança republicana ao tea party, vibrações os corações do governo-bashing, orçamento-corte fiel com seus ataques implacáveis na despesa federal do fugitivo. Ao cantor, um trilho $8 bilhões de alta velocidade que conecta Las Vegas a Disneylândia é do *"despesa desperdiçadoa carne de porco-tambor."* A Virgínia que o republicano estabelece "você cortou" o Web site para demonstrar como fácil é reduzir programas governamentais. E fez ao departamento de habitação e desenvolvimento urbano *a criança do cartaz* para o desperdício quando divulgou que a agência estava pagando abrigando para Ph.D.s. Mas longe das câmeras, o cantor puxa às vezes até *à calha da despesa*, incluindo a lei que mesma do estímulo filtrou em público. [...]

Enquanto o governo que a prova final sobre o débito continua- cortes de negócio do supercommittee do congresso assim chamado tem chafurdado para semana-Newsweek encontrou aproximadamente cinco dúzias dos republicanos o mais fiscal conservadores, dos caloiros do tea party como Allen ocidental aos candidatos presidenciais da anti-despesa como Rick Perry e Ron Paul, tentando devorar acima da largueza mesma repudiam publicamente, na tradição tradicional, orçamento-rebentando de *trazer em casa o bacon* para componentes locais.

Notes

[1] Translation from Brazilian Portuguese transcript of the free think-aloud protocol: "Gostei de algumas traduções desse novo aqui. Acho que foram boas." (P03_PEm_T2)

[2] Translation from Brazilian Portuguese transcript of the free think-aloud protocol: "Bom, a segunda tarefa, eu não vou falar que eu achei que foi mais fácil, mas foi mais simples perceber os erros, eu não sei se é porque o tipo da tradução mudou, ou se é porque eu já tinha percebido minhas dificuldades na primeira vez." (P06_PEm_T2)

[3] Translation from Brazilian Portuguese transcript of the free think-aloud protocol: "Esse realmente não tava precisando tanto de uma modificação pelo menos ao meu ver." (P06_PEm_T2)

[4] Systran was a hybrid MT system (rule-based and statistical engines) when data was collected.

[5] Google Translate provided statistical MT when data was collected.

[6] The participant refers to a different output but the same source text.

[7] Translation from Brazilian Portuguese transcript of the free think-aloud protocol: "Essa pós-edição foi mais fácil até porque eu já tinha feito a anterior com o mesmo texto, mas também o produto da tradução automática nesse caso foi bem melhor. Ele até apresentou algumas traduções melhores até do que eu fiz na anterior. E, e ele traz um texto até bem mais limpo da outra pós-edição, ficando bem mais fácil o trabalho da pós-edição nesse texto. E nesse eu fiz pouquíssimas alterações." (P02_PE_T2)

[8] Translation from: "na maioria das vezes, não é completamente arbitrária, e, em muitos casos, o significado é recuperado a partir dos significados dos constituintes individuais da expressão: a forma linguística serve como pista para a produção de inferências."

[9] The metaphors and their machine translation are italicized.

[10] The metaphors and their machine translation are italicized.

[11] The metaphors and their machine translation are italicized.